

Explainability, Fairness and Privacy in AI for Biometrics

Pedro David Carneiro Neto

Doctoral Programme in Electrical and Computer Engineering

Supervisor: Ana Filipa Sequeira, PhD

Co-Supervisor: Jaime dos Santos Cardoso, PhD

October 17, 2025

Explainability, Fairness and Privacy in AI for Biometrics

Pedro David Carneiro Neto

Doctoral Programme in Electrical and Computer Engineering

Approved by in public examination by the Jury:

President: Professor João Paulo Trigueiros da Silva Cunha

Referee: Professor Hugo Pedro Martins Carriço Proença

Referee: Professor Julian Fierrez Aguilar

Referee: Professor Luís Filipe Pinto de Almeida Teixeira

Referee: Ana Filipa Sequeira, PhD

October 17, 2025

Resumo

A adoção generalizada de sistemas de inteligência artificial (IA) está a ganhar, rapidamente, ritmo. Mesmo em ambientes, e aplicações menos seguras ou com elevado nível de risco, a IA está a incorporar a vida das pessoas. Alternativamente ao controlo, restrição ou proibição completa dos mesmos, deve-se procurar compreender a natureza, os riscos e as valências destes sistemas e como tirar o melhor partido deles.

Com o advento dos modelos de aprendizagem profunda e de grande escala, a segurança dos sistemas de reconhecimento facial, um ramo notoriamente delicado, ficou intrinsecamente conectada à sua explicabilidade. Com o passar do tempo, estas preocupações estenderam-se à privacidade e justiça.

Este trabalho doutoral focou-se em abordar vários aspetos relacionados com estas preocupações. Quer contribuindo com soluções e métodos novos, como realizando trabalho de investigação capaz de providenciar novas formas abordar, ou resolver um determinado problema. O primeiro dos grandes esforços foi alocado à explicabilidade, dando origem a uma nova abordagem a este tópico e novos métodos. Introduziu-se uma taxonomia para inteligência artificial explicável baseada em conceitos causais, que permite caracterizar modelos de IA explicável, além de facilitar a comparação entre os mesmos. Adicionalmente foram feitas contribuições sob a forma de novos métodos em múltiplas fases de um sistema de reconhecimento facial. Contribuiu-se para sistemas de deteção de ataques de apresentação e *morphing* com melhores resultados e de natureza mais explicável; métodos estado da arte para reconhecimento facial com máscaras; e um novo *score* probabilístico e interpretável para tornar os sistemas de reconhecimento facial mais usáveis.

Com o crescimento das preocupações com a privacidade, e devido à nova linha de investigação de reconhecimento facial com máscaras, decidiu-se por estudar a adaptabilidade destes modelos em contextos de reconhecimento facial com oclusões. Notou-se que um grande conjunto dos dados disponíveis para esta tarefa utiliza oclusões adicionadas de forma sintética, no entanto, não possuíam o aspeto desejado, nem realista. Por tal, desenvolveu-se uma estratégia para a criação de oclusões sintéticas com um posicionamento mais realista. Finalmente, realizou-se uma comparação, ao nível dos atributos, de conjuntos de dados reais e sintéticos, destacando algumas diferenças que podem apresentar-se como a causa do desnível na performance de modelos treinados nestes datasets.

O terceiro e último tópico de preocupação é a justiça. Esta temática foi abordada da perspetiva do viés no modelo e o viés nos dados. Para o primeiro, estudou-se como o viés se comportava para modelos de diferentes tamanhos, e se este comportamento se replicava quando modelos maiores eram quantizados para previsões menores. Em ambos os casos, o viés aumentou significativamente, no entanto, quantização com dados sintético pareceu diminuir a progressão do viés. Por outro lado, desafiou-se a noção de que as etnias são divididas em grupos discretos. Quando reinterpretados estes grupos para um espectro continuou, verificou-se como equilibrar conjuntos de dados alegadamente equilibrados.

Com o trabalho previamente mencionado, foi possível contribuir de forma significativa para

as linhas de reconhecimento facial e biometria com cerca de vinte publicações em revistas e conferencias científicas de renome. O trabalho foi também reconhecido com quatrocentas citações e vários prêmios, dos quais se destacam o prêmio de melhor artigo e o *EAB Max Snijder Award 2024*.

Palavras-chave: Inteligência artificial, biometria, reconhecimento facial, *deep learning*, visão por computador, justiça, dados sintéticos, inteligência artificial explicável

Abstract

The wide-adoption of artificial intelligent (AI) systems is quickly gaining traction, in fact, even in environments or applications with less safety or increased risk, AI is incorporating people's lives. Rather than controlling, restricting or completely forbid these systems, one must understand their nature, risks, strenghts and how to make the best use of them.

With the advent of large models and deep learning, the safety of face recognition systems, a very delicate field, has been intrinsically connected to its explainability. Over time, these concerns have extended to also privacy and fairness concerns.

This doctoral work focused on tackling several aspects of these concerns. Either by contributing with novel methods and solutions, or by performing investigation work capable of providing new insights or perspectives on how one can approach these problems. The first major efforts of this work have been directed towards explainability, a novel approach to this topic and well as new methods. With the introduction of a causality-based taxonomy that supports the categorization of explainable AI methods, it was created a framework to directly compare them. In addition, this work includes new methods, and experiments in multiple stages of a face recognition pipeline. For better and more explainable presentation and morphing attack detection system; for state-of-the-art masked face recognition trained in a constraint environment; and a novel interpretable probabilistic score to make face recognition systems more usable.

As privacy concerns rose, and masked face recognition created a new line of research, it was possible to explore works on occluded face recognition, and on understanding the adaptability of masked face recognition in these scenarios. For this task, datasets are mostly composed of synthetic occlusions, yet, they did not seem ideal nor realistic. Hence, a strategy to create synthetic occlusions with more realistic positioning was proposed. Lastly, on these concerns, attribute-based comparison between real and synthetic datasets have been conducted, highlighting some differences that might be causing the gap between models trained with these datasets.

The third and last topic of concern was fairness. This issue was adressed from the perspective of bias in the model and bias in the data. For the former, it was studied how bias behaves on different sized models, and if it replicates when larger models are quantized to have lower precision. These models have shown to be considerably more biased, and synthetic data helped mitigating the bias progression during quantization. On the other hand, the conventional idea that ethnicities are discrete attributes was challenged. By providing a continuous interpretation of these labels, it was possible to learn how to balance a balanced dataset to increase performance and fairness.

The work previously described managed to have meaningful contributions to face recognition and biometrics in general, with around twenty scientific publications in major computer vision and biometrics venues. This work has also been recognized by the community with four hundred citations and several awards, including the best paper award and the EAB Max Snijder Award 2024.

Keywords: Artificial Intelligence, biometrics, face recognition, deep learning, computer vision, fairness, synthetic data, explainable artificial intelligence

Acknowledgments

A PhD is a journey, a long and difficult journey. My particular journey was elevated by the presence of many friends, my family and fellow researchers. Because, as with any journey, a lonely journey is a sad and unworthy one. And the best things that come out a PhD are the experiences, the connections, and how much you grow and allow others to grow. The remaining part of the thesis will be about my work, so I would like to dedicate this part to everyone that was part of my life during this adventure.

If one believes that doing a PhD is hard, try to be the partner of someone that is pursuing a PhD, the true definition of insanity. And for that, I must thank my girlfriend (Ritinha, aka. floco de neve, aka. Ana Rita), and now house and family partner, for being on my side for every second of this degree. For supporting my decisions, for calling me out regarding my overworking nature, and for loving me with all her heart. I LOVE YOU. My parents (Alvaro e Antonieta), who motivated me endlessly since I was a baby, played a huge role in my success, they taught me endless lessons that I will take with me forever, while never leaving my side. I LOVE YOU. To my brother (Tiago Neto), who is a great source of strength and hides in his young age a natural wisdom that has, so many times, helped me clear my head and focus on important things. I LOVE YOU.

To Luis Vasconcellos, João Vasconcellos and Filomena Capelas, I wish to thank you for hosting me, for treating me as a family member. To Avó Geny, Avô Vasconcellos, Barbara Vasconcelos (which shared some of the same pain), Maria Capelas, Tio Nando, Tia Zé, André Capelas and Tio Paulo, a new extended family that adopted me as theirs. To Fátima, Carlos, Inês and André Caetano for being the cool family in the USA that keeps motivating me and my brother to be better and to aim for big things. You were crucial for the success of this journey and provided me a safe space towards this goal. The support of my old friends Arteiro and Rafa has continuously motivated me, and I am happy to be working next to them again. To Fonseca, one of my oldest friends, thank you for the support, the conversations and being my friend for more than 11 years. To Maria, Marcus (a fellow researcher), Catarina e André, the new friends that I expect to take for life. To Fina, Zé and Luis Frutuoso, to Berto, Fátima and Patricia, to Tó, Teresa and Joana. Thank you for the limitless support.

To everyone that is and that was a member of VCMI, I couldn't have chosen a better group. I became a "shining star" because of you. From the scientific discussions with Isabel, the ethics course with Tiago, the endless discussions over coffee with João Nunes and Eduardo, the "Sueca" games with Filipa and Joana, to the travels to eat "Leitão" in Agueda, those are some of the most important memories that I will take with me. Sara, my private Amsterdam guide, Wilson, João Pinto, Diogo Pernes, and so many other (older) VCMI members were crucial to make me mature during the initial phase of my journey. To Rafael, who believed that I am the right person to help him navigate in the same waters!

Darmstadt was a very important place in this journey. Not the place, not the awards that were won there, but the people. Naser, Marco, Fadi, Jan, Meiling. You made a great contribution to my personal and professional progression. One of my saddest realizations is that I won't be able to

visit you every year for a week full of beer, science and fun. That week was a meeting place for amazing people, from Kiran to Keerti and all the bright researchers on Biometrics. Had some of the best weeks with all of you!

To the biometrics summer school crew, Alex, Marco (again), Nélida and Ena. You make science look and sound cool, especially after some cocktails. To my previous master students, I have learned so much from you, from your questions, your ideas and specially from the hard discussions. Some chapters of this document originated from such discussions. Thank you for the amazing experience and letting me be part of your journey too.

To my current boss Renato, who believed in me and offered me a new opportunity while still writing this document. My current team that demonstrated a strong support and interest on this journey! I wish you the best for your future.

I couldn't finish this without thanking my Supervisors. Ana Filipa Sequeira and Jaime Cardoso. You believed that I could do the PhD since the first day that we met at the Summer Internships, those that I have pursued during the vacations of my Master's in Finland. And that faith made me believe that I could be better. Filipa was instrumental in every stage of this process, she was always available to discuss with me, to talk with me and to guide me. Jaime is one the brightest minds that I have met. A truly outstanding scientist that I admire with all my brain. If I had to start a new PhD the day after finishing this one, I would pick you again as my supervisors, not only because of your excellence, but because I believe that I still have so much to learn from you. Jaime trusted me with one of the toughest tasks ever, to teach. And for that opportunity I will be forever grateful.

And finally, a word to those that made me a Pawrent, those who clean my head, who cuddle me with their extensive fur, my daughters Bailey and Eevee! Every day becomes easier when it starts next to you, and happier when it ends next to you! I LOVE YOU. To my old pets Nina and Tonecas (Fanecas), you have taught me how to love a pet!

If this text fails to mention anyone dear to me, or important in this journey, then this is a reminder that this is just a piece of text that I have written the night before submitting the final version. Your mark in my life will never leave me!

Pedro moves his wand and says : Mischief managed

Pedro David Carneiro Neto

“It matters not what someone is born, but what they grow to be.”

Albus Dumbledore

“Wit Beyond Measure is Man’s Greatest Treasure”

Helena Ravenclaw

“The wise speak only of what they know.”

Gandalf

““All we have to decide is what to do with the time that is given to us.”

Gandalf

Contents

I	Prologue	1
1	Introduction	3
1.1	Context and Motivation	3
1.2	Objectives	4
1.3	Contributions	5
1.4	Dissemination	7
1.5	Collaborations	11
1.5.1	Research Projects	11
1.5.2	Synergies within the research group	12
1.5.3	Organisation of scientific events	13
1.5.4	Teaching	13
1.5.5	Supervision of dissertations and internships	14
1.6	Awards and Distinctions	15
1.7	Document Structure	15
2	Fundamental Concepts	17
2.1	Introduction	17
2.2	Biometric Systems	17
2.2.1	Overall Structure	17
2.2.2	Market Regulation and Security Concerns	22
2.3	Biometric Traits	24
2.3.1	General View	24
2.3.2	Face	25
2.4	Performance Evaluation	30
2.4.1	System Design	31
2.4.2	Data Quality	33
2.4.3	Recognition	33
2.4.4	Non Error Metrics	36
2.4.5	Users' Satisfaction	37
2.5	Conclusion	37
II	Explainable Biometrics	39
3	Explainable Artificial Intelligence	41
3.1	Context and Motivation	41
3.1.1	Towards a common terminology for xAI in Biometrics: Can we speak the same language?	41

3.1.2	A brief introduction to xAI	43
3.1.3	The causal relationships of a deep model	45
3.1.4	Work in xAI for Biometrics	53
3.1.5	Work in xAI computer vision that can benefit biometrics	62
3.2	Illustrating the Benefits of Interpretable Biometrics	64
3.2.1	Applications	65
3.2.2	The incompleteness of the current explanations	68
3.2.3	Unveiling biases	69
3.3	Conclusion: In which direction should we look in the future?	70
4	Inspecting Background Impact on the Model Behaviour	73
4.1	Context and Motivation	73
4.2	Related Work	75
4.3	Methodology	76
4.3.1	Binary classification (BC-PAD)	76
4.3.2	Multi-task classification (MT-PAD)	76
4.3.3	Adversarial training	77
4.3.4	Dynamic frame selection (DFS)	79
4.3.5	Evaluation Metrics	79
4.3.6	Experimental Setup	79
4.4	Datasets	79
4.5	Results and Discussion	81
4.5.1	Intra-dataset	81
4.5.2	Cross-dataset	83
4.5.3	One attack	83
4.5.4	Unseen attack	83
4.5.5	Visual Analysis	84
4.6	Summary and Conclusion	85
5	Identity Disentanglement for Morphing Attack Detection	87
5.1	Context and Motivation	87
5.2	Datasets	89
5.3	OrthoMAD	90
5.4	IDistill	91
5.5	Experimental Setup	94
5.5.1	OrthoMAD	94
5.5.2	IDistill	94
5.5.3	Evaluation Metrics	96
5.6	Results	96
5.6.1	Orthogonality	96
5.6.2	IDistill and OrthoMAD	96
5.7	Conclusion	98
6	Guiding the Model Vision for Masked Face Recognition	99
6.1	Context and Motivation	99
6.2	Related Work	101
6.3	Conditioned Triplet Loss	102
6.3.1	Methodology	102
6.3.2	Experimental Setup and Results	104

6.4	FocusFace	107
6.4.1	Methodology	108
6.4.2	Experimental Setup	111
6.4.3	Results	113
6.5	Conclusion	116
7	PIC: An Optimal Confidence Score for Interpretable Face Matching	117
7.1	Context and Motivation	117
7.2	Related Work	119
7.3	Methodology	120
7.3.1	Probabilistic Interpretable Comparison Score	120
7.3.2	Training PIC	121
7.3.3	Discussion	121
7.4	Experimental Setup	122
7.4.1	Databases	122
7.4.2	Evaluation Metrics	122
7.4.3	Face Recognition Models	123
7.4.4	Confidence Estimation Approaches	123
7.5	Results	125
7.5.1	Score Distribution Analysis	125
7.5.2	Single-Comparison Calibration Analysis	125
7.5.3	Multi-Comparison Analysis	127
7.6	Conclusion	129
III	Face Recognition with Synthetic Data	131
8	On the Generalization of Masked Face Recognition Models to Occluded Face Recognition	133
8.1	Context and Motivation	133
8.2	Related Work	135
8.2.1	Face recognition	135
8.2.2	Occluded face recognition	136
8.2.3	Masked face recognition	136
8.3	Methodology	137
8.3.1	Face verification by Masked Face Recognition models	138
8.3.2	Face verification by Occluded Face Recognition models	139
8.3.3	State-of-the-art face recognition	139
8.4	Databases	140
8.4.1	Masked face recognition benchmarks	140
8.4.2	Occluded face recognition benchmarks	142
8.5	Experimental Setup	143
8.5.1	Experiments	144
8.5.2	Evaluation metrics	144
8.6	Results	145
8.6.1	AR Face Database	145
8.6.2	Occ-LFW Database	146
8.6.3	Masked Face Verification	148
8.6.4	Results qualitative analysis	149

8.7	Conclusion	152
9	Towards Better Data for Occluded Face Recognition Evaluation	153
9.1	Context and Motivation	153
9.2	Database and evaluation criteria	155
9.2.1	OCFR-2022 evaluation benchmark	155
9.2.2	Evaluation criteria	157
9.2.3	Submission and evaluation process	158
9.3	Participants and submitted solutions	159
9.3.1	BIOSMC-CDTA Team	159
9.3.2	SiMiT Lab Team	159
9.3.3	UFPR-UNICO Team	160
9.4	Results and discussion	161
9.4.1	Non-occluded vs Occluded	162
9.4.2	Occluded vs Occluded	164
9.5	Conclusion	164
10	On the Domain Gap Between Synthetic and Real Datasets	165
10.1	Context and Motivation	165
10.2	Related Work	167
10.3	Methodology	168
10.3.1	Datasets	168
10.3.2	Experimental Design	169
10.3.3	Experimental Setup	172
10.4	Results	172
10.5	Conclusion	176
IV	Biases in Face Recognition	177
11	Inspecting the Biases Through Model Compression	179
11.1	Context and Motivation	179
11.2	Datasets	181
11.2.1	MS1MV2	181
11.2.2	BUPT-Balancedface and BUPT-Globalface	181
11.2.3	Synthetic data	182
11.2.4	RFW	182
11.3	Methods	182
11.4	Experimental Setup	183
11.4.1	Evaluation Metrics	183
11.5	Results	184
11.5.1	Bias on different architectures	185
11.5.2	Bias on different datasets	186
11.6	Conclusion	186
12	A Truly Balanced Dataset for Fair Face Recognition	189
12.1	Context and Motivation	189
12.2	Related work	192
12.3	Methodology	194

12.3.1	Balancing Training Data	195
12.4	Experimental setup	200
12.4.1	Datasets	200
12.4.2	Balancing Criteria	202
12.4.3	Ethnicity Score as an Utility Measure	203
12.4.4	Ethnicity Classifier	204
12.4.5	Implementation Details	205
12.4.6	Performance Evaluation	205
12.4.7	Hardware	206
12.5	Results	206
12.5.1	Single Ethnicity Analysis	206
12.5.2	Balancing in the Continuous Space	207
12.5.3	Protocol A	208
12.5.4	Protocol B	209
12.5.5	Protocol C	210
12.5.6	Comparison between Protocols	211
12.5.7	Different Architectures	211
12.5.8	Different Loss	212
12.6	Conclusion	213
V	Epilogue	215
13	Conclusion	217
	References	219

List of Figures

1.1	Overall summary of the contributions that resulted from author's doctoral work. Faded block are not included in the chapters of this thesis. Besides biometrics, other topics in computer vision and patter recognition that benefited from the author's contribution were not included.	6
2.1	Overall structure of a biometric recognition system. The system is divided into the Acquisition, Quality Assessment, Feature Extraction, Storage and Decision modules. Structure inspired by [40, 238, 406, 423]	18
2.2	Discrimination of the different tasks and operations that can be conducted by a recognition system. The enrollment (a), the identification (b) and the authentication (c). The decision module is not required by all of the operations, and the interaction between the user and the storage also differs.	19
2.3	General structure of a presentation attack detection system. Differently from recognition systems, these do not include a storage module, and thus, the decision is only based on the input.	20
2.4	Scheme representing the three different approaches to integrate presentation attack detection methods (green arrows) into recognition system. Attacks can also be performed at different stages of the recognition system (red arrows). Quality assessment module was omitted for simplicity. Diagram inspired on [172].	21
2.5	Images of the same individual captured on different light spectrum. It is possible to see the notables differences in the information captured by each technique. Images from [93]	26
2.6	Illumination variations in the Yale dataset [37] (Image from [384])	27
2.7	Pose variations in the ORL dataset [431] (Image from [384])	27
2.8	Face expression variations in the JAFFE dataset [319, 318] (Image from [384])	27
2.9	Age variations in the MORPH dataset [255] (Image from [384])	27
2.10	Examples of intra-subject variability that can affect the performance of face recognition systems. Sub-figures from [384]	27
2.11	Example of two inputs that could be used by a face recognition algorithm to compare an unmasked reference with a masked probe. The majority of the face is covered by the mask on the probe, and the to further increase the complexity of the task, the subject is wearing glasses in both images.	29
2.12	Examples of different types of presentation attacks. Sub-images (a)-(c) were collected from the ROSE-YOUTU dataset [297]. Sub-image (d) was collected from the Multispectral Latex Mask based Video Face Presentation Attack (MLFP) dataset [8].	30

2.13	Relation between the scores given to the impostors and genuine claims and its impact on the FAR and FRR error metrics. Besides it shows how the EER is graphically represented as the interception of the FAR and the FRR. Adapted from [470].	34
2.14	Receiver Operating Characteristic (ROC) curve regarding the FRR and the FAR. While this curve does not have the same structure as traditional ROC curves, it can be converted to a traditional curve by plotting 1-FRR on the y-axis. Adapted from [146].	35
3.1	The Ladder of Causality (adapted from [395]).	42
3.2	How we can explain black box models: a summary of current <i>post hoc</i> explainability approaches (adapted from [34]).	44
3.3	A figure illustrating the World's Causal Model which represents all the causal relationships that we observe and that rule the functioning of the world. Below we see a representation of the Causal Model of a Deep Neural Network. While the latter represents the causal relationships within the "rules" learnt by the DNN, the former represents the true causal relationships of the world. These two models are very likely to disagree, and the latter does not aim to explain the causal relationships of the world, just the specific deep learning model it represents.	45
3.4	Causality ladder repurposed to categorise different types of xAI approaches according to the understanding of the model provided by them.	46
3.5	Causal model representation of the training. The input, the label and the training rules directly affect the model, whilst the model affects the final prediction. The Inferred Label and the True Label are also separate, while the former is given to the model and inferred from the input, the latter is the cause of the input and is not always available.	50
3.6	Causal model representation of the inference process. The Model node is a causal system by itself and it is the system that we are trying to understand.	50
3.7	Proposed structure for a complete evaluation of the performance of biometric systems at four distinct levels. Quantitative, qualitative, security and trustworthiness evaluation. The latter is focused on the insights provided by explainable artificial intelligence methods.	66
3.8	Structure of a biometric system that leverages explainable artificial intelligence methods to improve its optimization function and to provide useful feedback to the user.	67
3.9	Simple draft of a diagram exemplifying how Grad-CAM output as heatmap-based images can be interpreted in a different manner by humans.	68
4.1	Architecture of the multi-task learning model. It receives an image (yellow box) and includes a CNN (blue figure), two output heads (green figures), where the first is a binary head and the second has 8 output nodes. The output node for the genuine samples on both heads ensures that the main goal remains on the detection of attacks.	76
4.2	Architecture of the adversarial learning model. It receives an input image (yellow box) and includes a CNN (blue figure), two feature vectors e_1 and e_2 (green boxes). These are used to reconstruct each other, decode (purple figure) the input and to classify the input. This architecture is deeply based on the Unsupervised Adversarial Invariance [240].	77

4.3	Architecture of the method for dynamic frame selection. It includes a CNN (blue figure) with shared weights that processes all the frames (yellow boxes) of a video and selects one (lower yellow box) based on a specific criteria.	78
4.4	Samples collected from the ROSE-YOUTU dataset [297] containing images from attacks and genuine captures. On the top row, cropped images are displayed. Whereas the bottom row contains the exact same images, but with all the background information included.	80
4.5	Explanations generated with Grad-CAM++ for different attacks of subject #23. -B indicates that the model was trained and tested on images with background. The (d) image was wrongly classified as genuine.	85
5.1	Overview of the architecture design for the integration of the new regularisation term. The regularisation loss is applied at the vector (Z_1 and Z_2), hence, their gradients do not affect the final classification layer. The system is composed of three components, a backbone network, two identity vectors and a final classification layer that uses the concatenation of both vectors.	90
5.2	Overall scheme of the architecture. Part A represents the training of the auto-encoder, whereas part B represents the training of the classification system. The orange line represents the backpropagation, the green one represents the calculation of the angle between vectors and the yellow represents the fusion of the two scores. Best viewed in color.	92
6.1	Triplet loss effect in the euclidean space	103
6.2	Proposed Triplet loss effect in the euclidean space	104
6.3	Examples of face landmarks obtained from one image with different types of masks added. These masks vary in shape and face coverage.	105
6.4	Examples of images from the real masked faces dataset: images of the same individual with (a) and without a mask (b) (from [106]).	105
6.5	Output of the Smooth Grad-CAM++ computed for each of the 512 features of the feature vector and normalized. Subfigures (a)-(f) computed from the cross-entropy model; (g)-(l) computed from the triplet loss model; and (m)-(r) computed from the mean squared error and triplet loss model.	108
6.6	Architecture of the proposed model. The backbone network receives masked and unmasked inputs, and computes the two embeddings for each input. The smaller embedding of each input is trained for mask detection with the cross-entropy (CE) loss. Whereas the other is trained for face recognition with the ArcFace loss. A third component of the loss is also included in the form of the mean squared error (MSE) between the two embeddings used for face recognition.	109
6.7	Images extracted from the validation dataset for masked face recognition: images of the same individual with (a) and without a mask (b) with a accurate augmentation; images of the same individual with (c) and without a mask (d) with a poor augmentation.	112

7.1	Probabilistic Confidence Interpretation Problem - Given two biometric samples, the face recognition system comes to the decision that the sample belonging to the same identity (genuine). Evaluating the correctness of this decision based on independent test data results in a probability of 72% that this decision is correct. However, most confidence estimation approaches (yellow) either overestimate (red) or underestimate (blue) the confidence. Contrarily, the proposed PIC approach provides a confidence value (green) reflecting the probability the decision is correct.	118
7.2	Score Distribution Analysis -The original (top) and the PIC (bottom) score distributions are shown for FaceNet. Based on the original score distributions, also the corresponding probabilities (see Sec. 7.3.1) for genuine and imposter are shown at the top. On the bottom, the optimal probabilistic interpretable comparison (PIC) scores are shown. Since the PIC score distributions build on a monotonic transformation, the order of the scores, and thus the performance, remains the same. . .	123
7.3	Confidence Calibration Curves (CCC) - The CCC for all dataset and FRS combinations are shown. Inconsistencies are due to the low number of samples for specific probability bins (e.g. for LFW and Morph). While most approaches have to deal with high under- and over-confident predictions, the proposed PIC produces close-to-ideal (black line) probabilistic confidence estimates in most cases.	124
8.1	Illustration of some occlusions that can highly increase the difficulty of the face verification task.	134
8.2	Several approaches can be followed in order to handle masks. These are the three most common approaches: Unmasking the input image, unmasking the template generated by a face recognition model, or using a model that is robust to masks. Blue templates do not contain mask information, whereas red templates contain mask information. Models robust to mask usually have some mask information leaked into their embedding, but in a low proportion when compared to the non-mask information. (Best viewed in color.)	138
8.3	Examples of the occluded face verification datasets. The first two rows correspond to Occ-LFW with synthetic occlusions, whereas the third corresponds to the AR database with real occlusions. The first row contains images from the Occ-LFW-1.0, and the second contains images from the Occ-LFW-2.0.	140
8.4	Examples of the validation datasets (LFW, CALFW, CPLFW, CFP-FP and AgeDB-30) augmented with face masks for masked face verification. The approach followed for mask augmentation is described by Huber <i>et al.</i> [221].	141
8.5	Study of the performance of the algorithms when exposed to no occlusions, masks and general occlusions.	148
8.6	Analysis of the performance variation of the algorithms in several versions of the LFW dataset that include different intensity of occlusion.	149
8.7	Some example pairs from the Occ-LFW-2.0 dataset that were wrongly classified as positive by the following models: FROM, ElasticFace - Arc+, KD - HG and EUM. Each couple of images placed horizontally represents a pair.	149
8.8	Some example pairs from the Occ-LFW dataset that were wrongly classified as negative by the following models: FROM, ElasticFace - Arc+, KD - HG and EUM. First row represents 1.0 dataset and the second the 2.0 dataset. Each couple of images placed horizontally represents a pair.	150

8.9	Some example pairs from the Occ-LFW dataset that were wrongly classified as positive by the following models: FROM; correctly classified as negative by the following: ElasticFace - Arc+, KD - HG and EUM. First row represents 1.0 dataset and the second the 2.0 dataset. Each couple of images placed horizontally represents a pair.	150
8.10	Some example pairs from the Occ-LFW dataset that were wrongly classified as negative by the following models: FROM; correctly classified as positive by the following: ElasticFace - Arc+, KD - HG and EUM. First row represents 1.0 dataset and the second the 2.0 dataset. Each couple of images placed horizontally represents a pair.	151
8.11	Some example pairs from the Occ-LFW dataset that were wrongly classified as positive by the following models: ElasticFace - Arc+, KD - HG and EUM; correctly classified as negative by the following: FROM. First row represents 1.0 dataset and the second the 2.0 dataset. Each couple of images placed horizontally represents a pair.	151
8.12	Some example pairs from the Occ-LFW dataset that were wrongly classified as negative by the following models: ElasticFace - Arc+, KD - HG and EUM; correctly classified as positive by the following: FROM. First row represents 1.0 dataset and the second the 2.0 dataset. Each couple of images placed horizontally represents a pair.	152
9.1	Examples of occluded face images of the OCFR-2022 benchmark comprising top of head (e.g. hats), upper (e.g. carnival masks), eye-based (e.g. sunglasses), and lower (e.g. mouth occlusions such as books, masks) face occlusions. Original non-occluded images are from the Labeled Faces in the Wild (LFW) dataset [218].	156
9.2	Examples of occluded face images comprising, from left to right, a combination of top of head and lower face (T+L); upper and lower face (U+L); and eye-based and lower face occlusions(E+L). Original non-occluded images are from the Labeled Faces in the Wild (LFW) dataset [218].	156
9.3	Comparison of the performance of the Baseline algorithms and the submitted solutions on the easiest and the hardest Protocols, 1 and 5 respectively. Side by side analysis of the performance of BLR-OP and OR-OP. The algorithms are displayed in the following order "Baseline UB", "Baseline", "SMT-OCFR 1", "SMT-OCFR 2", "AFOIRNet_1", "AFOIRNet_2", "AdaFace 4M", "AdaFace 12M".	162
10.1	An example of an annotated synthetic image. It is possible to observe some of the well defined attributes.	166
10.2	Comparison of the real and the synthetic datasets on individual attribute distribution. From the left to the right we have a comparison between BalancedFace and CPD-25, BalancedFace and Syn-GAN, GlobalFace and CPD-25, and GlobalFace and Syn-GAN. The first seven entries of each plot represent the most similar attribute distributions, whereas the seven bottom attributes represent the less similar distributions.	171
10.3	Distributions $Prop0_{d1}^{(a)} - Prop0_{d2}^{(a)}$, where $Prop0_{d'}^{(a')}$ represents the proportion of the Undefined prediction of MAC for attribute a' in dataset d' . The title of each plot represents the $d2$ set, and the label of each KDE plot represents the $d1$ set.	173

10.4	Cluster aggregation of the different identities on each of three datasets (Balanced-Face, GlobalFace and CPD-25). Clusters calculated with K-Means. Syn-GAN was removed from the comparison, since it consists of 500k images of distinct identities. Synthetic data has higher presence in cluster 2 ($> 60\%$) and lower in cluster 3 ($< 10\%$).	175
11.1	Representation on a potential source of biases created after quantizing a deep neural network. Less represented classes or classes prone to suffer from discrimination might be easier to forget, leading to errors focused on those classes.	180
11.2	The estimated distribution of the different ethnicity groups on the synthetic dataset used for quantization. Caucasians represent around 62% of all samples. Thus, demonstrating the unbalanced nature of these generated images.	185
12.1	Demonstration of the Sampling Effect: For each sampling intensity (using sampling strategy A) the sample with the lowest confidence of each label is shown. For instance, on No Sampling, it is possible to see an Asian image (according to the dataset label) that the ethnicity classifier model has a low confidence in the Asian class. It is also noticeable that as the sampling intensifies, the ethnicity traits become more prominent.	190
12.2	Ethnicity Shift in Demographic Groups: given the labels of the dataset, we show one sample of the identities that are the least and the most aligned with their label (introduced later in Section 12.3.1.1). It is visible the presence of noise in this label, as well as the high variability even within samples with high scores and the non-descriptive nature of discrete labels.	194
12.3	The evolution of the different scores $diag(ES_A)^{(y)}$, $diag(ES_B)^{(y)}$, and $diag(ES_C)^{(y)}$ (y-axis) as the number of identities in the dataset decreases (x-axis). It can be seen that for the first two protocols, balance is quickly achieved, whereas for the other only the last iteration is balanced.	201
12.4	Distribution shift of the ethnicity score on the different ethnicity groups when sampling identities with Protocol A. It can be seen that the Asian distribution is not affected, as no sample is removed. It would only be affected when the sampling removes up to 25% of the identities. Additionally, this shows that distributions shift from a flatter look into one that peaks near 1.	202
12.5	Utility metrics and Ethnicity Score relation. It seems that, despite difference across demographic groups, does not seem to exist a correlation between the Ethnicity Score and CR-FIQA utility score.	203
12.6	Identity Score ($IDS_A^{(y)}$) Distribution for the different demographic groups for both the original ethnicity labels and after a relabelling process.	204
12.7	Trade-off between bias and performance visually highlighted through the computation of a pareto frontier with all the dominant solutions. Different shapes on the points represent different dataset sizes, while the colour reflects the sampling strategy. Pareto frontier elements are connected through a blue line.	211

List of Tables

3.1	Methods that leverage explainability methods to inspect the predictions of biometric systems. Some of these methods, promote interpretability through the usage of certain techniques, whereas others focus mostly on explaining the predictions. Each method is placed in the respective rung in the xAI ladder. For the methods of the second rung they are further categorised into Model Prior (M. P.), Training Rule Prior (T. R. P.), Intervention on Input (I. I.) and Intervention on the Prediction (I. P.).	58
3.1	Methods that leverage explainability methods to inspect the predictions of biometric systems. Some of these methods, promote interpretability through the usage of certain techniques, whereas others focus mostly on explaining the predictions. Each method is placed in the respective rung in the xAI ladder. For the methods of the second rung they are further categorised into Model Prior (M. P.), Training Rule Prior (T. R. P.), Intervention on Input (I. I.) and Intervention on the Prediction (I. P.).	59
3.1	Methods that leverage explainability methods to inspect the predictions of biometric systems. Some of these methods, promote interpretability through the usage of certain techniques, whereas others focus mostly on explaining the predictions. Each method is placed in the respective rung in the xAI ladder. For the methods of the second rung they are further categorised into Model Prior (M. P.), Training Rule Prior (T. R. P.), Intervention on Input (I. I.) and Intervention on the Prediction (I. P.).	60
3.1	Methods that leverage explainability methods to inspect the predictions of biometric systems. Some of these methods, promote interpretability through the usage of certain techniques, whereas others focus mostly on explaining the predictions. Each method is placed in the respective rung in the xAI ladder. For the methods of the second rung they are further categorised into Model Prior (M. P.), Training Rule Prior (T. R. P.), Intervention on Input (I. I.) and Intervention on the Prediction (I. P.).	61
4.1	List of attacks present in the ROSE-YOUTU dataset [297].	80
4.2	Comparison of four different approaches with their versions with and without background. The columns represent the dataset used for both training and testing. The reported values represent the EER in %.	81
4.3	Comparison of four different approaches with their versions with and without background on the ROSE-Youtu dataset. APCER, BPCER and EER are displayed as %. In bold is the best result per column.	82

4.4	Comparison of all the seven different approaches explored in the ROSE-Youtu dataset. All of the approaches leveraged background information. In bold is the best result per column.	82
4.5	Comparison of four different approaches with their versions with and without background. Results for models trained on the ROSE-Youtu dataset and tested on the datasets of each column. The reported values represent the EER in %. . .	83
4.6	Evaluation of three approaches in the setting of one attack in the ROSE-Youtu dataset without cropping the background. In this setting, the attack in the first column is the only one used for training and testing. APCER, BPCER and EER are displayed as %. In bold is the best result per column.	84
4.7	Evaluation of three approaches in the setting of unseen attack in the ROSE-Youtu dataset without cropping the background. In this setting, the attack in the first column is excluded from the training and is the only one used for testing. APCER, BPCER and EER are displayed as %. In bold is the best result per column.	84
5.1	Results for the different types of morphing techniques included in the FRLL dataset. For morphing type, we present the performance of the model per training dataset. All the results are in percentage (%), and the best are in bold.	95
5.2	Results comparison with four models published in the literature. All the models were trained on the SMDD dataset, and evaluated on the dataset specified on the left column of the table. All the results are in percentage (%) and the best are in bold.	97
6.1	Results obtained for the synthetic masked face data (EER, FMR100, FMR10 in %).	106
6.2	Results obtained for the real masked face data (in the column “Mode”: U and M stands for unmasked and masked data, respectively; EER, FMR100, FMR10 in %).	107
6.3	Experiments on the distinct configurations of the proposed method on the test dataset (in the column “Selection”: Original stands for the use of the masked version of the unmasked image on the second run of the network, whereas Random stands for the use of a randomly picked masked image from the same subject of the unmasked image; in the column “Mode”: U and M stands for unmasked and masked data, respectively; EER, FMR100, FMR10 in %).	113
6.4	Number of parameters of each module of the proposed approach. Includes the information regarding the use of each module for inference. Training parameters refers only to trainable parameters during training.	114
6.5	Comparison of the FMR100 of the methods presented in the MFR competition [48], a baseline and the proposed approach. Methods evaluated on the competition dataset for U-M face verification. (FMR100 in %).	115
6.6	Comparison of the FMR100 of the methods presented in the MFR competition [48], a baseline and the proposed approach. Methods evaluated on the competition dataset for M-M face verification. (FMR100 in %).	115
7.1	Properties of biometric confidence estimation approaches.	120

7.2	Confidence Calibration Analysis - The ECE and MCE are shown for several dataset and FRS combinations at an FMR of 10^{-3} . ECE shows the average confidence calibration error, while the MCE presents the maximum calibration error. The best performance is marked in bold. Except for one case, the proposed PIC approach strongly outperforms the other confidence estimators in terms of interpretability. This holds for the average performance, as well as for the worst-case scenario.	126
7.3	Multi-Biometric PIC score recognition performance - The recognition performance of using a joint PIC score by combining a probe sample with 1/2/5 reference samples (RS) is shown for all database and FRS combinations. Generally, the joint PIC score leads to lower recognition errors at 10^{-3} FMR than in the single-biometric scenario.	128
7.4	Multi-Biometric Confidence Calibration Analysis - The expected calibration errors (ECE) are shown for a joint PIC score confidence by combining a probe sample with 1/2/5 reference samples (RS). This was done for all database and FRS combinations at an FMR of 10^{-3} . The ECE significantly decreases when more reference samples are used. Consequently, also the probabilistic confidence interpretation of PIC becomes more accurate when multiple samples are combined. . .	129
8.1	Face verification accuracy(%) on the AR dataset under protocols 1 and 2. “Sg” and “Sf” stand for “Sunglasses ” and “Scarf ”, respectively. The first group of methods represents general face recognition models; the second group of methods represents occluded face recognition models; the last group of methods represents masked face recognition models.	143
8.2	Face verification accuracy(%) on the LFW and its occluded versions. All the results, except for LFW, are reported for a verification with clear references and occluded probes. Besides FROM, all the other methods were evaluated by us. . .	145
8.3	Face verification TAR @ FAR = $1e - 3$ (%) on the LFW and its occluded versions. All the results, except for LFW, are reported for a verification with clear references and occluded probes. Besides FROM, all the other methods were evaluated by us. . .	146
8.4	Face verification accuracy(%) the masked versions of LFW, CALFW, CPLFW, CFP-FP, AgeDB-30. All the results are reported for a verification with unmasked references and masked probes. All the presented methods were evaluated by us. FROM evaluation was conducted on 96x112 images, since it was trained on images with those dimensions.	147
9.1	Occlusions	155
9.2	Evaluation protocols	157
9.3	Participants	157
9.4	Face verification results on three different protocols: protocol 0, protocol 1 and protocol 2 (in the BLR-OP setting). The values of EER, FMR100, FMR1000 are displayed in percentage (%). The rank orders the solutions from the best to the worst within that protocol.	160
9.5	Face verification results on three different protocols: protocol 3, protocol 4 and protocol 5. All protocols in the BLR-OP setting. The values of EER, FMR100, FMR1000 are displayed in percentage (%). The rank orders the solutions from the best to the worst within that protocol.	161

9.6	Face verification results on two different protocols: protocol 6 and protocol 7. All protocols in the BLR-OP setting. The values of EER, FMR100, FMR1000 are displayed in percentage (%). The rank orders the solutions from the best to the worst within that protocol. Includes information regarding the compactness of the models in terms of the number of learnable parameters and GFLOPs.	161
9.7	Final ranking of the solutions submitted. Solutions in bold surpassed the performance of the baseline (without access to the landmarks) in at least one protocol.	163
10.1	KL divergence between all the distribution learnt by KDE for all four datasets. Interesting to denote the lower information loss when real data is used to approximate synthetic data in comparison with the higher information loss when we swap P and Q.	175
11.1	Table comprising the results, evaluated on RFW, from the different models trained on MS1MV2 and their respective quantized versions for different bits and quantization strategies (real or synthetic data). The versions of the models quantized with synthetic data seem to display better fairness metrics at a comparable average performance.	184
11.2	Table comprising the results, evaluated on RFW, from two ResNet-34 models trained on BUPT-Balancedface (BL) and BUPT-Globalface (GL) and their respective quantized versions for different bits and quantization strategies (BL, GL or synthetic data). The versions of the models quantized with synthetic data seem to perform outstandingly well.	186
12.1	Results for an iResNet-34 (ElasticArcFace) when half of the samples from a single identity are removed. Three strategies are followed: Retaining 3500 random identities (Rand), the 3500 identities that minimise the score for that ethnicity (Min), and the 3500 identities that maximise this score (Max).	207
12.2	Results for iResNet-34 (ElasticArcFace) trained on several versions of the training set. Ranging from no sampling to three different sampling strategies, where two are based on Protocol A - A and A(R) . These latter strategies seem to hold significant advantages with respect to random sampling. Results are shown for several variations of the number of identities kept in the training set.	208
12.3	Results for iResNet-34 (ElasticArcFace) trained on several versions of the training set. Ranging from no sampling to three different sampling strategies, where two are based on Protocol B - B and B(R) . These latter strategies seem to hold significant advantages with respect to random sampling. Results shown for several variations of the number of identities kept in the training set.	209
12.4	Results for iResNet-34 (ElasticArcFace) trained on several versions of the training set. Ranging from no sampling to three different sampling strategies, where two are based on Protocol C - C and C(R) . These latter strategies seem to hold significant advantages with respect to random sampling. Results shown for several variations of the number of identities kept in the training set.	210
12.5	Results for iResNet-50 (ElasticArcFace) trained on several versions of the training set. Ranging from no sampling to four different sampling strategies, A, B, C and Random.	212
12.6	Results for iResNet-100 (ElasticArcFace) trained on several versions of the training set. Ranging from no sampling to four different sampling strategies, A, B, C and Random.	212

12.7 Results for iResNet-34 (CurricularFace) trained on several versions of the training set. Ranging from no sampling to four different sampling strategies, A, B, C and Random. These latter strategies seem to hold significant advantages with respect to random sampling. Results shown for several variations of the number of identities kept in the training set.	213
---	-----

Abbreviations

2D	Bidimensional
2.5D	Two-and-a-Half Dimensional
3D	Tridimensional
AI	Artificial Intelligence
AIMIA	AI-enabled medical image analysis
APCER	Attack Presentation Classification Error Rate
API	Application Programming Interface
APRP	Associação Portuguesa de Reconhecimento de Padrões
AUC	Area Under the Curve
BC-PAD	Binary Classification Presentation Attack Detection
BCE	Binary Cross-Entropy Loss
BFW	Balanced Faces in the Wild
BLR-OP	Baseline Reference and Occluded Probe
BIOSIG	Biometrics special interest group
BPCER	Bonafide Presentation Classification Error Rate
C3D	Convolutional 3D (networks)
CAE	Convolutional Autoencoder
CCC	Confidence Calibration Curves
CCD	Charge-Coupled Device
CE	Cross-Entropy Loss
CNN	Convolutional Neural Network
CVPR	Conference on Computer Vision and Pattern Recognition
DET	Detection Error Tradeoff
DFS	Dynamic frame selection
DL	Deep Learning
DPI	Dots Per Inch
DTC	Decision Threshold Confidence
EAB	European Association for Biometrics
ECCV	European Conference on Computer Vision
ECE	Expected Calibration Error
ECG	Electrocardiogram

EER	Equal Error Rate
ERBC	Error Rate-based Confidence
EU	European Union
EUSIPCO	European Signal Processing Conference
FAR	False Acceptance Rate
FCT	Fundação para a Ciência e Tecnologia
FMR	False Match Rate
FNIR	False Negative Identification Rate
FNMR	False Non-Match Rate
FR	Face Recognition
FRR	False Rejection Rate
FRS	Face Recognition Systems
FTCR	Failure to Capture Rate
FTER	Failure to Enroll Rate
FRL	Face Research London Lab
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GMean	Genuine Mean
GPU	Graphical Processing Unit
ICCV	International Conference on Computer Vision
IDR	Identification Rate
IEEE	Institute of Electrical and Electronics Engineers
IJCB	International Joint Conference on Biometrics
IMean	Impostors Mean
JAFFE	Japanese Female Facial Expression Dataset
KD	Knowledge Distillation
KDE	Kernel Density Estimation
KL	Kullback–Leibler Divergence
KNN	k-Nearest Neighbours
LED	Light-Emitting Diode
LFW	Labeled Faces in the Wild
LRC	Likelihood Ratio-based Confidence
LSTM	Long Short-Term Memory
M-M	Masked Reference and Masked Probe
MAC	Massive Attribute Classifier
MAD	Morphing Attack Detection
MCE	Maximum Calibration Error
MFR	Masked Face Recognition
MIDR	Missidentification Rate
MIL	Multiple Instance Learning
ML	Machine Learning
MSE	Mean Squared Error

MT-PAD	Multi-task Classification Presentation Attack Detection
MTCNN	Multi-Task Cascaded Convolutional Neural Networks
NIST	National Institute of Standards and Technology
NN	Neural Network
nm	Nanometers
OCFR	Occluded Face Recognition
OCFR2022	2022 Advanced Occluded Face Recognition Competition
OR-OP	Occluded Reference and Occluded Probe
ORL	Olivetti Research Laboratory (ORL) Face Dataset
PAD	Presentation Attack Detection
PFE	Probabilistic Face Embedding
PFES	Probabilistic Face Embedding Score
PI	Principal Investigator
PIC	Probabilistic Interpretable Comparison
RCNN	Region-based Convolutional Neural Network
RFW	Racial Faces in the Wild
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Curve
RR	Rejection Rate
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SMDD	Synthetic Morphing Attack Detection Development
SMFD	Synthetic masked face data
SOTA	State-of-the-Art
SVM	Support Vector Machine
TL	Triplet Loss
TPIR	True Positive Identification Rate
U-M	Unmasked Reference and Masked Probe
UI	User Interface
UPMC	Uncertainty-Propagation for Matching Confidence
UX	User Experience
WACV	Winter Conference on Applications of Computer Vision
XAI	Explainable Artificial Intelligence

Part I

Prologue

Chapter 1

Introduction

1.1 Context and Motivation

Artificial Intelligence (AI) has progressed, over the years, to a technology that is present in the daily lives of a large portion of the population and with a direct impact on their lives. Some applications, such as non-impacting recommendation systems or ranking algorithms, have smoother deployment life cycles due to their non-harmful nature [341]. On the other hand, other applications require the deployment of algorithms on legal use cases, medical scenarios, or other critical scenarios [115, 122, 248]. Given the black-box setting of current deep learning models, deploying them in those scenarios might pose a threat to the user, or other individuals [73].

Face recognition is a widely used form of biometric recognition, leveraging easily captured traits, such as the face to identify different individuals. These systems have outstanding accuracy, and, in several use cases, are capable of surpassing human performance [439, 472]. Their use has been widespread through different domains and applications [502]. For instance, one can see a face recognition system on a mobile device, which hold nearly zero risk to an individual. On the other hand, these can also be deployed in court with law enforcement entities leveraging face recognition to identify suspects. This is an application of considerable risk as it affects directly the life of individuals.

The existence of a black-box decision-making system extends several of these risks, but two of the most impacting are the inability to verify the decision of the model, and the presence of unwanted biases. This led to the development of two research directions: Explainability [367] and Fairness [140]. These have been highly discussed as essential to a safe deployment of machine learning models on life-impacting scenarios [418, 487].

One of the most prominent risks on Face Recognition systems is the presence of unfair behaviour [62, 521]. This, when not mitigated, can cause an increase in the error of certain demographic groups, leading to a higher rate of wrong predictions on such groups ¹ [62, 323]. Within the context of law enforcement the nefarious effects of such bias have led to the arrest of innocent

¹ Gender Shades Project

black individuals ²³. An early detection is important to mitigate these scenarios, however, deep learning models and its opaqueness are strong obstacles. At the time of writing this thesis, the most common strategy to assess any source of bias comes from the evaluation of a model on a balanced dataset [261, 503, 517]. Notwithstanding, these datasets are often not trivial to collect and a proper representation of all demographic groups is a complex task, thus leading to more challenges [481, 517].

In addition, the existence of a powerful black-box system implies that the predictions cannot be double-check by a human operator [367]. No only that, but even when the predictions are correct, the reasoning that led to them might be incorrect, hiding in each correct prediction a deeper problem within the model [97, 341]. In order to tackle this problem for high-stake decisions, the European Union introduced, through the General Data Protection Regulation (GDPR) [152], regulations on AI, such as the user's right to an explanation. This enforces the need to understand the reasoning of a deep learning model for each prediction when these affect lives.

The evolution of deep learning-based face recognition systems towards a more fair and explainable structure is necessary to ensure the wide adoption of this technology. In the future, it could provide a platform to ease several processes, ensuring a strong law enforcement deployment, secure authentication on banking infrastructures, and its application on other more trivial scenarios such as hotel check-in.

1.2 Objectives

The doctoral work conducted and described in this manuscript focused on improving face biometrics beyond raw performance through the application of pattern recognition and computer vision strategies. Explainability, privacy and fairness have been identified as some of the major drawbacks in the implementation and safe deployment of biometrics technology. For this, we individually tackle each of these aspects as follows:

- *Explainability*: Considering the recency of explainability literature, and some lack of clarity in the definition of certain concepts, we start approaching this domain with the proposal of a novel taxonomy strategy based on causality, while also providing a categorization scheme. Following this proposal, we advance with several explainability strategies. First, we study the impact of preprocessing inputs on a presentation attack detection problem with the help of visualization techniques. Afterwards, both on morphing attack detection and masked face recognition, we propose two strategies for each that leverage loss/training constraints to insert domain knowledge into the model. On the former we focus on better separability between morphed identities, whereas on the latter we formalize the irrelevance of the mask for identity prediction. Finally, we introduce a novel confidence score for face matching providing interpretable and calibrated benefits to face recognition systems.

² Case 1

³ Case 2

- *Synthetic Data:* As the collection of datasets has become more ambitious, concerns regarding the nature of the data, the ethical considerations and the absence of consent by the different subjects led to the retraction of several datasets. In this sense, synthetic data has gained traction as one of the best strategies to mitigate these privacy and ethical issues. Focusing on the current performance gap between models trained on real and synthetic data, we studied the current domain gap with regards to soft-biometric attributes. In addition, we also study the addition of synthetic occlusions and the interoperability of masked and occluded face recognition strategies. This represents an effort towards the creation of robust training/evaluation sets to feed future models.
- *Fairness:* It has been verified that an increase in the performance of a face recognition system does not distribute uniformly across demographic groups. Hence, some are neglected. To tackle this issue, we cover two topics. Following previous chapter insights, we study the bias of compressed models and the role of synthetic data on an increased robustness of such models. Afterwards, we question the considerations and assumptions used to create a balanced dataset, leading to a novel balancing strategy based on continuous labels. Our strategy led to fairer models without significant performance trade-offs. Whilst this has been developed for real data, it has been further tested on international competitions for fair face recognition using synthetic data.

1.3 Contributions

The contributions of the doctoral work presented in this thesis are mainly aligned with the previously introduced objectives. Hence, they can be sub-categorized into three groups which are highlighted in Figure 1.1. Most of these contributions resulted in the publication of state-of-the-art methods, findings regarding the interoperability of two domains or the identification of flaws in current methods. We aimed to develop general solutions when possible, and to advance science. For this, we open-sourced the majority of the contributions presented in this document, leading to some attention on GitHub and easy reproduction of our methods. Besides more than five hundred citations, these contributions have gathered more than sixty GitHub stars and ten forks. This highlights the scientific community recognition of the presented work and its ability to be reproduced effectively.

In this sense, these contributions are also described in three parts. The first part covers face biometric recognition, including presentation and morphing attack detection, and face recognition. In addition, these are framed in a context of explainable artificial intelligence, so each approach was designed with, at least, one concept of interpretability and explainability in mind. The second part introduces the concept of synthetic data, and from a broader interpretation, this can be fully generated images or digitally altered images to form occlusions. The main focus is not the generation, but the training and evaluation with this data. Finally, the third and last part aims to investigate and mitigate fairness concerns on a myriad of methods. These contributions are described below.

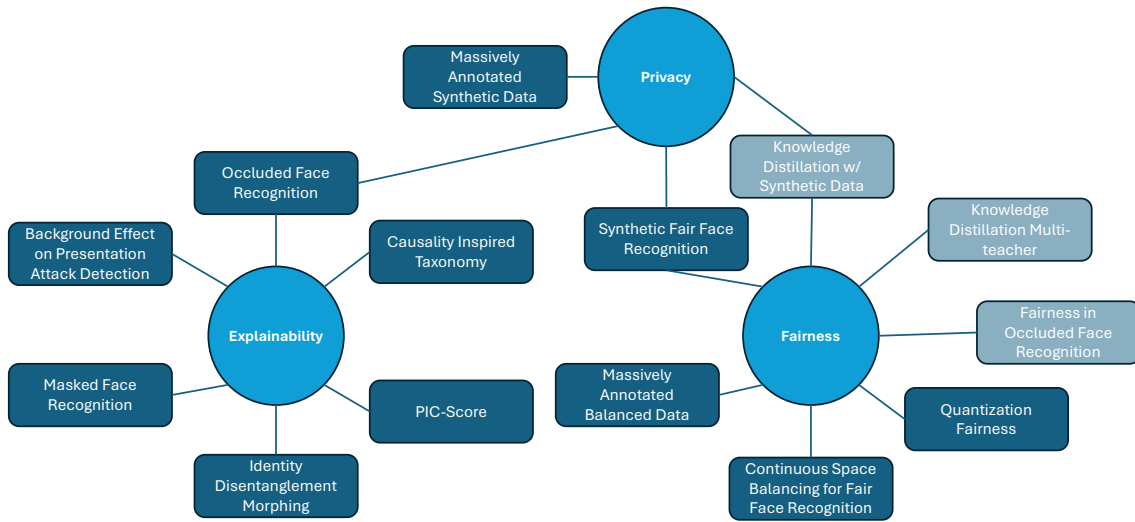


Figure 1.1: Overall summary of the contributions that resulted from author's doctoral work. Faded block are not included in the chapters of this thesis. Besides biometrics, other topics in computer vision and pattern recognition that benefited from the author's contribution were not included.

In Part II, *Explainable Biometrics*:

- a comprehensive survey of eighty one state-of-the-art methods that include explainable artificial intelligence for biometrics. This also integrates the revision and proposal of a new taxonomy inspired by causal theory (Chapter 3);
- the study of the implications of background on the performance and behaviour of presentation attack detection models. This included the inspection of relevant pixels with and without background (Chapter 4);
- two constraint-based approaches for the separation of identity information on morphed images (Chapter 5);
- a method with domain knowledge injection to guide the model into becoming invariant to masks on masked face recognition tasks (Chapter 6);
- an optimal interpretable score calibration method inspired by naive-bayes for face recognition applications that has no negative implications on the performance (Chapter 7).

In Part III, *Face Recognition with Synthetic Data*:

- a study on several masked face recognition and occluded face recognition models and their interoperability between both tasks (Chapter 8);
- a novel occluded face recognition dataset for evaluating face recognition models constructed from land-marked synthetically added occlusions (Chapter 9);

- the study of four different datasets, two synthetic and two real, to understand the synthetic-real performance gap from their soft-biometric distribution. These massive labels have been released (Chapter 10).

In Part IV, *Biases in Face Recognition*:

- the study of the fairness implications of quantizing face recognition models, and the finding that, while quantization led to an overall increase in bias, conducting the model tuning post-quantization with synthetic data mitigated the bias of a similar process with real data (Chapter 11);
- the introduction of continuous ethnicity labels and the finding that not all samples contribute equally to the balance of a dataset. Proposed a novel sampling strategy to balance existing datasets which has been further validated in synthetic face recognition competition [117] and real/synthetic data mixtures [364](Chapter 12).

These contributions can be further categorized into novel training processes/models, investigation works and novel datasets. Chapter 5, Chapter 6, Chapter 7 and Chapter 12 introduce novel methods and training strategies that help to push state-of-the-art methods. Chapter 3, Chapter 4, Chapter 8, Chapter 10 and Chapter 11 study different methods and approaches that lead to novel findings, the revision of the taxonomy or new research directions. Finally, Chapter 9, Chapter 10 and Chapter 12 introduce new datasets, novel training/testing protocols or an extension of the labels previously available. With exception of Chapter 5 and Chapter 10 which had some strong contributions of other authors, these contributions result mostly or completely from authors work.

1.4 Dissemination

The contributions of the author to fairness, explainability and synthetic-based face recognition are demonstrated in the following 17 publications (clustered by type and in reverse chronological order):

- International journal
 5. Neto, P. C., Damer, N., Cardoso, J. S. & Sequeira, A. F., (2025). Balancing Beyond Discrete Categories: Continuous Demographic Labels for Fair Face Recognition. Submitted to Transactions on Information Forensics and Security
 4. DeAndres-Tame, I., Tolosana, R., Melzi, P., Vera-Rodriguez, R., Kim, M., Rathgeb, C., ... Menotti, D. (2025). Second frcsyn-ongoing: Winning solutions and post-challenge analysis to improve face recognition with synthetic data. Information Fusion, 120, 103099.
 3. Neto, P. C., Caldeira, E., Gonçalves, T., Damer, N., Sequeira, A. F., & Cardoso, J. S. (2024). Disentangling morphed identities for face morphing detection. *Science Talks*, 10, 100331.

2. Neto, P. C., Gonçalves, T., Pinto, J. R., Silva, W., Sequeira, A. F., Ross, A., & Cardoso, J. S. (2022). Causality-Inspired Taxonomy for Explainable Artificial Intelligence. *arXiv preprint arXiv:2208.09500*. (Second round of Revision at ACM Computing Surveys)
 1. Neto, P. C., Pinto, J. R., Boutros, F., Damer, N., Sequeira, A. F., & Cardoso, J. S. (2022). Beyond masks: On the generalization of masked face recognition models to occluded face recognition. *IEEE Access*, 10, 86222-86233.
- Articles in international conference proceedings
12. Neto, P. C., Mamede, R. M., Albuquerque, C., Gonçalves, T., & Sequeira, A. F. (2024). Massively Annotated Datasets for Assessment of Synthetic and Real Data in Face Recognition. In *18th IEEE International Conference on Automatic Face and Gesture Recognition 2024*
 11. DeAndres-Tame, I., Tolosana, R., Melzi, P., Vera-Rodriguez, R., Kim, M., Rathgeb, C., ... & Menotti, D. (2024). FRCSyn Challenge at CVPR 2024: Face Recognition Challenge in the Era of Synthetic Data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
 10. Neto, P. C., Caldeira, E., Cardoso, J. S., & Sequeira, A. F. (2023, September). Compressed models decompress race biases: What quantized models forget for fair face recognition. In *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)* (pp. 1-5). IEEE.
 9. Neto, P. C., Caldeira, E., Gonçalves, T., Damer, N., Sequeira, A. F., & Cardoso, J. S. (2023, September). Unveiling the two-faced truth: Disentangling morphed identities for face morphing detection. In *2023 31st European Signal Processing Conference (EUSIPCO)* (pp. 955-959). IEEE.
 8. Neto, P. C., Sequeira, A. F., Cardoso, J. S., & Terhörst, P. (2023). PIC-Score: Probabilistic Interpretable Comparison Score for Optimal Matching Confidence in Single- and Multi-Biometric Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1021-1029).
 7. Huber, M., Boutros, F., Luu, A. T., Raja, K., Ramachandra, R., Damer, N., ... & Štruc, V. (2022, October). SYN-MAD 2022: Competition on face morphing attack detection based on privacy-aware synthetic training data. In *2022 IEEE International Joint Conference on Biometrics (IJCB)* (pp. 1-10). IEEE.
 6. Neto, P. C., Boutros, F., Pinto, J. R., Damer, N., Sequeira, A. F., Cardoso, J. S., ... & Menotti, D. (2022, October). OCFR 2022: Competition on occluded face recognition from synthetically generated structure-aware occlusions. In *2022 IEEE International Joint Conference on Biometrics (IJCB)* (pp. 1-9). IEEE.
 5. Neto, P. C., Gonçalves, T., Huber, M., Damer, N., Sequeira, A. F., & Cardoso, J. S. (2022, September). Orthomad: Morphing attack detection through orthogonal identity

disentanglement. *In 2022 International Conference of the Biometrics Special Interest Group (BIOSIG) (pp. 1-5). IEEE.*

4. Neto, P. C., Sequeira, A. F., & Cardoso, J. S. (2022). Myope models-are face presentation attack detection models short-sighted?. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 390-399).*
 3. Neto, P. C., Boutros, F., Pinto, J. R., Damer, N., Sequeira, A. F., & Cardoso, J. S. (2021, December). Focusface: Multi-task contrastive learning for masked face recognition. *In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) (pp. 01-08). IEEE.*
 2. Neto, P. C., Boutros, F., Pinto, J. R., Saffari, M., Damer, N., Sequeira, A. F., & Cardoso, J. S. (2021, September). My eyes are up here: Promoting focus on uncovered regions in masked face recognition. *In 2021 International Conference of the Biometrics Special Interest Group (BIOSIG) (pp. 1-5). IEEE.*
 1. Boutros, F., Damer, N., Kolf, J. N., Raja, K., Kirchbuchner, F., Ramachandra, R., ... Neto, P. C., Sequeira, A. F., Pinto, J. R., Saffari, M., & Cardoso, J. S. (2021, August). MFR 2021: Masked face recognition competition. *In 2021 IEEE International joint conference on biometrics (IJCB) (pp. 1-10). IEEE.*
- Abstracts in national conference proceedings
 4. Neto, P. C., Caldeira, E., Cardoso, J.S., Sequeira, A.F. (2023). Compressed Models Decompress Race Biases on Face Recognition. *In 29th Portuguese Conference on Pattern Recognition (RECPAD 2023), Portugal.*
 3. Sequeira, A.F., Neto, P. C., Cardoso, J. S. (2022). On the generalization of masked face recognition models to occluded face recognition. *In 28th Portuguese Conference on Pattern Recognition (RECPAD 2022), Portugal.*
 2. Mamede, R., Oliveira, M., Neto, P. C., Sequeira, A.F., Cardoso, J.S. (2022). Latent Noise Augmentation: On the applicability of noise as an augmentation in the latent space. *In 28th Portuguese Conference on Pattern Recognition (RECPAD 2022), Portugal.*
 1. Neto, P. C., Sequeira, A.F., Cardoso, J.S. (2021). caPAD - A context aware model for face presentation attack detection. *In 27th Portuguese Conference on Pattern Recognition (RECPAD 2021), Portugal.*

Besides the work conducted and presented in this manuscript, the author further published the following documents related to computer vision, pattern recognition for both biometrics and digital pathology domains:

- International journal

6. Neto, P. C., Montezuma, D., Oliveira, S. P., Oliveira, D., Fraga, J., Monteiro, A., ... & Cardoso, J. S. (2024). An interpretable machine learning system for colorectal cancer diagnosis from pathology slides. *NPJ precision oncology*, 8(1), 56
 5. Caldeira, E., Neto, P. C., Huber, M., Damer, N., & Sequeira, A. F. (2025). Model compression techniques in biometrics applications: A survey. *Information Fusion*, 114, 102657.
 4. Oliveira, S. P., Montezuma, D., Moreira, A., Oliveira, D., Neto, P. C., Monteiro, A., ... & Cardoso, J. S. (2023). A CAD system for automatic dysplasia grading on H&E cervical whole-slide images. *Scientific Reports*, 13(1), 3970.
 3. Montezuma, D., Oliveira, S. P., Neto, P. C., Oliveira, D., Monteiro, A., Cardoso, J. S., & Macedo-Pinto, I. (2023). Annotating for artificial intelligence applications in digital pathology: a practical guide for pathologists and researchers. *Modern Pathology*, 36(4), 100086.
 2. Neto, P. C., Oliveira, S. P., Montezuma, D., Fraga, J., Monteiro, A., Ribeiro, L., ... & Cardoso, J. S. (2022). iMIL4PATH: A semi-supervised interpretable approach for colorectal whole-slide images. *Cancers*, 14(10), 2489.
 1. Neto, P. C., Oliveira, S. P., Fraga, J., Montezuma, D., Monteiro, A., Monteiro, J., ... & Cardoso, J. S. (2021). CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance. *Scientific Reports*, 11(1), 14358.
- Articles in international conference proceedings
 6. Huber, M., Neto, P. C., Sequeira, A. F., Damer, N. (2025). FX-MAD: Frequency-Domain Explainability and Explainability-Driven Unsupervised Detection of Face Morphing Attacks. In *Proceedings of the Winter Conference on Applications of Computer Vision* (pp. 766-776).
 5. Neto, P. C., Colakovic, I., Karakatič, S., & Sequeira, A. F. (2024). How Knowledge Distillation Mitigates the Synthetic Gap in Fair Face Recognition. In *European Conference on Computer Vision Workshops (ECCVW) 2024*
 4. Caldeira, E., Cardoso, J. S., Sequeira, A. F., & Neto, P. C. (2024). MST-KD: Multiple Specialized Teachers Knowledge Distillation for Fair Face Recognition. In *European Conference on Computer Vision Workshops (ECCVW) 2024*
 3. Mamede, R., Neto, P. C., & Sequeira, A. F. (2024). Fairness Under Cover: Evaluating the Impact of Occlusions on Demographic Bias in Facial Recognition. In *European Conference on Computer Vision Workshops (ECCVW) 2024*
 2. Montenegro, H., Neto, P., Patrício, C., Rio-Torto, I., Gonçalves, T., & Teixeira, L. F. (2023). Evaluating privacy on synthetic images generated using gans: Contributions of the vcmi team to imageclefmedical gans 2023. *Challenge*, 8

1. Albuquerque, T., Moreira, A., Barros, B., Montezuma, D., Oliveira, S. P., Neto, P. C., Monteiro, J., Ribeiro, L., Goncalves, S., Monteiro, A., Pinto, I. M., & Cardoso, J. S. (2022, July). Quality control in digital pathology: automatic fragment detection and counting. *In 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 588-593). IEEE.*
- Abstracts in international congresses
 4. Neto, P. C., Oliveira, S.P., Montezuma, D., Oliveira, D., Ribeiro, L., Monteiro, A., Pinto, I.M., Cardoso, J. (2023). To err or to say “I don’t know”? A study on the usage of efficient mixed supervision with a rejection option to diagnose colorectal lesions on WSI. *European Congress on Digital Pathology (ECDP 2023), Hungary.*
 3. Albuquerque, T., Moreira, A., Barros, B., Montezuma, D., Oliveira, S. P., Neto, P. C., Monteiro, J., Ribeiro, L., Goncalves, S., Monteiro, A., Pinto, I. M., & Cardoso, J. S. (2022). Quality checkpoint in pathology specimens handling: an AI system to automate fragment detection and count. *European Congress on Digital Pathology (ECDP 2022), Hungary.*
 2. Oliveira, S. P., Neto, P. C., Montezuma, D., Cardoso, J. S. , Pinto, I. M. (2022). Computer-aided tool for CRC diagnosis: from the AI model to the clinical software prototype. *European Congress on Digital Pathology (ECDP 2022), Hungary.*
 1. Neto, P. C., Oliveira, S. P., Fraga, J., Montezuma, D., Cardoso, J. S. , Pinto, I. M. (2021). Computer-aided tool for CRC diagnosis: from the AI model to the clinical software prototype. *European Congress on Digital Pathology (ECDP 2021), Hungary.*
 - Abstracts in national conference proceedings
 2. Montenegro, H., Neto, P. C., Patricio, C., Rio-Torto, I., Gonçalves, T., Teixeira, L.F. (2023). Evaluating Privacy on Synthetic Images Obtained using Deep Generative Models. *In 29th Portuguese Conference on Pattern Recognition (RECPAD 2023), Portugal*
 1. Oliveira, S.P., Neto, P. C., Montezuma, D., Ribeiro, L., Monteiro, A., Pinto, I.M., Cardoso, J. (2021). A semi-supervised approach for colorectal cancer diagnosis from H&E whole slide images. *In 27th Portuguese Conference on Pattern Recognition (RECPAD 2021), Portugal.*

1.5 Collaborations

1.5.1 Research Projects

During his PhD the author participated and contributed to several research projects. Some of the work in this thesis have been developed in line with the interests of these projects, namely:

- CADPath: Computer-Aided Diagnosis in Pathology (active member) [Portugal 2020 Research and Technological Development Project]
- NewSpacePortugal: Agenda New Space Portugal (active member) [Recovery and Resilience Plan Project]
- XAIBIO: eXplainable Artificial Intelligence for BIOmetrics (wrote the proposal, PI of the project) [FCT Project]

1.5.2 Synergies within the research group

The author established several collaborations during his PhD period. First, in 2021 the author collaborated with João Ribeiro Pinto and Mohsen Saffari on a submission to the masked face recognition (MFR) competition at IJCB 2021 [48]. This participation led to a collaboration with the organizers, Naser Damer and Fadi Boutros at Fraunhofer IGD (Darmstadt, Germany), on the development of two state-of-the-art solutions for MFR during the pandemic period [360, 362]. Later this has been extended into a investigation work on the interoperability of the newly developed MFR models when evaluated against general occlusions [371]. Instigated by these findings, together with Naser Damer, Fadi Boutros and João Ribeiro Pinto, the author hosted a competition at IJCB 2022 on occluded face recognition [361].

Still in 2022 the author collaborated with Tiago Gonçalves on a submission on a synthetic morphing attack detection competition at IJCB 2022 [222]. This later led to the development of a novel morphing attack detection method submitted at BIOSIG 2022 [366] with Marco Huber, and an improved version submitted together with Tiago Gonçalves and Eduarda Caldeira to EUSIPCO 2023 [67]. In 2023, the author collaborated with Tiago Gonçalves, Luis Filipe Teixeira, Helena Montenegro, Cristiano Patricio and Isabel Rio-Torto on a submission for a competition on the detection of gan-generated content [344]. In that year, Philipp Terhörst collaborated with the author on the development of an interpretable and calibrated score for face recognition systems [372].

The collaboration with Eduarda Caldeira resulted in a investigation on the biases of compressed face recognition models, which won the best paper award at BIOSIG 2023 [363]. It further led to a survey on model compression in biometrics [68] and a novel knowledge distillation method which focus on creating an unbiased student from multiple biased teachers [66]. Since 2021, the author have been collaborating with Arun Ross, João Ribeiro Pinto, Wilson Silva and Tiago Gonçalves on the development of a new taxonomy for explainable artificial intelligence based on a causality perspective [367].

In 2024, the author led in collaboration with Marco Huber and Jan Niklas Kolf a submission to a competition on fair face recognition using synthetic datasets. The submitted solution was one of highest ranked solutions leading to a participation at CVPR 2024 [117]. Together with the organisers, the author has a paper at Information Fusion. This further led to a study on the domain gap between synthetic and real datasets, which was conducted in collaboration with Rafael M. Mamede, Carolina Albuquerque and Tiago Gonçalves [368].

Recently, the author collaborated on two fair face recognition works with Ivona Colakovic and Rafael M. Mamede. The former resulted in a study of knowledge distillation as a strategy to mitigate the synthetic gap [364], whereas the latter studies the effects of occlusions on bias while devising a new metric to assess the fairness of face recognition models in occluded scenarios [323]. Still on the topic of fairness, together with Naser Damer the author proposed, through a novel representation of ethnic features, a new strategy for balancing face recognition datasets [365].

Besides his works on biometrics, the author has collaborated, since 2020, with Diana Montezuma Felizardo, Sara P. Oliveira, Tomé Albuquerque, Domingos Oliveira and João Fraga on the development of computer-aided diagnosis systems for colorectal cancer. Leading to the development of novel multiple-instance learning, the release of three datasets and, annotation guidelines for pathologists and researchers [348, 369, 370, 383]. This was further supported by a fragment counting method [14]. Finally, these insights have been expanded to cervical cancer with novel methods and the release of a dataset [382].

1.5.3 Organisation of scientific events

During his doctoral studies, the author partaken in the organisation of several scientific events. These have allowed the author to establish contact with relevant researchers in the field. The events that the author took part as an organising member are:

- Publicity Chair for CDPATH at ICCV (2021)
- Publicity/Sponsorship Chair for AIMIA at ECCV (2022)
- Project Committee Member for VISUM Summer School (2022)
- Organiser of the OCFR2022 competition at IJCB (2022)
- Organising Committee Member/Sponsorship Chair for xAI4Biometrics at WACV (2022)
- Organising Committee Member/Publicity Chair for xAI4Biometrics at WACV (2023)
- Organiser of the Special Session Explainability and Interpretability in Biometric and Human-centric Information Processing at EUSIPCO (2023)
- Program Committee Member for xAI4Biometrics at ECCV (2024)

1.5.4 Teaching

Teaching is a relevant skill to acquire while pursuing a doctorate. Hence, the author had the opportunity to teach the following courses:

- 2024/2025
 - Lectures on Advanced Topics on Machine Learning to Master students (FEUP);
 - Lectures on Machine Learning to Master students (FEUP);

- Practical Classes on Programming to Bachelor students;
- 2023/2024
 - Machine Learning practical classes to Master and PhD students (FEUP);
 - Lectures on Advanced Topics on Machine Learning to Master students (FEUP);
 - Lectures on the Computer Vision Module promoted by Bank BPI to its trainees (Porto Bussiness School);
 - Algorithm Design practical classes to BSc students (FCUP/FEUP);
- 2022/2023
 - Machine Learning practical classes to Master and PhD students (FEUP);
 - Analytics in Informatics to BSc students (ISEP);

1.5.5 Supervision of dissertations and internships

The author further collaborated as co-supervisor of the following master dissertations related to his doctoral topic (in reverse chronological order):

5. Mafalda Oliveira, "Bias detection and mitigation for Biometrics Explainable Artificial Intelligence methods applied to Face Recognition", Master in Bioengineering, Universidade do Porto, 2024 - as co-supervisor, alongside Doctor Ana Filipa Sequeira (Supervisor).
4. Carolina Albuquerque, "An integrated and user-friendly platform for the deployment of explainable Artificial Intelligence methods applied to Face Recognition", Master in Bioengineering, Universidade do Porto, 2024 - as co-supervisor, alongside Doctor Ana Filipa Sequeira (Supervisor).
3. Eduarda Caldeira, "Adapting Biased Teachers for Fair Knowledge Distillation in Face Recognition", Master in Bioengineering, Universidade do Porto, 2024 - as co-supervisor, alongside Doctor Ana Filipa Sequeira (Supervisor). Winner of the APRP Best Thesis Award.
2. Ana Cardoso, "Explainable Artificial Intelligence - Getting insights from Deep Neural Networks for Interpretable and Fair Face Recognition", Master in Bioengineering, Universidade do Porto, 2023 - as co-supervisor, alongside Doctor Ana Filipa Sequeira (Supervisor).
1. Fábio Cunha, "Don't look away! Keeping the human in the loop with a interactive active learning platform", Master in Electrical Engineering and Computers, Universidade do Porto, 2023 - as co-supervisor, alongside Doctor Ana Filipa Sequeira (Supervisor).

In addition to these Master thesis, the author supervised dozens of summer, curricular and extra-curricular internships on computer vision topics.

1.6 Awards and Distinctions

The peer-reviewed publications previously mentioned, and the overall work of the author in the field of biometrics led to its recognition through awards and distinctions. These are listed below:

- The author of this thesis was granted the *EAB Max Snijder at the European Biometrics Awards 2024* organised by the *European Association for Biometrics (EAB)*. This award reflects the author contributions to the field of explainable AI and fairness in biometrics.
- IEEE Travel Grant for the IEEE Biometrics Summer School (2022).
- Best Paper Award for the work "Compressed models decompress race biases: What quantized models forget for fair face recognition" at BIOSIG (2023).

1.7 Document Structure

The following document is divided in five Parts and thirteen Chapters. Part I comprises this introductory Chapter and Chapter 2, which establishes some necessary concepts for a reader less experienced in biometrics.

Part II contains five chapters, all focusing on explainable artificial intelligence. While Chapter 3 introduces a novel taxonomy for explainable AI, the remaining Chapters 4, 5, 6 and 7 propose novel methods for increasing the explainability or interpretability of different face biometric systems. While the first two tackle presentation and morphing attack detection problems, the latter two focus on face recognition.

Part III includes three Chapters with a focus on synthetic data for face recognition. Chapters 8 and 9 discuss the use of synthetic data for masked and occluded face recognition with a focus on a clever use of the synthetic editions applied to real data. Chapter 10 discussed the gap between real and synthetic datasets through the analysis of a massive attribute classifier.

Part IV approaches two aspects of fairness in face recognition. Chapter 11 discusses the structural bias of smaller models, which is enhanced by the quantization process. Chapter 12 introduces a novel hypothesis on the true meaning of balancing face recognition data for different demographic, thus tackling the inherent bias of a dataset.

The document concludes in Part V with the conclusion chapter (Chapter 13)

Chapter 2

Fundamental Concepts

2.1 Introduction

Biometric systems are a niche of a broader set of pattern recognition use cases. These systems include several differences from the other pattern recognition systems. For instance, since the implementations of these systems are frequently used in critical scenarios, such as border control [275, 293], the impact of an erroneous prediction due to a mismatch or a statistical drift might cause significant harm. Furthermore, biometric systems require the storage of the personal information of the user for two different situations. First, to train and develop the model. Secondly, some methods need comparison data to evaluate and give the final prediction for inference. Hence, it is necessary to have guidelines to manage this data. This chapter presents the fundamentals to develop these systems and deploy them into production. In addition, it serves as a setup for the remaining topics discussed in this thesis regarding biometrics.

2.2 Biometric Systems

2.2.1 Overall Structure

2.2.1.1 Recognition Systems

Biometric recognition systems use pattern recognition-based algorithms and specialized hardware to assess the identity of someone. For that purpose, these systems leverage biometric traits information gathered at the moment of the identification and information previously stored during the enrollment phase. If the person did not go through the enrollment process, the gathered data should not match with stored biometric data. These processes are convenient in use cases that rely on user authentication. Alternative methods rely on cards, keys or other objects that can be stolen, cloned or misused. Since biometric traits are much harder to forge, access to private facilities, confidential or personal information, relies on them [131, 209].

Previously, researchers defined the general structure of a biometric recognition system [40, 238, 406]. To further complement this structure, Pinto *et al.* designed an informative diagram

regarding what he refers to be the five modules that compose a biometric system [423]. Accordingly to those researchers, acquisition, algorithm and storage are the main components of the system. It is also possible to discriminate the algorithm into quality assessment, feature extraction and decision modules. Despite including the above-mentioned modules, Figure 2.1 also includes the user as the owner and source of the biometric data to be acquired as described below [40, 238, 406, 423]:

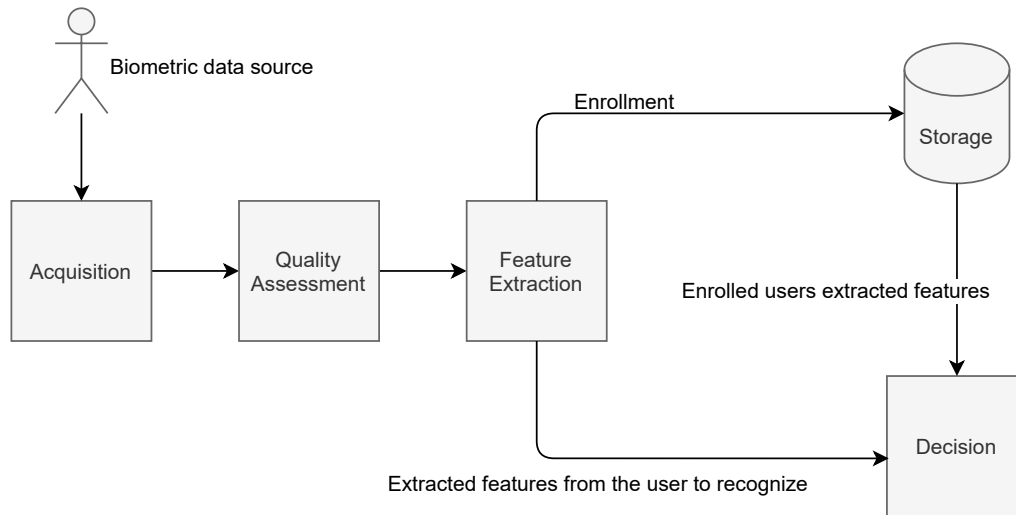
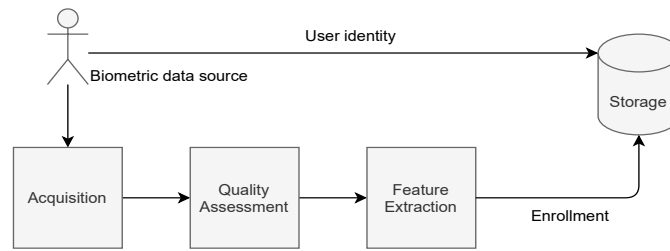
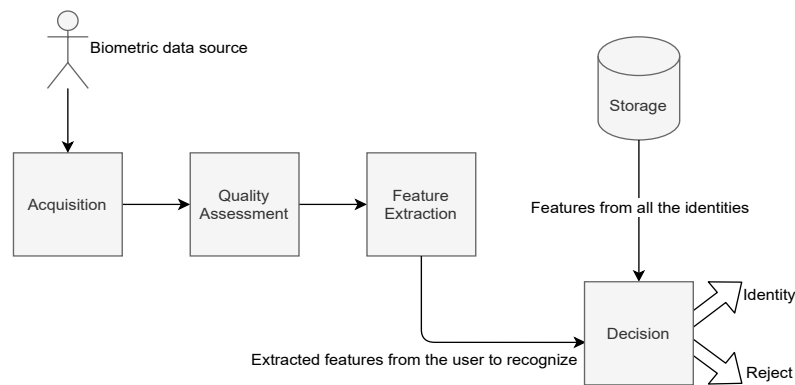


Figure 2.1: Overall structure of a biometric recognition system. The system is divided into the Acquisition, Quality Assessment, Feature Extraction, Storage and Decision modules. Structure inspired by [40, 238, 406, 423]

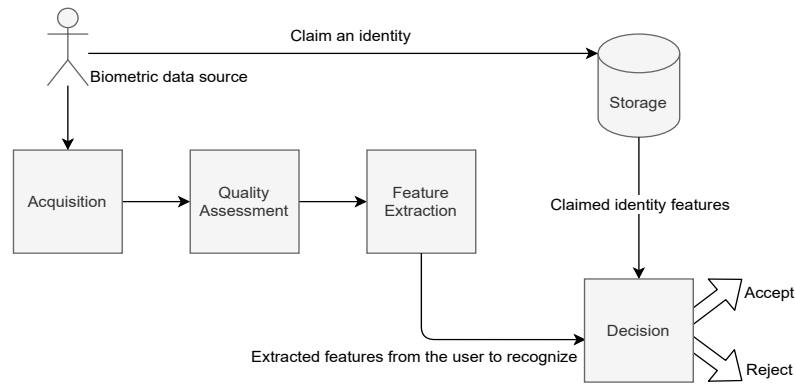
- *User*: The user is the central piece of every biometric system. He is the owner and the source of all the biometric data used to design the system. Users that do not belong to the system, such as not enrolled people, also provide their information, so that their identity can be inferred. Currently, due to the publication of the General Data Protection Regulation (GDPR) [152], the usage of personal data is more strict within the European Union. Therefore, the user and their consent with regards to his biometric traits information are, more than ever, crucial to the development and deployment of biometric systems.
- *Acquisition*: The acquisition module is the bridge between the user data and the algorithm. The process of collecting, reading and converting the data to a digital format is done in this module. It is important that it fits the system needs and that the process avoids excessive noise or other nuisances such as artefacts. This module might require, for some systems, specialized hardware to read and digitize the biometric information.
- *Quality assessment*: Despite the best efforts in the acquisition, it is frequent that the collected data contains noise. And thus, the quality assessment module aims to remove this noise and other artefacts from the collected samples. However, this is not always possible. Hence, some of the samples might be discarded.



(a) Enrolment



(b) Identification



(c) Authentication

Figure 2.2: Discrimination of the different tasks and operations that can be conducted by a recognition system. The enrolment (a), the identification (b) and the authentication (c). The decision module is not required by all of the operations, and the interaction between the user and the storage also differs.

- *Feature extraction*: The direct use of trait measurements to infer the identity of the user is not frequent and easy. Instead, pattern recognition methods, in the feature extraction module, process and map them to a new latent space. This latent space represents the extraction of the relevant features. These projections aim to have a high variance for a different subject and a low variance for the same. In other words, different identities should be distant, whereas

samples of the identity should be relatively close.

- *Storage*: The storage is composed of a database that includes the enrolled users' information and stores the data outputted by the feature extraction module. This data can be used in the future, for instance, for template matching. However, storing personal information is a complex topic that might involve several security measures.
- *Decision*: The decision module is responsible to return the result of the matching process. It receives both the features extracted from the feature extraction module, and it has access to the stored data from the enrolled users. And thus, this model leverages these two different input sources to identify the user.

A biometric recognition system comprises three different operations that leverage this general structure. They are known as enrollment, authentication and identification. Enrollment is the task of storing the identity of a user in the template database. Authentication accepts or rejects a user based on their identity claim. Finally, identification does not receive any identity claim. Instead, it tries to match the user's biometric features to all the stored templates. These operations are described in Figure 2.2.

2.2.1.2 Presentation Attack Detection Systems

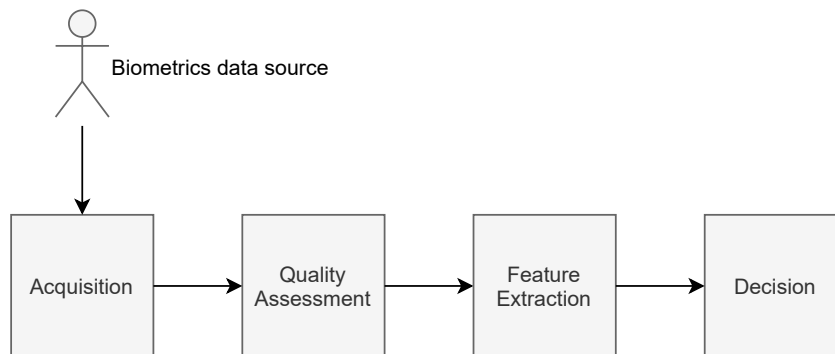


Figure 2.3: General structure of a presentation attack detection system. Differently from recognition systems, these do not include a storage module, and thus, the decision is only based on the input.

Liveness detection, more recently known as, presentation attack detection, aims to tackle and detect spoofing attempts [129]. Spoofing is the ability to trick a biometric system into authenticating/identifying a fake user as genuine. It generally requires an artefact, in other words, a forged version of the original biometric trait, that can be presented to the sensor. These attacks are frequently known as direct attacks [174], which is a subtype of the broader presentation attacks type.

Presentation attacks include other subtypes, such as the presentation of dead body parts, living traits acquired under coercion or the exploration of the performance of the overall system, for instance, zero-effort impostors [250].

Similarly to recognition, several traits can be used for this task. Despite the similarities, liveness detection does not require the storage of personal data. Thus, the overall structure of these systems is similar to recognition, with the particularity that there is no storage module (Fig 2.3). However, some works explore the potential of increasing the information given to the model, such as other personal details [91].

Due to the nature of both, the recognition and the liveness detection tasks can be executed together [306]. Moreover, it is safer when the authentication and the identification operations leverage both tasks [283, 497].

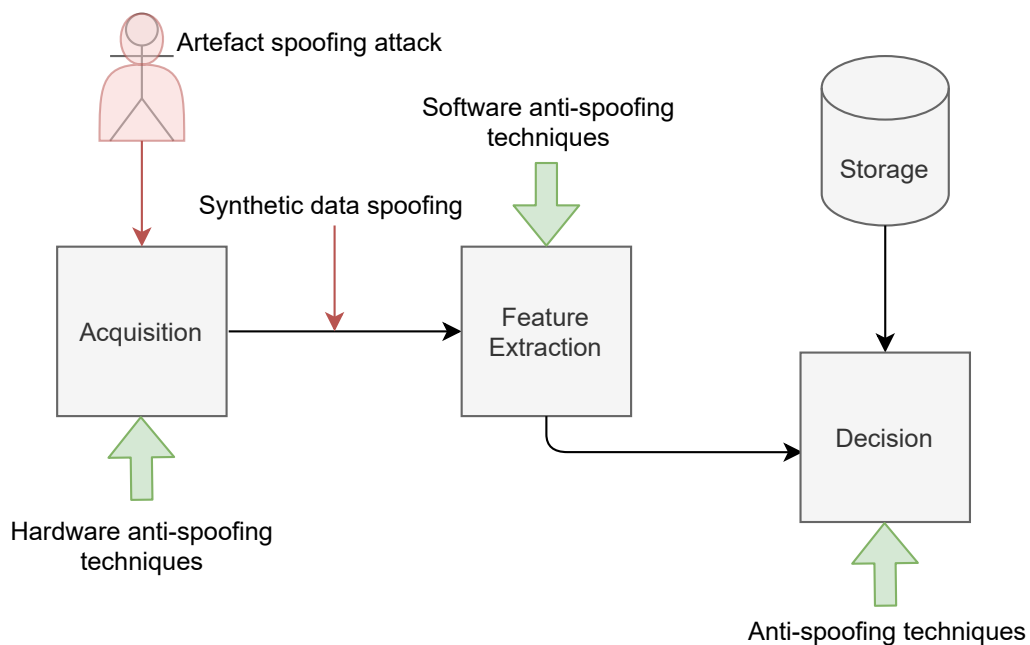


Figure 2.4: Scheme representing the three different approaches to integrate presentation attack detection methods (green arrows) into recognition system. Attacks can also be performed at different stages of the recognition system (red arrows). Quality assessment module was omitted for simplicity. Diagram inspired on [172].

Attackers have creative means to design and deploy their attacks. They might direct their malicious intentions towards potential fragilities in the sensor or directly into the algorithm. Thus, it is necessary to have defences and anti-spoofing methods at several modules. At the sensor level, which means during the acquisition module, the techniques have to be implemented directly into the hardware. However, during the feature extraction or the decision, the implementation of anti-spoofing techniques is software-based. These techniques, also seen in Figure 2.4, are further detailed below:

- *Sensor-level:* As mentioned before, these methods are applied to the hardware. In other

words, other devices are added to the naive sensor to capture additional information regarding a living trait. The captured information varies with the task and the sensor, and it can include blood pressure, fingerprint sweat or facial heat captures. While the work presented in this document does not focus on hardware-based methods, it is appropriate to be aware that they exist [130, 282] and that their information can be further leveraged on other modules [249, 356].

- *Feature-level:* After the acquisition step, and before providing the sample as input to the recognition's feature extractor, an auxiliary feature extractor can help to detect attacks. This second extractor converts the collected information into a set of features used to infer if the input is a spoofing attack or a live human. This method can be further divided into static and dynamic techniques. The former uses a single instance of the biometric trait, whereas the latter leverages a sequence of samples captured [96]. Most of the times, despite the benefits in performance, dynamic techniques require cooperation and interaction with the user. And thus, it becomes far more intrusive and less practical. It is also possible to have feature-level methods working with multi-modal inputs, for instance, accessing several traits for the recognition from a single sample. Moreover, these methods also protect against attacks based on synthetic data, which are employed after the acquisition module [72, 173].
- *Decision-level:* At decision time, liveness detection methods implemented at the feature and sensor levels have likely detected the vast majority of attacks. Hence, at the decision level, new methods are used as supplementary techniques to the previous. These methods combine scores from biometric modules, liveness detection modules, or the combination of both.

2.2.2 Market Regulation and Security Concerns

Market needs, costumer beliefs and environmental factors dictate the development of certain research areas. Propelled by consumers, the industry establishes which technologies shall be investigated and implemented in real situations. Due to potential security concerns or lack of trustworthiness, biometric systems are heavily impacted by this regulation. Despite the wide use of these technologies on smartphones and laptops, 79% of the users still prefer to rely on password-based protection on online payments [426]. The main reasons for this preference rely on lack of trust (45%) and education (35%) regarding biometric systems [426]. Despite security and usability advantages brought by these systems, some concerns have been directed towards other elements of the overall architecture. For instance, while a password can be easily replaced in the case of an attack on a database containing that information, biometric traits cannot. And thus, disclosure of biometric data presents itself as a substantial threat. Moreover, worries regarding the misuse of this technology by law enforcement entities have been rising lately [177].

In an attempt to mitigate these concerns regarding the security and trustworthiness of biometric systems, several international organizations and countries have defined some degree of regulation to biometrics personal data, such as the European Union (EU), with their General Data Protection

Regulation (GDPR) [152] limiting the processing of biometric data. Moreover, as other more specific regulations are being prepared [151], researchers have also started to address these concerns and to tackle the misuse of this data [234, 257]. Besides European Union, in the United States of America (U.S.A), at least three different states (Illinois ¹, Texas ² and Washington ³) have laws heavily regulating biometric systems. The market aims at increasing user's trust by diminishing the potential misuses of this technology.

The first step for a successful implementation of biometric systems is to explicitly ask for consent. This is key to foster trust on the user side. To do so, it is also important to carefully detail what are biometric systems and their purpose. These systems can only aim to a high degree of confidence after lecturing and educating consumers. However, regardless of all the setbacks and difficulties, the biometric systems market is expected to keep growing during the following years, as well as, the investment in this technology [305, 330, 381].

2.2.2.1 Bias Concerns

Some of human's behaviours and beliefs have hidden and unintentional biases against a demographic group that can be shown as highly harmful [212, 422]. These biases and their mitigation have been a current research area in the fields of sociology and psychology. However, lately it has been flagged by several researchers that artificial intelligence algorithms denote biases regarding minorities. Not only the collected dataset might reflect society biases, but also some loss functions might promote this behaviour. Due to the lack of fairness constraints, the learning process of machine learning algorithms promotes an increase in accuracy, even if it is at the expense of fairness. In other words, it is fine to sacrifice the performance on a minority of the dataset, if the overall performance increases, thus enlarging the gap between the performance on that minority data and on the most representative demographic group data.

Researchers and engineers have attempted to launch several models into production. However, they have been proven to contain crucial flaws regarding fairness [140, 465, 280]. Recent studies have delved into the effects of these biases, such as gender classification [62, 352, 376], face recognition tasks [4, 13, 98, 176, 194, 274, 285, 316] and the combination of both [413]. Moreover, besides detecting biases, some researchers have also developed methods and datasets that aim at mitigating these biases [180, 186, 532].

It is of utter importance to study, detect and mitigate these biases. They are threats to the trustworthiness of algorithms and a risk to the affected minorities. Lai *et al.* [290] studied the relation between risk, trust and bias. Some effects of the demographic biases have also been shown in the real world, leading to false accusations and innocent people arrests [1, 3]. These biases will be further discussed in subsequent chapters, also with novel mitigation strategies.

¹<https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57>

²<https://codes.findlaw.com/tx/business-and-commerce-code/bus-com-sect-503-001.html>

³<http://lawfilesext.leg.wa.gov/biennium/2017-18/Pdf/Bills/Session%20Laws/House/1493-S.SL.pdf>

2.3 Biometric Traits

2.3.1 General View

Biometric traits are characteristics of animals that can be used to identify, recognize and discriminate individuals within a species [10, 265]. Despite the fact that biometric traits are not exclusive to humans, most of the research focus is directed to the ones that apply to them. Biometric recognition follows the belief that the trait of an individual is closer to itself over time than to the same trait of other individuals [392]. These traits are also categorized with regards to the identity information included. For instance, if it contains enough information to provide the identity of the trait owner, it is called a hard trait. In contrast, soft traits require other traits or information to be able to do recognition [233].

Not only biometric systems, but also the biometric traits that are used in these systems have evolved since the initial studies [237]. Currently, electrocardiogram (ECG), face, fingerprint, palm print, iris, voice, gait and keystroke information are examples of the most used traits. Biometric traits fall into three different natures. The nature of a biometric trait is rather important since it might also influence the type of attacks on the system. Physiological traits, such as the ECG, capture the response of the body to the external environment and internal events. Anatomical traits, such as the face, fingerprint and palm print, register the information contained in specific body parts that are unique to the individual. Finally, behavioural traits, such as the gait and keystroke dynamics, capture the person's actions and unique behavioural patterns [5, 321].

In order to evaluate the quality of a trait, Jain *et al.* [232] proposed seven different criteria:

1. *Universality*: the trait should be present in all of the individuals using the system;
2. *Uniqueness*: the trait should possess sufficient personal information to identify the individual by flagging the inter-subject differences;
3. *Permanence*: while the previous criterion requires significant inter-subject variability, which represents the uniqueness of the trait, the trait must provide low intrasubject variability, in other words, it shall be stable enough between different instances of the same individual;
4. *Measurability*: the acquisition and digitisation of the trait should be easy and lead to a representation that can be comfortably processed and measured;
5. *Performance*: each biometric recognition system has its requirements and baselines with regards to the accuracy and performance of the system, and thus, the performance on the trait should meet or exceed those expectations;
6. *Acceptability*: the implementation of a system based on such trait should not raise any foreseeable reservation or insecurity leading to the unwillingness of the subjects to provide their trait information;
7. *Circumvention*: in order to protect the integrity of the system and prevent attacks, the trait should be as hard as possible to mimic or forged.

In the scope of these criteria, some biometric traits have been studied more in-depth than others in the past. For the work developed within this document, the focus was on a single trait, the face. This trait was selected due to its advantageous characteristics. The extraordinary performance levels achieved by biometric systems based on it makes it a popular choice for commercial solutions. Afterwards, its methods are exposed to potential biases and risks, motivating further research on their explainability. The following subsections discuss it with higher detail that is highly relevant for this PhD work.

2.3.2 Face

The face is one of the most well-studied traits within the biometrics scope. Its availability both in the wild and controlled environments was one of many factors that led to its recurrent study. Moreover, it is an effortless and inexpensive trait to capture since it does not require a high level of sensor specialization. Face has been one of the pioneer traits to leverage automatic feature extraction approaches.

In this subsection, we delve into different tasks that can use the biometric information present on faces in images or video captures.

2.3.2.1 Data Acquisition

Collecting data to construct a database for face recognition is one of the cheapest data collection processes. The requirements for the acquisition of the data are low in terms of material obligations and user effort. The basic requirements include a smartphone, camera or any other device capable of capturing images. However, there are more specialised techniques and devices that lead to data collected with superior quality. Whilst the acquisition is not the scope of this document, it is necessary to be aware of what are the current data acquisition methods and how they affect data quality. Moreover, there are links between the acquisition of the dataset and some implicit biases [281]. There are two main approaches to image acquisition that differ in the frequency spectrum that is used to capture the image. The first, which captures information in the infrared spectrum gathers physiological information [159], whereas the second, which captures the image in the visible spectrum gathers anatomical information [26, 384, 474].

- *Visible spectrum cameras:* The most frequent method for capture of the face trait is in the visible spectrum. Common commercial cameras capture their images through the use of light with a wavelength within the 400-750 nm. Due to the availability of devices that can capture information from visible light, it is inexpensive to use this method. Unfortunately, this method is not perfect and is affected by some drawbacks that can negatively affect its performance, such as low quality captures.
- *Infrared spectrum cameras:* Increasing the wavelength of the light captured by the sensor leads to the capture of significantly different images. In an attempt to provide supplementary information to recognition models, some sensors can also capture at four other wavelength

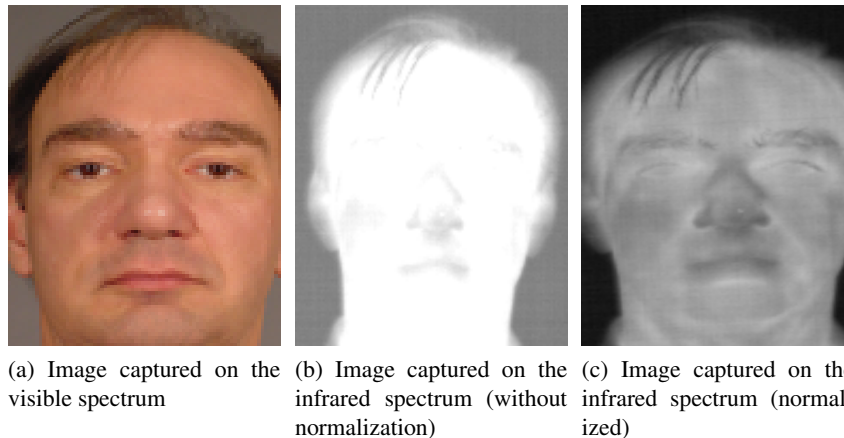


Figure 2.5: Images of the same individual captured on different light spectrum. It is possible to see the notables differences in the information captured by each technique. Images from [93]

ranges. These ranges are the Near-Infrared (750-1400 nm), Short-Wave Infrared (1400-3000 nm), Mid-wave Infrared (3000-8000 nm) and Long-wave Infrared (8000-15000 nm). The latter two are also responsible for what is called thermal images. The side-by-side comparison of these images with the visible spectrum is seen in Figure 2.5.

Moreover, depth can be included in the capture. While it is not as cheap and easy as the traditional 2D image capture, capturing 2.5D or even 3D images can be highly beneficial to face-based models. The fusion of information to build the 3D model is also a source of extra information to support both recognition and attack detection tasks [94, 451, 548, 554].

2.3.2.2 Data Variability

The development of biometric systems must account for potential changes and variations in the biometric data. Or at least the perception that the data varied. In other words, these systems have to be prepared to handle environmental and intrinsic factors, such as the ones in Figure 2.10. These factors can affect the systems in multiple ways, slightly modifying the parameters of their task [26, 60, 133, 134, 228]. Some of the more frequent factors that affect face-based biometric systems are:

- *Light Conditions:* Sensors prepared to capture images through the use of light within the range of the visible spectrum wavelength are susceptible to light variations. For instance, the light reflected by the subject's face might be closer to purple light instead of the most common white light. And thus, this directly impacts the light that reaches the sensor and the image given to the algorithm (Fig 2.6). Heat sources and high-temperature amplitude can impact some types of infrared sensors.



Figure 2.6: Illumination variations in the Yale dataset [37] (Image from [384])



Figure 2.7: Pose variations in the ORL dataset [431] (Image from [384])



Figure 2.8: Face expression variations in the JAFFE dataset [319, 318] (Image from [384])

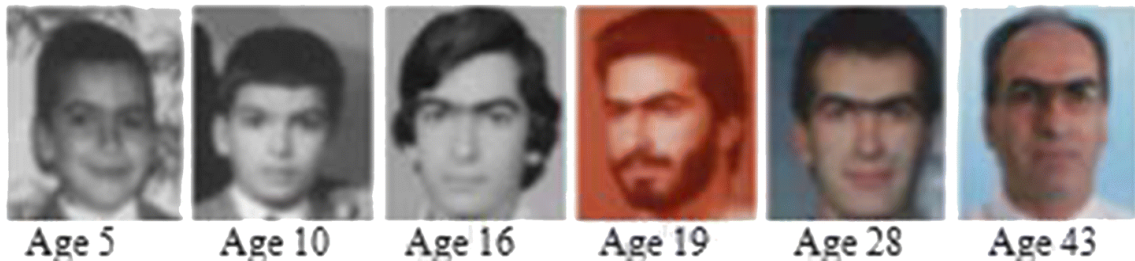


Figure 2.9: Age variations in the MORPH dataset [255] (Image from [384])

Figure 2.10: Examples of intra-subject variability that can affect the performance of face recognition systems. Sub-figures from [384]

- *Face expressions and emotions:* The expression on the face can also slightly modify some features of the face and their relative position, and thus hurt the performance of the algorithm. Some expressions are also responsible for creating new features (e.g. showing the teeth) or hiding other features (e.g. closing the eyes while laughing) as seen in Figure 2.8. While some authors argue that infrared images are less affected by this factor [164], they can be hurt by other related factors. For instance, if a smiling expression originates from a simulated action or if it originates from genuine and strong emotions. While the expressions might be quite similar on the visible spectrum (despite differences in the ocular muscles [101]), infrared sensors can capture different images.

- *Pose*: Head and neck relative positions are not static. Subjects vary their state by performing some muscular movements, such as rolling their heads. Due to this, two images that were captured on the same sensor, with the same expressions and light conditions and on the same minute can have considerable differences. Hence, this capture of facial features from different poses interferes with the recognition task. This can be seen in Figure 2.7.
- *Age*: Face features vary over the years, for instance, through the addition of wrinkles or changes to the hair colour. This decreases the chances of recognizing an individual, especially if the subject's template does not receive frequent updates. Figure 2.9 displays the evolution of a face over a span of almost 40 years.
- *Occlusions, Masks, Cosmetics and Tattoos*: Face images are often captured with some degree of occlusion, which hides some of the features. For instance, sunglasses and scarves. Due to the SARS-CoV2 pandemic, some captured images include face masks that hide almost 50% of its features. Several studies have already identified and investigated this problem [47, 103, 184]. Besides explicit and external occlusions, there are other elements that can interfere with the facial features. For example, tattoos, piercings and cosmetics can create new features or hide previous ones.
- *Aesthetic and Gender Surgeries*: The evolution of medicine, specifically aesthetic medicine poses a problem to face recognition systems. Transgender surgeries, which are frequently accompanied by hormonal treatments, might lead to significant changes in the subject's face. Moreover, other surgeries with aesthetic purposes are also frequent and a source of data variability.
- *Illness*: Some diseases can impact the external looks of a subject. Their treatments can lead to hair loss, extreme weight variations or skin tone changes. In some severe cases, this can increase the difficulty of the recognition task.
- *Sensor Quality and Stability*: The type and the characteristics of the sensor are also responsible for different quality images. Besides the precision of the colours or the overall resolution, the stability of the sensor is important to ensure the clarity of the images and avoid blurry photos. Different sensors produce different images, even under the same environmental conditions.

When constructing a dataset for a biometric system, it is important to carefully consider all these elements in order to incorporate them. Some are key to improve the capabilities of the system in a real world deployment scenario.

2.3.2.3 Recognition

One of the most intuitive and extraordinary uses of the face as a biometric trait is recognition. Most humans excel at this task. However, it is an innate skill. Humans can do it ever since they

can see and observe the environment. Nonetheless, they cannot explain how they do it or the most relevant elements for that inference, at least not in enough detail. In 1991 one of the first methods for automatic face recognition, known as EigenFace, was proposed by Turk *et al.* [486]. More than a decade later, in 2006, the first grand challenge for face recognition was launched [401]. The improvements were steady and consistent until 2012. The publication of the AlexNet article and the advancements brought forever changed artificial intelligence, computer vision and machine learning communities [286]. During the following years, the influence of deep learning kept growing and expanding to other domains that leverage the face as a biometric trait.

Within the scope of the recognition task, we can designate some independent challenges. Indispensable for the current recognition tasks is the face detection and face alignment problem. Nowadays, the best-performing algorithms for face recognition [126, 307, 500] still require faces to be cropped and aligned by other deep learning techniques [128, 546]. Hence, it is necessary to solve these tasks separately, for now, if one aims at solving face recognition as a whole. Another rather curious specification of recognition problems using faces is that this trait is not immutable. In other words, faces change with time, and sometimes they might also suffer significant changes (e.g. puberty). There is an independent research area that focuses on age estimation.

Recent events, such as the pandemic situation caused by the Sars-CoV-2 virus, led to new research directions. Researchers attempted to develop models for masked face detection and masked face recognition, and by association, face mask detection. These are even harder than traditional face recognition due to the limitations imposed on the input information as seen in Figure 2.11.

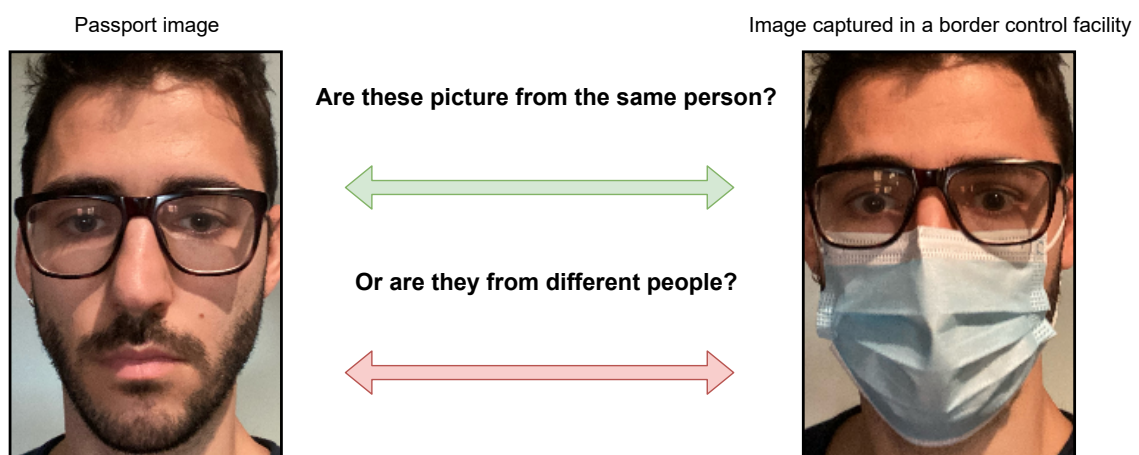


Figure 2.11: Example of two inputs that could be used by a face recognition algorithm to compare an unmasked reference with a masked probe. The majority of the face is covered by the mask on the probe, and the to further increase the complexity of the task, the subject is wearing glasses in both images.

2.3.2.4 Presentation Attack Detection

Notwithstanding the already difficult task of recognizing an individual through this feature-rich biometric trait, there are other parallel challenges to its success. The most prominent challenge, after the environmental variations, is known as presentation attacks. A presentation attack is a threat to the system that originated from spoofs or fake representations of a person presented to the sensor. Researchers have focused their attention on stopping these attacks through a defensive schematic called Presentation Attack Detection (PAD) [2]. The main motive of the attack might be to impersonate another person, for instance, to bypass the authentication system, or to obstruct the system and avoiding face detection.



Figure 2.12: Examples of different types of presentation attacks. Sub-images (a)-(c) were collected from the ROSE-YOUTU dataset [297]. Sub-image (d) was collected from the Multispectral Latex Mask based Video Face Presentation Attack (MLFP) dataset [8].

Defending a system from such attacks is an ever-changing task. Currently, there are a plethora of different attacks to be used against face-based biometric systems. They range from simple printed pictures of the target person to expensive and realistic masks (Fig. 2.12). It is impracticable and nearly impossible to develop a system with the information of all the potential attacks. Due to the recurrent new waves of novel attacks, it is required that developed systems can perform well in unseen and unexpected situations. Current approaches usually focus on detecting the attacks and their characteristics, which limits their capabilities to extrapolate to distinct scenarios. And thus, it is relevant to keep in mind the following questions: "Should the focus be on learning features that discriminate attacks or learning features that serve as proof of authenticity?" or "What makes a human face Human?".

2.4 Performance Evaluation

There are several different metrics and evaluation setups for biometric systems. These metrics evaluate the performance of the system, its overall complexity, computational cost, security and user satisfaction. Moreover, it is possible to follow guidelines for the development of biometric systems, which optimizes the evaluation of some of the metrics. Besides direct evaluations of the system, used data can also be subject to evaluation. The following subsections discuss the different evaluation metrics and their impact on the biometric system.

2.4.1 System Design

In order to facilitate the development and deployment of biometric systems that comply with the guidelines that aim at describing systems that behave well on realistic conditions, Bolle *et al.* detailed some relevant notions and aspects that researchers must account for [40]. Some of them are qualitative whereas the others are quantitative [236]. These six specifications are detailed below:

- *Accuracy*: In general, it is possible to interpret the biometric system prediction in a binary structure. It indicates if the input either matched or not at least one template from a list of possible templates. From this perspective, wrong predictions are of two types: False Non-Match and False Match. The former is caused by a significant intra-user variance, where the biometric trait of an individual is not recognized as the same on two different inputs. In contrast, the latter derives from a small inter-user variance (i.e. large similarity), where the traits of two different individuals are recognized as being the same. Due to the different nature and relevance of these errors, the accuracy of biometric systems is expressed in two distinct metrics based on these errors. False Non-Match Rate (FNMR) and False Match Rate (FMR) represent, as the name says, the percentage of False Non-Match and False Match occurrences. Verification systems, such as the authentication-based phone unlock feature, have less strict FNMR requirements. Asking a user to retry (by inserting the finger on the fingerprint sensor, or looking directly into the camera sensor) to prove their identity to their phone is largely more acceptable than allowing other people to access their phone. And thus, the FMR requirements are much more strict, often by an order of magnitude or more. Border control recognition systems can not afford to relax the requirements of any of the error measures. Due to time constraints and a potential inconvenience to the screened individuals, the FNMR must be low. Similarly, it would be harmful to fail to identify undesirable individuals, thus the FMR needs also to be low. It is then important to consider the application of the system that is being developed and designed, so that appropriate error thresholds and trade-offs can be defined.
- *Throughput*: The number of iterations that can be run on a biometric system per unit of time varies with three main factors. First, the capabilities and the potential of the hardware where the system is deployed. Afterwards, it also depends on the overall complexity of the algorithm and the communication API of the system. Finally, the integration of the system and the hardware also impacts the final performance, because in some cases, these systems do not take into consideration the specificity of the hardware where it will be deployed (i.e. concurrent, parallel and GPU programming). The processing time is also affected by the task at hand. For instance, authentication/verification tasks only require the match between one template and the input. On the other hand, identification systems have to try to match the input with all the templates in a database. And thus, there are special techniques and methods to improve the query time and to leverage both hardware and software. Despite the lack of concern with the template matching process, verification systems must ensure that

the production of a template from an image takes, in the majority of the cases, less than one second.

- *Cost*: As with any other complex system, developing a biometric solution is not cheap. Researchers and developers are consistently defied by a trade-off between the cost of the system and its performance, which can be thought of in terms of accuracy, processing capabilities and usability. Besides the costs that are common to the majority of computational systems, some intangible costs can be included. For example, a wrong prediction might imply further costs that cannot be expressed in the same way as, for instance, maintenance costs. Therefore, it is of utter importance to consider all the potential costs and trade-offs before deploying any biometric system.
- *Security*: As discussed in the previous sections, besides the benefits that these systems hold when compared to non-biometric based authentication methods, they are still at risk of being attacked [7, 239, 406]. Besides direct attacks, there are other risks, such as data leakage from a company that stores biometric information. It is necessary to keep the security of the system in mind when designing it and to make recurrent assessments of potential threats over time.
- *Privacy*: Through the capture of biometric traits, it is possible to construct a secure and stable authentication system. However, such information that links so irrefutably to user's identity can also be subject to abuse and misuse. For instance, instead of only leveraging face captures for authentication purposes, someone can also use these captures for the promotion of tracking methods. It is especially concerning when done without the knowledge and consent of the owner of that biometric information. Consequently, the privacy of the user must be incorporated into the biometric system by default and through the use of the appropriate development techniques.
- *Usability*: Whilst privacy and security are big concerns to developers, they are usually disregarded by the majority of users unless some negative incident related to them happens. What brings users to use a certain system is mostly related to its usability and easiness. The usability of biometrics systems can be seen from several perspectives of different types of users. The evaluation of usability is done in the form of its effectiveness, efficiency, satisfaction and learnability. Some of these evaluation metrics are not always related to the user providing its information. For instance, in a border control facility, face-based biometric information might be captured without requiring interaction with the user that owns the biometric information. On the other hand, users supervising biometric systems on border control facilities have to interact with the system. Hence, in this scenario learnability is relevant to the latter user but not for the one providing biometric information. Usability flaws lead to two different types of errors. Low-quality images or images incorrectly given (e.g. by a user that cannot interact as intended with the system) might cause a failure in the enrolment process. The frequency at which these failures occur is measured by the Failure

to Enroll Rate (FTER). On other occasions, problems at the sensor or other elements of the system might lead to a failure at the image capture of the biometric sample. The incidence of these scenarios is measure by the Failure to Capture Rate (FTCR). The usability of the system depends on the biometric trait, the sensor, the design of the user interface (UI) and user experience (UX) considerations.

2.4.2 Data Quality

After the increase in use of DL methods, partly due to the large amounts of data available, there was an urge to find and increase the size of the datasets. Recently, this urge cooled down and now there is an understanding that the increase of the number of samples is not useful if the quality of the samples is not guaranteed. Therefore, researchers started to improve their methods of quality assessment of biometric raw data. Sample quality is one of the factors that have an impact on the overall performance of biometric systems. Some artefacts, such as illumination, can impact the acquisition. Through the implementation of quality assessment methods, poor quality samples can be filtered at the enrollment or not considered during the identification or authentication. Multimodal approaches can benefit from the information regarding the sample quality [404].

Scientists working on biometrics believe that there should be a causal relationship between the quality of the biometric sample and its recognition performance [195]. Over the years there have been several different proposed methods for quality assessment on several different traits and with different purposes. Some focused on fingerprint [18, 88, 291, 452], others on iris [89, 258, 284] and also on face [175, 435, 545].

2.4.3 Recognition

2.4.3.1 General

Some of the metrics used on biometric systems are independent of the tasks. They can be used and provide useful information in any scenario. These metrics are:

- *Failure to Enroll Rate (FTER)*: Percentage of users that, after providing their biometric sample, have the biometric system incapable of extracting usable information from the sample.

$$FTER = \frac{\text{\# of biometric samples that failed to have their features extracted}}{\text{\# of biometric samples}} \quad (2.1)$$

- *Failure to Capture Rate (FTCR)*: Percentage of authentication or recognition attempts that are unable to be completed due to the system incapability to capture a sample with the required capability.

$$FTCR = \frac{\text{\# of attempts failed (system cannot capture a good sample)}}{\text{\# of attempts}} \quad (2.2)$$

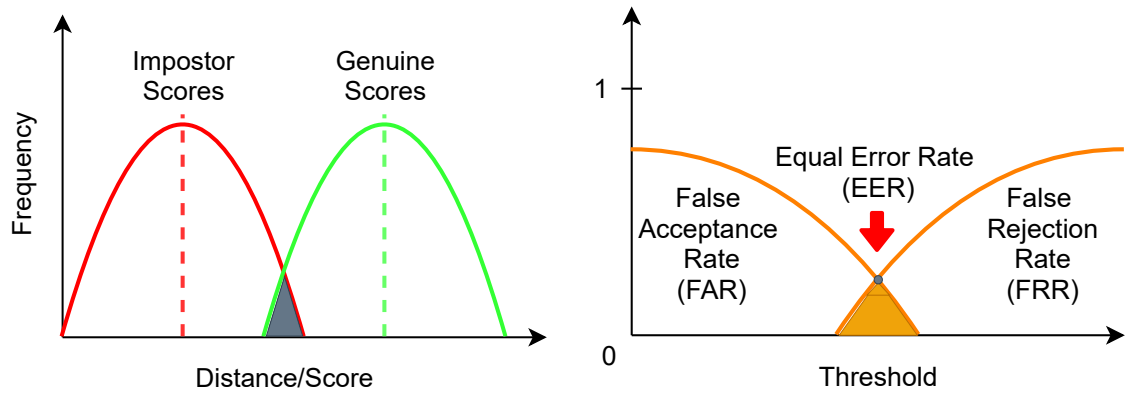
- **False Match Rate (FMR):** The proportion of templates extracted from different individuals that are considered to be from the same individual.

$$FMR = \frac{\text{\# of templates wrongly classified as the same person}}{\text{\# of template comparisons}} \quad (2.3)$$

- **False Non-Match Rate (FNMR):** The proportion of templates extracted from the same individual that are considered to be from different individuals.

$$FNMR = \frac{\text{\# of templates wrongly classified as different people}}{\text{\# of template comparisons}} \quad (2.4)$$

2.4.3.2 Verification/Authentication



(a) Distribution of the scores given to the model regarding impostor and genuine identity claims.

(b) Example of the distribution of the False Acceptance Rate (FAR) and the False Rejection Rate (FRR). Includes also the graphical representation and meaning of the Equal Error Rate (EER).

Figure 2.13: Relation between the scores given to the impostors and genuine claims and its impact on the FAR and FRR error metrics. Besides it shows how the EER is graphically represented as the interception of the FAR and the FRR. Adapted from [470].

The verification and authentication task has some specific metrics. These are mostly variations that come from the general evaluation metrics. As stated in previous sections, the errors on identification systems can be of two types false match and false non-match. On the other hand, there are also two success cases, the true match and the true non-match. Specifically for verification systems, the comparisons are one-to-one, because the user claims an identity. The system has to compare the template from the claimed identity to the template produced from the trait given by the user. A user claiming to be someone else is called an impostor, whilst if the claim is faithful the user is called a genuine. Whenever a system allows an impostor into the system (a false match), the error can be, specifically for verification systems, be called a false acceptance. Whereas when a genuine sees its claim rejected (a false non-match), the error is a false rejection. The goal of

verification systems is to separate into distinct distributions the scores of impostors and genuine as seen in Figure 2.13a. Metrics based on these errors are detailed below:

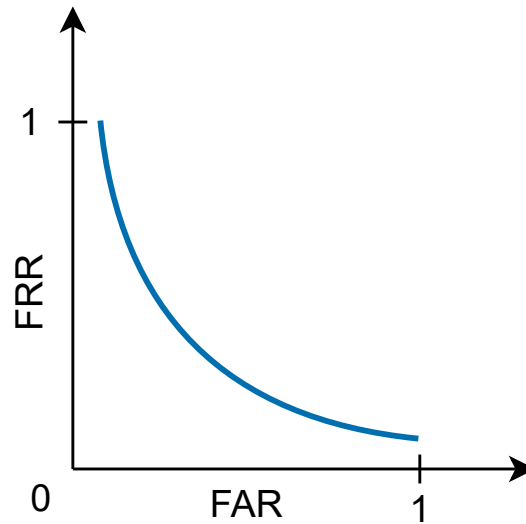


Figure 2.14: Receiver Operating Characteristic (ROC) curve regarding the FRR and the FAR. While this curve does not have the same structure as traditional ROC curves, it can be converted to a traditional curve by plotting $1 - \text{FRR}$ on the y-axis. Adapted from [146].

- *False Rejection Rate (FRR)*: Similarly to the FNMR, this metric represents the percentage of genuine users that are not allowed into the system. When a system is deployed it can be seen as the probability of rejecting a user that should not be rejected.
- *False Acceptance Rate (FAR)*: This metric represents the percentage of impostors wrongly allowed into the system. As the FMR, it can be interpreted as the probability of accepting a user that is falsely claiming to be someone.
- *Receiver Operating Characteristic (ROC) curve*: The ROC curve attempts to capture the relationship between the FAR and the FRR. This curve has many uses, but it is mostly used to understand the system trade-off between these two error rates and the implications of choosing a certain threshold. The FRR is plotted on the y-axis whereas the FAR is on the x-axis as seen in Figure 2.14.
- *Equal Error Rate (ERR)*: As seen in Figure 2.13b the equal error rate is represented as the point and error value at which FAR and FRR are equal. As the name says, at that point both FRR and FAR represent the EER. It is frequently used as a comparison measure of biometric authentication systems or presentation attack detection systems. The performance of the system improves as the EER decreases.

2.4.3.3 Identification

Differently from the lightweight comparison performed by the verification systems, the identification task requires several comparisons. Moreover, the number of operations usually grows with the number of users enrolled in the system. This is due to the fact that these systems do not receive an identity claim, thus instead of verifying one identity, they have to identify the correct one or at least include it on a ranking of the R most likely identities. Some of the specific metrics of these systems are described below:

- *True Positive Identification Rate (TPIR)*: Percentage of identity searches that return the correct identity within the ranking of the top R identities. The values of this metric and similar are highly dependent on the threshold value and the size of the list R .
- *Identification Rate (IDR)*: It is similar to the TPIR, however, for this metric the size of the ranking R is 1, thus considering only the identity with the highest similarity.
- *Reliability*: The reliability metric aims at evaluating how often the true identity is above the threshold. For this, the TPIR is calculated for a ranking, whereas its size R is the same as the number of enrolled subjects.
- *False Negative Identification Rate (FNIR)*: The proportion of identity searches where the identity of the enrolled user is not included in the ranking of the top R identities.
- *False Positive Identification Rate (FPIR)*: Percentage of identity searches that return at least one identity for a user that is not enrolled on the biometric system.
- *Selectivity*: For all the impostors' trials, the number of identities that are above the threshold is averaged. It is similar to the FPIR.
- *Reject Rate (RR)*: Percentage of identity searches that return no identity because all of them were below the threshold. Thus, the system rejects to identify the user.
- *Misidentification Rate (MDIR)*: Similar to FNIR but the list of the top R identities has size 1.

2.4.4 Non Error Metrics

Besides the error metrics mentioned before, there are other metrics that aim at measuring the performance of the system but do not represent any type of error. These are known as non-error metrics and usually represent computation time and memory costs.

- *Template capacity*: Recognition and verification systems have to store several templates for comparison. This metric indicates how many templates can be stored in the system to be used for these tasks. Higher capacity represents systems that can scale, at least from a memory point of view, to a higher number of enrolled users.

- *Average Matching speed*: The average time it takes to identify a user using the biometric system.
- *Average Enrollment Time*: The average time the system takes to process the information of a new user and store its data on the database/system.
- *Average Verification Time*: The average time it takes for the system to verify the identity claim of a user.

These metrics are more important if the system is implemented at a critical facility, such as border control facilities.

2.4.5 Users' Satisfaction

As said in previous sections, the wide adoption of biometric systems is highly dependent on the satisfaction and acceptance of users. For this reason, it is necessary to take these aspects into consideration when deploying a biometric system and to evaluate users' overall satisfaction over time. Researchers have been working on models that foster users' trust in these systems, but also on techniques to understand the source of the lack of satisfaction. These can be found all over the literature [146, 482, 483]. Some of the findings include the link between increased usage of biometrics systems and the overall acceptance, and the fact that users trust more on systems that provide more information. However, due to the fact that compromised biometric templates remain as such forever, some users are concerned about the misuse and lack of security of their data. Furthermore, the risk of having criminals targeting their biometric traits (e.g. aiming to steal eyeballs or cut fingers) is in the mind of the users. These findings were shown by several researchers on several independent studies [100, 119, 145, 147, 252, 349, 405].

2.5 Conclusion

Biometric systems are being implemented all over the world. Developers of such systems are attempting to foster the trust and confidence of current users and potential future users. From the fingerprint to the gait and face recognition, there are several applications that can improve the overall security of a wide range of systems. However, these improvements do not come without drawbacks, there are some security and privacy risks that must be mitigated and corrected. With the increase of users' usage, researchers will also become more aware of system flaws and become expeditious at fixing them.

Part II

Explainable Biometrics

Chapter 3

Explainable Artificial Intelligence

Author Contributions The research work described in this chapter was conducted in collaboration with João R. Pinto, Tiago Gonçalves and Wilson Silva, under the technical supervision of Ana F. Sequeira, Jaime S. Cardoso and Arun Ross. The author of this thesis contributed to this work on the problem conceptualisation, data curation, the proposal of the new taxonomy, the categorization of literature methods, and publication writing. Some parts of the chapter were originally published in, or adapted from:

- Neto, P. C., Gonçalves, T., Pinto, J. R., Silva, W., Sequeira, A. F., Ross, A., & Cardoso, J. S. (2022). Causality-Inspired Taxonomy for Explainable Artificial Intelligence. *arXiv preprint arXiv:2208.09500*. (Second round of Revision at ACM Computing Surveys)

3.1 Context and Motivation

As with any recent scientific field, the terminology of explainable artificial intelligence is surrounded by fog. There are several interpretations across the literature of the same terms. Hopefully, over time, the meaning of each term will converge to common ground. The first part of this Section is dedicated to defining these terms as they are used in the literature and to shed some light on how their definition can be improved. Once established the proposed interpretation of xAI terms, the second part of this Section discusses interpretability and explainability re-framed to a novel interpretation of these terms through the lens of causality. Finally, the third and last part of this Section presents existing biometric works that already include xAI concepts. In addition to this Section, this Chapter further includes two more sections to better detail the benefits of these approaches and to conclude.

3.1.1 Towards a common terminology for xAI in Biometrics: Can we speak the same language?

The definitions of interpretability and explainability found in the literature are not consistent between different authors. Some authors use them interchangeably, while others make a distinction

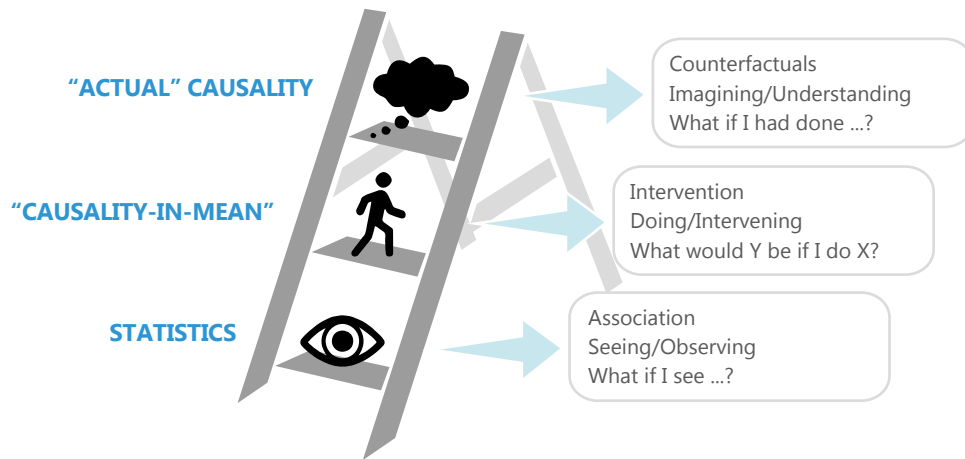


Figure 3.1: The Ladder of Causality (adapted from [395]).

between them [342]. Within the group of researchers that make a distinction, the understanding of these terms also varies. Hence, we aim at establishing a working definition to be used in the following Sections and Chapters of this document.

In our definition, explainability and interpretability are different but not independent concepts. Whose definition is linked to two main questions: “How?” and “Why?”. The former is related to interpretability, which aims at providing a clear understanding of the inner workings of a model and the reasoning process that leads to an output. The latter is the question that, given a model’s prediction, an explanation tries to answer [181]. However, it is possible to deduce the second question from the first, because if we fully understand the behaviour of a model, we also know the answer to the “Why?” question. The opposite is not true, as knowing the explanation of a certain prediction does not directly indicate the inner workings of that given model. We further introduce a third question, “Where?”, which is frequently associated with explainability, while not providing complete answers to the “Why?” question.

Due to the ability to answer to the “Why?” question, by answering the “How?” question, it can be assumed that knowing the answer to the first, it might be the case that we are moving towards the answer to the second, even if we do not have enough information to answer it completely. We believe that there is a strong and close relationship between xAI and causality [394].

Causality studies the relationship between causes and effects, and it aims to model the causal relationships that define an event or sequence of events [61]. Similarly, explainable artificial intelligence aims to understand the model inner mechanisms that led to a certain output when conditioned on a certain input. Pearl [394, 395] proposed the ladder of causality (Fig. 3.1). Within this framework, it is possible to characterise causality in three levels: (1) association, (2) intervention and (3) counterfactuals. Each of these levels aims to answer a certain type of questions, and they follow a hierarchical structure, in the sense that for answering the questions of a level i , we must also have the information to answer the questions of all previous levels [394]. Pearl [394] goes even beyond by defining the degree of causality associated with each level, he further establishes the mathematical notation for each level. For the Association level (1) the usual representation

is a conditional probability in the form of $P(y|x) = p$, which calculates the probability p of the event $Y = y$ given that the event $X = x$ was observed. Bayesian networks are effective and efficient at calculating these conditional probabilities [393]. Intervention level (2) operations can be described, leveraging the *do* operator, as $P(y|do(x), z)$, which calculates the probability of the event $Y = y$, given that we intervene to set X to x , and subsequently observed the event $Z = z$. The third and last level, the Counterfactuals, can be represented by $P(y_x|x', y')$, which states the probability $P(y|x)$ when we actually observed the events $X = x'$ and $Y = y'$. It is only possible to compute these if one has access to functional or Structural Equation models, or their properties.

The connection between these two scientific fields can be used to re-frame xAI, and we suggest this change in the following sections.

3.1.2 A brief introduction to xAI

Some of the terms used to define explainable artificial intelligence methods have overlapping definitions and concepts. Some authors aim at dividing them into pre-, in-, and post-model [268], others in methods that are either specific or agnostic to the deep model, and even categorizing them as post hoc or not.

By dividing xAI into three smaller sub-topics, researchers allow for the creation of different goals for each task. The pre-model analysis focuses on data knowledge and understanding. It significantly leverages visualization techniques that allow researchers to understand the distributions of the input samples [493]. This is an approach frequently used in Business Intelligence (BI) use cases. Traditionally, this technique led to the improved capability of creating handcrafted features for simpler models; nowadays, it is quite relevant to detect some biases in the data, which can be described through the analysis of the skewness in the dataset [516]. Understanding the data used for training allows for increased confidence in the posterior decisions and explanations.

On the other hand, there are also techniques that can be applied to the model and its outputs during and after training. When the training process is conditioned by some constraint, these are known as in-model techniques and they focus on the direct integration of interpretability into the model through the usage of these conditioning priors. Some families of machine learning models already include inherently/intrinsically interpretable mechanisms in their design. Hence, their use is also considered to be an in-model approach. Some of these interpretable models are rule based [59, 425, 506], per-feature based [161, 204] and some simple versions of linear models such as linear regressors. They are, however, limited by the semantic meaning of the original features and the size/complexity/depth of the model. Nonetheless, it is possible to increase, to a certain degree, the interpretability of a complex model. Constraints to these models, frequently expressed in the form of regularisation terms (such as L1 regularisation [190]), are the most common approach to increase the interpretability of those models. Other approaches include constraining the model to have certain properties, such as monotonicity [455]. These constraints can, sometimes, be applied or enforced in specific modules. While, it is known that achieving global holistic interpretability is an extremely difficult task, analysing the model in modules for global modular interpretability is much more manageable [340]. The exploration of these in-model approaches

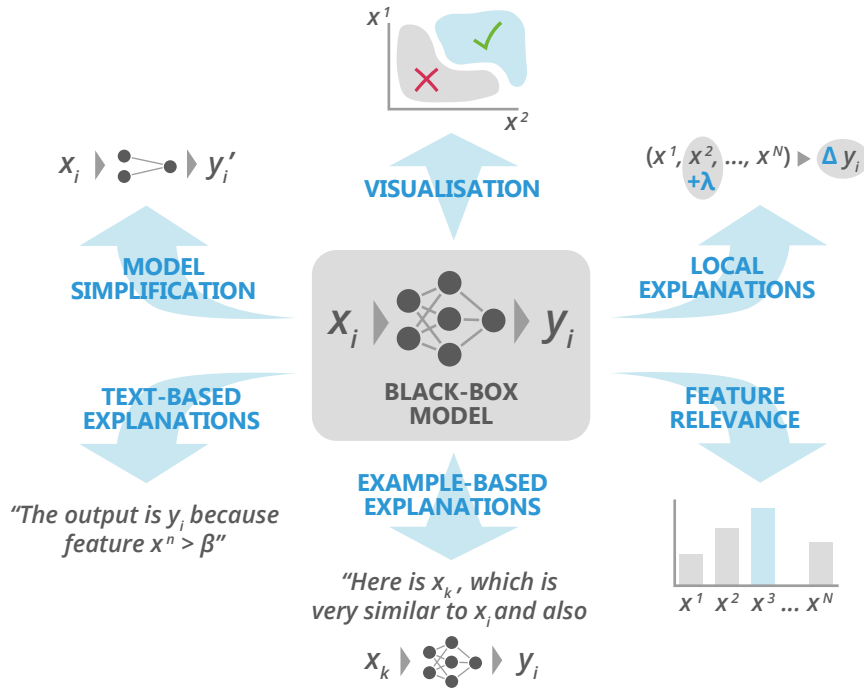


Figure 3.2: How we can explain black box models: a summary of current *post hoc* explainability approaches (adapted from [34]).

within the context of biometric systems is rather limited due to the potential sacrifice of predictive power. Instead, the focus has been on *post hoc* methods, which is a different term to post-model approaches. It is, however, important to note that, if the constraints applied to the model represent meaningful priors, it can be the case that a better solution can be found, since the search space is reduced to exclude worse solutions [366, 519].

Aligned with the efforts towards continuing the development of more accurate models, the focus has been partially redirected towards post-model interpretability. This approach is frequently known as explainability since these methods generate explanations to a given prediction of a trained model. Generally in-model approaches are model-specific and post-model approaches are model-agnostic, however, this not always the case. For instance, gradient-based methods, which leverage gradient information to identify areas of the image that contribute to the final decision, are specific to deep neural networks [31, 458, 462, 464]. One of the common approaches to post hoc interpretability adds perturbations to the input in order to assess the impact on the output [160]. As discussed in the following Sections, these methods do not have a rich semantic meaning. Hence, to enrich the explanations produced, some methods map the latent features of a model to the input space [139, 539] or to a representation of semantic concepts [35].

Lately, there has been research on a post-model approach that closely resembles ideas from in-model methods. While training a simpler model is somewhat complicated if the task is too complex, it is possible to teach a simpler model to mimic the behaviour of the more complex (deeper) model [27]. Due to the lower parameter count, simpler models present a more interpretable view

of the decisions made by the complex model. This is known as model distillation or knowledge distillation. This and other compression techniques, such as compression [49] and pruning [302] can help to reduce the complexity of an overly complex neural network architecture [69].

In biometrics, the prevalence of pre- and in-model approaches is rather limited when compared to post-model methods. Due to the high requirements for accuracy and other performance metrics, the usage of post hoc methods is much more convenient, requires less engineering effort, and provides an acceptable trade-off between accuracy and these efforts. Nonetheless, while it provides immediate convenience, it is not closer to a proper path to xAI than the other two approaches. Moreover, the current taxonomy of xAI is rather limited as it falls short at categorizing the different approaches to explainable AI. This is result of the current focus on techniques that aim to explain the prediction instead of providing answers to the reasoning behind the model's behavior. Hence, his taxonomy fails to include several in-model approaches that improve the interpretability of the system through the introduction of domain knowledge within the training process. As such, the following Section proposes a revised taxonomy for Explainable AI and how it encapsulates the methods seen in Figure 3.2.

3.1.3 The causal relationships of a deep model

Causality's main goal is to understand and learn the causal model of the *world*. Once this model is found, it is possible to make predictions and informed guesses with a tremendous degree of certainty. It is possible to grasp the implication of the smallest details on the unfolding of a certain event. As previously mentioned, the ladder of causality represents a simplification of the steps one must take towards getting a causal model of the world. In practice, the last step on the ladder represents a mental representation of the *world* that allows humans to replay, recreate and even imagine new events on their heads. One can formulate the question of "What would happen if I stop taking my antibiotics now?", and humans are capable of answering through imagination.

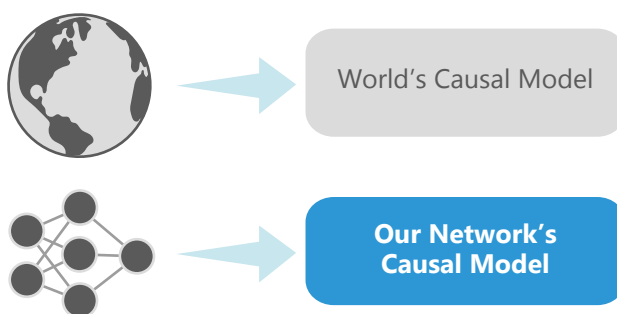


Figure 3.3: A figure illustrating the World's Causal Model which represents all the causal relationships that we observe and that rule the functioning of the world. Below we see a representation of the Causal Model of a Deep Neural Network. While the latter represents the causal relationships within the "rules" learnt by the DNN, the former represents the true causal relationships of the world. These two models are very likely to disagree, and the latter does not aim to explain the causal relationships of the world, just the specific deep learning model it represents.

Explainable artificial intelligence shares the same goal as causality approaches, in the sense that it aims to understand the model and provide explanations for the relation between two or more events (for instance, input and prediction). This means that, in a sense, we can try to frame xAI within the context of causality. And, for that, it is necessary to answer the following question "Which causal relationships should our causal model represent?". Since we are only trying to understand a specific deep model, our causal system does not need to represent the *world*. Instead, we want to find a causal system that models the causal relations of our deep model, as show in Figure 3.3.

When compared to the previous approach, there are significantly more advantages to this approach. First, it is possible to measure the transition between the different steps of causality. It is also possible to quickly detect spurious connections learnt by our model, through a comparison between its causal model with reality. For instance, if a model uses the skin tone for a task that should not use that specific feature, the causal model of the deep model shall provide this information. Hence, hidden biases can be detected and explained in a straightforward manner. In other words, we can look for misalignments between the causal relationships of *reality* and the causal relationships of a deep model. So, in a biased model, we would find, for instance, a causal relationship between gender and lower wages, and through a careful comparison with the *world* causal relationships, we can deem this specific relationship as biased.

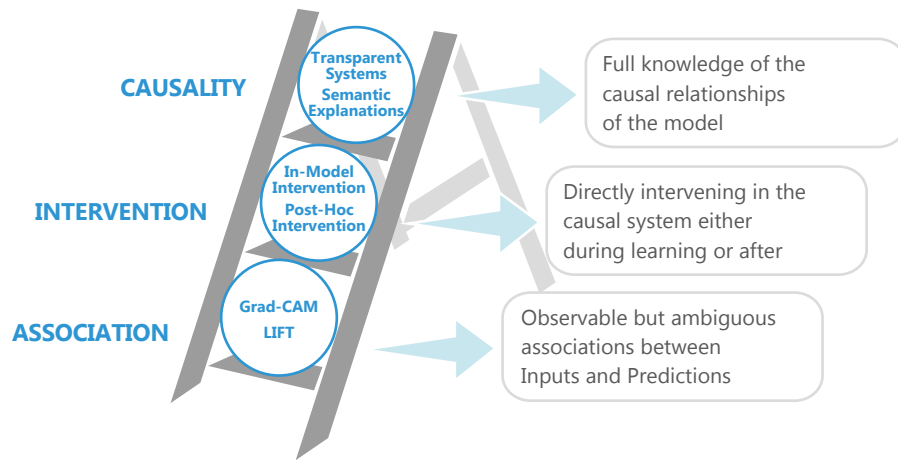


Figure 3.4: Causality ladder repurposed to categorise different types of xAI approaches according to the understanding of the model provided by them.

For this we revisited the ladder of causality (Fig. 3.1) and propose a similar structure for the ladder of explainable artificial intelligence as seen in Figure 3.4. This structure for categorising different xAI approaches shows how they contribute to answering the questions that we introduced in Section 3.1.1. Furthermore, it highlights the importance of each technique and the journey ahead. On the first step of the ladder, we can see observational explanations, such as visualisation maps, which are quite powerful at locating elements that were relevant for the decision, but that do not hold semantic value or any information regarding what was in fact important for the decision. This presents itself closely linked with the associations at the first step of the ladder of causality.

Similarly, pre-model approaches, that rely on the understanding of the data, are unaware of the inner mechanisms of the model, hence, these approaches cannot, by definition, go beyond the first step on the ladder.

The second step on the xAI ladder encompasses some of the more advanced approaches currently used in the literature. Furthermore, it is the case that an approach might be between two different rungs. For instance, one might insert prior causal knowledge to condition the training process. Thus, it is possible to know the inner workings of the model regarding that prior, which means that it is possible to reduce the set of possible explanations while partially knowing a subset of the causal relationships of our deep network. As such, the rung climbing is not a discrete process, but instead a continuous one. Some of the most common methods on the second rung are case-based explanations, semantic explanations, LIME (which closely resembles the *do(.)* operation [394] by intervening in the input and analysing its effects. But in fact, the majority of applications is in between the first and second rung, where they can perfectly answer the "Where?" question regarding the importance of the variables, but cannot provide a strong and complete answer the "Why?" question. Climbing from the second to the third rung is the final goal of researchers working on xAI. This last rung requires a complete understanding of the inner causal relationships of the deep network. When this happens, one can fully answer the "How?" question, and can further answer to the "Why?" and "Where?" questions. Moreover, knowing these relationships means that producing counterfactuals is not a difficult task, for instance, for a skin-tone biased model we can quickly verify that a change in the skin tone would impact the final decision. And for this, we are not required to provide a new empirical sample to the model so we can empirically check the result (as happens with LIME for instance).

The implications, nuances and adaptations of this new framework to categorise are vast, and require a careful analysis and continual study. As such, we provide in the following sections a few notes on each rung.

3.1.3.1 The First Rung: Observational methods as a first order explanation

As with the association within the causal domain, in xAI, visualisation approaches are the most common. This tendency can be explained by the fact that most of the visualisation approaches require little computational effort, are less dependent on the model architecture and work well for several tasks. This new definition categorises visualisations as first order explanations, meaning that these are the most immediate explanations one can generate to quickly inspect the reasons behind a prediction. Furthermore, they provide less information than second and third order explanations.

One of the first strategies for the visualization of deep learning models was proposed by Zeiler *et al.* [539]. This was mainly motivated by the scarce insights that researchers had about the functioning and behavior of the models that led them to achieve high performances. In their approach, the authors used a multi-layered deconvolutional network [540] to project the feature activations into the input space. Besides, they also proposed several occlusion tests to evaluate the sensitivity of the classifier. They occluded portions of the input image and revealed the feature

activations, thus showing the parts of the inputs that were relevant to the classification. Later, Simonyan *et al.* [458] proposed two techniques that focus on the respective class of the input to generate the feature activations. This is achieved through the computation of the gradient of the class score concerning the input image. In the first use case, the authors built their work on top of Erhan *et al.* [149]. Given a score, and a learned model, the idea is to generate an image that maximizes the score, outputted by the model. This synthetic image is thus representative of the class appearance learned by the model [458]. The second use case applies a similar strategy to generate a class *saliency map*. In this approach, the idea is to generate the saliency map that maximizes the class of a given input image. The result of this method is a mask that has pixels with high intensities in the locations that are related to that class.

Springenberg *et al.* [464] revisited the work behind deconvolution and performed experiments with a model composed solely of convolutional layers. Known as *guided backpropagation*, this approach proposed a modified way to handle backpropagation through rectified linear (ReLU) nonlinearity: instead of removing the negative values corresponding to negative entries of the top gradient (*i.e.*, deconvolution) or bottom data (*i.e.*, backpropagation), they remove the values for which at least one of these entries is negative. According to the authors, this technique prevents the backward flow of negative gradients, which often correspond to the neurons which decrease the activation of the higher layer unit one aims to visualize. Sundararajan *et al.* [469] argued that most attribution methods did not respect two fundamental axioms (*sensitivity* and *implementation invariance*), thus representing an important flaw of such methods. Taking this premise into account, the authors proposed a method called *integrated gradients*, which does not require any modification to the original network architecture, while being related only to the gradient operator. Following a similar research line, Shrikumar *et al.* [454] created DeepLIFT. This method is based on the backpropagation of the contributions of all the neurons in the network to all the features of the input and works by comparing the activation of each neuron to a given *reference activation* and assigning contribution scores according to this difference. DeepLIFT also allows practitioners to give separate consideration to positive and negative contributions, hence allowing them to find other dependencies.

Bach *et al.* [29] studied the influence of pixel-level contributions through the introduction of the *layer-wise relevance propagation* (LRP) technique which considers that a deep model can be divided into several layers of computation which may be responsible for extracting features or performing classification. Following the work of Zhou *et al.* [553] on *class activation mapping* (CAM), Selvaraju *et al.* [441] developed the *gradient-weighted class activation mapping* (Grad-CAM), which allows the practitioner to generate saliency maps with respect to a specific class (*i.e.*, class-discriminative), a property that may help researchers unveiling some visual concepts related to a given class.

Frequently, these visualisations have not been seen as explanation methods, instead, they have been seen as a tool that can be used to provide additional information to the end-user [80, 433]. In a sense, it can be seen as a parallel of distillation methods frequently applied to complex models, but in this case, used on complex high-order explanations. Even more than traditional explanations,

the interpretation of outputs from a visualisation method is highly subjective and user-dependent, despite some attempts to objectively evaluate them [432]. The former aligns with the perspective of assuming visualisations as approaches to achieve first order explainability, whereas the second cements it through another parallelism between visualisations and associations, which are also subject to very different interpretations. Hence, currently, saliency maps and other methods of feature visualisation are a bridge between the limited perception of humans on any number of dimensions above two or three, and the complete user-specific semantic explanations.

The practicality of the visualisations as a tool for first order explanations led to a wide-spread implementation of this. And while researchers agree that using visualisation techniques has a positive effect on our understanding of the deep model, the amount of information contained within these explanations is rather limited and vague. For instance, we remove crucial information from these explanations, such as the importance of colour, transmitting a more open-ended problem to be discussed among humans. One of the most consensual statements between experts in explainable artificial intelligence is that humans can better understand counterfactual explanations [64, 490]. In other words, a counter-example of the one given as input, which introduces a notion of contrastiveness [303]. This behaviour is representative of the relationship between counterfactual examples and causality [63, 514]. In other words, to move towards more complete explanations it is necessary to keep climbing the ladder.

Despite their limitations and their current use, it is expected that in the near future explanation methods will keep on leveraging methods from the first rung. Their practicality is difficult to defeat, and several tasks and problems are not concerned with higher-order explainability, since they only focus on small assessments of the model's behaviour.

It is worth noting that while this first rung considers both post-model approaches (e.g. Grad-CAM) and pre-model approaches (e.g. data visualisation), these can still be differentiated as some of them already have some information regarding the model's architecture and gradient propagation. Hence, it can be said that while not providing an answer to the "Why?" question, they provide stronger insights into the behaviour of the model. The following sub-section keeps on climbing the ladder.

3.1.3.2 The Second Rung: Higher Order Explanation with Priors and Semantic Information

The climb from the first to the second rung presents itself as a task with a diverse set of pathways. Differently from the direct explanation presented in previous chapters regarding interventions, this operation can take many forms when applied to artificial intelligence. In a sense, we can describe any of the operations that will be discussed as an intervention to the deep neural network that we want to describe in causal terms. Nonetheless, there are fundamental differences between them. Moreover, some of these interventions require careful analysis, as confounding bias is a threat that remains present [461]. In contrast with the previous rung, the methods in this rung require their user to be, at least, partially aware of the causal relationships of the problem. Moreover, we further differentiate between the different elements in an AI system. Our overview of the system's causal

structure can be seen in Figure 3.5. It is important to distinguish between the inferred label and the true label. The former is defined by an human annotator and is inferred from the input itself. Moreover, it has a direct impact on the model. The latter is the true label, it is not always available and is the cause of the input, as discussed in medicine, the disease causes the image alteration and not the opposite [74]. Hence, the latter has an indirect effect on the former too.

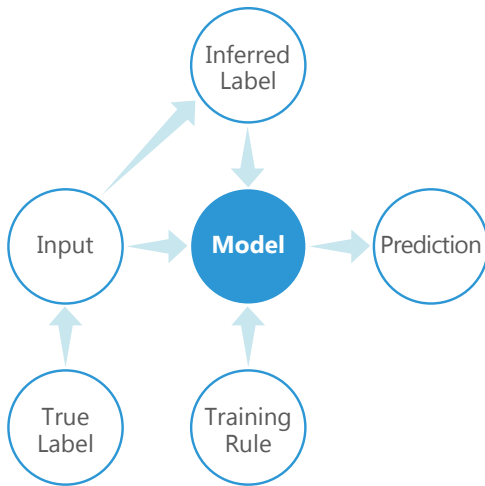


Figure 3.5: Causal model representation of the training. The input, the label and the training rules directly affect the model, whilst the model affects the final prediction. The Inferred Label and the True Label are also separate, while the former is given to the model and inferred from the input, the latter is the cause of the input and is not always available.

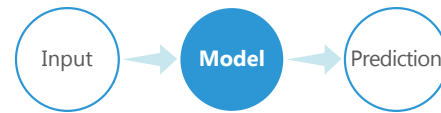


Figure 3.6: Causal model representation of the inference process. The Model node is a causal system by itself and it is the system that we are trying to understand.

In the second rung we can have both post-hoc and in-model methods. Within each of these two groups it is also possible to find distinct strategies to do interventions, or to partially find a causal representation of a deep learning model. For the post-hoc interventions, we highlight interventions in the input and interventions in the prediction.

- *Intervention in the Input:* As shown in Figure 3.5, the input given to the model has a direct arrow pointing towards the model. In other words, there is a causal effect between the input and the model, which means that changing the input will have consequences on the model's behaviour. It is possible that, by intervening on the Inputs by modifying them to the desired values, we can observe and learn the causal relationships between the model and the prediction. This type of interventions allows for the interpretability assessment to occur after training. It can be particularly useful if one aims to detect the implications of skin tone on the final prediction. A transformation in the skin tone of the image might reveal some hidden causal relationships between the model and the prediction, that should not

exist. The relevance of this type of interventions is growing with the current developments of generative artificial intelligence methods, such as Diffusion Models [396, 550].

- *Intervention in the Prediction:* Explaining the model is the main goal of xAI, nonetheless, in some scenarios explaining the decision might represent a good proxy. As such, one of the most common strategies to explain these predictions relies on finding similar samples in the input space that produced similar predictions. This behaviour might help the human to focus on the common areas of both inputs. The most relevant implementations of these interventions are usually related to medical images [343]. An extension of these case-based explanations can be used as a proxy to find counterfactual samples. Setting the prediction as P , it is possible to find the nearest input, X , with a different value for P . The difference between both inputs highlights some reasons that distinguish both classes.

On the other hand, for the in-model methods, we highlight the introduction of causal information in the model. This can be achieved through training or strong model constraints. The main advantage of these approaches is that they provide partial control of the model, which translates into an *expected behaviour* that can be explained through the causal "rules" directly given by researchers. In other words, as seen in Figure 3.6 acting on the Training Rule affects the model in a manner that can be predicted, and acting on the model also leads to predictable behaviour. This means that both approaches reveal a part of the causal model that will represent the final version of our system.

- *Model Priors:* Deep neural network architectures are, in the majority of the cases, shared across applications and tasks. This is possible due to their high expressiveness and capability to represent a myriad of complex functions. Nonetheless, this representativeness reinforces the opaqueness of the system. However, it is possible to insert certain decision choices in the model that reduce the function space, and lead to a predictable behaviour under certain circumstances. An example is the rotation equivariance on certain filters of the network [75, 193]. Another approach aligned with this strategy is the ProtoPNet [81, 135] which learns prototypes that can be used afterwards to explain model predictions.
- *Training Rule Priors:* In a similar fashion, these priors reveal certain parts of the causal system that represents the learnt deep neural network. However, these constraints are usually more relaxed than when applied to the model itself. Some examples are orthogonality constraints through the minimization of the inner product of two vectors [366], occlusion robust losses that promote similar embeddings between occluded and non-occluded samples [360]. Due to their relaxed nature, these priors are seen as an approach to promote a desired behavior, usually a behavior data researchers already know to be true. For that reason, and since they do not completely enforce the constraint, the interpretability of the model is usually lower than with Model Priors.

The range of approaches that fit the second rung of the causal explainability ladder is wide and might benefit from one or several of the previously described strategies. For instance, it is

possible to have strong model priors, and leverage them through interventions on the predictions to construct a naive counterfactual. A naive counterfactual can be achieved if a system is capable of providing to a user a close and plausible alternative to the original input that belongs to a different class. It is frequently achieved through a search on the latent space [267]. The term naive highlights the ignorance related to the inner relations of the model, and a focus on searching on the latent space. Some methods promote a certain structure on the latent space, and thus, this search can be somewhat controlled for certain variables. This introduces the capability to have more human friendly explanations. However, this is not easy to devise in practice due to humans' limited range of perception and control over the majority of the variables [327]. Thus, what a human searches for in a counterfactual explanation might not be fully represented in the provided explanation, leading to a trial and error search.

3.1.3.3 The Third Rung: Imagining/Extrapolating

Constructing and understanding the causal model that represents a deep neural network is not trivial. At the time of writing this article, there was no approach capable of fully explaining a deep neural network. As such, the third rung is yet to be achieved. The majority of the efforts have been devoted to the first two rungs, as such explanations suffice for the vast majority of tasks. Nonetheless, in critical domains, such as biometrics and healthcare applications, an incomplete knowledge of a deep learning model might prove itself to be a challenge in terms of security and user acceptance.

The third rung of the explainability ladder is the last rung in the latter, which means that explanations provided by these methods are the highest possible degree of explanation. One very relevant example to compare this rung with the second rung is the construction of counterfactual samples. Differently from a naive counterfactual, a true counterfactual (or just counterfactual) originates from the complete understanding of the models' inner working mechanisms. Instead of searching on the latent space, one can wonder "What would be my prediction had my input image been given with a different head pose?". Moreover, this highlights the capabilities for detection and mitigation of biases, through simple queries, such as "Had the ethnicity been different, would it represent a worse prediction?". And we do not need to directly compute these values, as we can directly try and find the relationship between skin tone and certain predictions.

Locating problems within a deep learning model is a particularly difficult task. However, when the causal model of that neural network is known, it is possible to quickly debug it and look for unwanted/spurious relationships through the comparison of that causal model and a partial model of the "world". And while some relationships might be spurious in real life, they can represent causal knowledge to the learnt deep neural network. Hence, highlighting information wrongly acquired by that same network.

3.1.4 Work in xAI for Biometrics

Recently, the number of applications of methods that merge techniques of explainable artificial intelligence and biometrics has grown. As such, biometrics is a great case study for analyzing these techniques and how they fit within the proposed taxonomy. In this review, we analyzed 81 different papers published between 2018 and 2024. These studies have focused on distinct modalities and applications of biometrics with varying degrees of complexity. The curation of the paper list follows two main steps. First, a comprehensive keyword-guided search was conducted on papers indexed on Google Scholar. The keywords used are related to biometrics and explainable artificial intelligence. Subsequently, we manually scrapped the papers published at the main Computer Vision and Biometrics conferences, which were selected from the Google Scholar conference ranking¹. From the selected research documents, 53.1% were published between 2022 and 2023, highlighting progress and interest in the field. In comparison, only 37.0% of these studies were published between 2020 and 2021. A small proportion (approximately 9 %) was published between 2018 and 2019. As shown in Table 3.1, the face is the most frequent modality (76.5% of the papers), with applications in recognition, morphing, presentation, and deepfake attack detection. Owing to its easier acquisition and familiarity, the face is an obvious target for explainability research. Iris recognition and presentation attack detection have also had several studies focusing on bringing explainability to these methods.

Nearly half of the studies analyzed did not include any mechanism to promote interpretability in their designs. They rely exclusively on explanation-based methods that are used in the vast majority of situations to produce visualizations and observational information of a potential explanation for a prediction. Thus, 37 of the 81 papers were placed on the first rung of the xAI ladder. For instance, Trokielewicz *et al.* published studies on recognition [485] and presentation attack detection [484] on cadaver irises. Both approaches rely exclusively on Grad-CAM to produce visualizations, similar to Hu *et al.* [215] for regular iris recognition. Nonetheless, not all visualizations are equal, and the approach used is an important element. Fu *et al.* [167] introduced differential activation maps that were tweaked to explain demographic biases from the perspective of face recognition systems. The activation maps were grouped according to the ethnicity or gender of the individual, and the mean and variation information about them was extracted. These were used to construct a final map that illustrates the deviation from the Caucasian class.

Although extensively used for two-dimensional inputs, visualization methods are not restricted and can be applied to different inputs. For instance, Pinto and Cardoso [403] studied different components of an ECG signal to determine the most relevant component for a recognition task. Four distinct visualization methods were used to dissect ECG signals. Similarly, Lim *et al.* [301] proposed the use of visualization techniques, such as Deep Taylor, to study the differences in rhythm and pitch between DeepFake and human voices, and LRP to uncover characteristics not seen through human visual perception. Deepfake applications rely mostly on approaches that belong to the first rung. Mazaheri and Roy-Chowdhury [331] proposed the use of CAM-based methods

¹https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computervisionpatternrecognition

to detect inconsistencies in a manipulated image. Their method achieved state-of-the-art results for both DeepFake and face-manipulation detections. In the attack detection domain, Seibold *et al.* [440] and Xu *et al.* [522] presented methods for detecting morphing and deep fake attacks, respectively. The former proposed Focused Layer-wise Relevance Propagation (FLRP), which is an extension of Layer-wise Relevance Propagation (LRP), to explain the predictions of the network to a human. This new method focuses on intermediate layers to generate a visualization map so that it can capture the artefacts that characterize morphing attacks. Its use has proven to be particularly beneficial in situations where the backbone neural network displays uncertainty. The novel contrastive solution proposed by Xu *et al.* [522] for deep fake detection can be generalized for distinct generators of attacks. More importantly, through the use of heat maps and Uniform Manifold Approximation and Projections (UMAP), it is possible to provide a simple visualization of the model decision. Genovese *et al.* [179] introduced a method to simulate facial ageing. While the proposed generative method does not include any interpretability element, the authors proposed a novel method to explain the behaviour of generative adversarial (GAN) models. The method known as the Cross-GAN Filter Similarity Index (CGFSI) uses extracted filters from two generators, classifies them as belonging to the same or different categories and finally gives a similarity score for both generators. Sharma and Ross [450] tackled the problem of iris presentation attack detection (PAD). They proposed the integration of an Optical Coherence Tomograph (OCT) within the systems that aim to detect presentation attacks. Besides the promising results, the performance of the model was studied from the explainable artificial intelligence perspective. The t-distributed stochastic neighbour embedding (t-SNE) and the Grad-CAM algorithm are used to create a visualization of the prediction in a smaller latent space and to generate a heatmap to spatially highlight the most relevant areas for the prediction.

In contrast to CAM-based methods, LIME-based approaches intervene in the input through slight variations. These interventions allow for an effective assessment of the importance of certain features and how local changes in these features affect the final performance of the system. Although not as effective in images as it is in explaining interpretable features, Rajpal *et al.* [419] proposed an LIME-based explanation for face recognition systems. This approach was tested on four distinct neural network architectures and three different datasets. However, the proposed methodology is not comparable to the current literature because of the chosen architectures and experimental design. Both Bousnina *et al.* [43] and Lu and Ebrahimi [317] explored RISE-based [398] to explain face verification algorithms. The former generates randomly masked samples from the probe image and compares all samples to the reference image. Based on these comparisons, similar or dissimilar regions were highlighted for positive and negative matches, respectively. In the latter work, Lu and Ebrahimi [317] proposed a variation of Rise called S-RISE. Their approach consisted of comparing an unmasked probe with a masked reference image and a masked probe with an unmasked reference image. Based on these comparisons, it was possible to generate explanations that were evaluated quantitatively through insertion and deletion processes. Huber *et al.* [225] expanded on patch-replacement approaches as a form of intervention on the input. Their work focused on distinguishing between the contributions to the decision of the dif-

ferent features in the feature vector, and based on those individual contributions, it was possible to generate explanations that aligned with the final prediction. For positive matches, only similar features between the two images matter to find an explanation, whereas in negative cases, only dissimilar features matter. They further use this information and propose a new dataset called Patch-LFW, which allows them to assess explanations through the questions "What would be my explanation if I swapped this part of the face?". A similar patch-replacement strategy was studied by Knoche *et al.* [276], who also proposed an intervention on the prediction approach that was further expanded in Chapter 7. The latter approach generates an interpretable and probabilistic score for the final decision of a verification system based on genuine and impostor distributions.

Interventions on the input have also been shown in fingerphoto verification and post-mortem iris recognition. Ramachandra and Li [420] tackled the complex task of fingerprint recognition through a mobile camera photo, and additionally, to the state-of-the-art performance, this study included several visualization techniques, such as Grad-CAM and gradient attribution, as well as interventions on the input such as LIME and Occlusion Sensitivity [539]. Boyd *et al.* [58] expanded the explainability approaches to post-mortem iris recognition proposed by Kuehlkamp *et al.* [288] that is aware of the areas that are altered by the lack of blood flow. Their new proposal included human annotators that indicated which areas of the two distinct iris images matched. Based on these annotations and features, the final algorithm was able to determine which areas were a strong match between two images, in accordance with human beliefs. Leveraging this information allows this method to be perfectly aligned with the expectations of the final users, which are forensic examiners.

While the majority of the previous approaches do not necessarily reflect a change in the training process of the deep neural network, some other methods have focused on introducing regularization, priors, and assumption constraints to their learning rule. Specifically, in Masked Face Recognition, acknowledging that the mask is not relevant to the final prediction, the losses should constrain the model output feature vector to be invariant to a potential mask. Huber [221] presented an approach in this direction for designing a knowledge-distillation framework. Similarly, Franco *et al.* presented two studies [162, 163] that targeted the mitigation of biases within the face recognition system. Following the Demographic Parity constraint [70] the authors designed several regularization terms based on the Tikhonov/Ridge regularizer that were initially introduced by ONeto *et al.* [385]. Considering that a face has two distinct components, Han *et al.* [201] developed a training rule in which a face generates a kernel that is subsequently disentangled into two distinct kernels. The first commonality component represents the shared characteristics among the subjects in the reference set used for optimization. The second component represents the special features of a certain person and is composed of the residuals between the difference of the initial kernel and the commonality kernel. This method showed a significant improvement over conventional CNNs on the three different datasets. Yin *et al.* [535] presented one of the first approaches for interpretable face recognition. The proposed algorithm was optimized using two novel losses: Feature Activation Diversity (FAD) loss and Spatial Activation Diversity (SAD) loss. The former promotes filter robustness against occlusions, whereas the latter encourages the inclusion of

semantic information. Finally, predictions can be analyzed by studying the locations of the filter responses.

Williford *et al.* [512] presented one of the first face recognition approaches that included interventions on both the input and predictions as well as a trained rule prior. They investigated a new form of explanation, based on triplets and an inpainting game. The main goal was to identify regions that were similar between the probe and the positive and dissimilar regions between the probe and the negative. Furthermore, during training, two other approaches to promote interpretability were studied: subtree Excitation Backpropagation (EBP) and Density-based Input Sampling for Explanation (DISE). The former computes the gradients of a triplet for every node of the original network and ranks them into a subtree. These are subsequently used to construct a saliency map that supports these explanations. The latter leverages these outputs to perform prior-based sampling of perturbations in the input image. This is an extension of a previous method known as Randomized Input Sampling for Explanation (RISE) [398], which did not leverage any prior for the sampling. Later, Liu *et al.* [304] addressed the heterogeneous face recognition problem using adversarial training to disentangle different face modalities. Through this approach, their method, called heterogeneous face interpretable disentangled representation (HFIDR), can introduce semantic and interpretable information in the latent space. In this representation, information about identity and information related to modality are explicitly separated. Thus, conditioned on prediction, the author managed to reconstruct the original image in a different modality space. Recently, Terh r st *et al.* [477] used pixel quality indicators to explain face recognition systems. One of the most novel proposals of this approach is the adaptability of the method to different face recognition systems via the propagation of an input image 100 times through stochastic variation of the original network. The authors also proposed inpainting strategies to mitigate the impact of low-quality pixels in a face image. Although not completely interpretable pixel-quality can refer to obstacles in the recognition process, such as occlusions.

Early studies on explainable AI applied to presentation attack detection (PAD) aimed to understand what makes an attack different from a genuine image. While it might be easier to check, given an appropriate clue, if an image is indeed an attack, it is not trivial to define the differences between attacks and genuine images. Jourabloo *et al.* [254] attempted to model spoof information induced by each attack. For this, it was necessary to make certain assumptions regarding the properties of noise, such as its ubiquity. Through this method, it was possible to decompose the original image into noise and facial features. Liu *et al.* [310] approached the facial PAD problem from a similar perspective. Their disentanglement approach was flexible enough to not only detect spoof and live samples through the disentangling of spoof traces but also to reconstruct a live version of that same input or synthesize a new spoof image with the same style and a different reference live image. Wang *et al.* [507] method is another example of this type of approach. By decomposing the input image into a multi-scale frequency image with 12 channels, Fang *et al.* [157] achieved state-of-the-art results in cross-dataset scenarios. This ability to generalize aligns with the fact that the frequency representation is invariant to the dataset and that attacks display distinct frequency patterns. The architecture of the detection network was further designed using a hier-

archical attention mechanism to support the detection of these different patterns [185]. Yang *et al.* [531] proposed a method that works with both temporal and spatial features. Specifically on the spatial features, the author splits a single image into several patches of the same image, which are processed by a shared-weight neural network. Subsequently, an attention module weighs the importance of each patch and provides a final classification. This training strategy favours an interpretable notion of the patches that influence the decision. Similarly, Bian *et al.* [38] introduced features that can be interpreted by design. The proposed method was designed to work with both facial and environmental cues. However, the final decision uses a concatenation of interpretable and obscure features, making it difficult to trace the motives for the decision. Pan *et al.* [387] proposed a framework with an integrated attention mechanism and generator for Grad-CAM visualization. The authors proposed the use of visualizations to optimize the training of the attention mechanism to which they designed a set of experimental studies and comparisons. Additionally, their method can generate textual explanations from a set of predefined text options.

Chen and Ross [83] presented an attention-based approach to tackle iris presentation attack detection. When compared to previous attention-based their approach achieved significant improvements, which are reinforced by Grad-CAM generated visualizations. On the other side, Fang *et al.* [155] developed a patch-based approach assisted by attention mechanisms to solve a similar task. The proposed approach has a loss that works at the pixel level as an auxiliary to the main task of detecting presentation attack inputs. As such, it is also possible to obtain a pixel score or, if intended, a patch score considering a group of pixels. It works especially well when combined with spatial attention blocks, as demonstrated in the Score-CAM visualizations. Joshi *et al.* [253] leveraged the stochastic nature of Monte Carlo dropout to generate uncertainty maps on the segmentation of the fingerprint regions of interest.

In fact, despite the visible growth of xAI applied for biometrics, some important elements are often discarded or ignored. No work explored the effects of causality on the production of explanations, nor the fact that rarely there is only one interpretation for a given explanation/visualization. Moreover, there is no focus on abnormal or rare events that can be extremely useful to produce explanations. The usage of semantic explanations and the analysis of the human-friendliness of the explanations produced are nearly nonexistent. Thus, it is important to look into what is being done in other areas of computer vision regarding xAI, and how that work can be transposed or adapted into biometric applications. Furthermore, the implementation of these methods must be applied across the most diverse biometric traits and tasks. There is still a long path ahead for xAI methods applied to biometrics, and to introduce some consensus and uniformity between the needs of these methods. Despite a categorization of methods in the different rungs, not all the methods in a given rung are equally explainable, and it is important to pay attention to the methods that focus on going beyond simpler strategies, such as LIME.

Table 3.1: Methods that leverage explainability methods to inspect the predictions of biometric systems. Some of these methods, promote interpretability through the usage of certain techniques, whereas others focus mostly on explaining the predictions. Each method is placed in the respective rung in the xAI ladder. For the methods of the second rung they are further categorised into Model Prior (M. P.), Training Rule Prior (T. R. P.), Intervention on Input (I. I.) and Intervention on the Prediction (I. P.).

Method	Year	Modality	xAI Ladder	Approach to xAI
Face Recognition				
Huber <i>et al.</i> [225]	2024	Face Verification	Rung 2 - I. I. + I. P.	Decision-based Patch Replacement and Individual Dimension Contribution to Decision
Li [298]	2023	Face Recognition	Rung 2 - M. P.	Biologically Inspired Architecture + Attribute Weight for Decision
Correia <i>et al.</i> [99]	2023	Face Verification	Rung 2 - M. P.	Attention + Explainability Masks Construction
Bousnina <i>et al.</i> [43]	2023	Face Verification	Rung 2 - I. I.	RISE-Based Similarity/Dissimilarity Maps
Knoche <i>et al.</i> [276]	2023	Face Verification	Rung 2 - I. I. + I. P.	Patch Replacement
Lu and Ebrahimi [317]	2023	Face Recognition	Rung 2 - I. I.	S-Rise
Rajpal <i>et al.</i> [419]	2023	Face Recognition	Rung 2 - I. I.	LIME Maps
Neto <i>et al.</i> [372]	2023	Face Recognition	Rung 2 - I. P.	Interpretable Match Score
Rodriguez <i>et al.</i> [428]	2023	Quality Face Recognition	Rung 2 - T. R. P.	Environmental Attributes
Terh�r�st <i>et al.</i> [477]	2023	Face quality for Face Recognition	Rung 2 - I. I. + I. P.	Pixel-level Quality Maps
Fu <i>et al.</i> [167]	2022	Face Recognition Bias	Rung 1	Differential Activation Maps
Han <i>et al.</i> [201]	2022	Face Recognition	Rung 2 - T. R. P.	Personalized Kernels + Feature Maps
Roy <i>et al.</i> [430]	2022	Heterogeneous Face Recognition	Rung 2 - M. P. + I. I.	Invariant Features
Winter <i>et al.</i> [513]	2022	Face Recognition	Rung 2 - T. R. P. + I. P.	Local Patch-based Features with Eexplainable Boosting Machine (EMB)
Mery <i>et al.</i> [337]	2022	Face Verification	Rung 1	Saliency maps AVG contours
Fu <i>et al.</i> [166]	2022	Face quality for Face recogniton	Rung 1	Score-CAM[501] + Activation Maps
Franco <i>et al.</i> [162]	2022	Face Recognition	Rung 2 - T. R. P.	Tikhonov regularizers
Neto <i>et al.</i> [362, 360]	2021	Masked Face Recognition	Rung 2 - T. R. P. + I.I.	Mask-Invariant Loss + Grad-CAM Visualisation
Huber <i>et al.</i> [221]	2021	Masked Face Recognition	Rung 2 - T. R. P. + I.I.	Mask-Invariant Knowledge Distillation
Franco <i>et al.</i> [163]	2021	Face Recognition	Rung 2 - T. R. P.	Tikhonov regularizer
Liu <i>et al.</i> [304]	2021	Heterogeneous Face Recognition	Rung 2 - T. R. P. + I. P.	Modality/Identity Disentangled Representation + Cross-modality Synthesis
Anghelone <i>et al.</i> [21]	2021	Face Recognition	Rung 1	Heatmap-based Visualisation
Jiang and Zeng [246]	2021	Face Recognition	Rung 2 - T. R. P.	Face Components Representation and Self-Attention based Reconstruction + Components Similarity Matrix

Table 3.1: Methods that leverage explainability methods to inspect the predictions of biometric systems. Some of these methods, promote interpretability through the usage of certain techniques, whereas others focus mostly on explaining the predictions. Each method is placed in the respective rung in the xAI ladder. For the methods of the second rung they are further categorised into Model Prior (M. P.), Training Rule Prior (T. R. P.), Intervention on Input (I. I.) and Intervention on the Prediction (I. P.).

Method	Year	Modality	xAI Ladder	Approach to xAI
Williford <i>et al.</i> [512]	2020	Face Recognition	Rung 2 - I.I. + T. R. P. + I. P.	DISE + subtree EBP + Inpainting game
Zee <i>et al.</i> [538]	2019	Face Recognition	Rung 1	Guided Backpropagation & CAM
Yin <i>et al.</i> [535]	2019	Face Recognition	Rung 2 - T. R. P. + I. I.	Spatial and Feature Activation Diversity Losses
Face Morphing				
Caldeira <i>et al.</i> [67]	2023	Face Morphing Detection	Rung 2 - M. P. + T. R. P.	Interpretable Score + Identity Disentanglement
Dargaud <i>et al.</i> [113]	2023	Face Morphing Detection	Rung 1	PCA-based Visualisation
Dwivedi <i>et al.</i> [143]	2023	Face Morphing Detection	Rung 1	Ensemble of Visualisations
Neto <i>et al.</i> [366]	2022	Face Morphing Detection	Rung 2 - T. R. P.	Orthogonal Identity Disentanglement
Myhrvold <i>et al.</i> [353]	2022	Face Morphing Detection	Rung 1	LRP
Seibold <i>et al.</i> [440]	2021	Face Morphing Attack Detection	Rung 1	Focused LRP
Face PAD				
Prasad <i>et al.</i> [407]	2023	Face PAD	Rung 2 - T. R. P.	Intermediate Depth Estimation
Pan <i>et al.</i> [387]	2022	Face PAD	Rung 2 - T. R. P.	Grad-CAM + Textual Explanations
Fang <i>et al.</i> [157]	2022	Face PAD	Rung 2 - T. R. P.	Multi-Scale Frequency Decomposition + Hierarchical Attention + t-SNE
Bian <i>et al.</i> [38]	2022	Face PAD	Rung 2 - T. R. P.	Auxiliary Learning with Interpretable Cues
Wang <i>et al.</i> [507]	2022	Face PAD	Rung 2 - T. R. P.	Live Features Disentanglement + t-SNE [489]
Sequeira <i>et al.</i> [445]	2021	Face PAD	Rung 1	Grad-CAM Visualisation
Chen <i>et al.</i> [84]	2021	3D Mask Face PAD	Rung 1	Grad-CAM Visualisation and t-SNE
Neto <i>et al.</i> [359]	2021	Face PAD	Rung 1	Grad-CAM Visualisation
Aghdaie <i>et al.</i> [9]	2021	Face PAD	Rung 2 - T. R. P.	Attention + Attention Maps Visualization
Yu <i>et al.</i> [536]	2020	Face PAD	Rung 2 - M. P. + T. R. P.	Central Difference Convolution + Depth-Map Estimation
Pinto <i>et al.</i> [402]	2020	Face PAD	Rung 1	Grad-CAM Visualisation
Wang <i>et al.</i> [509]	2020	Face PAD	Rung 2 - T. R. P.	Temporal Depth-Map Estimation + t-SNE
Shao <i>et al.</i> [448]	2020	Face PAD	Rung 2 - T. R. P.	Meta-learning with Depth Regularization + Attention Map Visualization
Liu <i>et al.</i> [310]	2020	Face PAD	Rung 2 - T. R. P.	Spoof-traces disentanglement
Yang <i>et al.</i> [531]	2019	Face PAD	Rung 2 - T. R. P.	Patch-based Attention + Grad-CAM + t-SNE
Jourabloo <i>et al.</i> [254]	2018	Face PAD	Rung 2 - T. R. P.	Spoof Noise Modeling + t-SNE

Table 3.1: Methods that leverage explainability methods to inspect the predictions of biometric systems. Some of these methods, promote interpretability through the usage of certain techniques, whereas others focus mostly on explaining the predictions. Each method is placed in the respective rung in the xAI ladder. For the methods of the second rung they are further categorised into Model Prior (M. P.), Training Rule Prior (T. R. P.), Intervention on Input (I. I.) and Intervention on the Prediction (I. P.).

Method	Year	Modality	xAI Ladder	Approach to xAI
Liu <i>et al.</i> [309]	2018	Face PAD	Rung 2 - T. R. P.	rPPG Signal Extraction
Face DeepFakes				
Mathews <i>et al.</i> [328]	2023	Face DeepFake Detection	Rung 1	Grad-CAM Visualisations
Silva <i>et al.</i> [456]	2022	Face DeepFake Detection	Rung 1	Attention Maps
Mazaheri and Roy-Chowdhury [331]	2022	Face Manipulation Detection	Rung 1	CAM
Korshunov <i>et al.</i> [279]	2022	Face DeepFake Detection	Rung 1	t-SNE
Xu <i>et al.</i> [522]	2022	Face DeepFake Detection	Rung 1	UMAP and Heatmap Visualisation
Face Emotion/Expression Recognition				
Nam <i>et al.</i> [355]	2023	Face Expression Recognition	Rung 2 - T. R. P.	Spatio-Temporal Attention Maps + Facial cues
Rathod <i>et al.</i> [421]	2022	Children Emotion Recognition	Rung 1	Grad-CAM + Grad-CAM++ + SoftGrad
Cesarelli <i>et al.</i> [76]	2022	Face Emotion Recognition	Rung 1	Grad-CAM Visualisations
Zhu <i>et al.</i> [556]	2022	Face Emotion Recognition	Rung 1	LRP
Araf <i>et al.</i> [25]	2022	Face Emotion Recognition	Rung 1	Grad-CAM Visualisations
Fingerprint PAD/Recognition				
Rai <i>et al.</i> [416]	2023	Fingerprint PAD	Rung 1	Heatmap Visualisation
Ramachandra and Li [420]	2023	Fingerphoto Verification	Rung 2 - I. I.	Grad-CAM + Occlusion Sensitivity Maps + LIME + Gradient Attribution
Chowdhury <i>et al.</i> [95]	2020	Fingerprint Recognition	Rung 1	Grad-CAM Visualisation
Iris Recognition				
Boyd <i>et al.</i> [58]	2023	Post-mortem Iris Recognition	Rung 2 - I. I.	Visual Patch-matching
Kuehlkamp <i>et al.</i> [288]	2022	Post-mortem Iris Recognition	Rung 1	Segmentation Masks + CAM Visualisations
Hu <i>et al.</i> [215]	2020	Iris Recognition	Rung 1	Grad-CAM Visualisation
Trokielewicz <i>et al.</i> [485]	2018	Cadaver Iris Recognition	Rung 1	Grad-CAM Visualisation
Iris PAD				
Sharma and Ross [450]	2021	Iris PAD	Rung 1	Grad-CAM Visualisation and t-SNE
Fang <i>et al.</i> [155]	2021	Iris PAD	Rung 2 - T. R. P.	Patch-based Supervision + Attention + Score-Weighted CAM [547]
Chen and Ross [83]	2021	Iris PAD	Rung 2 - T. R. P.	Channel and Position Attention + Grad-CAM
Sharma <i>et al.</i> [449]	2020	Iris PAD	Rung 1	Grad-CAM Visualisation and t-SNE

Table 3.1: Methods that leverage explainability methods to inspect the predictions of biometric systems. Some of these methods, promote interpretability through the usage of certain techniques, whereas others focus mostly on explaining the predictions. Each method is placed in the respective rung in the xAI ladder. For the methods of the second rung they are further categorised into Model Prior (M. P.), Training Rule Prior (T. R. P.), Intervention on Input (I. I.) and Intervention on the Prediction (I. P.).

Method	Year	Modality	xAI Ladder	Approach to xAI
Trokielewicz <i>et al.</i> [484]	2018	Cadaver Iris PAD	Rung 1	Grad-CAM Visualisation and Guided Backpropagation
Others				
Diaz <i>et al.</i> [132]	2023	Signature Verification	Rung 2 - M. P. + T. P.	Universal Background Model(UBM) + Explainable Features
Aquino <i>et al.</i> [24]	2022	Accelerometer User Identification	Rung 1	Grad-CAM
Lim <i>et al.</i> [301]	2022	Voice DeepFake Detection	Rung 1	Deep Taylor + LRP
Algermissen and Hörnlein [15]	2021	Footstep Person Recognition	Rung 1	Grad-CAM
Alshazly <i>et al.</i> [19]	2021	Ear Recognition	Rung 1	Guided Grad-CAM
Chandaliya and Nain [77]	2021	Face Age Estimation	Rung 2 - T. R. P.	Attention
Joshi <i>et al.</i> [253]	2021	Fingerprint ROI Segmentation	Rung 2 - T. R. P. + I. P.	Monte Carlo Dropout + Uncertainty Map
Pinto and Cardoso [403]	2020	ECG Biometric Identification	Rung 1	Gradient-SHAP + DeepLIFT + Saliency Maps
Ahmed <i>et al.</i> [11]	2020	Ethnicity Estimation	Rung 1	Grad-CAM
Genovese <i>et al.</i> [179]	2019	Face Aging	Rung 1	Cross- GAN Filter Similarity Index

3.1.5 Work in xAI computer vision that can benefit biometrics

The scenario of interpretable methods for biometrics has grown and highlights an increasing research interest in this domain. However, other computer vision domains had the opportunity to further develop and deeply explore different approaches to xAI. These approaches, independently of the rung where we place them, are, in most cases, yet to be applied to biometrics. Given its impact on medical image and other critical domains, it is undeniable that they might present themselves as attractive research directions for anyone looking towards improving the explainability of biometric systems. Hence, the following paragraphs delve into some roads already crossed in the computer vision literature. As seen in Table 3.1, one of the most common training rule constraints on second-rung approaches is the usage of attention mechanisms. These mechanisms have the advantage of being more interpretable, as they provide the importance of each element for their prediction. One of the most commonly used attention-based systems is the Transformer [494]. This architecture led to advances in both computer vision and natural language processing domains. However, it is usually data-hungry, expensive to run on inference and not yet ready for embedded devices. Additionally, the recent usage of transformer-based foundation models further highlights the course of complexity and the opaqueness of these systems as they scale.

In 2019, Chen *et al.* proposed a novel deep neural network, *prototypical part network* (ProtoPNet), that contains an inner reasoning process that may be considered interpretable due to its transparency [82]. This model is trained in an end-to-end fashion and it is optimized to learn *prototypes* (*i.e.*, in this case, prototypes are latent representations) that are specific to a given class. To achieve this, the authors proposed to use generalized convolution by including a *prototype layer* that computes the squared L_2 distance instead of the inner product. Moreover, they also proposed to constrain each convolutional filter to be identical to some latent training patch. And thus, allowing us to interpret the convolutional filters as visualizable prototypical image parts. When the model is shown a new image, it dissects the image to find parts that may look like a prototypical part of some class and makes its prediction based on a weighted combination of the similarity scores between parts of the image and the learned prototypes. This approach is fitting to identification and authentication problems in the biometrics domain. Due to the capability to create patch-related explanations, it becomes easier to provide an understandable explanation to the user. Besides, this pipeline could help algorithm developers to assure that their models respect the “right to explanation” [259] in the sense that this reasoning process is transparent to the end-user. Additionally it could be possible to find common and identity-related face elements. On the strongest challenges of these methods is their classification-based design, which makes them inappropriate for open-set tasks, such as face verification. Other works have also explored the use of prototypical networks [203, 459, 557, 357, 338].

Inspired by the Recognition-By-Components cognitive psychology theory proposed by Biederman, Saralajew *et al.* [437] proposed an architecture called Classification-By-Components network (CBC). The proposed network is optimized to learn generic characterization components for object detection. Moreover, it follows a class-wise reasoning approach with three distinct types

of reasoning (positive, negative and indefinite) that is learnt to solve the classification task. The combination of the reasoning types branch into a probabilistic classifier and the decomposition of objects into generic components provides a precise interpretation of the decision process. This approach is even shown to be able of explaining the success behind an adversarial attack, which can be particularly useful for critical biometric systems. It could be explored if other forms of attacks to biometric systems can benefit from this approach. Other relevant application could be related to an extension of patch-based iris recognition work [58].

Qi *et al.* [409] proposed a novel Explanation Neural Network (XNN), which maps the output embedding into an explanation space. Their XNN algorithm is called Sparse Reconstruction Autoencoder (SRAE), and the produced explanation embedding is capable of retaining the prediction power of the original feature embedding. This explanation space can be understood and visualized by humans when employed a visualization approach proposed by the authors. Recently, Utkin *et al.* [488] proposed a similar approach based on autoencoders. Current works in biometrics, such as the one presented by Caldeira *et al.* [67] could leverage this approach to further boost the explainability of the autoencoder in their proposed approach.

In 2021, Booth *et al.* [41] presented a novel model inspection method called BAYES-TREX. Instead of relying on potential abnormal behaviour in the test set, this method finds in-distribution examples with specified prediction confidence. This probabilistic method can reveal misclassifications that have high confidence and display the boundaries between the classes through ambiguous samples. Moreover, it can also predict the behaviour of the network when exposed to a sample from a novel class. The importance of the prediction score in verification tasks is of higher relevance. Some methods present scores that are difficult to interpret, and this type of approaches aligns with the proposal of Knoche *et al.* [276] to create an interpretable matching score.

Aiming at including some semantic meaning into the inspectability methods of deep neural networks, Losch *et al.* [314] proposed Semantic Bottlenecks (SB). These are integrated into pre-trained models and align their channel outputs with visual concepts. The proposed method is rather impressive since even with a reduction of the number of channels by two orders of magnitude, the results are still on par with the state-of-the-art for segmentation challenges. SB have two variations, the unsupervised SBs (USB) and the concept-supervised SBs (SSB). Despite the similar performance, the latter shows a higher degree of inspectability.

Another challenge related to the training of deep neural networks is the learning and proper visualization of concepts that are important for a given task. There are several methods that attempt to see inside the hidden layers of DL models, however, they are applied *a posteriori* (*i.e.*, post-hoc analysis), which may lead to misleading conclusions regarding the properties of the latent space [90]. Therefore, to avoid relying on post-hoc analysis of neural networks, Chen *et al.* [90] proposed a mechanism called *concept whitening* (CW). This mechanism applies a whitening operation to the latent space (*i.e.*, decorrelation and normalization) thus forcing the axes of the latent space to be aligned with known concepts of interest. The authors showed that this approach may be a reliable alternative to the use of batch normalization without hurting predictive performance. This methodology may be of interest in the field of PAD since it could open a different path for

the comprehension of the concepts related to *bona fide* and attack signals. Moreover, it could be used to align with soft-biometrics attributes, which would help the system to provide a semantic and meaningful explanation based on terms that the end user is familiar with.

Due to being, by nature, more interpretable than more complex models, the interest in simple models is gaining traction. However, training these models frequently leads to worse performance. Hence, researchers have investigated the possibility to change how these models learn. One of the most used approaches, is the distillation of knowledge learned by a complex model (teacher) into a smaller one (student) [308, 16]. It has been already been leveraged in biometrics, not as a promoter of interpretability, but as an auxiliary process for a different task [221, 50]. Hence, even if the training of inherently interpretable models is less stable, it is possible to optimise it by gathering information from a more complex, opaque model.

Other methods that are being explored and can contribute to the increase of interpretability of biometric methods include counterfactual explanations and influence functions. And there is some work in that direction [256, 518, 229, 197]. However, they are still growing and maturing. In their current form, they are similar to the naive counterfactuals described in previous chapters as a second rung approach. It is clear that xAI was already properly introduced into computer vision systems. Nonetheless, when comparing the variability of approaches presented in this Section with the ones shown in Table 3.1 that biometric systems are far from being interpretable. Hence, this work also aims at motivating the need to accelerate the research of these methods so that biometric applications can reach the same level of interpretability shown by other computer vision fields.

There has been also attempts to build explainability pipelines capable of handling different types of machine learning models and distinct applications. IBM, presented “The AI 360 Toolkit”², which is one of the most advanced pipelines for model and application agnostic explanations. Despite the potential of these systems, they lack the domain information and design that is frequently needed for complete explanations. Hence, while it can be seen as a good starting point, it must not be seen as the definitive solution.

Conducting a careful analysis of the methods presented in this section, it is possible to categorise them accordingly to the xAI ladder proposed in previous sections. As such, it shows the robustness of this new taxonomy across multiple domains. Nonetheless, it is also visible that the field is still far from being saturated and further efforts are needed to achieve the third rung where researchers can understand the causal relations in their models.

3.2 Illustrating the Benefits of Interpretable Biometrics

Interpretable biometrics are still being used exclusively for academic purposes. However, leveraging these techniques can impact lives outside the academic domain and, for instance, support vulnerable groups against discriminatory behaviour. Its integration with the systems that are currently deployed is crucial to the progression towards a more transparent artificial intelligence.

²<https://developer.ibm.com/articles/the-ai-360-toolkit-ai-models-explained/>

Hence, this Section discusses the situations where interpretable biometric systems should be deployed, the benefits that shall arise from their usage, the limitations of the current approaches, the needs for innovation, and finally, it wraps up with some potential research directions.

3.2.1 Applications

The applications of biometric systems are vast, and thus, the applications of their interpretable counterpart are also vast. This raises the question of how it is possible to create a single interpretable system for all the tasks. Realistically, this is a question that shall not be answered for now. Before being solved holistically, interpretable biometrics must be tackled at each smaller subdomain. Progression into a more general approach occurs iteratively through the explanation of the different methods. It can, for instance, explore the validation of an interpretable approach on several tasks or traits. It is however essential to keep in mind that although it would be desirable to be an interpretable system that fits the most variate scenarios, in practice, it is known that there are no free lunches. Moreover, to achieve the third rung it is important to progress towards more complex methods within the second rung instead of focusing on the more simple approaches seen frequently. Besides the benefits previously mentioned, it is possible to discuss the impact of these systems in more areas. For this analysis, interpretable methods for biometrics are deconstructed into three areas where they can be integrated.

3.2.1.1 Improving evaluation

Metrics have been one of the most crucial elements for the raise of deep learning systems. Not only they are responsible for the understanding of how the systems are performing, but they are also used to establish comparisons between different methods. Moreover, measures such as the accuracy and model's confidence have been the golden standard to increase the public acceptance and deployment of these systems. Metrics are useful for and provide a significant portion of information, nonetheless, they also lack significant details.

Understanding how many predictions were correct and by which margin is needed for a quick analysis. However, a correct prediction might be correct for the wrong reasons, and a wrong prediction might be wrong for even worse motives. Hence, providing a single prediction with no explanation does not suffice to fully understand these hidden motives, it is necessary to show why. As such, the evaluation of deep neural network systems is far from complete, leading to flawed assessment of their performance. Leveraging our knowledge of the causal structure, attained through xAI methods, of the model that we have just built holds a higher potential for performance assessment. It allows for two types of evaluation. Retrospective evaluation where after getting a prediction, we try to find the "cause" of that prediction through our knowledge of the model. And also *apriori* evaluation, where researchers can inspect the learnt rules and infer if they are aligned with the expected rules.

We propose an evaluation of deep learning systems that can be divided into four major focal points as displayed in Fig. 3.7. First, it is essential to quantitatively evaluate our system, and for

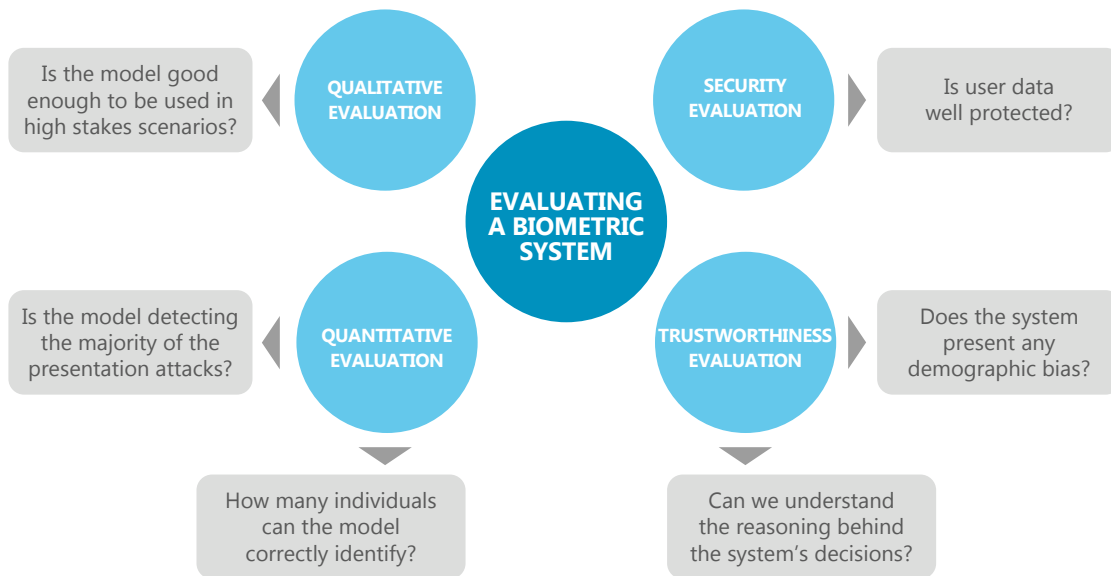


Figure 3.7: Proposed structure for a complete evaluation of the performance of biometric systems at four distinct levels. Quantitative, qualitative, security and trustworthiness evaluation. The latter is focused on the insights provided by explainable artificial intelligence methods.

this we use our typical metrics such as accuracy, F1-Score and others. It is of no use to explain a "dumb" model. Secondly, one should assess in a qualitative manner the performance of its model. In that sense it is important to understand why the model takes the decisions that it takes, how he reaches the decision and which factors are important. This leads to a third evaluation process where we, now fuelled with the knowledge of the "why's" of our system, inspect for bias and trustworthiness problems. It is worth noting that Huber *et al.* [223] has shown that certain Rung 1 approaches suffer from biases themselves. Finally, it is essential to understand the implications of explainability on the privacy of the users and the privacy of the system itself. For instance, Matulionyte [329] explored the implications that having a transparent system has on trade-secrets and how it can be possible to navigate in the trade-off between these two concepts. Additionally, counterfactual explanations have severe risks if not done properly. It could be possible that by proving such an explanation for a biometric recognition system we compromise the safety and privacy of another user.

3.2.1.2 Providing feedback

Whilst expressing different information, metrics and losses share a lot of similar details. The former, as previously described, supports the evaluation of deep learning models. While the latter provides information for the model optimization. Usually, the minimization of the latter improves the metrics. However, as already mentioned, a correct prediction might rely on incorrect cues. Hence, it is also important to have explainable artificial intelligence methods at the optimization level.

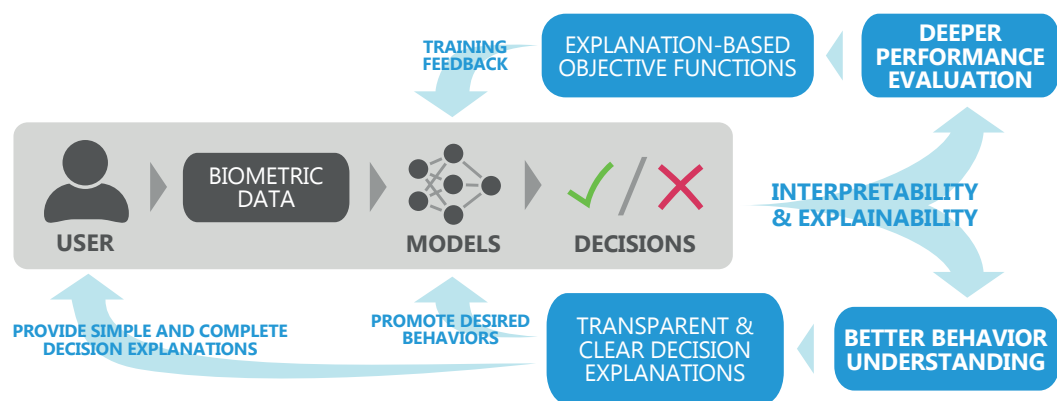


Figure 3.8: Structure of a biometric system that leverages explainable artificial intelligence methods to improve its optimization function and to provide useful feedback to the user.

The integration of xAI and optimization processes can be done in two ways, as shown in Fig. 3.8. First, it is possible to promote certain behaviours or properties through the implementation of constraints to the model. These constraints can be applied to specific layers or to the output of any layer. The second approach is to design objective functions that incorporate feedback from the explanations produced from the predicted batch. This latter approach is significantly more challenging. Nonetheless, there are works aiming to tackle this fusion. For instance, the framework proposed by Pan *et al.* [387] tunes the weights of the attention mechanism through the visual explanations produced in the previous forward pass. Although it is a young field, these strategies have the benefit to directly affect the causal model learnt by the system. And, since we can control these strategies we can promote certain information to be learnt by our system leading to a lower knowledge gap between researcher and deep neural network.

3.2.1.3 Protecting privacy

Some approaches of explainable machine learning focus on image retrieval tasks. For instance, when explaining why two images do not match, it is possible to find examples of mated and non-mated images to create some sort of counterfactual explanation. However, this might expose users' private information. Silva *et al.* [457] leveraged interpretability methods to guide the search for the appropriate image on a similar task on the medical image domain. Further research on the medical domain also focused on the privacy-preserving component of image retrieval [345]. A potential extension for biometric applications was already briefly addressed [346]. We argue that privacy-preserving methods are of extreme importance due to the nature of biometric data. Furthermore, it ensures that users' trust and public opinion, which are crucial for the applicability of biometric systems, remain untouched. Hence, the exposure and potential leakage of important and private information is minimised. The integration of these mechanisms in real-world situations depends on the desired degree of privacy and the efforts required to change the system.

3.2.2 The incompleteness of the current explanations

Rung 1 approaches such as saliency maps have been widely used to determine which parts of the image are being used by DL-models to perform a decision. However, it is important to acknowledge that knowing where the model is looking within the image does not tell the user what it is doing with that part of the image. A representative example of this is that we can obtain saliency maps for different classes, even if they have no relation to the true label. From all the "possible" explanations given to each class, only the predicted class maps are given to the user. This might create a false sensation of confidence in the decision, which might not be reflected in the overall explanations. For instance, the case where the explanations for multiple (or all) of the classes are identical. Moreover, these methods focus on presenting to the user the meaningful pixels, allowing the user to decide the reason behind the prediction based on those pixels. This raises a problem as humans are independent and subjective animals, and thus, from a simple heatmap N different users can find N different motives for the highlight of those pixels (Fig. 3.9). In a sense, Grad-CAM and similar methods give a small hint to the user regarding to where the model was looking at. The reason why those were the meaningful pixels remains unanswered and the prediction remains unexplained. Still from Figure 3.9, the model might have hidden biases against a demographic group, which are not carefully described by the observational explanation.

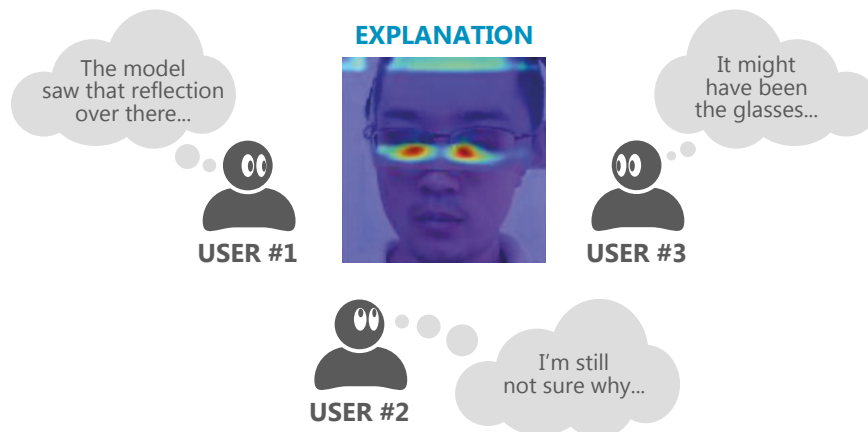


Figure 3.9: Simple draft of a diagram exemplifying how Grad-CAM output as heatmap-based images can be interpreted in a different manner by humans.

Several models have been designed to generate textual explanations as part of their training process [181]. This approach has demonstrated interesting results in systems of visual question answering [22] and in algorithms for image classification [210]. To achieve human-understandable explanations, the modules that generate explanations are trained on large data sets of human-written explanations, thus generating models that explain their decisions using natural language [211]. This approach has been extended to multi-modal explanations, in [388]. In this use case, the model has an extra module (based on attention mechanisms) that generates visual explanations. Rio-Torto *et al.* [424] recently explored parameter efficient strategies to generate textual explanations for chest x-rays, leveraging the recently popular transformer architecture. There is, however, an

important detail that we should take into account: the explanations are generated based on the decision, after the decision of the network has already been made. Hence, given that the explanations are learned in a supervised manner and that the generated explanations are conditioned by the output of the network, should we consider these as real explanations? What differs these models from general classification models? How can we be sure that the textual explanation that was generated truly represents the reasoning process used by the model to output a prediction? Additionally, the current concern regarding hallucinations of large language models highlights some potential problems of explaining an obscure system through another obscure system [525, 245].

3.2.3 Unveiling biases

Biometric systems have at their disposal inputs that frequently encode demographic information or sensitive attributes. Not only that, but it has been shown that these attributes are also encoded in the latent space of these models. And while this information is useful for increased performance, one must be careful with the way it is used. On very important topic, especially in face recognition, is the measurement and mitigation of racial bias. This bias is characterised by an intra-ethnicity profile, in other words it relates to the capability of correctly matching the individuals within an ethnic group. However, if these embeddings are used for different tasks, for instance, as a secondary input to a credit approval system, the hidden demographic information can lead to an increased biased inter-ethnicity hindering the necessity to explain the information contained in the latent space.

While they can be measured, it is hard to directly infer the existence of biases simply by analysing the trained model. As such, several models are flagged as biased several days, weeks or even months after being released. This further reinforces the need for explainable AI, and through the proposed taxonomy we can see that the more causal relations of a model we understand, the easier it is to compare them to the causal relations of the *real world*. Through this comparison we can easily find relations that are not aligned and that could represent a source of bias, such as the usage of the skin tone to predict crime re-incidence or credit assessment. Kei *et al.* also argues that explanations can provide much-needed help in this matter [266]. It is important to detect and act quickly on those cases. Frequently edge cases are related to under-represented minorities in the dataset [62]. Moreover, some datasets with specific data from some minority include lower-quality samples [491]. Hence, they are the most exposed to potential harms by algorithmic mistakes. Explanations can serve as a tool to mitigate these nefarious effects and impel machine learning into fairness [138]. Terhörst *et al.* [477] and Boutros *et al.* [52] are examples of these quality assessments that can be used to construct balanced datasets.

With the rise of deep learning-based biometric systems developed with direct applications to police forces and law enforcement, unveiling bias becomes mandatory [17, 408, 6, 171]. Hence, these use cases have a high requirement for appropriate explanations and not incomplete ones. For instance, if a face presentation attack detection algorithm suffers from a demographic bias that is prominent on darker skin tones, it cannot be detected by visualization methods. Nor can it explain that the skin tone played a role in the decision. Moreover, if the explanation was given to

a human, he would most likely focus his attention on the shape of the areas around the highlighted pixels, search for blurriness or artifacts. Hence, for the safe deployment of these systems one must attempt to learn the hidden rules that guide the model towards a decision. However, despite not being a complete explanation nor unveiling the hidden mechanisms of a deep neural network, rung 1 approaches are still useful for preliminary verification of the deep learning system.

3.3 Conclusion: In which direction should we look in the future?

Throughout this Chapter we have covered several topics in explainable AI. More importantly, we have covered a novel taxonomy for explainable artificial intelligence that benefits from the experience and insight of causality. The proposed taxonomy borrows concepts from the Ladder of Causality, leading to our proposed Ladder of xAI. In this, we have proposed a division of xAI method in three different rungs, where we consider that every deep learning model learnt has an inherent causal model that can represent its inner rules. The first rung, considers the simplest approach to xAI, visualisation methods, which bring observational information to the user. However, these methods are the equivalent of correlation for causality, as they answer a different question and are not guaranteed to be related to the model inner workings. On the second rung the number of different approaches grows significantly. First, we have approaches based on inserting prior to either the model or the training rule. Model priors, due to their strictness partially reveals the final causal representation of the learnt model, whereas training priors, due to its relaxed nature leads to a less effective revelation of that same model. Additionally, it is possible to make interventions at input and prediction levels. When there is an intervention at the input level, the researcher aims to know "How will this change affect my model predictions?". Future methods on the third rung will allow to answer this question immediately, and further answer to the "Why is this the behaviour of my model?". However, we are still far from reaching the third rung with deep learning-based methods.

The xAI ladder was applied to 81 different research papers published between 2018 and 2024. These papers proposed a myriad approaches from the first two rungs for different biometric modalities and tasks. From Face Recognition to Post-Mortem Iris Recognition and DeepFake Detection, the range of tasks that benefited from explainable approaches is outstanding. Moreover, there seems to be a growth trend. As such, it is important that researchers have a common language to address, categorise and assess explainable methods. Despite the current demonstration and application with biometrics works, the proposed taxonomy is sufficiently general to be applied to any domain of computer vision. Hence, other areas such as medical applications can benefit from this framework and converge to common ground. Additionally, this framework can be extended to fit the necessities of a specific domain as it is flexible and not bonded to a single strategy.

Additionally we have further discussed some points regarding explainable AI and how it could be used to improve different areas of biometrics. These discussed areas are frequently not seen as fields that can benefit from xAI, and for this we decided to illustrate some benefits that can be drawn from xAI approaches. In this sense, we have discussed the evaluation of models, the

feedback given and the privacy protection of both user data and the trade-secrets behind an artificial intelligence model. Furthermore, we have discussed a few topics on unveiling biases with explainable AI. With the proposed taxonomy and strategy to approach xAI, the unveiling of biases begins during development and model assessment, reducing the amount of models deployed with hidden biases.

Chapter 4

Inspecting Background Impact on the Model Behaviour

Author Contributions The research work described in this chapter was conducted under the technical supervision of Ana F. Sequeira and Jaime S. Cardoso. The author of this thesis contributed to this work on the problem conceptualisation, data curation, the preparation and conduction of experiments, the results discussion, and publication writing. Some parts of the chapter were originally published in, or adapted from:

- Neto, P. C., Sequeira, A. F., & Cardoso, J. S. (2022). Myope models-are face presentation attack detection models short-sighted?. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 390-399).
- Neto, P. C., Sequeira, A.F., Cardoso, J.S. (2021). caPAD - A context aware model for face presentation attack detection. *In 27th Portuguese Conference on Pattern Recognition (RECPAD 2021), Portugal*.

4.1 Context and Motivation

Presentation attacks are one of the weaknesses and dangers posed to face recognition systems (FRS). Apart from a few exceptions [260], most of the methods used in presentation attack detection (PAD) rely on tight face crops [273]. This process removes surrounding and contextual information that does not belong to the face. Thus, the information provided as input for these systems is limited, leading to myopic (i.e. short-sighted) face PAD models. This pre-processing step has several advantages, including the capability to process several faces per frame and first generation face PAD methods were designed to rely only on the face region, thus remaining compatible with FRS. However, on the other hand, it removes spatial and contextual information present in the original image. In comparison, the human visual cortex can process this spatial and contextual information to identify some attacks targeting human inspection. In some cases, the cropping of the face can reveal itself as obstacle for a human-based decision, for instance on replay attacks, when the resolution of the replay device is high enough. Similarly, machine vision systems can most likely learn to leverage that information when it is available. Furthermore, they can decide if it is more important to focus on the contextual information or on the face itself.

In this chapter, we aim to understand the behaviour and the decisions of deep learning models applied to PAD, when trained with and without background information. In this sense, we theorize that the reasoning of the model might not converge to the same visual cues, leading to distinct performance, and a variation in the explanations. Hence, we designed a set of experiments, where we perform a comparative study of a state-of-the-art supervised binary classification model and its combination with an adversarial approach on three datasets of face images with and without crops. The latter approach focuses on disentangling relevant information and noise on two separate embeddings, which can be used to reduce noise modelling effects on the generalization of these models. On our evaluation, we selected the ROSE-Youtu [297] dataset due to its attack type variability that is not easily found on other datasets. Leveraging this property, we designed a multi-task model, with and without adversarial training, that is optimized to distinguish between the different attacks, leading to the propagation of more fine-grained information. Finally, we evaluate a novel strategy based on the principles of multiple instance learning, which leads to a smart and dynamic frame selection scheme, where the model is responsible for selecting the frame with the highest chance of containing an attack. Differently from the previous approaches, this method requires more processing power since it goes through all frames in a video.

Analysing the quantitative performance of these different models trained under different settings and datasets provides only information regarding the correctness of the final decision. In fact, it disregards the reasons that led to such decision. In order to properly assess the differences in the model's behaviour under the presence or absence of background, one can resort to methods on the first rung of the xAI Ladder (Chapter 3). These methods highlight the regions of interest of an image, and such regions can be compared for different settings. For instance, "if we increase the proportion of the background, is the correct decision of my model still based on face-related pixels?" is one of the most relevant questions to describe the impact of that additional spatial information. In this direction, we decided to use Grad-CAM++ [79] as our visualization approach.

This study contributes to establishment of methods from the first rung of the xAI ladder as tools to understand, to some extent, differences in the behaviour of distinct models. In this scenario, this explainability approach potentiates the understanding of the background impact on the decision, on the relevant regions and on the overall generalization of PAD methods. We further show that when background is used, artifacts and visual cues, which could be relevant for human-based decisions, become important to these models. With this in mind, the major contributions of this work are: i) the evaluation of a state-of-the-art supervised binary classification model and its combination with an adversarial strategy on two datasets under the presence or absence of background; ii) a multi-task face PAD approach that leverages background and achieves state-of-the-art results on the ROSE-Youtu; iii) a frame selection strategy on the ROSE-Youtu dataset. It was not the focus of this study to see if specific models in the literature perform better with and without background, instead, it focuses on studying a set of different and simple approaches that are effective on PAD in order to analyse the consistency of their quantitative performance and changes on the regions of interest.

Besides this introduction and the conclusion, this chapter contains four major sections. The first, Section 4.2, discusses related work on this topic. It is followed by Section 4.3, which establishes the proposed methodology and experimental setup. Section 4.4 describes the datasets utilised on training and test scenarios. Finally, Section 4.5 details the results and discusses their implications.

4.2 Related Work

Previous works on face PAD do not explore background information. Hence, they frequently integrate a pre-processing step on their pipeline which detects and crops faces on images. This step is conducted through the usage of face detection systems, which can be deep learning-based, such as MTCNN [546] and RetinaFace [125], or traditional techniques, such as Viola-Jones cascade detector [496].

The literature displays a myriad of different solutions to tackle PAD, where one can highlight reinforcement learning approaches [65], 3D-CNNs [296], a two stages approach relying on blinking [202] and several other colour-based methods [297, 42]. Some works included background details [492, 36, 380, 527, 260], however, without performing comparative studies on different models regarding the performance with and without background. This study can be found in other works that leverage conventional machine learning models instead of end-to-end deep learning [528, 278]. In addition, no study performed this comparison through the lens of xAI Ladder’s first rung approach to understand non-quantitative impact on the model’s performance.

Due to the nature of the ROSE-Youtu dataset, which contains three-dimensional and two-dimensional attacks, there are fewer methods tested on this dataset than on others. The variability of attacks included in the dataset significantly increase the difficulty of finding a model capable of performing well on all of them. For this reason, even methods that achieve almost zero error on other datasets, have worse performance on the ROSE-Youtu [141]. To the best of our knowledge, there is no other method inspired by multiple instance learning applied to face PAD. Deepfake detection has benefited from a similar approach [299]. The adversarial approach followed in our experiments was described first at [240] and its capabilities to work with face PAD systems was evaluated later by Jaiswal *et al.* [241].

The utilization of methods from the first rung of the xAI ladder as explainability/visualization tools on face presentation attack detection has been previously explored [444, 446]. This work studied the visual cues that impacted a decision, concluding that even if a model is sufficiently accurate, it can still base its decisions on parts of the image that do not correspond to real face features. On the other hand, the categorisation of an attack does not need to be linked to face features. No other method utilised these visualisation techniques to infer the behavioural changes of a deep learning model applied to PAD. Nor has been explored the visual cues identified by a deep learning model as relevant under the presence of background information.

4.3 Methodology

Despite the varied attacks threatening biometric systems, real-world systems are frequently designed to distinguish between attack and non-attack scenarios, thus, not requiring discrimination between different types of attack. Thus, this problem is normally addressed as a binary classification problem between genuine and impostor images. In this chapter, we discussed and experimented with not only a binary, but also a multi-task setting. It is worth noting that this multi-task approach has no impact on the inference computational requirements. The different approaches are described in the following subsections.

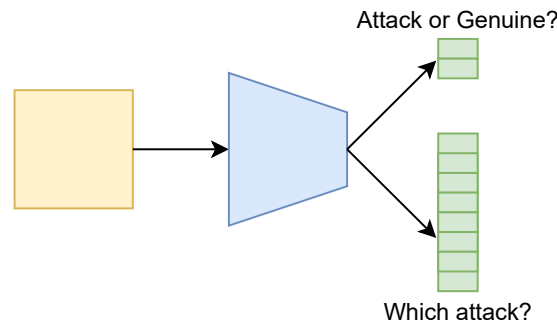


Figure 4.1: Architecture of the multi-task learning model. It receives an image (yellow box) and includes a CNN (blue figure), two output heads (green figures), where the first is a binary head and the second has 8 output nodes. The output node for the genuine samples on both heads ensures that the main goal remains on the detection of attacks.

4.3.1 Binary classification (BC-PAD)

As in traditional PAD, we designed this first approach to classify face images between impostor and genuine images. For this, we used a MobileNet v2 [434] backbone that is paired with a linear layer that outputs two softmax-activated values. Due to this binary setting we minimize the binary cross-entropy loss (BCE) as seen in Eq. 4.1.

$$BCE(y, p) = -(y \log(p) + (1 - y) \log(1 - p)) \quad (4.1)$$

4.3.2 Multi-task classification (MT-PAD)

More than distinguishing between a genuine and impostor images, it is possible to distinguish between different attacks. This empowers the model through a more complex learning process to identify more discriminative features, which can lead to better global and generalisation performance. On the previously described backbone, we further included a second softmax-activated linear layer, which classifies the input into one of eight distinct classes. One class corresponds to the genuine class, whereas the remaining seven represent different attacks. This configuration, as seen in Figure 4.1, configures our training strategy as a multi-task learning process. The outputs of

the first task layer are assessed with BCE (Eq. 4.1). The second utilises cross-entropy to penalise classification errors (Eq. 4.2). Finally, the training process is controlled by a linear combination of both losses with equal weights as seen in Eq. 4.3.

$$CE(y, p) = - \sum_{c=1}^M y_{o,c} \log(p_{o,c}) \quad (4.2)$$

$$Loss_{Multi}(y_1, p_1, y_2, p_2) = BCE(y_1, p_1) + CE(y_2, p_2) \quad (4.3)$$

4.3.3 Adversarial training

The presence of background and other artifacts can lead to an increase of noise, irrelevant information, and dataset specific elements can hurt the training process. One event caused by such information could be an overfitting to a specific dataset or training data. In essence, the model should refrain from using such information to make a prediction. Jaiswal *et al.* [240] introduced Unsupervised Adversarial Invariance, a method for the disentanglement of prediction-relevant information and other nuisances. To do so, this approach aims to separate this information into two embeddings e_1 and e_2 . The first shall hold sufficient information for the prediction, whereas the second comprises the remaining information. This method has three fundamental ideas: 1) e_1 suffices to make a correct prediction; 2) e_2 when joint with a noisy e_1 should suffice to reconstruct the original image; 3) there is no mutual information between e_1 and e_2 . To ensure an alignment with such ideas, the loss to optimise this framework requires four terms as follows:

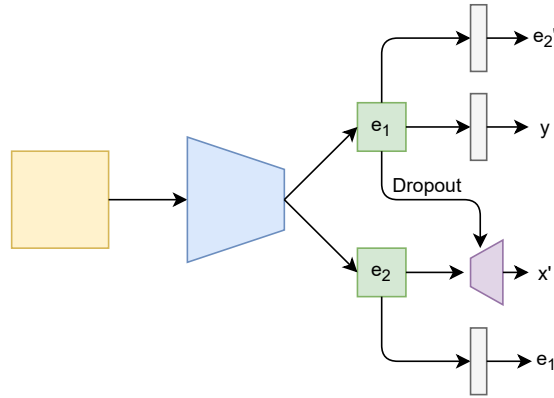


Figure 4.2: Architecture of the adversarial learning model. It receives an input image (yellow box) and includes a CNN (blue figure), two feature vectors e_1 and e_2 (green boxes). These are used to reconstruct each other, decode (purple figure) the input and to classify the input. This architecture is deeply based on the Unsupervised Adversarial Invariance [240].

- 1) e_1 is the input of a linear layer that is optimised with BCE, to minimise the classification error. The multi-class framework can be adapted in this setting too;
- 2) The mean squared error is used to measure the square of the distance between the reconstructed image x' and the original image x as seen in Eq. 4.4.

$$MSE(x, y) = \sum_{i=1}^D (x_i - y_i)^2 \quad (4.4)$$

x' is given by a decoder network that receives e_2 and a noisy e_1 as inputs. Noise is applied through dropout;

- 3) To minimise mutual information, it is necessary to ensure that e_1 cannot be predicted by e_2 and vice-versa. To do so, we minimise the negative MSE between e_2 and e'_2 , and between e_1 and e'_1 , where e'_1 and e'_2 are reconstructions of e_1 and e_2 from the opposite vectors (Eq. 4.5).

$$Loss_{Adv}(e_1, e_2, e'_1, e'_2) = -MSE(e_1, e'_2) - MSE(e_2, e'_1) \quad (4.5)$$

The overall architecture is represented in Figure 4.2. The classification and reconstruction terms are given by Eq. 4.6, where the term α controls the magnitude of the reconstruction term. We initialise α as 0.025 and increase it by 0.025 after each epoch.

$$Loss_{min}(y, y_{gt}, x, x') = BCE(y, y_{gt}) + \alpha MSE(x, x') \quad (4.6)$$

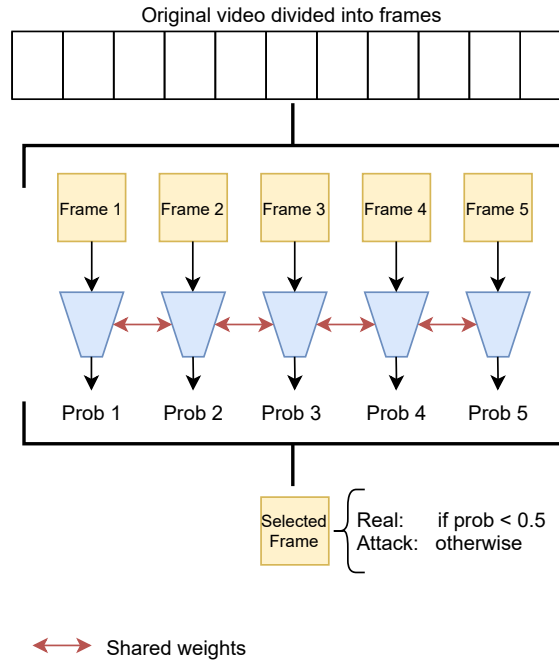


Figure 4.3: Architecture of the method for dynamic frame selection. It includes a CNN (blue figure) with shared weights that processes all the frames (yellow boxes) of a video and selects one (lower yellow box) based on a specific criteria.

4.3.4 Dynamic frame selection (DFS)

In PAD scenarios, the majority of the the input format includes a temporal dimension, i.e. videos. Hence, it is necessary to select which frames should be utilised by the deep neural network to infer the likelihood of an attack. However, this ignores the fact that not all frames might contain sufficient information for the prediction, and in fact, we might not be selecting the optimal frames to optimise our network. Inspired by Multiple Instance Learning frameworks, we propose a dynamic frame selection system. Under the assumption that having a frame where an attack is detect suffices to categorise the video as an attack, we designed a two-stage method. First, all frame are passed through a deep neural network, which assesses the likelihood of each of those frames containing an attack. As this is done without gradient computations it is far more efficient than training with all the frames. If the video's label indicates an attack, we select from all frames the three with the lowest probability of being an attack. On the other hand, for genuine videos we select the ones with the highest probability of being an attack. These selected frames are used, on that epoch, to optimise the network. For each epoch, we repeat both processes. Figure 4.3 shows the training scheme for this method, where one of the most interesting aspects is its integrability with previous approaches when required.

4.3.5 Evaluation Metrics

We assess and compare the different models and their performance on PAD tasks using the following metrics: the *Bona fide Presentation Classification Error Rate (BPCER)* (the proportion of bona fide presentations erroneously classified as attacks), and the *Attack Presentation Classification Error Rate (APCER)* (the proportion of presentation attack wrongly classified as bona fide) [231]. Finally, we also computed the *Equal Error Rate (EER)*, which is the error at the operation point where the APCER and BPCER have the same value. For the APCER and the BPCER we used a threshold of 0.5. These metrics have been previously described in Chapter 2.

4.3.6 Experimental Setup

The different methods have been implemented in PyTorch and optimised under similar settings. Weight updates leveraged Adam optimiser with a fixed learning set to $10e^{-3}$. The batch size was set to 32, and it did not use any data augmentation pipeline. The backbone, a MobileNet v2 was pre-trained on ImageNet [123]. Each input image was resized to 224x224 with three colour channels, and the images used for training, validation and testing were the same described in the protocol of each dataset. For the qualitative analysis through methods of the first rung of the xAI ladder, we leveraged Grad-CAM++ [79].

4.4 Datasets

For the experiments described in this chapter, we have used the following datasets: ROSE-Youtu [297]; NUAA [473]; and Replay-Attack [92]. These are detailed in this section as follows.

ROSE-Youtu [297]: Contains, in its public version, 3350 videos with 20 different subjects. On average, video clips have a duration of 10 seconds. For each of the subjects, it contains around 150 to 200 videos captured from five mobile devices (all with different capture resolutions) and five lighting conditions. The front-facing camera was used with a distance of about 30 to 50 centimeters to the face.

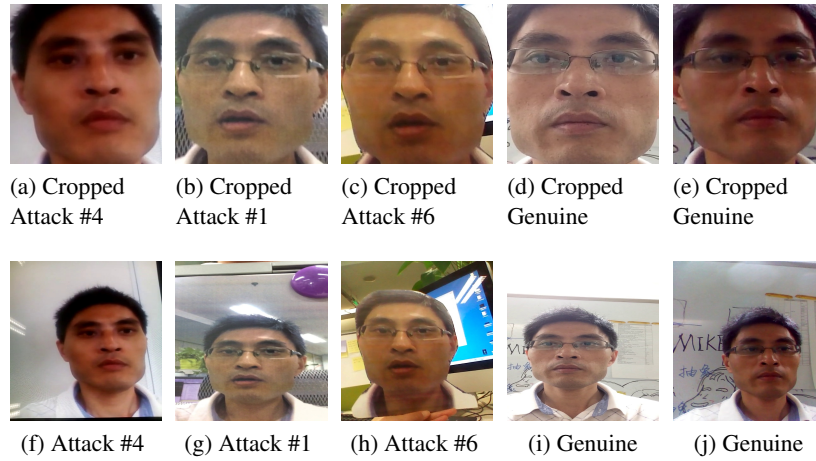


Figure 4.4: Samples collected from the ROSE-YOUTU dataset [297] containing images from attacks and genuine captures. On the top row, cropped images are displayed. Whereas the bottom row contains the exact same images, but with all the background information included.

Videos are categorised into one of eight possible types, which relate to the type of attack or a genuine instance, as described in Table 4.1. The first label is reserved for genuine videos, whereas the remaining seven distribute between print attacks (1-2), replay attacks on Lenovo and Apple laptops (3-4), and paper mask attacks with different crops (5-7). These latter attack add three-dimensional information to the attack.

Table 4.1: List of attacks present in the ROSE-YOUTU dataset [297].

Attack	Description
-	Genuine (bona fide)
#1	Still printed paper
#2	Quivering printed paper
#3	Video which records a Lenovo LCD display
#4	Video which records a Mac LCD display
#5	Paper mask with two eyes and mouth cropped out
#6	Paper mask without cropping
#7	Paper mask with the upper part cut in the middle

The dataset was prepared as two distinct copies, where the first has no face cropping, and the second includes a face pre-processing with MTCNN [546] information. The first 10 subjects have been used for training, and the remaining ten for testing. Examples of both copies are seen in Figure 4.4.

NUAA [473]: as one of the first public datasets for training and evaluating the performance of face PAD methods, this database simulates a general printing attack. It contains real and presentation samples for 15 subjects, where for each, the capture pipeline extended over three sessions. Capture relied on generic and cheap webcams. In total, it includes 5105 real and 7509 presentation attack face images in color space with 640×480 pixels. Training and testing protocol are already defined, leaving 1743 real and 1748 presentation attack face images for training, and the remaining for testing.

Replay-Attack [92]: consists of 1300 video clips of photo and video face presentation attacks to 50 different clients. All videos were captured by either having a real client accessing a laptop through a built-in webcam, or an impersonator showing a photo or video recording of the same client for at least 9 seconds. These captures occurred under different light conditions.

4.5 Results and Discussion

This section further details experiments and respective results for the different methods and datasets described in the previous sections. It is divided into Intra- and cross-dataset experiments, as well as one and unseen attacks on Rose-Youtu. Finally, a visual analysis is conducted through Grad-CAM++.

4.5.1 Intra-dataset

Due to the lack of variability on the attacks present in both NUAA and Replay-Attack datasets, we restricted the training on these datasets to comprise the binary approach (BC-PAD), and the adversarial approach in a binary setting (Adversarial BC-PAD). As shown in Table 4.2, both models effectively detect attacks in the presence of background, becoming slightly less performing when trained on images without background.

Table 4.2: Comparison of four different approaches with their versions with and without background. The columns represent the dataset used for both training and testing. The reported values represent the EER in %.

Method	Background	NUAA	Replay-Attack
BC-PAD	No	2.91	0.00
	Yes	0.00	0.00
Adversarial BC-PAD	No	3.03	0.33
	Yes	0.00	0.00

The variety of attacks present in the ROSE-Youtu dataset motivates a multi-task learning approach. This allows for an analysis of multiple approaches under these two different background-related scenarios. Table 4.3 showcases the results for the previously used methods (BC-PAD and Adversarial BC-PAD), in addition to a multi-task version of both (MT-PAD and Adversarial

MT-PAD). Adversarial approaches did not have a positive impact, however, for all testing configurations, the presence of background seems to benefit performance. Multi-task approaches further enhanced performance.

Table 4.3: Comparison of four different approaches with their versions with and without background on the ROSE-Youtu dataset. APCER, BPCER and EER are displayed as %. In bold is the best result per column.

Method	Background	APCER	BPCER	EER
BC-PAD	No	0.49	2.20	1.32
	Yes	0.25	2.03	0.73
MT-PAD	No	1.34	1.17	1.26
	Yes	0.15	0.40	0.24
Adversarial BC-PAD	No	1.42	2.71	1.76
	Yes	0.52	1.29	0.76
Adversarial MT-PAD	No	1.18	2.93	1.91
	Yes	0.29	1.11	0.60

Performance gains in the presence of background were visible for both multi-task and single task classification scenarios. Hence, a set of experiments leveraging all the proposed methods was conducted on ROSE-Youtu in the presence of background. In addition to the previous experiments, Table 4.4 includes dynamic frame selection on multi-task scenarios with and without adversarial training (DFS MT-PAD and Adversarial DFS MT-PAD), and on binary classification scenarios (DFS BC-PAD). Perhaps related to the small size of these datasets, both DFS and Adversarial approaches exhibit lower performance. In addition, the combination of both further degrades performance, which could be linked to changes in the inherent data noise caused by DFS. Multi-task classification displays the best performance across all metrics, highlighted by the low EER (nearly $3\times$ smaller than the second lowest EER).

Table 4.4: Comparison of all the seven different approaches explored in the ROSE-Youtu dataset. All of the approaches leveraged background information. In bold is the best result per column.

Method	APCER (%)	BPCER (%)	EER (%)
BC-PAD	0.25	2.03	0.73
MT-PAD	0.15	0.40	0.24
Adversarial BC-PAD	0.52	1.29	0.76
Adversarial MT-PAD	0.29	1.11	0.60
DFS BC-PAD	0.54	4.68	1.62
DFS MT-PAD	0.31	2.23	0.69
Adversarial DFS MT-PAD	2.15	1.78	1.78

4.5.2 Cross-dataset

These methods were further validated in cross-dataset scenarios, where their performance was assessed on NUAA and Replay-Attack, after being trained on ROSE-Youtu. The results, Table 4.5, showcase a clear advantage on models that have been trained with background, and further highlight the capabilities of the multi-task model. These results, which are aligned with the one son a single dataset, show consistently quantitative benefits of training in the presence of background.

Table 4.5: Comparison of four different approaches with their versions with and without background. Results for models trained on the ROSE-Youtu dataset and tested on the datasets of each column. The reported values represent the EER in %.

Method	Background	NUAA	Replay-Attack
BC	No	22.04	29.43
	Yes	13.45	12.29
MT	No	23.61	26.13
	Yes	3.89	13.91
Adv.+BC	No	28.31	26.03
	Yes	18.11	17.12
Adv.+MT	No	35.66	26.15
	Yes	23.85	19.46

4.5.3 One attack

The best performing approach had its experiments extended to evaluate both One and Unseen attacks. The first, which concerns this section, focuses on training and testing the model with only one attack type. This allows to understand the intrinsic hardness of each attack. Table 4.6 shows that nearly all models successfully overfit to their attacks. Nonetheless, attack #4, video records from an Apple laptop display, is significantly tougher than the others. One reason sustaining this result might relate to the higher resolution of the laptop’s display. Thus, leading to attacks with higher levels of realism.

4.5.4 Unseen attack

Unseen attacks present a more challenging configuration, where one attack is removed from training, and it becomes the only attack to be used on testing. This tests the generalization capabilities of the model, as well as the capability to extract from the other attacks sufficient information to make predictions on novel attacks. In line with the previous results, Table 4.7 highlights a particular difficulty on attack #4, followed by attack #3 (also a video recording attack). Showing a high difficulty at generalizing to the different video recordings, causing a significant rise on the number of attacks incorrectly detected. This behaviour is reflected in all three training strategies.

Table 4.6: Evaluation of three approaches in the setting of one attack in the ROSE-Youtu dataset without cropping the background. In this setting, the attack in the first column is the only one used for training and testing. APCER, BPCER and EER are displayed as %. In bold is the best result per column.

Attack	MT			Adversarial MT			DFS MT		
	APCER	BPCER	EER	APCER	BPCER	EER	APCER	BPCER	EER
#1	0.00	0.20	0.05	0.20	0.29	0.27	0.50	0.22	0.50
#2	0.00	0.02	0.02	0.00	0.07	0.00	0.00	0.00	0.00
#3	0.00	0.16	0.11	0.00	0.29	0.25	0.00	0.45	0.00
#4	0.30	1.27	0.71	1.16	1.98	1.46	0.50	1.34	1.01
#5	0.00	0.02	0.00	0.00	0.09	0.05	0.00	0.45	0.00
#6	0.00	0.07	0.00	0.00	0.16	0.11	0.00	0.00	0.00
#7	0.00	0.05	0.00	0.00	0.11	0.00	0.00	0.00	0.00

Table 4.7: Evaluation of three approaches in the setting of unseen attack in the ROSE-Youtu dataset without cropping the background. In this setting, the attack in the first column is excluded from the training and is the only one used for testing. APCER, BPCER and EER are displayed as %. In bold is the best result per column.

Attack	MT			Adversarial MT			DFS MT		
	APCER	BPCER	EER	APCER	BPCER	EER	APCER	BPCER	EER
#1	1.00	0.85	0.95	1.00	6.90	2.87	1.00	6.90	2.30
#2	0.00	0.49	0.25	0.00	1.18	0.27	0.00	4.90	0.67
#3	3.09	3.41	3.27	1.44	3.07	2.61	2.14	10.24	5.79
#4	13.82	3.88	7.57	12.66	6.61	9.15	17.59	16.70	17.09
#5	0.00	1.98	0.65	0.25	0.58	0.45	0.50	4.90	1.49
#6	0.00	0.89	0.33	0.10	0.58	0.10	0.00	4.01	1.00
#7	0.35	3.63	1.77	2.32	7.68	5.20	0.51	9.58	1.78

4.5.5 Visual Analysis

In order to understand the behavioral changes in the model’s inspection of the input image, we have produced Grad-CAM++ visualizations of the model’s prediction for different attacks with and without background. In the replay attack shown in Figures 4.5a and 4.5d, one can observe that the latter has a higher focus on facial features, whereas the former leverages visual cues such as reflections (caused by the laptop monitor). In spite of its focus on facial features, Figure 4.5d represents a sample wrongly predicted by the model, whereas when background is visible the respective model correctly identifies the attack. Figures 4.5b and 4.5e highlight that, even in the presence of background, the model focuses on facial features when deemed relevant. Finally, changes in focus elements also include print attack under the presence of background, as seen in Figures 4.5c and 4.5f. The former image highlights the capability of its model to identify spacial cues, such as the pin keeping the printed image in place. The latter provides a less clear visualization with a focus on facial cues.

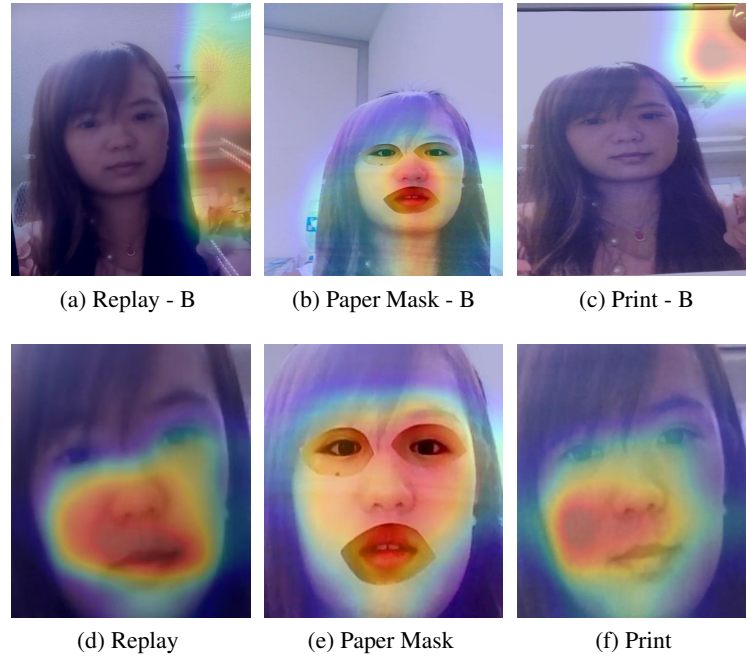


Figure 4.5: Explanations generated with Grad-CAM++ for different attacks of subject #23. -B indicates that the model was trained and tested on images with background. The (d) image was wrongly classified as genuine.

The previous examples demonstrate some changes in the model's behavior under the presence of different spatial cues, leading to questions regarding the parallel between human and machine decision process in these scenarios, as the behavior in the presence of background better mimics human decision process.

4.6 Summary and Conclusion

This chapter focused on the study of the behavioral changes of a deep learning model for presentation attack detection, trained on images with and without background crops. Besides studying the quantitative effects of such crops on a plethora of different models and training strategies, it leveraged visualization techniques to inspect a possible shift in the pixels deemed relevant for the classification task. In addition, some of the proposed methods surpassed state-of-the-art alternatives in the quantitative analysis on the ROSE-Youtu dataset.

As previously discussed, CAM-based methods represent associative knowledge, and hold no significant or causal explanatory meaning. Nonetheless, in this chapter we have shown that they remain powerful tools for the inspection of different models, comparative studies and higher understanding of the decision process (at least to identify unexpected behavior).

Chapter 5

Identity Disentanglement for Morphing Attack Detection

Author Contributions The research work described in this chapter was conducted in collaboration with Eduarda Caldeira, Tiago Gonçalves, Marco Huber and Naser Damer, under the supervision of Ana F. Sequeira and Jaime S. Cardoso. The author of this thesis contributed to this work on the problem conceptualisation, data curation, the preparation and conduction of experiments, the results discussion, and publication writing. Some parts of the chapter were originally published in, or adapted from:

- Neto, P. C., Gonçalves, T., Huber, M., Damer, N., Sequeira, A. F., & Cardoso, J. S. (2022, September). Orthomad: Morphing attack detection through orthogonal identity disentanglement. *In 2022 International Conference of the Biometrics Special Interest Group (BIOSIG) (pp. 1-5). IEEE.*
- Neto, P. C., Caldeira, E., Gonçalves, T., Damer, N., Sequeira, A. F., & Cardoso, J. S. (2023, September). Unveiling the two-faced truth: Disentangling morphed identities for face morphing detection. *In 2023 31st European Signal Processing Conference (EUSIPCO) (pp. 955-959). IEEE.*

5.1 Context and Motivation

Face recognition (FR) systems have had large-scale adoption in the most diverse scenarios [360, 371, 502]. Deep learning (DL) techniques have taken this and other biometric recognition systems towards above-human performance. While it also benefited biometric systems adoption, DL methods led to two problems. First, the same approaches that improve the recognition power of these systems are the same to be used to design novel and dangerous attacks [111]. Some attacks can take the form of adversarial noise addition or be developed with FR systems in mind. The latter comprises both morphing [334] and presentation attacks [359]. Besides the attacks, deep learning methods are notorious for their black-box behaviour, which compromises the understanding of both the inner workings of the model and the reasoning behind a decision. Furthermore, FR and face attack detection systems are consistently designed using problem-agnostic tools, which do not leverage domain knowledge. For a wider adoption and to be able to deploy these systems on critical scenarios, it is necessary to guarantee that their thought process is, at least to some extent,

explained. One can explain a decision using a post-hoc approach [372], or directly interfering with the training behaviour of the model as stated in Chapter 3.

Face morphing attacks, which merge two images from different identities into a single image capable of misdirecting the recognition system, have progressed significantly. In other words, this attack aims to increase the number of false positives of the FR system, granting access to two distinct users. When left undetected, the fusion of two images might allow two different people to pass border control with the same passport, for example. Due to this threat, researchers focused their attention on the development of robust morphing attack detection (MAD) systems [111, 112, 222]. Usually, they are designed to detect if the input image is an attack or a *bona fide* sample and do not include any information regarding the fused identities in their training. The recent blossoming of synthetic data generation methods, such as generative adversarial networks (GAN) and diffusion models, led to the creation of synthetic datasets with a diverse number of identities represented [105, 109]. Although these identities are usually represented only once, it suffices to increase identity diversity.

In this chapter we introduce two distinct interventions to MAD systems to introduce domain knowledge in their training process, thus, shedding some light on their inner workings. Both approaches follow a similar strategy of imposing additional restrictions that reflect some desired property of the final trained model. OrthoMAD, the first of these, aims to disentangle the two identities used to create a morph through the minimization of the mutual information between two vectors (representative of each identity). This orthogonality constraint, is relatively strong, and might not fully capture the nuances between both identities. Hence, we have additionally proposed IDistill, which focuses on relaxing this constrain, while leveraging knowledge distillation from an auto-encoder to capture the nuances of both identities. In addition, to further increase the interpretability of our model, we create a more transparent decision scheme, where both identities vectors are processed independently, and forced to be in a similar latent space, allowing us to trace the decision back to the corresponding vector. Both methods are represented in the second rung of the xAI ladder with OrthoMAD adding prior to the training rule and IDistill further adding a prior on the model structure.

The main contributions of this chapter are: 1) two novel approaches for identity disentanglement with distinct levels of relaxation; 2) the exploration of the angle between identity vectors learned by an auto-encoder; 3) an interpretable strategy to identify morphing through the analysis of these vectors; 4) two state-of-the-art solutions which can be easily integrated in different backbones and training strategies.

This chapter has been divided into five large sections. The first, Section 5.2 concerns the datasets required to train, validate and test the proposed methods. The following two, Sections 5.3 and 5.4, explore the methodological choices for OrthoMAD and IDistill, respectively. Section 5.5 presents the experimental setup for reproducibility of both methods. Finally, Section 5.6 discusses the results of each approach, and how they compare with other state-of-the-art approaches.

5.2 Datasets

To train and test our methods, we have chosen four different datasets. One is exclusive for testing, comprising different types of morphing strategies, whereas the others are designed for training. These datasets are described below:

1. **FRL**: The Face Research London Lab dataset [121] was used to produce the FRL-Morphs dataset [438], which is a common benchmark for evaluating morphing attack detection methods. Five different morphing techniques are used in the dataset, including StyleGAN2 [263, 495], WebMorph [120], AMSL [373], FaceMorpher [414], and OpenCV [322]. Each of the five methods uses 204 genuine samples and more than one thousand morphed faces made from high-resolution frontal faces. We used this database only for evaluation purposes because it lacks distinct train, validation or test sets. Its variety of morphing techniques allows for a thorough evaluation of the generalization of each model to different attacks.
2. **SMDD**: The Synthetic Morphing Attack Detection Development (SMDD) [109], is a novel dataset that uses synthetic images to create a dataset of morph and *bona fide* samples. It started by generating 500k images of faces using a random Gaussian noise vector and the official open-source version of StyleGan2-ADA [263]. The quality estimation method known as CR-FIQA [51] was exploited to choose 50k of these photos for analysis due to their high quality, and 25k of them were set as *bona-fide* samples. The non-*bona fide* photos were randomly paired with five other non-*bona fide* images, and five thousand of those images were chosen as key morphing images. They were, afterwards, morphed using the OpenCV [322] method, which resulted in 15k different attack samples. The original 25k images that were used to construct the morphs are also publicly available. The evaluation protocol of this dataset was already defined, with 85% of the samples used for training and 15% for test.
3. **LMA-DRD**: The LMA-DRD dataset [112] uses the images published with the VGGFace2 [71] dataset to create morphing attack. Despite having more than 3 million images available, only a few were selected to be fused into attacks. The selected images shared some common traits, such as a frontal pose, high resolution, and a neutral expression. Using the parametrization introduced in [415], the images were morphed with the OpenCV approach. *Bona-fide* images were selected using the previously mentioned criteria. This dataset has two versions, a digital - LMA-DRD (D) - and a printed and scanned - LMA-DRD (PS). We used the train set for training (96 morphs and 121 *bona-fide* images), and the identity-disjoint test set for testing (88 morphs and 123 *bona-fide* images).
4. **MorGAN**: Exploiting CelebA [311] images, the authors selected a pool of images with frontal pose, and utilized OpenFace [32], a FR solution, to find the most similar pairs of faces, so that morphed images become more effective [312]. These images were then fused with the landmark-based OpenCV, or the GAN-based approach. These two approaches led

to the creation of two branches of the original MorGAN [111] dataset, MorGAN-LMA and MorGAN-GAN. The train and evaluation protocol is common to both sets, with 500 morphing attacks and 750 *bona fide* samples each. Resolution wise this dataset includes images of very low resolution (64x64).

5.3 OrthoMAD

Following the premise that morphed faces hold information regarding the two identities used in the morphing process, we have devised a strategy to separate the information of those identities. The validation for our initial assumption comes from the fact that morphed samples manage to trick FR systems into accepting that images as belonging to more than one identity. Our theory is that if information about two identities is present, we can separate information, which is not common to both identities. This would be identity exclusive information.

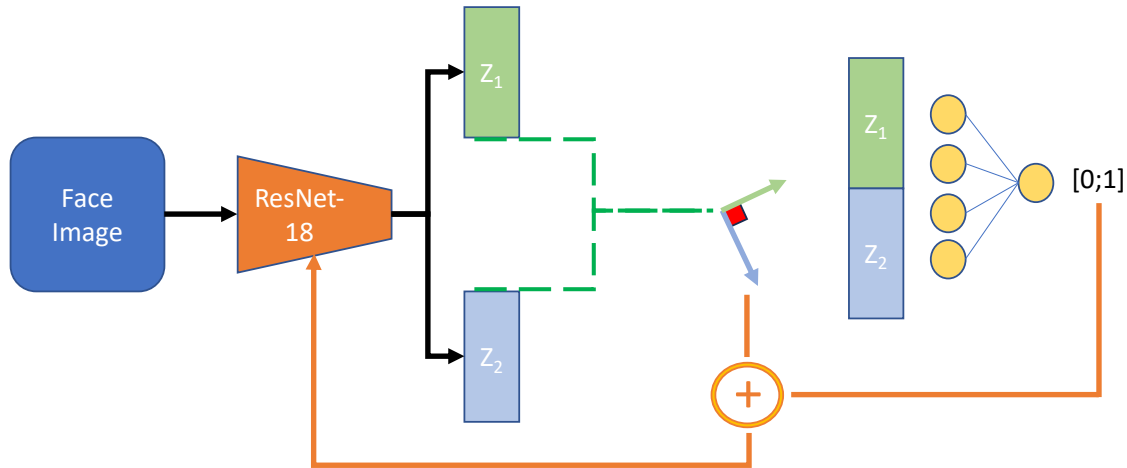


Figure 5.1: Overview of the architecture design for the integration of the new regularisation term. The regularisation loss is applied at the vector (Z_1 and Z_2), hence, their gradients do not affect the final classification layer. The system is composed of three components, a backbone network, two identity vectors and a final classification layer that uses the concatenation of both vectors.

Hence, we adapt a backbone convolutional neural network to output two distinct feature vectors, extracted from a common set of layers. These two vectors, z_1 and z_2 , represent independent features. Thus, we assume that no mutual information must be shared between them. As such, we establish an additional training rule promoting the absence of such information. As orthogonal vectors hold no mutual information, we exploit this property. In addition to the training loss, we introduce a regularization term that computes the inner product between Z_1 and Z_2 . The optimization process will also be guided by the minimization of this term. We calculate the inner product as shown in Equation 5.1. We assume that the information held by each vector will correspond to identity information. The overall structure is seen in Figure 5.1.

$$Reg = (z_1^T z_2)^2 \quad (5.1)$$

As the minimization process might not be as smooth with the original inner product expression, we square it. The magnitude of this regularization is controlled by a α scalar as shown below. Both z_1 and z_2 were defined as 32-d vectors, but their dimensions can vary for different tasks, datasets and applications. For the classification layer that originates the final prediction we give as input the vector z , a 64-d vector that corresponds to the concatenation of both z_1 and z_2 , as seen in Equation 5.2.

$$z = \text{concat}(z_1, z_2) \quad (5.2)$$

The final prediction, $\hat{y}$, is a scalar that comes from a weighted linear combination of the features in the vector z , and further activated with a sigmoid function as described in Equation 5.3, where w represents the weight vector of this linear combination.

$$\hat{y} = \frac{1}{1 + e^{-(w^T z)}} \quad (5.3)$$

To compute the correctness of $\hat{y}$ we use the binary cross-entropy (BCE), described in Equation 5.4, which minimizes the negative of the logarithm of the probability of the ground-truth class.

$$L_{BCE} = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \quad (5.4)$$

Both the regularization term and the BCE loss are combined in a single loss function (Eq. 5.5). While the BCE is not controlled by any scaling parameter, α sets the magnitude of the regularization term.

$$L_{OrthoMAD} = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) + \alpha (z_1^T z_2)^2 \quad (5.5)$$

The presented method cannot ensure that z_1 and z_2 contain identity-related information. Nonetheless, it is expected that enforcing the ability to detect an attack from such vectors leads to the inclusion of identity information, or at least, information regarding the original, non-morphed sample. Some of these gaps are addressed by the method described in the subsequent sections of this chapter.

5.4 IDistill

OrthoMAD proposes the separation of deep latent features in two vectors without mutual information. However, it has a few gaps: 1) the system optimization to promote identity information

on each vector is weak, and it cannot be guaranteed that the information is present; 2) orthogonality is a strong constraint that might not reflect the reality, as morphing is frequently performed with identities sharing a high similarity factor; 3) the interpretability of the last decision is rather limited due to the concatenation of both vectors prior to the decision. Hence, we propose a follow-up method (IDistill), which mitigates such flaws, in an attempt to identify its impact on the final performance.

To address the first issue, we considered the best strategy to infuse identity information in such vectors. Hence, we decided to leverage knowledge distillation (KD) to drain this information from an auto-encoder, trained on *bona-fide* images. This leads to a more flexible regularization term. In future iterations it can directly leverage features from FR systems instead of an auto-encoder. The overall architecture of this solution can be seen in Figure 5.2, and each component is described below.

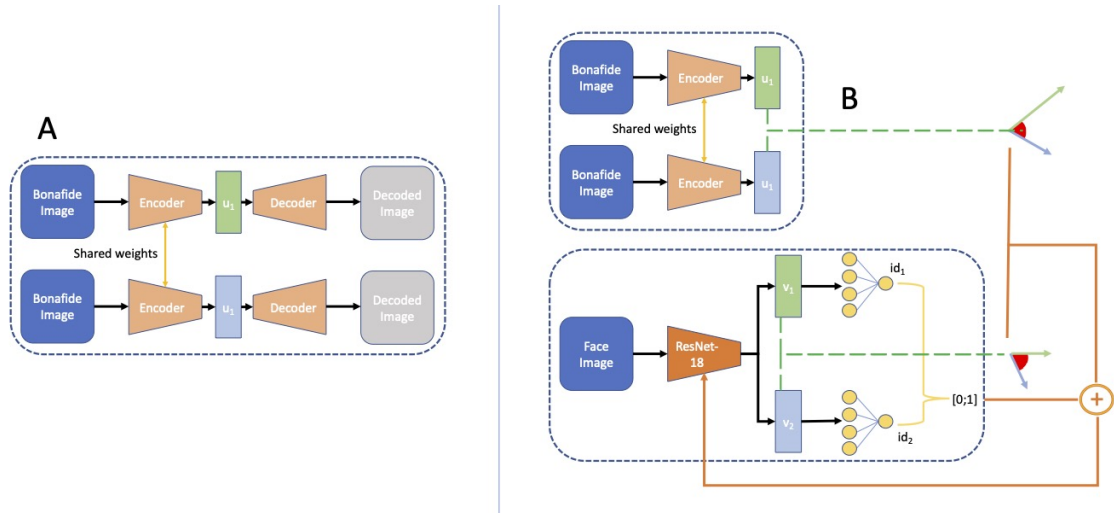


Figure 5.2: Overall scheme of the architecture. Part A represents the training of the auto-encoder, whereas part B represents the training of the classification system. The orange line represents the backpropagation, the green one represents the calculation of the angle between vectors and the yellow represents the fusion of the two scores. Best viewed in color.

To create a representation of what we refer as "identity information" or "image-specific information", we relied on a reconstruction task. For this, we train an auto-encoder to reconstruct *bona-fide* samples. This encoder aims to create a minimalist representation of a face I . Alternatively to the usage of the auto-encoder, we could have leveraged a pretrained face recognition system. The decision to follow with the auto-encoder yields three reasons: 1) Fang *et al.* [153] has shown a difference in the reconstruction performance from abnormal and normal face images; 2) Besides being large (512-d), the latent vector face recognition systems might not contain all the information necessary for the reconstruction. 3) Encoder-Decoder have been explored for face de-morphing [33]. The proposed auto-encoder is based on the U-Net architecture [429] and the size of the latent representation of the image was chosen as 128. As in other reconstruction tasks, the auto-encoder receives the image I , and its encoder branch maps I to a latent representation u .

This representation is later used by the decoder branch to build $\tilde{I}$, a reconstructed version of I . The reconstruction error is assessed by the mean squared error (MSE) loss function as seen in Eq. 5.6, where H and W represent the height and width of the image, respectively.

$$L_{MSE} = \frac{1}{HW} \sum_i^H \sum_j^W (I_{ij} - \tilde{I}_{ij})^2 \quad (5.6)$$

The architecture of the morphing classifier is based on a ResNet-18 [205] where the last fully-connected layer is replaced with two distinct branches, each containing a fully-connected layer. These branches output two 128-d vectors, v_1 and v_2 . We assume that, in the hypothesis that these vectors encode information regarding each identity, one shall be clear of such information given a *bona-fide* sample. Pushing a vector to be filled by zeros might not align with the representation learned, so we introduce a fully-connected layer that computes, individually for each vector, if they contain identity information. The output of that vector is computed as described in Eq. 5.7, where θ represents the weights of the layer, and v is either v_1 or v_2 .

$$id = \frac{1}{1 + e^{-\theta^T v}} \quad (5.7)$$

Considering that the produced score hold information regarding the presence of encoded identity information on a vector v , the final prediction for an image I is designed as follows. Given I , the backbone architecture produces v_1 and v_2 , which will result in the identity probabilities id_1 and id_2 , respectively. The probability of I containing information of only one identity is given by $id_1 * id_2$. Consequently, the morphing prediction probability, $\tilde{y}$ is given by:

$$\tilde{y} = 1 - id_1 * id_2 \quad (5.8)$$

For the classification task, we have introduced the Binary Cross-Entropy, L_{BCE} (Equation 5.4) at the level of the final fused prediction $\tilde{y}$.

To better align the information of the latent vectors v_1 and v_2 with the information extracted by the auto-encoder, we introduce a knowledge distillation term. When the input sample is an attack, this term aims to extract, using the morphing classifier, vectors from the morphed image that form an angle between them that is close to the one formed by the latent vectors u_1 and u_2 extracted with the encoder from the two images that originated the morphed image (Eq. 5.11). Thus, we are promoting a similar structure between the latent space of the auto-encoder and of the morphing classifier, ensuring that relative angles remain similar. This term only considers the angle formed between these vectors, since their identity intensity, or the expression of each feature in the vector, might be diminished in the morphing process. For *bona-fide*, we expect one vector to hold identity information, while the other does not. As such, we designed a term that first selects the vector v with the highest cosine similarity (S_{cos}) to u . Following this choice, we can maximize the similarity between u and the selected vector v , while approximating the id of this vector to 1, and the other id to 0 (Eq. 5.10).

$$Ver_{term} = S_{cos}(v_1, u) > S_{cos}(v_2, u) \quad (5.9)$$

$$L_{KD_1} = \begin{cases} (1 - id_1)^2 + (id_2)^2 - S_{cos}(v_1, u) & \text{if } Ver_{term} \\ (1 - id_2)^2 + (id_1)^2 - S_{cos}(v_2, u) & \text{otherwise} \end{cases} \quad (5.10)$$

$$L_{KD_2} = [S_{cos}(u_1, u_2) - S_{cos}(v_1, v_2)]^2 \quad (5.11)$$

$$L_{KD} = \gamma L_{KD_1} + (1 - \gamma) L_{KD_2} \quad (5.12)$$

Both losses are incorporated in a single equation as follows:

$$Loss = L_{BCE} + L_{KD} \quad (5.13)$$

5.5 Experimental Setup

In this section we present the necessary details to replicate the experimental design followed in this Chapter. For both methods, we applied the same preprocessing strategy to the datasets. Each image had a probability of being flipped horizontally as augmentation, and every crop of the face was resized as a 224×224 image. The following subsections discuss specific details of each method, as well as evaluation metrics.

5.5.1 OrthoMAD

OrthoMAD experiments aimed at validating the idea of regularizing these identity vectors as a strategy to enhance performance. Hence, the majority of the experiments focus on comparing a similar network with and without regularization. For this assessment, we trained both models in the five different types of morphing approaches found in the FRLL dataset. With exception to the regularization, all other hyperparameters are shared across both models. The batch size was set to 16, the learning rate to 10^{-5} , and for the regularized version α was set to 100.

5.5.2 IDistill

For IDistill, the first step required the training of an auto-encoder model. Using a learning rate of 1×10^{-4} , a batch size of 32, and Adam [272] we trained this model for 300 epochs. It trained exclusively on *bona-fide* samples.

We further trained the MAD model with a joint loss (Eq. 5.13) utilizing a learning rate of 1×10^{-4} , a batch size of 16 and Adam optimizer. Furthermore, to align with the auto-encoder, both v_1 and v_2 are 128-d, although this can vary. The training utilized the synthetic dataset SMDD, which allowed this regularization term to utilize the original samples used to create the morphing samples.

Table 5.1: Results for the different types of morphing techniques included in the FRLl dataset. For morphing type, we present the performance of the model per training dataset. All the results are in percentage (%), and the best are in bold.

Test	Train	Model	EER	BPCER @ APCER =	
				1%	20%
FRLl - Style-GAN2	MorGAN-LMA	ResNet-18	28.72	99.01	60.81
		OrthoMAD (Ours)	51.47	97.46	81.51
	MorGAN-GAN	ResNet-18	53.60	99.59	87.07
		OrthoMAD (Ours)	17.34	59.81	14.32
	LMA-DRD (D)	ResNet-18	62.60	100.00	97.87
FRLl - WebMorph	MorGAN-LMA	ResNet-18	31.50	99.91	63.74
		OrthoMAD (Ours)	71.19	100.00	97.38
	LMA-DRD (PS)	ResNet-18	29.21	99.59	53.02
		OrthoMAD (Ours)	11.62	35.18	7.77
	SMDD	ResNet-18	6.54	13.74	3.76
FRLl - AMSL	MorGAN-LMA	ResNet-18	37.10	99.09	60.68
		OrthoMAD (Ours)	8.35	45.61	2.53
	MorGAN-GAN	ResNet-18	23.25	89.59	25.71
		OrthoMAD (Ours)	14.41	45.12	9.90
	LMA-DRD (D)	ResNet-18	93.85	100.00	100.00
FRLl - FaceMorpher	MorGAN-LMA	ResNet-18	23.58	97.37	29.40
		OrthoMAD (Ours)	88.62	100.00	100.00
	LMA-DRD (PS)	ResNet-18	10.65	60.28	7.20
		OrthoMAD (Ours)	16.29	85.33	16.29
	SMDD	ResNet-18	15.23	70.92	9.50
FRLl - OpenCV	MorGAN-LMA	ResNet-18	54.16	99.63	84.36
		OrthoMAD (Ours)	0.91	0.91	0.00
	MorGAN-GAN	ResNet-18	44.59	90.94	63.72
		OrthoMAD (Ours)	7.91	97.01	1.19
	LMA-DRD (D)	ResNet-18	50.67	95.86	75.86
FRLl - Average	MorGAN-LMA	ResNet-18	30.43	89.70	40.46
		OrthoMAD (Ours)	83.35	100.00	100.00
	LMA-DRD (PS)	ResNet-18	27.26	87.21	37.19
		OrthoMAD (Ours)	34.43	65.97	27.26
	SMDD	ResNet-18	14.80	65.05	10.89
FRLl - FaceMorpher	MorGAN-LMA	ResNet-18	5.48	25.45	1.23
		OrthoMAD (Ours)	34.20	97.29	54.90
	MorGAN-GAN	ResNet-18	37.15	97.95	55.72
		OrthoMAD (Ours)	35.27	94.10	56.71
	LMA-DRD (D)	ResNet-18	39.60	99.59	76.51
FRLl - OpenCV	MorGAN-LMA	ResNet-18	30.19	83.87	38.13
		OrthoMAD (Ours)	40.02	99.42	70.21
	LMA-DRD (PS)	ResNet-18	34.12	99.83	71.84
		OrthoMAD (Ours)	2.95	5.32	2.37
	SMDD	ResNet-18	0.98	2.37	0.08
FRLl - OpenCV	MorGAN-LMA	ResNet-18	16.29	64.61	11.62
		OrthoMAD (Ours)	27.92	93.44	38.82
	MorGAN-GAN	ResNet-18	40.29	99.34	79.11
		OrthoMAD (Ours)	29.07	97.21	43.48
	LMA-DRD (D)	ResNet-18	66.75	99.83	96.56
FRLl - Average	MorGAN-LMA	ResNet-18	33.98	99.59	60.68
		OrthoMAD (Ours)	49.63	100.00	85.17
	LMA-DRD (PS)	ResNet-18	42.17	99.09	73.87
		OrthoMAD (Ours)	1.22	11.8	0.41
	SMDD	ResNet-18	0.73	0.73	0.32
FRLl - Average	MorGAN-LMA	ResNet-18	28.35		
		OrthoMAD (Ours)	24.57		
	MorGAN-GAN	ResNet-18	39.78		
		OrthoMAD (Ours)	20.08		
	LMA-DRD (D)	ResNet-18	62.69		
FRLl - Average	MorGAN-LMA	ResNet-18	29.94		
		OrthoMAD (Ours)	66.56		
	LMA-DRD (PS)	ResNet-18	28.68		
		OrthoMAD (Ours)	13.20		
	SMDD	ResNet-18	7.66		

5.5.3 Evaluation Metrics

To evaluate the performance of the morphing detection, we evaluated our algorithm using different metrics, commonly used in the literature: the Attack Presentation Classification Error Rate (APCER) (i.e., morphing attacks classified as *bona-fide*); and the *bona-fide* Presentation Classification Error Rate (BPCER) (i.e., the *bona-fide* samples that are classified as morphing attacks). We evaluated these metrics at two different fixed APCER values (1.0% and 20.0%). The equal error rate (EER), which is the BPCER and APCER at the decision thresholds where they are equal, was also evaluated.

5.6 Results

5.6.1 Orthogonality

To evaluate if the promotion of orthogonality was providing sufficient guidance to improve the MAD model, we train this model on several datasets, while evaluating on a myriad of different morphing creation techniques. In the vast majority of these datasets, there is no sufficient information to train the auto-encoder of IDistill, as such, in this subsection we only focus on the constraint imposed by OrthoMAD.

In Table 5.1 we highlight the results of training these models on a certain dataset, and testing on a different one. For testing we leveraged FRLI and calculated results independently for each morphing technique. For training, we used every other dataset individually. The entries of the table, in the majority of the scenarios (22 out of 25), showcase a better performance of OrthoMAD when compared to a non-constraint model. In some cases, it is visible that the EER goes above the 50% line. Being a binary task we could simply flip the classes and improve the performance. However, we choose to display these results as is to highlight that some of these datasets might lead to this behavior and instability.

A secondary analysis on the overall performance of each dataset as a training dataset highlights SMDD as the one that leads to models that are more robust to the different morphing techniques. OrthoMAD displays a larger percentage of errors when the training data is smaller. This might convey the lack of varied identity information in those datasets, as well as variability in the other factors such as lighting. The better results support the use of SMDD for further experiments.

5.6.2 IDistill and OrthoMAD

The literature on face morphing attack detection is large, however, is also disperse. In other words, the datasets used for benchmarking and training are not always the same, and as such, direct comparisons are not trivial. The combination of FRLI and SMDD has been found in at least two different documents in the literature. The first [109] introduces the SMDD datasets and evaluate three different methods from the literature: Inception [471], PW-MAD [112] and MixFacenet [44]. Their results vary and there is not one that beats the others consistently across the different FRLI morphing methods. OrthoMAD and IDistill were also evaluated using the

exact same protocol [366] and achieved state-of-the-art results on three out of the five morphing approaches.

Table 5.2: Results comparison with four models published in the literature. All the models were trained on the SMDD dataset, and evaluated on the dataset specified on the left column of the table. All the results are in percentage (%) and the best are in bold.

Test	Model	EER	BPCER @ APCER =	
			1%	20%
FRL-Style-GAN2	Inception	11.37	72.06	6.86
	PW-MAD	16.64	80.39	13.24
	MixFacenet	8.99	42.16	4.41
	OrthoMAD (Ours)	6.54	13.74	3.76
	IDistill (Ours)	1.96	8.51	0.08
FRL-WebMorph	Inception	9.86	53.92	2.94
	PW-MAD	16.65	80.39	13.24
	MixFacenet	12.35	80.39	7.84
	OrthoMAD (Ours)	15.23	70.92	9.50
	IDistill (Ours)	4.01	14.41	0.33
FRL-OpenCV	Inception	5.38	38.73	0.98
	PW-MAD	2.42	22.06	0.49
	MixFacenet	4.39	26.47	1.47
	OrthoMAD (Ours)	0.73	0.73	0.32
	IDistill (Ours)	2.46	6.14	0.16
FRL-AMSL	Inception	10.79	72.06	4.90
	PW-MAD	15.18	96.57	5.88
	MixFacenet	15.18	49.51	11.76
	OrthoMAD (Ours)	14.80	65.05	10.89
	IDistill (Ours)	4.00	21.10	2.85
FRL-FaceMorpher	Inception	3.17	30.39	0.49
	PW-MAD	2.20	26.47	0.00
	MixFacenet	3.87	23.53	0.49
	OrthoMAD (Ours)	0.98	2.37	0.08
	IDistill (Ours)	2.05	4.26	0.16
Average	Inception	8.11		
	PW-MAD	10.62		
	MixFacenet	8.96		
	OrthoMAD (Ours)	7.66		
	IDistill (Ours)	2.89		

Since we follow the protocol introduced by Damer *et al.* [109], the comparison between our methods and the ones in the literature focuses on the above mentioned approaches. The results of our methods, OrthoMAD and IDistill, are displayed in Table 5.2. As seen, IDistill has been able to surpass MixFacenet and Inception in all the test databases, and PW-MAD in four out of five databases. OrthoMAD has better results on two databases. A careful analysis of the results highlights an important notion that IDistill is fairly more consistent, as such, the improvements on the databases where it surpasses the literature are much wider than the loss in the performance on the two other databases. Looking at the most extreme examples, in FRL-OpenCV the EER of IDistill is only 1.73 percentual points larger than the value obtained by OrthoMAD, while IDistill decreases the EER in 11.22 percentual points when tested in the FRL-WebMorph dataset, which constitutes a much more relevant difference in performance. While looking beyond EER it is also possible to see a wide improvement on the BPCER@APCER at both 1% and 20%. Moreover, on FRL-OpenCV the higher EER of IDistill is mitigated by a lower BPCER@APCER = 20%.

When compared to OrthoMAD, IDistill presents an architecture with the same computation

cost on inference, but significantly more interpretable, since OrthoMAD does not guarantee that the information yield by both vectors is related to identity. We are capable of identifying not only attacks, but justify utilizing the information of which vectors contain identity information, and which do not. Due to the approximation between the auto-encoder latent space and the IDistill latent space, it might also be possible to reconstruct parts of the identity utilizing the decoder. While not used in this study, the information regarding the intensity of the vectors extracted by the morphing classifier, v_1 and v_2 might also allow to infer the morphing percentage associated with each fused identity, which might be useful in future works.

These results show that even with a less parameterized backbone network, our regularization terms were capable of achieving state-of-the-art performance in the task of morphing attack detection.

5.7 Conclusion

In this Chapter we have presented two novel methods for face morphing attack detection that are interpretable, compact and perform at the state-of-the-art level. The proposed IDistill method was trained utilising a two step scheme based on the training of an auto-encoder to reconstruct *bona-fide* images, and a distillation step integrated on the standard training of a morphing classifier, utilizing the encoder as teacher and the first part of the classifier as student. On the other hand, the proposed OrthoMAD method focused on establishing a constraint regarding the orthogonality of two vectors.

IDistill relaxed the orthogonality constraint from OrthoMAD, and presents a more consistent and reliable solution to ensure that the identity information is, in fact, separated in two individual vectors. Moreover, it dismisses the concatenation of these vectors, ensuring an interpretable analysis of the scores produced by each and their contribution to the final prediction. As future work on the interpretability capabilities of this study, it would be interesting to explore the reconstruction capabilities utilizing the identity vectors and the decoder model from the auto-encoder architecture. Another possible direction is to verify whether the intensities of the vectors extracted by the morphing classifier allow to quantify the morphing percentage of the identities that were fused to generate each attack sample.

Overall, IDistill surpasses the performance of the previous methods published in the literature, while ensuring the advantages previously mentioned. In some scenarios the performance is drastically better. There is much work to be done on the topic of face morphing attack detection, nonetheless, IDistill is a step forward towards the integration of interpretable approaches that are competitive with fully black-box systems.

Chapter 6

Guiding the Model Vision for Masked Face Recognition

Author Contributions The research work described in this chapter was conducted in collaboration with João R. Pinto, Naser Damer, Fadi Boutros and Mohsen Saffari, under the supervision of Ana F. Sequeira and Jaime S. Cardoso. The author of this thesis contributed to this work on the problem conceptualization, data curation, the preparation and conduction of experiments, the results discussion, and publication writing. Some parts of the chapter were originally published in, or adapted from:

- Neto, P. C., Boutros, F., Pinto, J. R., Damer, N., Sequeira, A. F., & Cardoso, J. S. (2021, December). Focusface: Multi-task contrastive learning for masked face recognition. *In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) (pp. 01-08). IEEE.*
- Neto, P. C., Boutros, F., Pinto, J. R., Saffari, M., Damer, N., Sequeira, A. F., & Cardoso, J. S. (2021, September). My eyes are up here: Promoting focus on uncovered regions in masked face recognition. *In 2021 International Conference of the Biometrics Special Interest Group (BIOSIG) (pp. 1-5). IEEE.*
- Boutros, F., Damer, N., Kolf, J. N., Raja, K., Kirchbuchner, F., Ramachandra, R., ... Neto, P. C., Sequeira, A. F., Pinto, J. R., Saffari, M., & Cardoso, J. S. (2021, August). MFR 2021: Masked face recognition competition. *In 2021 IEEE International joint conference on biometrics (IJCB) (pp. 1-10). IEEE.*

6.1 Context and Motivation

SARS-CoV-2 pandemic and related hygienic concerns intensified the necessity to adopt contactless biometric recognition systems. One of the most common contactless biometric recognition systems that meet high accuracy requirements is face recognition. It has been widely adopted in several scenarios such as automated border control, access control of private areas and convenience applications [192, 315]. Nonetheless, the pandemic was also responsible for a drastic increase in the use of face masks. These masks, enforced by several governments, are a form of face occlusion and limit the performance of face recognition systems. Some recent studies on the effects of the face mask on face recognition accuracy have indicated a significant degradation in the accuracy of these models [106, 377, 378, 243].

The new scenario, imposed by the pandemic situation, motivated the development of face recognition approaches robust to face masks. Through this, it motivated research methods, competition and other forms of scientific contributions. In this Chapter, we introduce two distinct approaches that have been explored for masked face recognition (MFR). The first, originated from a submission to one of most relevant competitions on the topic and part of the International Joint Conference on Biometrics (IJCB 2021) [48]. The second, called FocusFace, extended the first by retaining the core idea and improving on nearly all factors surrounding the design of the final solution. In the former we extend the triplet loss to incorporate information about masked versions of the anchor sample, whereas on the second we extend ArcFace [126]. loss to incorporate template approximation on masked and unmasked samples. Current works focus on the addition of synthetic masks to the original dataset [23] or restraining the input to contain only the periocular area of the face [300]. Our approach promotes the focus on uncovered areas of the face.

With the first method, we introduce domain information regarding the mask invariance. In other words, by approximating masked and unmasked anchors, we promote similar templates on both, and through that we reduce the contribution of masked areas to the template. In the subsequent work, by combining cross-entropy and ArcFace loss into a multi-task loss for joint mask detection and face recognition we promote a behavior similar to the one shown by periocular methods. However, we do not restrain the available information to be exclusively from the periocular area. Hence, we argue that both our methods learn where to focus in a end-to-end fashion.

Both methods are trained with masked and unmasked versions of an image. For each of those images, the network predicts the corresponding identity and, in FocusFace, whether the person is wearing a mask. The focus on unmasked areas is promoted by minimizing the distance between the masked and unmasked embeddings produced by the network in a contrastive learning fashion. The constrain on the invariance to the mask acts as a strategy to include domain information in the training process. Accordingly to the xAI Ladder proposed in Chapter 3, this information increases our knowledge regarding the behavior of the model, thus improving its interpretability. Moreover, as mentioned by Boutros *et al.* [48], the vast majority of current methods are focused on training models from scratch. However, ours can also be utilized on top of state-of-the-art face recognition methods.

The proposed methods are trained and validated on a dataset using synthetic masks [126, 218], and evaluated on a dataset that contains real masks [106, 104]. The latter was also used to evaluate the solutions of the International Joint Conference on Biometrics (IJCB 2021) competition on masked face recognition. After validating the invariance constraint on the first method, FocusFace verification results support the performance gains of our approach with an FMR100 (the lowest false non-match rate (FNMR) for false match rate (FMR) ≤ 1.0 %) improvement over the regular ArcFace loss.

This Chapter is further divided into four major sections. Section 6.2 discusses related work focused on masked face recognition. Afterwards, we dedicate Section 6.3 to the methodology and results of the first method proposed in this Chapter, and Section 6.4 for the second. Finally, the conclusions are drawn in Section 6.5.

6.2 Related Work

Recent research on face recognition made significant improvements in the verification performance on non-occluded faces [45, 46, 126]. Before the COVID-19 pandemic, the works on occluded faces were rather limited and scarce, and focused on sunglasses or scarves [386, 463]. Following the wide adoption of face masks during the pandemic situation, several studies examined the effects of wearing a mask on face recognition [182]. Two important studies were conducted by the National Institute of Standards and Technology (NIST). The first evaluated pre-COVID-19 algorithms on masked face recognition [377], whereas the second evaluated post-COVID-19 algorithms [378]. These studies are a component of the ongoing Face Recognition Vendor Test (FRVT), and indicated that competitive algorithms on unmasked faces, which fail to authenticate less than 1% of the individuals, have their failure rate increased up to 50% in the presence of masks. Despite the conclusions of these studies indicating that face masks hurt the performance of face recognition, their validity is limited due to the exclusive use of synthetic masks. Hence, it might not capture the real effect of wearing masks. Damer *et al.* [106] have additionally studied the limitations caused by face masks on two academic face recognition algorithms and one commercial solution. In their study, the evaluation dataset was collected over three sessions with 24 participants. Both Damer *et al.* [106] and a Department of Homeland Security (DHS) study [243] showed the performance degradation that occurs whenever face masks are worn. The first has shown that the SphereFace method can have its FMR100 increased from 0.065% to 27.35% when comparing conventional face recognition to masked face recognition. Wearing a mask have been additionally shown to significantly affect the verification performance of human operators [102], as well as other components of face recognition systems like face quality estimation [168] and presentation attack detection [156].

The most active research area related to facial images and face masks is the detection of the latter. Several articles focused on this task have been published [313, 410]. Nevertheless, these studies present virtually no improvement to the task of masked face recognition. Neither focused on evaluating the impact of the mask on face recognition. The work presented by Li *et al.* [300], which proposed an attention-based method to train a model on the periocular area of masked faces, has shown interesting improvements. However, its evaluation was limited by the exclusive use of simulated masks. Other works with limited testing, either due to the use of simulated masks or small datasets, have been published [23]. The experiments focused mostly on fine-tuning pre-trained models, such as FaceNet [439], with simulated masks.

In an attempt to mitigate these limitations, Boutros *et al.* [48] have promoted a competition within the 2021 International Joint Conference on Biometrics (IJCB). The works that joined the competition were evaluated on a dataset containing real masks and a significant number of samples (1,511,113 masked-masked pairs and 800,207 unmasked-masked pairs) [106, 104]. The vast majority of the competing solutions relied on retraining the ArcFace model with datasets containing simulated masks. One of the approaches proposed multi-task learning to jointly detect masks and identify the correct sample identity [347]. That multi-task model, despite learning to capture

masked related patterns, did not have a mechanism to promote the use of that information to improve face recognition by avoiding masked areas. There have also been methodologies based on modifications to the triplet loss [47].

Motivated by the endeavours of the solutions proposed to the IJCB 2021 competition and by the recent evaluations on the harmful effect of face masks on face recognition performance, we present in this Chapter a novel approach to improve masked face recognition. Differently from previous approaches, our multi-task method is evaluated on a large test set and contains mechanisms to explore the learnt features related to the mask in order to improve masked face recognition. The proposed solution can be used on top of existing solutions or trained from scratch, thus it does not require the sacrifice of conventional face recognition performance. The details are presented in Sections 6.3 and 6.4.

6.3 Conditioned Triplet Loss

In this section, we describe the methodology behind the solution submitted to the IJCB 2021 competition on masked face recognition, and a set of results that go far beyond those described in the competition document. It comprises the construction of a synthetic masked face dataset based on the VGGFace2 [71] and proposes a solution to address the challenge of MFR. This solution is based on a proposed loss that combines the traditional triplet loss and the mean squared error (MSE) intending to improve robustness in the comparison between masked and unmasked samples.

One should keep in mind that this is a description of the submitted solution, with its flaws and domain gaps, which have been posteriorly solved by FocusFace. Its main contribution is the validation of the mask invariance constraint. Thus, paving the way for future solutions.

6.3.1 Methodology

Due to the difficulty of selecting triplets, and several other restrictions, performing direct optimization of embedding predictions is often trickier than learning a model capable of performing classification. Hence, we propose a constrained triplet loss, specially crafted for masked face recognition (MFR), to be used after the classification optimization. In our approach, we aim to redirect the focus of an already trained model towards unmasked regions of the face image. While classification training is performed with a single image, our optimization process requires four images: a positive sample, a negative sample, an anchor and a masked version of that anchor.

We start by describing the approach followed to train these models for classification. Afterwards, we discuss our proposed loss that improves the representations learned for MFR.

Classification Training: Face recognition can be explored in two scenarios. Close- and open-set problem solving. The first, is solved under the assumption that the classes used for training remain the same for validation and testing. On the other hand, open-set formulation assumes that test images might originated from a unseen class. In this first stage, we treat the problem with a

close-set, where the model is optimized to associate a face with an identity present in the training data. For this, we explore cross-entropy to train the model on 80% of the data and validate it on the remaining 20%. Due to its closed-set formulation, the data split process did not ensure mutual exclusive identities within train and validation sets. This training process is similar to the one designed by Cao *et al.* [71].

Embedding Optimization: Through embedding optimization we aim to learn the representations of the inputs, instead of classifying them. From the previously trained network, we remove the final classification layer and replace it with a fully-connected layer outputting a 512-d vector. While this layer is trainable, every other weight in the network is frozen. This layer is trained with a triplet loss, as described in Equation 6.1.

$$TripletLoss = \sum_{a,p,n \in \mathcal{T}} \max(0, \alpha - \|x_a - x_n\|_2^2 + \|x_a - x_p\|_2^2) \quad (6.1)$$

$$x_i = W' \frac{\phi(l_i)}{\|\phi(l_i)\|_2} \quad (6.2)$$

Equation 6.2 describes the euclidean normalization and projection, parameterized by W' , of the frozen layer's output, $\phi(l_i)$, to an embedding space with 512-d.

Triplet loss verifies the distances between the anchor (a) and the positive (p), and between the anchor (a) and the negative (n). It penalizes the network if the last one is smaller by a given margin (α). This margin helps the model to define the degree of separability between positives and negatives.

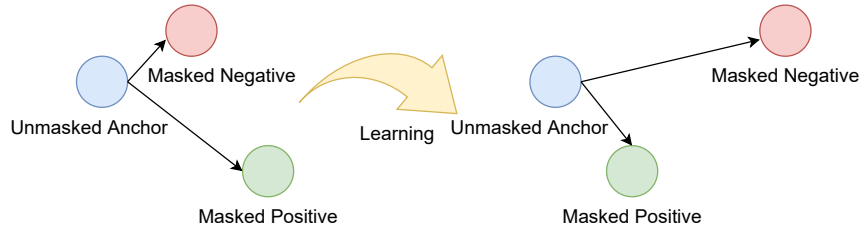


Figure 6.1: Triplet loss effect in the euclidean space

Figure 6.1 illustrates the expected distribution of the embeddings before and after optimization with the triplet loss. The anchor is a randomly selected image. From the same identity of the anchor, we select a different image, our positive. And from a different identity we select our negative. As in the majority of face verification challenges, the reference image is unmasked, we set the anchor to be unmasked and the remaining samples to be masked.

We propose to a change to this interpretation. To promote a mask invariance behavior, we aim to map both the original anchor (a), and a masked version of that anchor (a_m) to the same embedding. This can be done by minimizing the distance between the vectors produced by both versions of the anchor, as demonstrated in Equation 6.3.

$$TripletLoss_{prop} = \sum_{a, a_m, p, n \in \mathcal{T}} TripletLoss(a, p, n) + MSE(a_m, a) \quad (6.3)$$

The new loss implies a slight change in the spatial distribution of the embeddings post-optimization. Such change is illustrated in Figure 6.2.

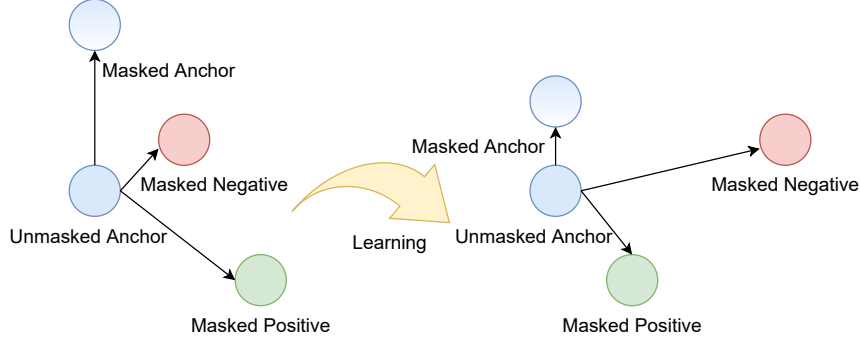


Figure 6.2: Proposed Triplet loss effect in the euclidean space

6.3.2 Experimental Setup and Results

Evaluation Metrics: To report the results we present the *false non-match rate (FNMR)*; the *FMR100* and *FMR10* which are the lowest *FNMR* for a *false match rate (FMR)* $< 1.0\%$ and $< 10.0\%$, respectively; the *equal error rate (EER)*; and the *area under the receiver operating characteristic curve (AUC)*. We also report the *genuine mean (GMean)* and *impostors mean (IMean)*, which represent the mean distances between the embeddings of the same individual and from different people respectively.

Face Data: The development of face recognition methods requires large and diverse datasets. When the Covid-19 pandemic started and it became evident that the use of face masks had a negative impact on FR systems there was no ready-to-use data for research. The creation of synthetic data allows leveraging existing data to fit the problem. Still, using real data is crucial as the ultimate test to the models and the community started to also collect face samples of individuals using face masks.

Synthetic masked face data (SMFD): Here we describe the synthetic masked face data (SMFD) creation process. Adding a facial mask requires information regarding facial landmarks. Moreover, the position and inclination of the face affect the positioning of the mask. Due to the lack of large-scale pairs of masked and unmasked identities, in this work, we synthetically generate different types of masks and adjust them on the unmasked samples of the VGGFace2 dataset [71]. The dataset includes 3,310,000 face images from 8,631 train identities and 500 disjoint identities for the test, and includes a diverse set of samples regarding the various poses, ages, and ethnicity. Mask generation is carried out using the proposed algorithm by NIST [374]. The algorithm exploits the Dlib C++ toolkit to obtain 68 facial landmarks for each image; afterward, using the

extracted facial landmarks and interpolation between the points, various synthetic masks are generated. The details of the landmark extraction and mask generation are described in [374, 271]. Due to variability regarding shape (Wide vs. Round) and face coverage (High, Medium, Low) we obtained six possible combinations. Figure 6.3 shows the result of applying the mask generation algorithm on a randomly selected sample from VGGFace2. Both the shape of the mask and its colour are randomly selected, for each image, while generating the masks.

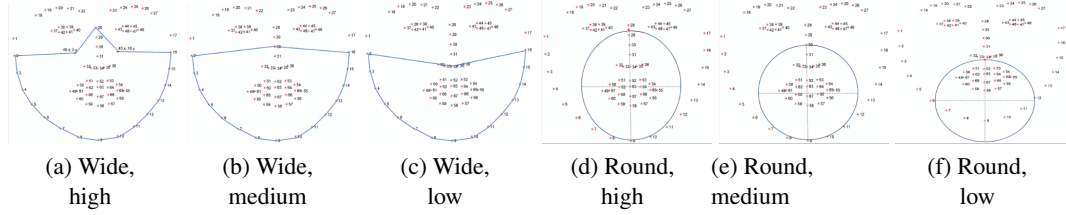
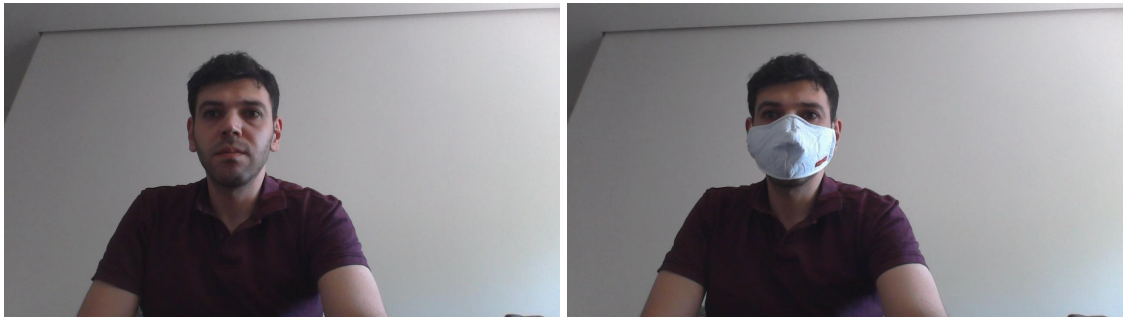


Figure 6.3: Examples of face landmarks obtained from one image with different types of masks added. These masks vary in shape and face coverage.

Real masked face dataset: We evaluated our proposed solution using masked face recognition competition dataset (MFRC-21) [48] (Fig. 6.4). MFRC-21 dataset was collected on 3 different days from 47 subjects. The first day is considered as a reference session, while the second and third sessions are considered probe sessions. In each session, 3 (two with mask and one without) videos are recorded using a webcam, while the subjects are requested to look directly into their camera. An overlapping database, and the same capture and frame selection procedure is describe in [104, 106]. In total, the references contain 470 unmasked images and 940 masked images. The probes contain 940 unmasked images and 1880 masked images. We evaluate our proposed solution under two evaluation scenarios. The first is between unmasked references and masked probes (U-M) and the second is between masked references and masked probes (M-M).



(a) Example from the test set of a reference image without a mask (b) Example from the test set of a probe image with a mask

Figure 6.4: Examples of images from the real masked faces dataset: images of the same individual with (a) and without a mask (b) (from [106]).

Implementation Details: To implement the mask FR model, we exploit ResNet-50 architecture [205] to extract the features from masked/unmasked facial samples. We trained the model by

Table 6.1: Results obtained for the synthetic masked face data (EER, FMR100, FMR10 in %).

Method	GMean	IMean	AUC	EER	FMR100	FMR10
VGG Face[71, 439]	0.505	0.325	0.951	11.8	38.2	13.5
CE Loss	0.528	0.426	0.941	13.2	38.5	21.5
CE + TL	0.601	0.320	0.977	7.8	28.9	11.9
CE + TL + MSE (Ours)	0.596	0.319	0.985	6.2	18.5	4.1

making use of cross-entropy (CE). The training of this model required around 150 thousand iterations. It trained with an initial learning rate of 0.1. It was decreased by a factor of 10 whenever the validation accuracy decreased. Stochastic gradient descent was used, with a batch size of 400, momentum of 0.9, and 0.0005 weight decay. We did not use any face alignment of the VGGFace2 dataset, giving as input of the model 224x224 images. After achieving convergence, it was fine-tuned with triplet loss and the combination of triplet loss with the mean squared error. Triplets were randomly generated at training time, and thus, it was unlikely to have them seen by the network more than once. Furthermore, we did not use any triplet mining. The models trained for 65 thousand iterations with a batch size of 200. And thus, 13 million triplets were created and used for the weight updates. The margin hyper-parameter α in Triplet loss is empirically determined as 0.2.

Results and discussion: We evaluated our method on two distinct datasets. One evaluation used synthetic masked face data (SMFD), with all the identities used for testing being disjoint from the training identities. We also evaluated the model with real masked face data (RMFD). Evaluating on these datasets allows us to infer the generalization capabilities of the model for unknown identities and images with different characteristics from the gallery images (e.g. real masks). The results are provided using the already mentioned metrics: GMean, IMean, AUC, EER, FMR100, and FMR10.

Our method is evaluated through a detailed step-wise ablation study that allowed us to understand the impact of the proposed modification of the triplet loss. Hence, we evaluate the model, in both datasets, with different training frameworks. This allows us to capture information regarding the impact of individual components of the model, such as, training with triplet loss, or not optimizing the embeddings. Besides, we also included the results of the method referred to in the tables as “VGG Face” [71, 439] consisting of an Inception-ResNet pre-trained on the original VGGFace2 datasets. In Table 6.1, one can observe stable performance on synthetic mask data with just cross-entropy training (CE). However, optimizing the produced embeddings with triplet loss (CE+TL) led to significant improvements in performance, lower mean distance for impostors, and higher mean distance for genuines. Moreover, our proposed adapted triplet loss resulting from the addition of the MSE constraint (CE+TL+MSE) lead to higher improvements across all metrics, for example, approaching the AUC to 0.99.

In Table 6.2 one can observe that the model’s performance on real masks degrades when compared to synthetic masks. Regardless of that, for the targeted comparison mode - U versus M -

Table 6.2: Results obtained for the real masked face data (in the column “Mode”: U and M stands for unmasked and masked data, respectively; EER, FMR100, FMR10 in %).

Method	Mode	GMean	IMean	AUC	EER	FMR100	FMR10
VGG Face[71, 439]	U-M	0.523	0.426	0.769	29.419	90.587	58.959
	M-M	0.616	0.461	0.847	23.552	68.979	38.159
CE Loss	U-M	0.610	0.475	0.931	11.687	32.041	12.852
	M-M	0.702	0.503	0.936	9.002	16.628	8.791
CE + TL	U-M	0.647	0.396	0.943	11.213	34.744	11.874
	M-M	0.699	0.414	0.945	10.806	26.457	11.249
CE + TL + MSE (Ours)	U-M	0.649	0.383	0.957	9.799	28.252	9.678
	M-M	0.699	0.390	0.959	9.292	23,507	9.035

the results show the superior performance of our proposed loss. Furthermore, it is possible to conclude that the model is competent in the task despite being trained only on synthetic data. It should be noted that, the M-M mode experimental results do not keep up with the U-M one, in other words while our method improves the U-M, it offers no improvements for M-M recognition. This can be due to the fact that the embedding optimization process is made in a way that the model is trained to minimize the distance between masked and unmasked (U-M) genuine pairs, thus aiming at making it greater than the distance between impostor pairs. However, main application scenarios commonly contain unmasked references, such as automatic border control with an unmasked passport-stored reference and a possible masked live probe.

Besides the quantitative evaluation of the proposed method, we aimed at verifying that our constraint was, as expected, promoting a focus on unmasked areas of the face. In order to infer the importance of each pixel for the overall embedding, we used the Smooth Grad-CAM++ method. The output of this gradient-based method was computed for each of the 512 features. Afterwards, all the map outputs were summed and divided by 512, thus generating a final map with the normalized importance of the pixels to the overall embedding. Figure 6.5 displays the outputs for three of the studied methods, from the top to the bottom we have CE, TL and TL+MSE methods. While the first is already capable of ignoring the masks, it still constructs the embedding of the unmasked images based on the chin area. Between the other two models, it is not clear the difference between their outputs. However, for the proposed model, the masked/unmasked pair on the first two columns seems to share a similar heatmap, highlighting what was promoted by the additional constraint.

6.4 FocusFace

The model proposed and described in the previous section validates the ability to learn how to focus more on unmasked regions of the face. However, it is flawed in several senses. First, the use of a softmax-based loss is suboptimal for FR, leading to a final model with suboptimal performance. Furthermore, it does not perform any alignment on the train, validation and test

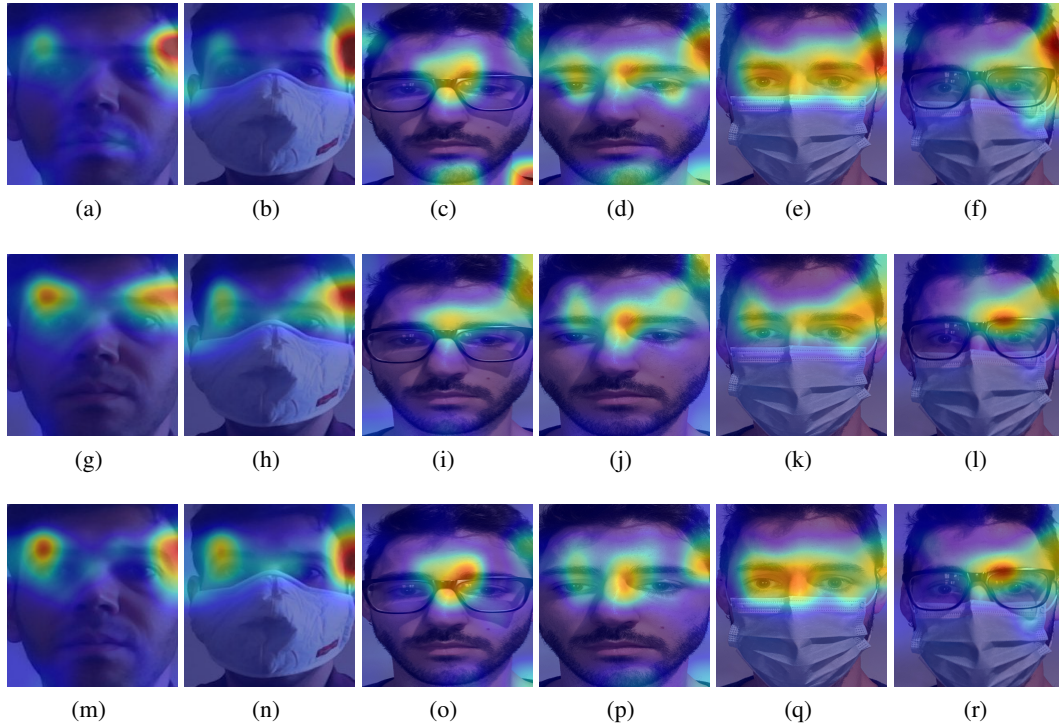


Figure 6.5: Output of the Smooth Grad-CAM++ computed for each of the 512 features of the feature vector and normalized. Subfigures (a)-(f) computed from the cross-entropy model; (g)-(l) computed from the triplet loss model; and (m)-(r) computed from the mean squared error and triplet loss model.

images, further hurting the final performance and the optimization process. Triplet loss is also outdated, and expensive to run. Finally, the previous model explored a cascade setting with two separate training processes. With this subsequent work we aimed at mitigating those flaws and contribute further for novel methods in masked face recognition.

6.4.1 Methodology

The proposed approach is composed of two distinct components designed to work together. First, the solution delves into a multi-task learning approach so that the network becomes aware of the existence of a face mask. Secondly, mask awareness promotes a more stable learning of a regularization loss between the features of both masked and unmasked images. Our approach is further detailed in the following paragraphs.

6.4.1.1 Multi-task architecture

The proposed architecture contemplates two distinct tasks. First, the most important of both is the recognition and classification of the image identity. The other is responsible for the detection of a face mask in the input image. The architecture uses a convolutional neural network to produce two distinct branches, one per task. They are optimized for their respective tasks.

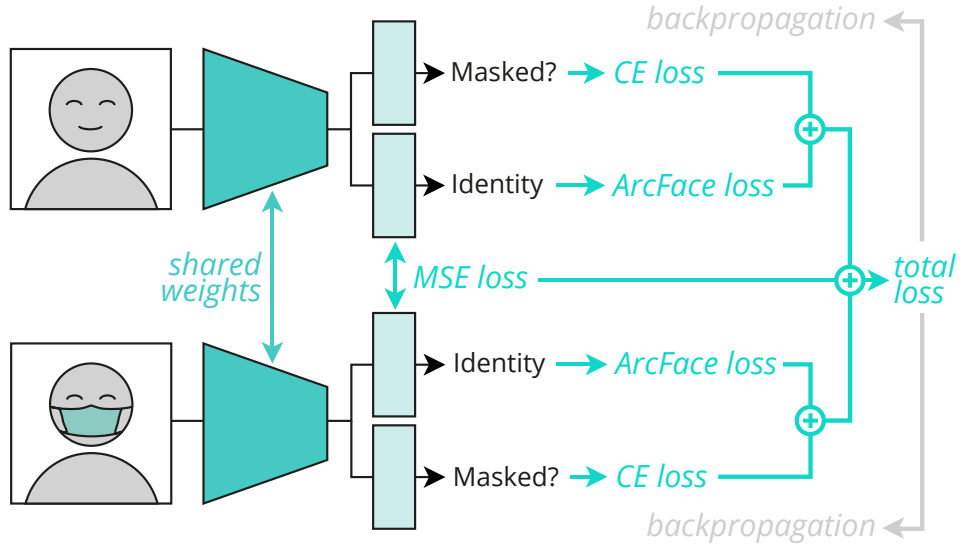


Figure 6.6: Architecture of the proposed model. The backbone network receives masked and unmasked inputs, and computes the two embeddings for each input. The smaller embedding of each input is trained for mask detection with the cross-entropy (CE) loss. Whereas the other is trained for face recognition with the ArcFace loss. A third component of the loss is also included in the form of the mean squared error (MSE) between the two embeddings used for face recognition.

By adding a component to the network to improve its capabilities to distinguish between masked and unmasked individuals, it is created a mechanism to promote mask awareness. This awareness is intended to improve the convergence stability and the features extracted. For this, the branch dedicated to the detection of the mask is composed of a fully connected layer with one output neuron. The optimization for this task is carried by minimizing the cross-entropy loss. The formulation of this loss is seen in Equation 6.4, where N represents the number of samples in the mini-batch, n is the class number, and y_{it} represents the output of the network for the sample i and target class t . In our specific case, we are optimizing a special case of this loss known as the binary cross-entropy loss.

$$L_{CE} = -\frac{1}{N} \sum_{i \in N} \log \frac{e^{y_{it}}}{\sum_{k \in n} e^{y_{ik}}} \quad (6.4)$$

For the recognition (i.e. identity classification) task, the state-of-the-art ArcFace loss was used. ArcFace showed impressive results when compared to the conventional softmax-loss due to the explicit enforcing of a higher intraclass similarity and diversity for inter-class samples. Through the l_2 normalization of both the weights and the feature vector, this loss becomes equal to the geodesic distance margin penalty in a normalised hypersphere. The formulation for this loss can be seen in Equation 6.5, where m represents the margin, s the scale, and θ_{it} the angle between features, x_i and weights, W_t , for sample i and target class t . Both m and s are hyperparameters of the loss.

$$L_{arc} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(\theta_{it}+m))}}{e^{s(\cos(\theta_{it}+m))} + \sum_{j=1, j \neq t}^n e^{s \cos \theta_{ij}}} \quad (6.5)$$

6.4.1.2 Masked embedding regularization

The multi-task configuration of the network ensures that the network learns useful features to be aware of the mask presence. However, there are no guarantees that those features are leveraged by the main recognition branch. In fact, in practice, it can negatively impact the network performance whenever not carefully handled. Due to this, the proposed method introduces a regularization term for training to promote mask awareness on the main branch.

This new term aims to give two images from the same individual to the same network. One of the images is augmented with synthetic masks, whereas the other has no mask augmentation. At loading time if one of the images is horizontally flipped, the other is also flipped. The produced face embeddings of both images are given to a loss function that calculates the distance between them. The formulation of this loss is seen in Equation 6.6, where F stands for the feature vector size, xm_{ij} represents the value for masked feature vector i at position j , whereas xu_{ij} represents the same but for the unmasked feature vector.

$$L_{MSE} = \frac{1}{N} \frac{1}{F} \sum_{i=1}^N \sum_{j=1}^F (xu_{ij} - xm_{ij})^2 \quad (6.6)$$

By enforcing smaller distances between masked and unmasked images, the network focus is promoted towards common features to both images. This also improves the capabilities of the model at the verification of masked references and masked probes. The selection of the augmented image was subject to careful evaluation. For this, the experiments consisted of training a model where the masked image was the masked version of the unmasked image and another where the masked image was a random masked version of any other image of the individual.

6.4.1.3 Global architecture

The overall architecture of the model comprises the previous two modules and a backbone convolutional neural network as seen in Figure 6.6. For the backbone, the selected architectures were the 50 and 100 layers versions of the ResNet [205]. The last layer of these was replaced by two parallel fully-connected layers that produce two distinct embeddings. The first embedding with a size of 512 contains the facial features extracted for the verification task, whereas the second with a size of 32 is responsible for the mask detection.

The modules introduced in this architecture can be used on top of already trained backbones. This leads to faster convergence times overall. To show that, we included in the experiments the pretrained ResNet-100 backbone called ElasticFace [46].

To train the entire architecture end-to-end, it is necessary to combine the losses. First, the multi-task loss must be formulated, as seen in Equation 6.7. The network runs twice, one with masked images and the other with unmasked images. Initially, this loss is computed independently

for both runs. Afterwards, it is combined on the final loss. The hyperparameter λ is responsible for controlling the impact that the mask detection has on the loss.

$$L_{masked/unmasked} = L_{arc} + \lambda L_{CE} \quad (6.7)$$

The masked and unmasked branches are combined on Equation 6.8, which represents the formulation of the final loss. In this loss, the component regarding the L_{MSE} has its impact controlled by the hyperparameter α , whereas the impact of the combination of the two branches (masked and unmasked) is controlled by β . It is important to note that the weights are shared between the networks that are used to computed both masked and unmasked inputs, and that its optimization occurs simultaneously for all loss components.

$$L_{comb} = \alpha L_{MSE} * \beta (L_{masked} + L_{unmasked}) \quad (6.8)$$

6.4.2 Experimental Setup

6.4.2.1 Training details

The backbone architectures to be trained for masked face recognition are the ResNet-100 and ResNet-50 [205]. The selection of these architectures was motivated by the wide use of the first in the literature and state-of-the-art face recognition methods [20, 126, 142, 220, 468]. Similarly to the literature [46], on ResNet-100 we set the parameter s to 64, and on the ResNet-50 we set it to 30. In both architectures, m was set to 0.5. Due to GPU memory constraints, experiments that trained the backbone from scratch used a mini-batch size of 480. If the backbone was pretrained and frozen the mini-batch size increased to 512. The models are implemented using Pytorch [390] and ran on one Linux machine with 4 Nvidia Tesla V100 34gb GPUs. All the experiments were trained with Stochastic Gradient Descent (SGD) with an initial learning rate of 1e-1, a momentum of 0.9 and weight decay of 5e-4. The learning rate is decreased by a factor of 10 at 34k and 64k iterations for models with pretrained backbones, and at 60k and 96k for the others. While loading the images at training, the image is horizontally flipped with a probability of 0.5. The networks expect 112x112x3 sized images and produce 512-d and 32-d embeddings. The latter is only computed at training time. The images are pre-aligned and cropped using the Multi-task Cascaded Convolutional Networks (MTCNN) [546]. The images are also further normalized to have pixel values between -1 and 1. Regarding the hyperparameters of the loss, λ , α and β are set to 0.1, 1/3 and 1/2, respectively.

6.4.2.2 Training dataset

Similarly to recent works in the literature [20, 126, 220, 335], we used the MS1MV2 dataset [126]. The MS1MV2 is a refined version of the original MS-Celeb-1M dataset [198]. This refined dataset contains 5.8M images of 85k distinct identities. We augmented this dataset with face masks to twice its original size while not interfering with the identities. For these augmentations, we used

the MaskTheFace open-source code [23]. The mask generation process is offline, thus not interfering with training times. MaskTheFace has five types of masks from which we used four (surgical, N95, KN95, cloth). The type of the mask and its colour were selected randomly from all the available options. A similar methodology was applied to all the probes in the Labeled Faces in the Wild (LFW) dataset [218]. Due to its wide use in the literature, the LFW dataset is used for the validation of the algorithm after each epoch.

6.4.2.3 Evaluation benchmarks and metrics

We report the verification performance through the Equal Error Rate (EER), the area under the curve (AUC), FMR100 and FMR10. The latter two are the lowest FNMR for an $\text{FMR} \leq 1.0\%$ and $\leq 10.0\%$, respectively. We further report the mean of genuine predictions (Gmean) and the mean of impostor predictions (IMean) to analyse the distribution of the scores and how distant are the means of these two categories. We evaluate our approach in two distinct settings. Firstly, the solution was evaluated in the U-M setting, which has unmasked references and masked probes. Afterwards, the evaluation in the M-M setting, where both probes and references are masked, was also conducted. The validation of the model and the model selection was carried on the LFW dataset. Figure 6.7 shows some examples of the images that are included in the dataset. The metric used to select the final model was the FMR100 on the U-M setting in the validation dataset.



Figure 6.7: Images extracted from the validation dataset for masked face recognition: images of the same individual with (a) and without a mask (b) with a accurate augmentation; images of the same individual with (c) and without a mask (d) with a poor augmentation.

6.4.2.4 Test dataset

While for sanity validation of the model performance the LFW dataset with synthetic masks was used, the values reported for the experiments are with respect to another database. The proposed solution, trained on synthetic data, is evaluated on a dataset containing real masks. This dataset was the one used in the masked face recognition competition [48]. It was collected during 3 different days from 47 subjects. The images captured on the first day are used as references, whilst the probes were captured during the second and third sessions.

In each session, three videos were recorded using a webcam, while the subjects looked directly at the camera. Two of the videos included masks, whereas the third did not. An overlapping

database, and the same capture and frame selection procedure is described in [106, 104]. The references contain 470 unmasked images and 940 masked images. The probes contain 940 unmasked images and 1880 masked images. These distinct captures allowed for the evaluation of the U-M and M-M configurations.

Table 6.3: Experiments on the distinct configurations of the proposed method on the test dataset (in the column “Selection”: Original stands for the use of the masked version of the unmasked image on the second run of the network, whereas Random stands for the use of a randomly picked masked image from the same subject of the unmasked image; in the column “Mode”: U and M stands for unmasked and masked data, respectively; EER, FMR100, FMR10 in %).

Method	Pretrained	Selection	Mode	GMean	IMean	AUC	EER	FMR100	FMR10
ResNet-50	No	Original	U-M	0.561	0.333	0.988	5.377%	7.285%	4.519%
			M-M	0.625	0.341	0.983	5.346%	6.546%	4.642%
ResNet-100	Yes	Original	U-M	0.571	0.357	0.987	4.917%	5.986%	3.922%
			M-M	0.627	0.365	0.983	5.005%	6.109%	3.882%
ResNet-100	Yes	Random	U-M	0.567	0.354	0.984	5.371%	5.986%	4.991%
			M-M	0.625	0.361	0.984	5.581%	6.184%	5.056%
ResNet-100	No	Original	U-M	0.624	0.373	0.992	4.594%	5.750%	2.582%
			M-M	0.675	0.383	0.992	4.329%	5.509%	2.836%
ResNet-100	No	Random	U-M	0.621	0.371	0.988	5.164%	5.917%	4.197%
			M-M	0.671	0.382	0.991	4.757%	6.302%	3.668%

6.4.3 Results

In this section, we present the results achieved by our method and how it compares with the literature. Comparisons focus mostly on variations of the proposed method, on the method presented in the Section 6.3 of this Chapter (CE + TL + MSE) and the solutions submitted to the competition hosted by Boutros *et al.* [48].

We tested several variations of the proposed method. From using a smaller backbone network to adding the proposed modules on top of pretrained methods, and distinct image selection approaches to the regularization module. Four of these variations are capable of beating the baseline method presented at the masked face recognition competition [48] for the U-M setting. For M-M there is model capable of outperforming the baseline. Table 6.3 shows the results of the several variations of the proposed method on the test dataset. The best performance is achieved by a ResNet-100 backbone trained from scratch for masked face recognition with pairs of images where the masked image corresponds to the masked version of the unmasked image. On the other hand, we show that with little training, it is possible to improve the performance of already existing models without changing the features learned by them for conventional face recognition. The majority of the solutions presented in the competition use ResNet-100, the training process required at least tuning 65 million parameters from the backbone and almost 44 million parameters from the

ArcFace layer. By freezing the parameters of the backbone (except the last layer), we reduce the parameters that require training by roughly 47.6%. All of this, while still beating the baseline results. When the backbone is frozen our approach trains $13,648,960 + 43,899,904 = 57,548,864$ parameters, from which $43,899,904 + 802,880 = 44,702,784$ are removed on inference. Our approach parameters are shown in Table 6.4, as well as their use on inference.

Table 6.4: Number of parameters of each module of the proposed approach. Includes the information regarding the use of each module for inference. Training parameters refers only to trainable parameters during training.

Module	#Parameters	Inference
ResNet-100 w/o Embedding Layer	52,309,568	Yes
Face Recognition Embedding Layer	12,846,080	Yes
Mask Detection Embedding Layer	802,880	No
ArcFace Layer	43,899,904	No
Mask Detection Fully-Connected Layer	66	No
Total (Training w/o pretraining)	109,858,498	-
Total (Training w/ pretraining)	57,548,930	-
Total (Inference)	65,155,648	-

As expected, when selecting the masked image for the regularization, using the masked version of the unmasked image leads to slightly better results. However, it is worth noting, that while it performs worse than our best model, using random masked images from the same individual still leads to better results than the baseline for both M-M and U-M.

In order to compare to the competition solutions, we adopted a selection criterion based on the loss function used, the feature vector size, the input size and the dataset used. Using this criterion we selected the Top-3 solutions. The solutions were named in the competition as A1_Simple (Top-1), VIPLFACE-M (Top-2) and MaskedArcFace (Top-3). These solutions used augmented versions of the MS1MV2 dataset, trained on ArcFace loss, expect inputs size of $112 \times 112 \times 3$ and return a feature vector of 512-d.

In Table 6.5 it is possible to see how the proposed method performs against the methods in the literature and the baseline on U-M face verification. The proposed approach achieved better results than the baseline and Section 6.3 method. However, it had a minor disadvantage when compared to the competition solutions. Nonetheless, the performance of those solutions is close to the proposed method, and there are explanations for the lower results. The best performing solution, the only one with a performance improvement ≥ 0.1 percentage points, used a model that has a roughly 34% larger number of parameters [48]. The second-best solution used a distinct synthetic mask generation technique, which can heavily impact the behavior of the final model. Finally, the third-best solution, accordingly to the description given in the MFR competition paper [48], seems to also apply synthetic masks to the test set. Allied with the non-deterministic behavior, distinct random seeds that lead to different random initialization, and potentially distinct mask

Table 6.5: Comparison of the FMR100 of the methods presented in the MFR competition [48], a baseline and the proposed approach. Methods evaluated on the competition dataset for U-M face verification. (FMR100 in %).

Method	FMR100
CE + TL + MSE	28.252%
Baseline [48]	6.009%
Top-1 solution [48]	5.538%
Top-2 solution [48]	5.681%
Top-3 solution [48]	5.687%
Ours	5.750%

augmentations (due to randomness on mask generation) these factors can explain the performance difference between methods. Moreover, the second and third-best performing solutions have only minor improvements of 0.069 and 0.063 percentage points, respectively.

Table 6.6: Comparison of the FMR100 of the methods presented in the MFR competition [48], a baseline and the proposed approach. Methods evaluated on the competition dataset for M-M face verification. (FMR100 in %).

Method	FMR100
CE + TL + MSE	23.507%
Baseline [48]	5.925 %
Top-1 solution [48]	5.771%
Top-2 solution [48]	5.759%
Top-3 solution [48]	5.825%
Ours	5.509%

Differently from U-M, in Table 6.6 it is possible to see that the proposed method outperforms the other solutions in M-M. Not only that but it outperforms every solution by larger margins than the ones shown in the previous table. When compared with our method, 1_Simple, VIPLFACE-M and MaskedArcFace have performance drops of 0.262, 0.250 and 0.316 percentage points, respectively. Which means that the proposed approach performs better by a margin at least ≥ 0.250 percentage points when compared to the remaining approaches. While this seems a secondary evaluation, it is possible to exploit the potential of our method on M-M by adding synthetic masks to the reference image.

Propelled by the results shown on both tasks and by the effectiveness of the implementation of these modules without changing the weights of the backbone network (pre-trained approach), we further argue in favor of the proposed method. Since the backbone is frozen during the training

and only new embeddings are created it is possible, with just an increase of 13,648,960 parameters (roughly increasing the network size by 20%), to compute simultaneously conventional face recognition, masked face recognition and mask detection embeddings. Thus, by leveraging the predictions of the masked detection embedding it is possible to decide between using the other two embeddings. Hence, avoiding any degradation to conventional face recognition, something that is not evaluated by the other methods in the competition. All of this, while still beating baseline performance.

6.5 Conclusion

In this Chapter, we proposed and evaluated a triplet loss-based approach and a novel multi-task approach to improve the verification accuracy on masked face recognition. The presented work was motivated by recent efforts on improving the verification accuracy of these models since there has been an increase in the use of face masks. The proposed solution is designed to work either on top of existing face recognition models or trained from scratch. It has been shown that the multi-task model outperforms the baseline. In addition, when running on top of existent methods, its training and convergence are considerably lighter and faster.

The main goal of our FocusFace model is to promote the focus on areas that are not covered by face masks. To accomplish this, we include an additional embedding optimized to detect face masks. Hence optimizing the network to recognize face masks. To the multi-task module, we add a contrastive learning approach to enforce the network into constructing face embeddings that do not contain any mask related information. We explore the combination of three loss functions and two distinct input batches per iteration.

By experimenting with several variations of the architecture and comparing with distinct approaches, we demonstrated the effectiveness of our approach on improving the masked face verification performance. Moreover, we show that thanks to the capability of the proposed model to detect masks and to be implemented on top of current face recognition models, there is no need to sacrifice verification performance on conventional face recognition. Future work includes several variations to the contrastive learning module, a pre-processing step that adds synthetic masks to test data, and the study of larger and more accurate backbone systems.

The regularization term included in both solutions behaves as a mechanism to add domain information to the model's training. Hence, it fits in the second rung of the xAI ladder (Chapter 3). This strategy helps promoting a behavior on the network, thus, reducing the space of candidate solutions.

Chapter 7

PIC: An Optimal Confidence Score for Interpretable Face Matching

Author Contributions The research work described in this chapter was conducted in collaboration with Philipp Terh rst, under the supervision of Ana F. Sequeira and Jaime S. Cardoso. The author of this thesis contributed to this work on the problem conceptualisation, data curation, the preparation and conduction of experiments, the results discussion, and publication writing. Some parts of the chapter were originally published in, or adapted from:

- Neto, P. C., Sequeira, A. F., Cardoso, J. S., & Terh rst, P. (2023). PIC-Score: Probabilistic Interpretable Comparison Score for Optimal Matching Confidence in Single-and Multi-Biometric Face Recognition. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1021-1029).

7.1 Context and Motivation

Biometric recognition systems, such as face recognition, have a growing effect on our daily life [502]. Since these systems are increasingly involved in critical decision-making processes, such as in forensics and law enforcement, it is important for these applications to act on reliable decisions [226, 367]. While human operators can intuitively state how sure they are about a decision and if they can carry out justifiable actions based on this decision [533], current biometric systems do not possess such reliable confidence estimates.

For face recognition (FR), to decide if two faces belong to the same person, a feature representation for each sample is created by a face recognition model. These are known as templates and comparing two templates with a similarity function, such as cosine similarity, results in a comparison score that describes the similarity between the two faces. If the comparison score is above a given threshold, the matching decision is genuine (same identity). Otherwise, the decision is imposter (different identities) [235].

The score uncertainty describes the uncertainty in the score depending on the uncertainty of the data and the model. Contrarily, decision confidence refers to the confidence that the made decision

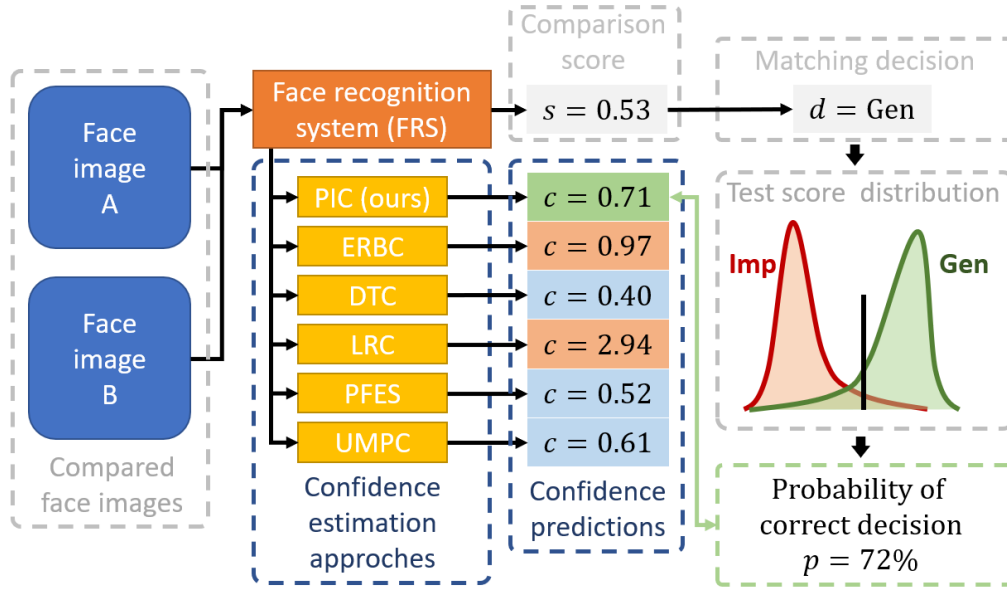


Figure 7.1: **Probabilistic Confidence Interpretation Problem** - Given two biometric samples, the face recognition system comes to the decision that the sample belonging to the same identity (genuine). Evaluating the correctness of this decision based on independent test data results in a probability of 72% that this decision is correct. However, most confidence estimation approaches (yellow) either overestimate (red) or underestimate (blue) the confidence. Contrarily, the proposed PIC approach provides a confidence value (green) reflecting the probability the decision is correct.

is correct [226, 397]. A low-confidence decision is therefore more likely to be wrong than a high-confident one. Consequently, confidence estimation might prevent high-cost mistakes e.g. in front of a court. Previous works proposed several approaches to state the confidence of a model's decision. However, these confidence measures lack interpretability (see Figure 7.1), meaning that those are hardly interpretable and thus, do not reflect the probability that the matching decision is correct.

In this Chapter, we propose a probabilistic interpretable comparison (PIC) score that accurately reflects the probability that the score originates from samples of the same identity. Additionally, the PIC score provides a natural way of combining several comparisons from multiple samples originated from the same network without losing its probabilistic interpretability. Our proposal fits in the Causality Ladder, proposed in Chapter 3, as a form of an intervention to the prediction. The experiments were conducted on five state-of-the-art face recognition systems (FRS) and four publicly-available datasets. Comparing the proposed approach against all available biometric confidence estimation methods, the results demonstrate that PIC results in much more accurate, stable, and interpretable confidence estimates. Moreover, PIC scores from multiple instances lead to a strong boost in recognition performance and probabilistic interpretability.

In contrast to previous works, the proposed PIC score unifies several beneficial properties:

- **Interpretability:** The PIC score accurately reflects the probability that the comparison belongs to a genuine (same person) comparison. This allows critical decision-making pro-

cesses to act upon reliable decisions or, in case of low confidence, allow to ask a more confident system or human operator for the decision.

- **Optimality:** Despite its simplicity, PIC is derived from Bayes' theorem and thus, provides optimal matching confidences given suitable training data. This might be useful, e.g. in law enforcement, when approaches with a theoretical foundation are preferred.
- **Universality:** Since PIC operates on score level, it is not limited to face and can be applied to any biometric modality and recognition model without changing its single-biometric performance.
- **Combinability:** Contrarily to standard comparison scores, a joint PIC score can naturally be computed when multiple samples are given. This leads to stronger multi-biometric recognition performance without losing its interpretability and is highly beneficial when e.g. dealing with multiple video frames.
- **Integratability:** PIC is easily integrateable. It can be easily added on top of an existing biometric system without retraining that system. Moreover, it avoids the need for data- and computationally-expensive experiments to determine the wanted decision threshold due to its interpretable nature.

7.2 Related Work

Confidence estimation in biometric recognition is a relatively new but important field. It aims at estimating the confidence that a decision is correct [226, 397]. While for biometric attribute estimation [478] model calibration methods for classification [196] can be easily adapted due to the output of probability values, this becomes more challenging for zero-shot representation learning tasks, such as biometric recognition. To the best of our knowledge, there are only five recent confidence estimation methods for face recognition. In [541], Zeinstra et al. proposed to use the likelihood ratio between genuine and imposter as a confidence measure for forensic use cases. Huber et al. [226] proposed a more intuitive, but effective solution, by utilizing the distance between the score and the decision threshold as a confidence measure. To get probabilistic confidence estimates, in [227], the corresponding error rate for the decision threshold, lying closest to a given comparison score, was interpreted as the decision confidence. Other approaches [226, 453] require a specialized network that produces probabilistic face embeddings with corresponding feature uncertainties. In [226], Huber et al. propagated these uncertainties through an approximated decision function to obtain matching confidence. In [453], Shi and Jain used these uncertainties to compute the probability that two samples share the same face embedding.

Table 7.1 compares the properties of existing biometric confidence estimation methods. Only the proposed PIC approach is jointly (probabilistic) interpretable, optimal, universal, combinable, and integrable. While all of these confidence estimation methods work well in differentiating

between lower and higher confidence, our experiments will demonstrate that the produced confidence estimates can not be well interpreted as the probability that a decision is correct. To fill this gap, we propose the PIC score approach.

Table 7.1: Properties of biometric confidence estimation approaches.

Method	Interpretable	Optimal	Universal	Combinable	Integrable
DTC [226]	—	—	+	—	+
LRC [541]	—	—	+	+	+
ERBC [227]	+	—	+	+	+
PFES [453]	+	—	—	+	—
UPMC [226]	—	—	—	—	—
PIC (Ours)	+	+	+	+	+

7.3 Methodology

7.3.1 Probabilistic Interpretable Comparison Score

To make the PIC score probabilistic interpretable, we define the score $s_{PIC}(\bar{s}) = P(g(s)|\bar{s})$ as the probability that the set of comparison scores $\bar{s}$ originates from the genuine distribution $g(s)$. To be precise, given are n distributed standard comparison scores $\bar{s} = \{s_1, s_2, \dots, s_n\}$ and we want to compute the probability $P(g(s)|\bar{s})$ that these scores were drawn from the genuine distribution $g(s)$ rather from the imposter distributions $f(s)$. The probability can be modeled by the Bayes rule

$$P(g(s)|\bar{s}) = \frac{P(\bar{s}|g(s)) \cdot P(g(s))}{P(\bar{s}|g(s)) \cdot P(g(s)) + P(\bar{s}|f(s)) \cdot P(f(s))} \quad (7.1)$$

where $P(g(s))$ and $P(f(s))$ are the prior probabilities that the scores belong to the genuine or imposter distribution. Assuming that the scores are drawn independently, the likelihood functions $P(\bar{s}|g(s))$ and $P(\bar{s}|f(s))$ are given by

$$P(\bar{s}|g(s)) = L_g(\bar{s}) = g(s_1) \cdot g(s_2) \dots g(s_n) \quad (7.2)$$

$$P(\bar{s}|f(s)) = L_f(\bar{s}) = f(s_1) \cdot f(s_2) \dots f(s_n). \quad (7.3)$$

For simplicity, we further assume that genuine and imposter comparisons happen equally often $P(g(s)) = P(f(s))$ and obtain the multi-instance biometric solution $P(g(s)|\bar{s})$ that we define as the PIC score

$$s_{PIC}(\bar{s}) = P(g(s)|\bar{s}) = \frac{L_g(\bar{s})}{L_g(\bar{s}) + L_f(\bar{s})}. \quad (7.4)$$

For a single-instance scenario ($\bar{s} = s_1$), this simplifies to

$$s_{PIC}(s_1) = P(g(s)|s_1) = \frac{g(s_1)}{g(s_1) + f(s_1)}. \quad (7.5)$$

If a system decides for genuine, $s_{PIC}(s_1) = P(g(s)|s_1)$ is the probability that the decision is correct. Contrarily, if the decision is imposter, then the probability for a correct decision is given by $1 - s_{PIC}(s_1)$. The simplification assumption of equal prior probabilities aims at keeping the confidence unbiased by having an equalized weighting between both kinds of decision errors. However, it can be optimized for particular applications (e.g. border control) by specifying these prior probabilities to operational conditions.

7.3.2 Training PIC

So far, the derivation assumes that we know $g(s)$ and $f(s)$ in advance. The training processes of PIC involve learning these probability density distributions, e.g. with kernel density estimation (KDE). Given the training data $\mathcal{D} \in \{s_i, y_i\}_{i=1, \dots, N}$ consisting of pairs with comparison scores $\{s_i\}$ and if these belong to genuine or imposter comparisons, the training data scores are split in genuine and imposter scores $\mathcal{D}_g$ and $\mathcal{D}_f$. Then, for each of the score sets a probability density distribution is learned via KDE. For genuine, this is given by

$$g(s) = \frac{1}{|\mathcal{D}_g| \cdot h} \sum_{s_i \in \mathcal{D}_g} K\left(\frac{s - s_i}{h}\right). \quad (7.6)$$

As the kernel $K(x)$, we used a Gaussian Kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (7.7)$$

and selected the bandwidth h with the Scott's rule for one dimension $h = N^{-\frac{1}{5}}$. These resulted in training the probability density distributions $f(s)$ and $g(s)$ needed for the PIC score calculation.

7.3.3 Discussion

A comparison score describes the similarity between two samples. A higher score refers to a higher chance of belonging to the same identity and vice versa. From this perspective, it makes sense to interpret the probability of samples originating from the genuine distribution (or not) as a comparison score. For a single comparison, as long as $\frac{f(s)}{g(s)}$ is monotonic, the order of PICS scores, and thus its single-comparison recognition performance is identical to the ones of the standard comparison score. However, since the derivation of $s_{PIC}(\vec{s}) = P(g(s)|\vec{s})$ was already done for the case of multiple comparisons, its probabilistic interpretation, and the fusion process is optimal. Since inference with Gaussian KDE can be slow with large training data, and thus the PIC score calculation, we recommend creating look-up tables for $g(s)$ and $f(s)$ to speed up the score calculation. Lastly, the PIC scores avoid the need for data- and computationally-expensive experiments to determine the wanted decision threshold. Since the scores already reflect the probabilities for errors, the threshold t for a false match rate (FMR) can be chosen by $t = 1 - \text{FMR}$.

7.4 Experimental Setup

7.4.1 Databases

The experiments were conducted on four publicly-available face recognition datasets with various properties. The Adience dataset [144] consists of 26k images from over 2k different subjects. The images of the Adience dataset possess a wide range of challenges such as low image quality and very young faces. LFW is a dataset [218] containing 13k face images of over 5k identities that were captioned from news images. The ColorFeret database [399] consists of 14k face images from over 1k different individuals with a variety of face poses (from frontal to profile) and facial expressions under well-controlled conditions. Lastly, the Morph [255] dataset consists of 55k frontal face images from over 13k subjects in high resolution. For training, we applied a subject-exclusive test (50%) train split (50%). The 50/50% train split is respective to the number of genuine and impostor samples. As such, the identities of both sets are selected in a way that the sum of the combination of samples per identity is similar in both. Since we do not want to add prior knowledge about the evaluation process in the training step, we use the simplified version of the PIC score with $P(g(s)) = P(f(s))$ as described in Equation 7.4.

7.4.2 Evaluation Metrics

For evaluating the recognition performance, we follow the international standard for biometric verification evaluation [230] by reporting the face verification error in terms of false non-match rate (FNMR) at fixed false match rate (FMR). In the experiments, we focus on reporting the FNMR at 10^{-3} FMR as recommended by the best practice guidelines for automated border control of the European Border Guard Agency Frontex [165].

To evaluate the probabilistic interpretability of the confidence scores, we adapt the widely-used Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) [196, 354]. In contrast to error-vs-reject curves from quality assessment [480] which evaluates the order of confidence prediction and neglects their probabilistic interpretation, this work focuses on the widely-used confidence estimation metrics ECE and MCE. Dividing n samples into M equally-spaced bins B_m based on their confidence, the ECE

$$ECE = \sum_{m=1}^M \frac{|B_m|}{n} |p_{true}(B_m) - p_{pred}(B_m)| \quad (7.8)$$

describes the average error between the model's predicted confidence p_{pred} and its true confidence p_{true} given the test data. Since reliable confidence estimates are absolutely necessary in high-security applications, the MCE

$$MCE = \max_{m \in \{1, \dots, M\}} |p_{true}(B_m) - p_{pred}(B_m)| \quad (7.9)$$

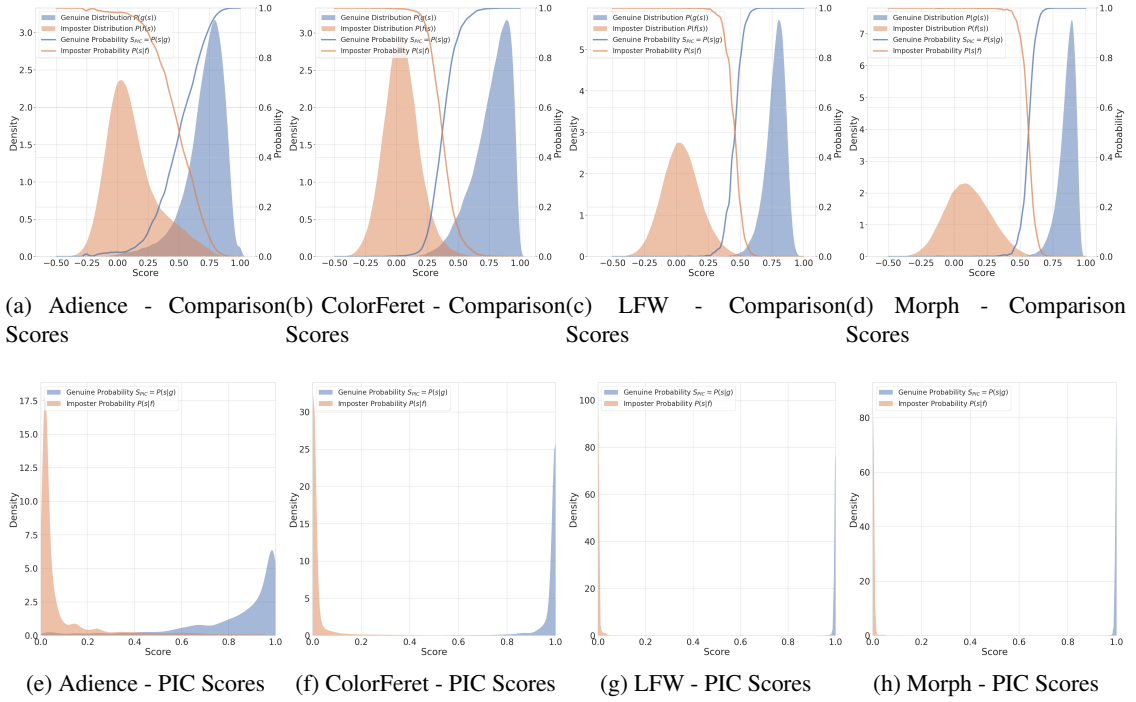


Figure 7.2: Score Distribution Analysis -The original (top) and the PIC (bottom) score distributions are shown for FaceNet. Based on the original score distributions, also the corresponding probabilities (see Sec. 7.3.1) for genuine and imposter are shown at the top. On the bottom, the optimal probabilistic interpretable comparison (PIC) scores are shown. Since the PIC score distributions build on a monotonic transformation, the order of the scores, and thus the performance, remains the same.

measures the worst-case deviation between the predicted and true confidence. For a perfect confidence estimator and suitable test data, the MCE and ECE are both zero.

7.4.3 Face Recognition Models

To ensure high compatibility of confidence estimation methods with a wide variety of recognition systems, the experiments were conducted on five state-of-the-art face recognition models pre-trained and released by the corresponding authors¹. This includes the models FaceNet [439], PFE [453], ArcFace [126], MagFace [335], and QMagFace [479].

7.4.4 Confidence Estimation Approaches

In the experiments, we compare the proposed probabilistic-interpretable comparison (PIC) score against all the other decision confidence estimation methods for face recognition that we are aware of. This includes score-based decision confidence methods such as the distance to (decision) threshold confidence (DTC) metric [226] (BMVC22), the likelihood ratio-based confidence

¹For FaceNet, the authors never made the model publicly-available. Instead, a third-party implementation was used: <https://github.com/davidsandberg/facenet>

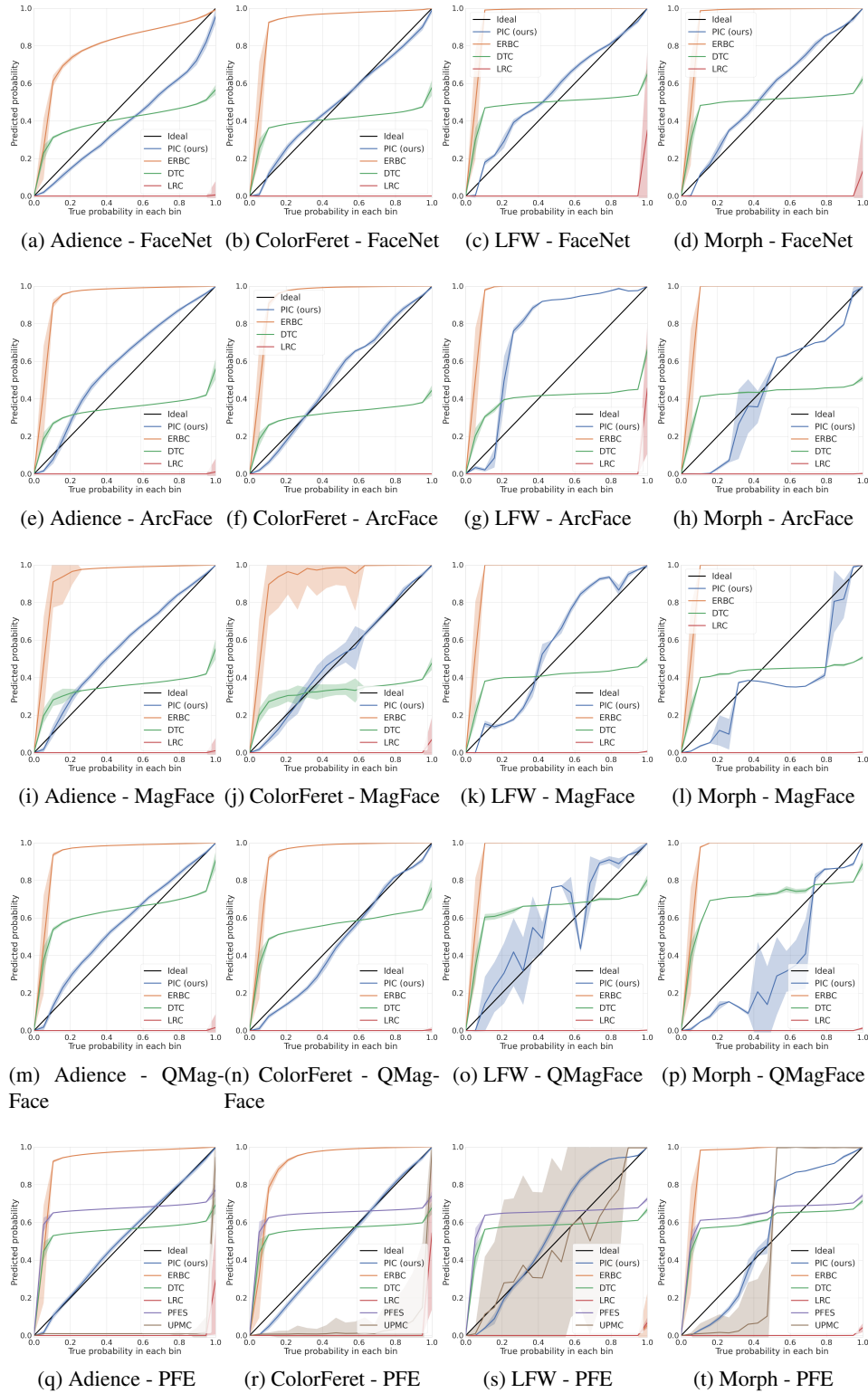


Figure 7.3: **Confidence Calibration Curves (CCC)** - The CCC for all dataset and FRS combinations are shown. Inconsistencies are due to the low number of samples for specific probability bins (e.g. for LFW and Morph). While most approaches have to deal with high under- and over-confident predictions, the proposed PIC produces close-to-ideal (black line) probabilistic confidence estimates in most cases.

(LRC) score [541] (BTAS18), and the error rate based confidence (ERBC) [227] (ICPR22). Other approaches need probabilistic face embeddings to make use of feature uncertainties for predicting decision confidence. This includes uncertainty-propagation for matching confidence (UPMC) [226] (BMVC22) and the probabilistic face embedding score (PFES) [453] (ICCV19).

7.5 Results

7.5.1 Score Distribution Analysis

To understand how the original comparison score is transformed to gain a probabilistic interpretation, Figure 7.2 shows the original genuine and imposter score distributions for FaceNet², as well as the corresponding PIC score distributions. In the top row, the original score distributions for the different datasets are shown. For the Adience and ColorFeret datasets, the genuine and imposter score distributions strongly overlap due to the challenges of these datasets, such as low image quality and strong pose differences. LFW and Morph show mostly frontal and well-illuminated faces and thus, the score distributions show significantly less overlap. Consequently, on these "easy" datasets, LFW and Morph, fewer samples for the probabilities in the overlap areas exist. This will lead to some unstable results in the confidence calibration analysis, as we will see in Section 7.5.2. Additionally to the score distributions, the optimal probability that a sample with a score of s belongs to genuine or imposter (see Sec. 7.3.1) is shown. Intuitively, both probabilities equalize when the number of genuine and imposter scores for s are the same.

In the bottom row, the score distributions for the proposed PIC scores are shown. The score refers to the probability that a sample belongs to a genuine comparison. Since the order of the scores remains the same for the original and the PIC scores, this will lead to similar single-biometric recognition performances. For a single sample, the PIC score adds the optimal probabilistic interpretability for a generic biometric system. However, contrary to the standard comparison score, a PIC score for multiple comparisons can be calculated as demonstrated in Section 7.5.3.

7.5.2 Single-Comparison Calibration Analysis

To analyze the probabilistic interpretability of different confidence estimation approaches, we introduce confidence calibration curves (CCC). A CCC compares the true confidence of each sample (x-axis) in a given test set bin-wise with the average predicted confidence (y-axis). For a perfectly calibrated confidence estimator, the CCC shows a linear bisectrix line. To compute a CCC, the true confidence of each test set sample is calculated and divided into $b = 30$ bins. For each bin, the mean and standard deviation of the predicted confidences are computed. Please note that in the LFW or Morph datasets, the number of samples for single bins might be low. Since these datasets are less challenging (see Section 7.5.1) and thus, provide fewer samples for specific probabilities, the performance becomes more unstable in these cases.

²Since the distributions for the other FRS lead to similar observations, we refer to the supplementary.

Table 7.2: **Confidence Calibration Analysis** - The ECE and MCE are shown for several dataset and FRS combinations at an FMR of 10^{-3} . ECE shows the average confidence calibration error, while the MCE presents the maximum calibration error. The best performance is marked in bold. Except for one case, the proposed PIC approach strongly outperforms the other confidence estimators in terms of interpretability. This holds for the average performance, as well as for the worst-case scenario.

Database	FRS	ECE [%] at 10^{-3} FMR						MCE [%] at 10^{-3} FMR					
		PIC (ours)	ERBC	DTC	LRC	UPMC	PFES	PIC (ours)	ERBC	DTC	LRC	UPMC	PFES
Adience	FaceNet	2.90	26.19	23.45	24.68	-	-	15.55	56.40	41.56	97.62	-	-
	PFE	1.14	36.65	40.00	15.04	4.36	50.19	6.36	85.07	48.87	93.06	90.83	57.73
	ArcFace	1.48	38.53	20.89	19.31	-	-	18.00	83.53	50.46	98.70	-	-
	MagFace	1.38	40.13	22.43	18.73	-	-	13.00	83.90	50.56	98.69	-	-
	QMagFace	1.50	39.25	31.79	18.73	-	-	13.29	86.79	46.96	98.15	-	-
ColorFeret	FaceNet	0.95	40.83	26.68	14.12	-	-	8.33	84.97	44.88	99.56	-	-
	PFE	0.96	36.81	41.64	7.17	3.69	50.74	4.07	75.30	45.78	92.33	80.91	54.83
	ArcFace	0.96	46.44	16.99	4.68	-	-	10.91	83.48	54.82	99.40	-	-
	MagFace	1.06	46.44	18.91	4.79	-	-	8.01	82.91	53.09	92.40	-	-
	QMagFace	1.05	45.39	34.00	4.02	-	-	6.32	84.71	41.29	98.83	-	-
LFW	FaceNet	0.37	40.76	30.43	11.01	-	-	15.76	91.91	39.79	92.52	-	-
	PFE	0.07	0.59	41.42	0.59	0.91	51.36	22.24	94.50	48.87	93.07	49.39	56.37
	ArcFace	0.89	40.58	19.15	12.04	-	-	54.99	91.22	47.24	92.26	-	-
	MagFace	0.09	52.61	20.63	0.15	-	-	23.29	92.87	49.76	98.89	-	-
	QMagFace	0.07	53.25	33.60	0.16	-	-	29.89	92.35	52.85	99.45	-	-
Morph	FaceNet	0.23	41.76	30.25	7.84	-	-	11.83	90.90	40.44	92.56	-	-
	PFE	0.18	50.88	44.02	0.94	1.34	49.50	32.34	90.68	49.18	95.61	49.88	53.48
	ArcFace	0.06	48.41	21.61	0.10	-	-	18.06	91.19	48.24	99.14	-	-
	MagFace	0.23	47.76	20.99	1.00	-	-	34.93	92.36	48.71	99.19	-	-
	QMagFace	0.83	47.92	38.91	1.10	-	-	36.13	91.40	55.83	98.59	-	-

Figure 7.3 shows the CCC plots for all datasets and face recognition system combinations at an FMR of 10^{-3} . The ideal case with an optimal confidence estimation is shown as a black line. The ERBC approach shows strongly overconfident behavior in all cases since it is based on the error-rates of the whole system rather on single comparisons. The DTC approach simply uses the distance between the score and the threshold as a confidence estimator. Consequently, it overestimates less confident decisions and underestimates highly confident ones. The LRC solution is based on the likelihood ratio between genuine and imposter. Since this approach does not have a probabilistic interpretation, it strongly underestimates confidence. The PFES and UMPC both require uncertainties per feature to make a confidence estimation. Consequently, they could be only applied to PFE since this is the only utilized FRS able to state the uncertainty per feature. UPMC strongly underestimates confidence for Adience and ColorFeret. For LFW, the confidence estimates work well on average but are quite unstable. This might be explainable through the training data. Since the confidences are based on the uncertainties of the FRS. The training data of the FRS might contain many well-illuminated and frontal faces, similar to LFW and unsimilar to Adience and ColorFeret. For PFES, a similar behavior to DTC is observed, a mixture of strong over- and underestimating confidence. Contrarily, the proposed PIC approach produces stable and accurate confidence estimations that are often close to the optimal solution (black line).

To analyze the probabilistic interpretability quantitatively, Table 7.2 shows the ECE and MCE of the confidence estimation methods for all database and FRS combinations at a FMR of 10^{-3} . The ECE shows how much a confidence estimator is off on average and thus, states how reliable a method can state confidence. To also cover the worst-case scenarios, the MCE shows how much a confidence estimator is off in the worst-case. For nearly all cases, the proposed PIC approach leads to significantly smaller ECE and MCE values demonstrating its effectiveness for estimating probabilistic interpretable confidence values.

7.5.3 Multi-Comparison Analysis

7.5.3.1 Recognition Performance

The proposed PIC approach is able to naturally combine multiple samples into a single comparison score. This is known as a multi-biometric fusion and aims to fuse information from multiple sources to improve recognition performance whilst addressing some of the limitations of single-biometric systems, such as poor data quality [480] or overlap between identities [460]. In this section, we will show that PIC significantly increases recognition performance with multiple samples. In the next section, we will then demonstrate that the joint PIC scores also increase the probabilistic interpretability.

Table 7.3 shows the recognition performance of the proposed PIC approach when combining multiple samples. The recognition error FNMR is shown for a fixed FMR of 10^{-3} for all database and FRS combinations. In this multi-biometric context, a given probe sample is compared to 1/2/5 reference samples to calculate a joint PIC score. The results demonstrate that the PIC score

can be efficiently used for multi-biometric recognition scenarios. However, since the main contribution of this Chapter lies on probabilistic confidence estimation, it is not compared against (non-interpretable) score fusion approaches.

Table 7.3: **Multi-Biometric PIC score recognition performance** - The recognition performance of using a joint PIC score by combining a probe sample with 1/2/5 reference samples (RS) is shown for all database and FRS combinations. Generally, the joint PIC score leads to lower recognition errors at 10^{-3} FMR than in the single-biometric scenario.

Database	FRS	FNMR [%] at 10^{-3} FMR		
		1 RS	2 RS	5 RS
Adience	FaceNet	71.00	65.81	58.77
	PFE	18.47	30.22	19.43
	ArcFace	10.10	6.43	2.57
	MagFace	9.75	5.00	2.59
	QMagFace	9.78	5.09	2.28
ColorFeret	FaceNet	12.22	6.09	4.47
	PFE	6.60	8.66	9.08
	ArcFace	4.22	2.18	2.39
	MagFace	3.92	1.86	1.77
	QMagFace	3.24	1.47	1.42
LFW ³	FaceNet	0.79	1.02	2.05
	PFE	0.08	0.25	0.24
	ArcFace	4.38	2.31	1.88
	MagFace	0.05	0.16	0.28
	QMagFace	0.05	0.18	0.27
Morph	FaceNet	0.54	0.15	0.07
	PFE	0.89	0.70	0.52
	ArcFace	0.05	0.05	0.03
	MagFace	0.96	0.48	0.45
	QMagFace	0.96	0.50	0.42

In general, the recognition error for multiple reference samples is significantly lower than in the single-biometric case with one sample. The only exceptions are mostly on the LFW dataset which is not well-suited for multi-biometric recognition analysis. For instance, 80% of the identities in LFW have only one image. Consequently, doing multi-sample comparisons results in an insignificant low number of genuine scores for the evaluation which makes the interpretation of the results less meaningful. Besides this, the joint PIC score significantly increases the multi-biometric recognition performance in all other cases demonstrating its effectiveness for multi-biometrics besides confidence estimation.

7.5.3.2 Calibration Performance

In the following, we will show that the probabilistic confidence interpretation of the proposed PIC approach still remains for multi-biometric scenarios. Table 7.4 analyses the probabilistic interpretability of the proposed PIC approach for the multi-biometric scenario.

³The identity distribution of LFW does not allow creating many multi-sample comparisons. Thus, is not well-suited for analyzing multi-biometric recognition. However, we reported these results for the sake of completeness.

For all database and FRS combinations, it shows the expected calibration errors (ECE) at a decision threshold for a 10^{-3} FMR. Similar to before, a given probe sample is compared to 1/2/5 reference samples to calculate a joint PIC score confidence. In general, confidence estimates of the proposed PIC approach improve when combining multiple samples. In nearly all cases, the ECE decreases for more reference samples. This demonstrates that, similar to the recognition performance, also the probabilistic confidence estimation of the joint PIC approach increases significantly.

Table 7.4: **Multi-Biometric Confidence Calibration Analysis** - The expected calibration errors (ECE) are shown for a joint PIC score confidence by combining a probe sample with 1/2/5 reference samples (RS). This was done for all database and FRS combinations at an FMR of 10^{-3} . The ECE significantly decreases when more reference samples are used. Consequently, also the probabilistic confidence interpretation of PIC becomes more accurate when multiple samples are combined.

Database	FRS	ECE [%]		
		1 RS	2 RS	5 RS
Adience	FaceNet	2.90	1.80	0.44
	PFE	1.14	0.51	0.00
	ArcFace	1.48	0.26	0.11
	MagFace	1.38	0.27	0.15
	QMagFace	1.50	0.31	0.10
ColorFeret	FaceNet	0.95	0.40	0.36
	PFE	0.96	0.63	3.03
	ArcFace	0.96	0.24	0.06
	MagFace	1.06	0.19	0.13
	QMagFace	1.05	0.17	0.17
LFW	FaceNet	0.37	0.09	0.11
	PFE	0.07	0.27	0.00
	ArcFace	0.89	0.07	0.01
	MagFace	0.09	0.03	0.04
	QMagFace	0.07	0.03	0.14
Morph	FaceNet	0.23	0.07	0.02
	PFE	0.18	0.16	0.00
	ArcFace	0.06	0.01	0.00
	MagFace	0.23	0.01	0.00
	QMagFace	0.83	0.03	0.00

7.6 Conclusion

Since mistakes are coming at high costs in critical decision-making processes, such as forensics or law enforcement, it is necessary to accurately state the matching confidence of a biometric system. Previous works on confidence estimation in biometrics can well differentiate between low and high-confident decisions but produce confidence estimates that are not interpretable as the probability that the decision made is correct. To fix this issue, we proposed the PIC score, a probabilistic interpretable comparison score. The conducted experiments demonstrate that the proposed approach outperforms all related approaches in terms of probabilistic interpretability

and can also be applied in multi-biometric recognition scenarios. The proposed PIC score jointly achieves interpretability, optimality, universality, combinability, and integratability. Moreover, the score accurately states the probability that the compared samples belong to the same identity (interpretability). Since it was derived from the Bayes' theorem, it provides optimal matching confidences given suitable training data (optimality). It can be applied to any biometric modality and system without changing its single-biometric performance (universality). In contrast to the standard comparison score, multiple samples can be efficiently combined into a joint PIC score (combinability) leading to a significant gain in recognition performance and interpretability. Lastly, the PIC solution can be easily integrated into existing biometric systems and avoids the need for data- and computationally-expensive experiments to determine the desired decision threshold.

Part III

Face Recognition with Synthetic Data

Chapter 8

On the Generalization of Masked Face Recognition Models to Occluded Face Recognition

Author Contributions The research work described in this chapter was conducted in collaboration with João R. Pinto, Fadi Boutros and Naser Damer, under supervision of Ana F. Sequeira and Jaime S. Cardoso. The author of this thesis contributed to this work on the problem conceptualisation, data curation, the preparation and conduction of experiments, the results discussion, and publication writing. Some parts of the chapter were originally published in, or adapted from:

- Neto, P. C., Pinto, J. R., Boutros, F., Damer, N., Sequeira, A. F., & Cardoso, J. S. (2022). Beyond masks: On the generalization of masked face recognition models to occluded face recognition. *IEEE Access*, 10, 86222-86233.
- Sequeira, A.F., Neto, P. C., Cardoso, J. S. (2022). On the generalization of masked face recognition models to occluded face recognition. *In 28th Portuguese Conference on Pattern Recognition (RECPAD 2022), Portugal*.

8.1 Context and Motivation

Face recognition methods have evolved over time and are now performing at a human level [472]. Despite the fact that the evaluation of these methods relies on datasets that do not address challenges such as occlusions since these datasets are mostly composed of clear images [218]. In real world applications, there are no guarantees on capturing pictures that show the face free of obstructions. Hence, the training and evaluation of these models are somewhat lacking as they do not include some of the challenges. The recent SarS-Cov2 pandemic led to the mandatory use of facial masks, which became a frequent obstruction to face images. For this particular occlusion, Damer *et al.* [103] have shown that the performance degradation affects both humans and face recognition systems.

Parallel to the search for better face recognition models, researchers have also been interested on occluded face recognition (OCFR) architectures. Frequently, OCFR models are less discriminative on clear images. Thus, these models also present a trade-off between performance under occlusions and without occlusions. One solution to mitigate such problem is to start by detecting the presence of an occlusion, which, on the other hand, increases the computational requirements. The variety of potential occlusions is also growing as part of technological advances. Examples of this evolution are the popularization of smartphones and the usage of headphones. However, the vast majority of the occlusions are either for aesthetics, or random elements, such as high brightness or hands in front of the face [544].

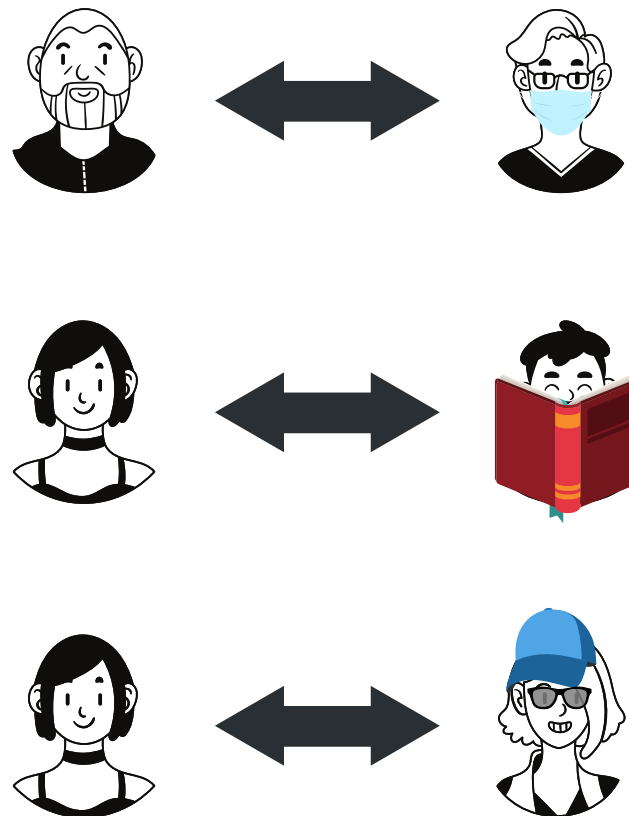


Figure 8.1: Illustration of some occlusions that can highly increase the difficulty of the face verification task.

Due to the pandemic, researchers have directed their focus towards models capable of verifying the identity of someone that is wearing a mask (also known as masked face recognition). Mask-based occlusions obscure the majority of facial features below the periocular area. These efforts led to the publication of two competitions [48, 124] and several distinct models based on diverse approaches [47, 221, 300, 347, 360]. Compared to standard face recognition models, the majority of the published models for masked face recognition (MFR) advanced the state of the art. However, as seen in Figure 8.1, masks are not representative of all the different occlusions and other types of occlusion might cover additional parts of the face. Hence, it remains to be seen if MFR models can also bring benefits to a more general occluded face recognition problem. The importance of

models that are robust to both problems is proportional to the growth of face recognition adoption across different technologies and scenarios. For instance, in border control it is necessary to prevent the increase in false positives due to specific occlusions, which could lead to security issues. On the other hand, false negatives should remain low in smartphone's face recognition authentication so that the users' experience does not get negatively impacted.

In recent literature, researchers aimed at improving the current robustness of general face recognition for masked face recognition. Thus, knowing that masked faces are a uniform type of face occlusion, the main question posed in this Chapter is “can the solutions developed for masked face recognition enhance the performance of face recognition for more general occlusion?”. Towards that goal, this Chapter presents several contributions to both masked and occluded face recognition fields. It mainly focuses on understanding the capabilities of models trained for a specific occlusion (face masks) to generalize to others. Validating the performance of a masked face recognition approach in occluded datasets poses two main contributions. First, since these algorithms are fairly lighter than occluded face recognition methods, there is a clear advantage regarding computational costs. Moreover, since masked face recognition approaches leverage knowledge existent in traditional face recognition methods, the growth of traditional face recognition datasets might fuel a performance increase in both masked and occluded face recognition. The main outcomes of the Chapter show that there is an interoperable deployability of MFR methods on OFR datasets. In fact, MFR methods can be effectively deployed for general OFR and achieve reasonable performance.

8.2 Related Work

8.2.1 Face recognition

Advances in face recognition models have been driven by a combination of three elements. One of the elements is the steep increase in computational power. Graphical processing units (GPUs) potentiated the training of larger and more complex models in a reasonable time [466]. This led to deeper models capable of approximating the real function with extraordinary precision. The other is the introduction of novel architectures, such as ResNets [205], that pushed the limits of the State-of-the-Art. Finally, an unprecedented proliferation of the internet led to remarkable growth in the data available and how it can be collected [71], with recent works aiming at replacing such data with privacy-friendly synthetic data [55, 49, 110].

Recently, the design of training losses, aiming at enhancing discriminative ability of the extracted face embeddings, has assumed a preponderant role. Face recognition models highly improved their accuracy and generalization capabilities through the usage of contrastive losses. For instance, triplet loss does not train the model to classify an identity; instead, it optimizes the model to have positive pairs with smaller distances than negative pairs. However, these types of losses have been outperformed by the recent angular margin-based softmax losses [126]. For instance, Center loss [524] uses the learned centres of the features of each identity to decrease intra-class

variance. SphereFace [307] (Angular Softmax) adds angular constraints to each identity and normalize its weights, thus creating a hypersphere space. CosFace [500] and ArcFace [126] propose variations of these previous losses. The former introduces an additive margin for the cosine space, whereas the second presents an additive margin for the angular space. MagFace [335] rethinks the margin of loss to be given as a function of the quality of the samples. Finally, ElasticFace [46] proposes a relaxation of the margin constraint by randomly picking its value from a Gaussian distribution.

8.2.2 Occluded face recognition

As stated previously, the performance of state-of-the-art face recognition models degrades significantly when a face is occluded [544]. Research on how to deal with occlusions led to two main directions: a) predicting the occluded features [549, 515, 555, 530, 208]; and b) removing the occluded features. The former paradigm approached deep learning for the first time in a work proposed by Zhao *et al.* [549]. Their use of a LSTM-Autoencoder to reconstruct the corrupted features led to the loss of identity information, which degraded the results. Ignoring the occlusions is a tough task for deep learning systems due to the difficulty of locating them. To mitigate it, MaskNet, introduced by Wan *et al.* [498], presented a middle branch to assign smaller weights to hidden units affected by the corrupted features. However, the training of the middle branch is too relaxed, leading to the inclusion of irrelevant information. The recent state-of-the-art includes two related approaches. The first was proposed by Song *et al.* [463] and it uses a binary mask to clean the corrupted features. It learns a mapping of each block of the face image (divided into $K \times K$ blocks) to the feature mask. The detection of occluded blocks and the processing of the image is conducted in two separate steps. FROM [411], proposed by Qiu *et al.*, removes the need for two different inference stages for training and testing. Their Mask Decoder predicts, in a dynamic fashion, the correct feature masks. Hence, this method is much more efficient at training and testing. Wang and Guo [505] proposed an approach to deal with occlusions and pose variance. Their model, DSA-Face, is based on sparse attention mechanisms, and their results indicate a performance that is slightly above the one delivered by traditional face recognition models. However, it is worth noting that it presents a higher complexity than most known models for face recognition. With a more direct application, Biswas *et al.* [39] proposed a method to tackle Child Sexual Exploitation Material. They state that abusers have their eyes covered in the majority of the content and that traditional face recognition systems fail to recognize their identity. Hence, they propose a model based on Perceptual hashing to help solve this task and mitigate the humanitarian problem of Child Sexual Exploitation.

8.2.3 Masked face recognition

The effects of the COVID-19 pandemic were felt worldwide. Which led to the obligation of using a face mask in the majority of the developed countries [183]. These masks occlude a vast part of the face. Hence, due to the lack of these features and the change in the nature of the representation,

face recognition systems cannot maintain performance [107, 103, 244]. This drove researchers to focus on developing face recognition solutions that aim at maintaining face recognition performance despite the presence of masks. Such solutions were even part of performance competitions organized to consolidate the reporting of advances in this field [48, 124].

Anwar and Raychowdhury [23] showed the benefits of adding synthetic mask augmentations to the training data to increase the algorithm performance on real masks. Boutros *et al.* [47, 124] proposed an efficient solution that can be integrated on top of any face recognition model. In their solution, when the person wears a mask, a neural network attempts to unmask the embedding produced by the face recognition system. This neural network is trained with a self-restrained triplet loss that puts more focus on the highly affected genuine pairs, when compared to impostor training pairs. The solution introduced in Chapter 6 encompasses a model trained with a modified triplet loss and ArcFace loss. This model feeds two versions of the same image to the face recognition model. The first contains a mask, whereas the second does not, and the network is constrained to minimize the distance between the two embeddings produced. The multitask optimization for recognition and mask detection was also present in other works [347], while others include multitask learning for face analysis [391].

Geng *et al.* [178] present a generative model to augment the current datasets with masked images. The proposed generative adversarial network (GAN) is specially crafted to retain information regarding the identity of the subject in the image. Mandal *et al.* [324] attempted to approach the problem with a domain adaptation solution. Huber *et al.* [221] intended to approximate embeddings produced by masked and unmasked images. Towards that, he used template level knowledge distillation to a student model, from a teacher model trained on ElasticFace loss. Moreover, to further improve the results, it includes an auxiliary term that relies on the ElasticFace loss to retain identity information within the embedding. Similarly, Li *et al.* [295] proposes a knowledge distillation approach accompanied with image-level face completion. The operation of other sub-components of face recognition systems has also been affected by mask wearing. Particular attention was paid to presentation attack detection [154, 158] and face image quality assessment [169].

With the rise of masked face recognition as an urgent challenge in face recognition, its solutions have evolved as a separate research direction from occluded face recognition. This work tries to link these two conceptually similar research directions by probing the scope of their interoperable deployability.

8.3 Methodology

The methodology can be divided into three main steps. As previously stated, one of the contributions of this Chapter is the assessment of the performance of MFR algorithms in OFR problems. Hence, in the first step, we evaluated the performance of these algorithms in both masked and occluded face recognition datasets. The algorithms used were carefully selected from recent literature. Afterwards, we repeated this process with a state-of-the-art occluded face recognition

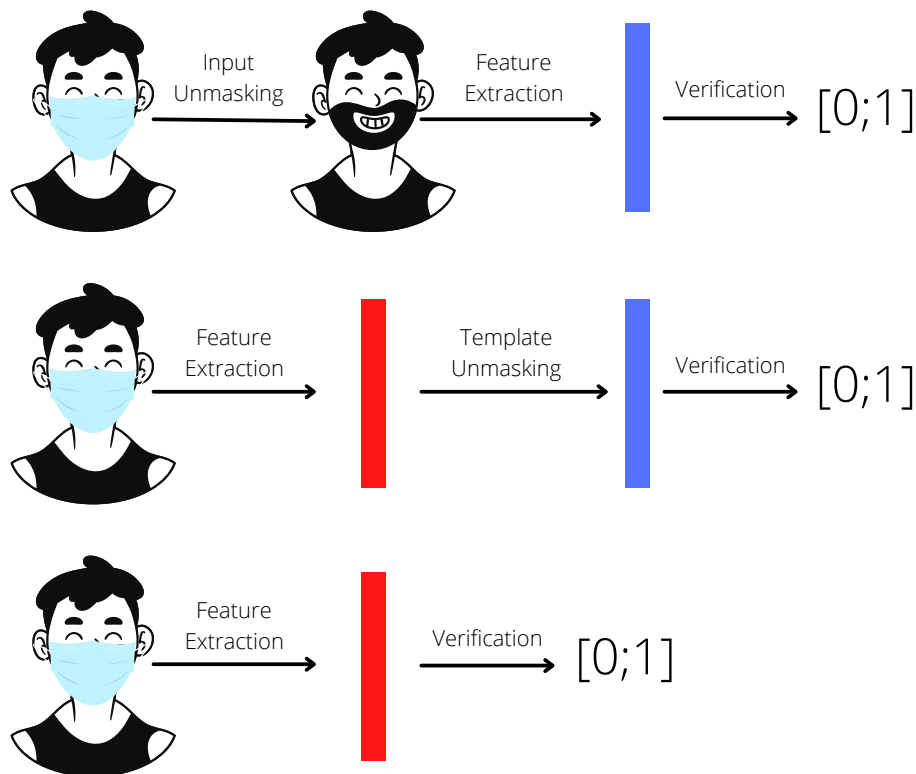


Figure 8.2: Several approaches can be followed in order to handle masks. These are the three most common approaches: Unmasking the input image, unmasking the template generated by a face recognition model, or using a model that is robust to masks. Blue templates do not contain mask information, whereas red templates contain mask information. Models robust to mask usually have some mask information leaked into their embedding, but in a low proportion when compared to the non-mask information. (Best viewed in color.)

algorithm. To further understand the need for both OFR and MFR algorithms, we repeated the evaluation process with algorithms designed for general face recognition.

8.3.1 Face verification by Masked Face Recognition models

From 2020 to 2023, the number of different models for masked face recognition has grown exponentially. Therefore, we decided to select a few of these models, where the selected ones must include sufficiently distinct approaches so that the evaluation is diverse. Thus, in the experiments conducted for this research, three masked face recognition models are used. In the context of masked face recognition, models can be evaluated for their performance on masked reference to masked probe (M-M) or unmasked reference to masked probe (U-M) configurations. The former is when both the probe and the reference have the masks, whereas the other has only the probe masked. Approaches that can be followed for masked face recognition are illustrated in Figure 8.2.

FocusFace, described on Chapter 6, presents a multitask neural network that is contrastively optimized to create similar embeddings for masked and unmasked images. FocusFace achieved

competitive results in the masked face recognition competition [48] database on M-M face verification and beat the baseline on U-M. The second model (KD) [221] approached this problem through knowledge distillation at the template level, from a teacher network (ElasticFace-Arc) [46], which receives unmasked input, to a student network that receives a mixture of masked and unmasked faces of the same identities. This enforces the student network to learn how to build embedding from masked images similar to the embedding constructed by the teacher network from unmasked images, while maintaining the discrimination learned by the teacher model on unmasked training data. It is complemented by auxiliary supervision through regularization with ElasticFace loss. Finally, the last model (EUM) [47] presents a model that operates on top of pre-trained face recognition systems. This model is trained with the self-restrained triplet loss proposed also in [47]. It aims to receive the embedding produced by a face recognition model from a masked input, and map it into an embedding that behaves similarly to the one produced by the same face recognition model from an unmasked input. In other words, it attempts to unmask face embeddings. The self-restrained triplet loss ensures that it automatically adapts the loss, putting more weight on the intra-class compactness when required, as this is more affected by the masks, as demonstrated in [104]. We trained all the presented models on the MS1MV2 dataset [198, 126] with RGB images aligned (through similarity transfer) and cropped to 112×112 , accordingly to the instructions on the original papers. We decided to train and evaluate these models with a ResNet-100 [205] backbone architecture. Hence, all three models use the same backbone.

8.3.2 Face verification by Occluded Face Recognition models

Differently from masked face recognition, occluded face recognition requires that its models generalize across occlusions situated at diverse locations on an image. For instance, a scarf presents a challenge similar to a mask. On the other hand, sunglasses create different problem. These nuances led to models and architectures with distinct properties compared to those from methods described in the previous section. Research in this field has been slower than on masked face recognition over the past two years. Hence, the evaluation of occluded face recognition models is focused on a single model. It significantly outperforms others designed for the same task. FROM [411] extends the approach followed by the previous state-of-the-art model PDSN [463]. They present a dynamic approach to correct feature masks called Mask Decoder. This latter mechanism removes the need for two different train steps and greatly reduces inference time. In contrast to previous models, FROM receives aligned images cropped to 112×96 . The model's backbone network is initially trained with the CASIA-WebFace dataset for general face recognition. Afterwards, the entire model is fine-tuned end-to-end with the occluded version of the CASIA-WebFace dataset.

8.3.3 State-of-the-art face recognition

For general face recognition algorithms, three different models were chosen from top performing models. They differ from each other in the loss function used for network optimization. The other

parameters, such as the training set and the backbone network, remain the same. All use ResNet-100 [205] trained with the MS1MV2 dataset [198, 126] as their backbone. ArcFace [126] presents an angular margin-based softmax loss to increase the discriminative power of the trained model; Boutros *et al.* [46] proposed ElasticFace [46], a loss similar to ArcFace that draws its margin at each iteration from a Gaussian distribution. The ElasticFace can further draw from the Gaussian distribution while considering the samples proximity to their class centers during training, namely the ElasticFace+. Similarly, Meng *et al.* [335] suggested improvements to the loss margin. Their method, MagFace, presents a novel approach to make the sample quality correlate to the margin parameter. All of these models were trained with RGB 112×112 cropped and aligned images.

None of these methods was specifically designed to handle occlusions. Hence, their performance can be expected to decrease with the presence of the different occlusions. Their use serves as a baseline for the performance of occluded and masked face recognition methods. Masked face recognition algorithms are expected to have, at least, a similar performance to these models when confronted with occlusions.

8.4 Databases



Figure 8.3: Examples of the occluded face verification datasets. The first two rows correspond to Occ-LFW with synthetic occlusions, whereas the third corresponds to the AR database with real occlusions. The first row contains images from the Occ-LFW-1.0, and the second contains images from the Occ-LFW-2.0.

8.4.1 Masked face recognition benchmarks

For masked face recognition, we utilize the masked face benchmarks (proposed in [221]) based on five of the most widely used face verification benchmarks. Hence, performance comparisons



Figure 8.4: Examples of the validation datasets (LFW, CALFW, CPLFW, CFP-FP and AgeDB-30) augmented with face masks for masked face verification. The approach followed for mask augmentation is described by Huber *et al.* [221].

between distinct models are direct and straightforward. The used datasets are: Labeled Faces in the Wild (LFW) [218]; Cross-Age LFW (CALFW) [552]; Cross-pose LFW (CPLFW) [551]; AGEDB-30 [351]; Celebrities in Frontal-Profile (CFP) [443].

Labeled Faces in the Wild (LFW) introduced in 2007, this dataset has been widely used to compare the performance of face recognition models on face images in unconstrained environments. Moreover, it already includes a protocol, which we strictly follow, for the evaluation of face verification performance on 6,000 image pairs [218].

Cross-Age LFW (CALFW) and Cross-pose LFW (CPLFW) are a more challenging version of the original LFW dataset. They present themselves as a renovation of LFW, which is already suggested by their “Cross-Age” and “Cross-Pose” prefix. CALFW introduces images with an age gap, whereas CPLFW presents variations to the pose of the same individual; thus, these two versions increase the intra-class variation to better simulate a real-world face verification scenario.

AgeDB-30 is a subset of the larger AgeDB [351]. While the latter has no restrictions on the age difference between images of the same individual, AgeDB-30 imposes a 30-year interval. It is widely used for evaluating the age-invariant face verification performance and contains 16,488 different images of 568 identities. The evaluation protocol followed is also well defined in the literature.

Celebrities in Frontal-Profile (CFP) was created to improve the evaluation capabilities on frontal to profile face verification “in the wild”. In a sense, the proposed evaluation can be considered similar to occluded face recognition, since, in some extreme poses, many features are completely occluded. It contains 500 subjects, each with 10 frontal and 4 profile images. There are two main evaluation protocols, namely frontal to profile (FP) and frontal to frontal (FF). In this article, only the first protocol is used.

As seen in Figure 8.4, these datasets were not used as originally proposed. They were modified with face masks as proposed by Huber *et al.* [221] for performance validation on masked face verification. The added masks follow the exact protocol described in the original proposal, and its code is publicly available¹.

8.4.2 Occluded face recognition benchmarks

Occ-LFW Database Song *et al.* [463] presented an approach to synthesizing a dataset of occluded faces with occluders that include sunglasses, scarf, face mask, hand, eye mask, eyeglasses, book, phone, and cup. While it can be argued that these are everyday items that can appear in front of a face, it can also be said that the generated occlusions are not really close to being natural. Nonetheless, it was also the approach followed by Qiu *et al.* [411]. Hence, also followed in this methodology. The images are synthesized by adding a random occluder image on top of a random location of a face. In some scenarios, the full occluder might be absent from the image, since its center (location of the image selected to add the occlusion) might have a smaller distance to the margin than the size of the occluder.

The dataset selected to be occluded is LFW, since it was also used by Qiu *et al.* [411]. It originated three different versions of the LFW. These variations differ from each other due to the different scales of the occluder and the percentage of the face occluded. According to the convention, we name the datasets as follows: Occ-LFW-1.0, Occ-LFW-2.0, and Occ-LFW-3.0. The first has, on average, >18% of the image area occluded. The second and the third have >40% and >50% of occluded area, respectively. The same random seed was used, and the code for generating the occlusions is publicly available². Some of these images can be seen in the first two rows of Figure 8.3.

AR Database The last dataset used is the AR Face Dataset [326]. It contains more than 4,000 face images, which vary in facial expression, illumination and occlusions. These images are from

¹<https://github.com/fdbtrs/Masked-Face-Recognition-KD>

²<https://github.com/haibo-qiu/FROM>

Table 8.1: Face verification accuracy(%) on the AR dataset under protocols 1 and 2. “Sg” and “Sf” stand for “Sunglasses ” and “Scarf ”, respectively. The first group of methods represents general face recognition models; the second group of methods represents occluded face recognition models; the last group of methods represents masked face recognition models.

Type	Method	Protocol 1		Protocol 2	
		Sg	Sf	Sg	Sf
FR	CosFace [500, 411]	99.44	100.00	98.06	100.00
	SphereFace [307, 411]	88.06	96.25	87.50	95.28
	ArcFace [126, 411]	98.33	100.00	95.00	99.72
	ElasticFace - Arc[46]	100.00	99.61	99.22	98.82
	ElasticFace - Arc+[46]	99.87	99.34	99.22	98.69
	ElasticFace - Cos[46]	99.87	99.61	99.22	98.95
	ElasticFace - Cos+[46]	100.00	99.48	99.22	98.69
	MagFace [335]	96.73	96.99	92.41	93.33
OFR	Stringface [87]	-	-	82.00	92.00
	LMA [332]	-	-	96.30	93.70
	SRC [515]	87.00	59.50	-	-
	NMR [526]	96.90	73.50	-	-
	MLERPM [510]	98.00	97.00	-	-
	SCF-PKR [529]	95.65	98.00	-	-
	MaskNet [498]	90.90	96.70	-	-
	RPSM [511]	96.00	97.66	84.84	90.16
	PDSN [463]	99.72	100.00	98.19	98.33
	FROM [411]	100.00	100.00	100.00	100.00
MFR	FocusFace [360]	96.86	99.22	94.90	98.56
	FocusFace - Pretrained [360]	99.22	99.48	98.16	98.82
	ElasticFace-Arc-Aug [221, 46]	99.87	99.61	99.08	98.95
	KD - HG [221]	99.61	99.61	98.95	98.95
	KD - LG [221]	99.08	99.61	98.95	98.82
	EUM - ResNet100 [47]	100.00	99.22	98.69	98.30

126 individuals. It presents two different types of real occlusions: sunglasses or a scarf. Occluded images are part of the probe set and the occluded face verification on this dataset is conducted through two different protocols. Some examples can be visualized in the last row of Figure 8.3.

8.5 Experimental Setup

In this section, the performed experiments are presented in detail, as well as the evaluation metrics used to assess the performance of the different models.

8.5.1 Experiments

Each experimental setup targets a specific dataset and task. For instance, experiments on the AR dataset are targeted for occluded face recognition. Each of these datasets has well-defined protocols for evaluation. Thus, the proposed experiments already benefit from them. For occluded face recognition, the evaluated models were: four distinct versions of ElasticFace [46], MagFace [335], two versions of FocusFace [360], Knowledge Distillation model(KD) [221], and Embedding Unmasking Model (EUM) [47]. For masked face recognition, in addition to the previously mentioned models, FROM [411] was also evaluated. FocusFace³, ElasticFace⁴, KD⁵, EUM⁶ and FROM⁷ have their code publicly available.

The occluded face verification performance experiments on the AR dataset were conducted on two different protocols [498, 511]. Protocol 1 uses six images per person to form the reference set, whereas protocol 2 uses only one. This dataset includes several previous results published in the literature and is considered to be saturated, especially after the work by Qiu *et al.* [411]. For each protocol, the performance of each method is evaluated individually for both scarves and sunglasses allowing for an analysis of upper and lower occlusions.

Experiments conducted on the Occ-LFW dataset are reported individually for each scale value. The defined protocol uses clear images as reference and occluded ones as probe. Further experiments were also conducted on the original LFW dataset as a baseline for the model's performance. The order of the pairs remains the same for all the dataset versions. The experiments conducted for this database are similar to those conducted for the masked validation datasets. The only difference remains in the fact that only unmasked to masked verification was conducted. The most interesting protocol and the one to be used in practice is when the reference is unmasked and the probe is masked. Hence, for masked face recognition, the masked versions of the LFW, CALFW, CPLFW, CFP-FP and AgeDB-30 were used as defined in [221].

We further extend the experiments by selecting a couple of pairs that were either misclassified by all the algorithms or only by the occluded face recognition algorithm. To this, we also add some pairs that were only correctly classified by the occluded face recognition algorithm with the hope that these can help to understand the difficulties and strengths of each approach.

8.5.2 Evaluation metrics

The evaluation of the models' verification performance is reported accordingly to two different widely used metrics. These metrics report a quantitative performance value. The first metric is the accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8.1)$$

³<https://github.com/NetoPedro/FocusFace>

⁴<https://github.com/fdbtrs/ElasticFace>

⁵<https://github.com/fdbtrs/Masked-Face-Recognition-KD>

⁶<https://github.com/fdbtrs/Self-restrained-Triplet-Loss>

⁷<https://github.com/haibo-qiu/from>

where TP, TN, FP and FN are true positives, true negatives, false positives and false negatives, respectively. However, since the accuracy is somewhat lacking in scenarios where false positives are worse than false negatives, the second metric adopted is TAR (True Accepted Rate) under FAR (False Accepted Rate), which is defined as:

$$TAR = \frac{TP}{TP + FN} \quad (8.2)$$

$$FAR = \frac{FP}{FP + TN} \quad (8.3)$$

We acknowledge the evaluation metrics in the ISO/IEC 19795-1 [325] standard, however, to enable the comparability with previous works, we follow the evaluation metrics defined in the utilized benchmarks.

8.6 Results

This section presents the quantitative results of all the conducted experiences, and a discussion/interpretation of the meaning and the reasons behind certain results.

Table 8.2: Face verification accuracy(%) on the LFW and its occluded versions. All the results, except for LFW, are reported for a verification with clear references and occluded probes. Besides FROM, all the other methods were evaluated by us.

Type	Method	LFW	Occ-LFW-1.0	Occ-LFW-2.0
OFR	FROM [411]	99.38	-	94.70
FR	ElasticFace - Arc[46]	99.80	99.49	90.88
	ElasticFace - Arc+[46]	99.82	99.23	90.98
	ElasticFace - Cos[46]	99.82	99.40	91.66
	ElasticFace - Cos+[46]	99.80	99.22	92.06
MFR	FocusFace [360]	99.08	94.95	76.96
	FocusFace - Pretrained [360]	99.53	98.41	85.86
	ElasticFace-Arc-Aug [221, 46]	99.73	99.32	92.33
	KD - HG [221]	99.68	99.30	91.47
	KD - LG [221]	99.68	99.22	91.77
	EUM - ResNet100 [47]	99.82	99.23	91.13

8.6.1 AR Face Database

As one of the most used databases for research, it is possible to report the results of a myriad of methods over the years. As seen in Table 8.1, the performance has been steadily increasing for occluded face recognition models. However, recent general face recognition models have shown impressive results, especially after considering that they were not trained to handle occlusions.

Table 8.3: Face verification TAR @ FAR = $1e-3$ (%) on the LFW and its occluded versions. All the results, except for LFW, are reported for a verification with clear references and occluded probes. Besides FROM, all the other methods were evaluated by us.

Type	Method	LFW	Occ-LFW-1.0	Occ-LFW-2.0	Occ-LFW-3.0
OFR	FROM [411]	98.63	96.17	76.53	70.23
FR	ElasticFace - Arc[46]	99.67	98.43	74.67	60.43
	ElasticFace - Arc+[46]	99.70	99.10	76.30	59.50
	ElasticFace - Cos[46]	99.70	99.03	80.30	65.50
	ElasticFace - Cos+[46]	99.67	98.67	78.16	62.26
MFR	FocusFace [360]	98.30	87.97	48.80	31.96
	FocusFace - Pretrained [360]	99.26	95.86	62.47	45.26
	ElasticFace-Arc-Aug [221, 46]	99.60	98.46	78.40	59.00
	KD - HG [221]	99.63	98.73	76.37	62.37
	KD - LG [221]	99.60	98.76	78.13	60.10
	EUM - ResNet100 [47]	99.23	98.53	75.20	60.87

For instance, ElasticFace is only surpassed by FROM, which achieves a perfect score. PDSN is capable of achieving a better result, only on protocol 1, and when the occlusion is a scarf.

While masked face recognition models achieve lower results than FROM, they are highly competitive with PDSN and surpass previous models trained for occluded face recognition. Moreover, these models are lighter than FROM and PDSN as they require less computational resources and have fewer parameters. Although these results can be interpreted with optimism, the performance of general face recognition methods is similar to the one displayed by masked face recognition models. Thus, it is still unclear how much of this performance is due to an improvement of the former.

8.6.2 Occ-LFW Database

The experiments of the various models on the original LFW datasets show that the majority of masked and general face recognition models surpass the performance of FROM, which was trained for occluded face recognition. As the best performing Occluded Face Recognition method, and considering that it was also evaluated in masked face recognition datasets, FROM is the ideal benchmark for the comparison with the MFR methods. These results are displayed in Table 8.2. Despite the good performance of these methods, it is interesting to note that some masked face recognition, such as FocusFace without pre-training, has lower performance than FROM on general face recognition. FROM performance on the Occ-LFW-1.0 dataset was not reported, hence, it was not included. On the other methods, there is a small degradation in the performance. Nonetheless, two methods are capable of achieving better performance on this dataset than the one attained by FROM on the original LFW.

The performance gap is completely removed, and FROM becomes the best performing method

when evaluated on the Occ-LFW-2.0 dataset. This is due to a faster degradation in the performance of general and masked face recognition methods when the size of the occlusions increases. Nonetheless, ElasticFace and KD achieve unexpectedly good results on a dataset with more than 40% of the image area occluded.

To further understand how this performance degradation affects the results in a real-world evaluation, we also collected the TAR when the FAR was set to $1e-3$. Considering that on a real application the false acceptance rate must be set to a considerably low value, due to its impact on the practicality of the system, the TAR measures how well it performs when the threshold is selected to optimize the FAR value. Table 8.3 displays the results of each of the evaluated methods. It is interesting to note that on this specific metric the performance of a general face recognition method, ElasticFace, surpasses both masked and occluded face recognition models for Occ-LFW-1.0 and 2.0. The same remarkable performance is attained by KD and EUM.

These impressive and unexpected results are enlightening. In a sense, they argue against the need for models specifically designed for occluded face recognition, since these have a worse performance when the maximum number of false positives is limited. Moreover, the extra parameters, complexity and requirements for these models diminish their interest. On the other side, the success of masked face recognition models indicates a potential research direction to be followed by occluded face recognition methods.

Table 8.4: Face verification accuracy(%) the masked versions of LFW, CALFW, CPLFW, CFP-FP, AgeDB-30. All the results are reported for a verification with unmasked references and masked probes. All the presented methods were evaluated by us. FROM evaluation was conducted on 96x112 images, since it was trained on images with those dimensions.

Type	Method	LFW	CFP-FP	AgeDB-30	CALFW	CPLFW
OFR	FROM [411]	98.47	93.01	80.18	89.35	85.70
FR	ArcFace [126]	99.45	95.50	94.68	94.76	89.83
	MagFace [335]	99.33	94.00	94.12	94.15	88.67
	ElasticFace - Arc[46]	99.40	95.29	95.38	94.42	90.40
	ElasticFace - Arc+[46]	99.52	95.44	95.68	94.62	89.90
	ElasticFace - Cos[46]	99.52	95.43	95.58	94.62	90.18
	ElasticFace - Cos+[46]	99.37	96.27	95.47	94.78	90.37
MFR	FocusFace [360]	98.82	86.69	93.68	92.50	83.37
	FocusFace - Pretrained [360]	99.10	91.04	94.25	93.80	86.77
	ElasticFace-Arc-Aug [221, 46]	99.58	95.91	96.73	95.58	91.08
	KD - HG [221]	99.73	96.94	96.69	95.75	91.67
	KD - LG [221]	99.67	96.83	97.07	95.90	91.98
	EUM - ResNet100 [47]	99.43	95.07	95.47	94.90	89.93

8.6.3 Masked Face Verification

On the other hand, the diversity of different occlusions used to train the occluded face recognition model, does not lead to performance comparable to the performance of masked face recognition models. Not only that, but the performance of the general face recognition models is also superior on all the five datasets tested as displayed in Table 8.4. One important element to consider is that as seen in the previous sections, some approaches that start from a pre-trained general face recognition model and are afterwards tuned for masked face recognition perform well on both masked and occluded datasets. It is also independent of the tuning approach followed.

Furthermore, the performance difference between occluded masked face recognition models and the others is more noticeable on CALFW and AgeDB-30. Hence, it shows that either the model trained for face recognition under occlusions is unable to handle large age gaps, or that the masks make that task significantly harder for that model.

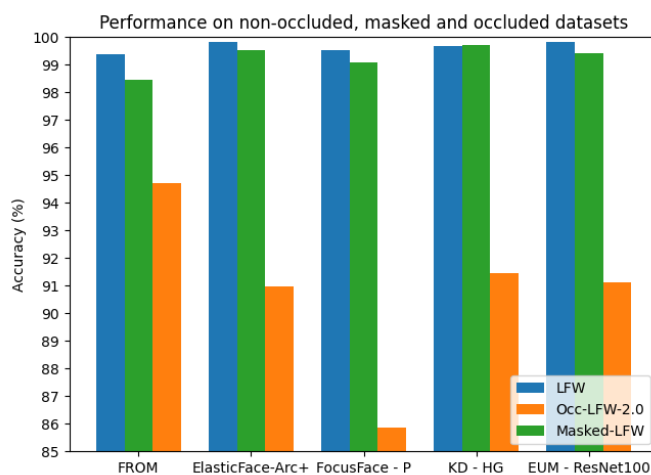


Figure 8.5: Study of the performance of the algorithms when exposed to no occlusions, masks and general occlusions.

Despite the presence of occlusions similar to face masks on the occlusion dataset, the performance of FROM drops significantly on these masked benchmarks. Although it does not vary by more than 1 percent point on the masked LFW when compared to the original LFW, it is a much wider gap than the other models. This performance drop is better seen in Figure 8.5, which compares the performance of five algorithms in three distinct databases, one for each task. It also shows the impact of Occ-LFW-2.0 in the performance of these algorithms. While the performance drop is significant, we will argue in the following sections that this version of LFW dataset includes unreasonable occlusions. Moreover, if we analyze Figure 8.6, which studies the performance of the algorithms on progressively more occluded datasets, it is possible to see that the majority of MFR algorithms is comparable to FROM in Occ-LFW-1.0. But as the occlusions increase in size their performance decreases faster.

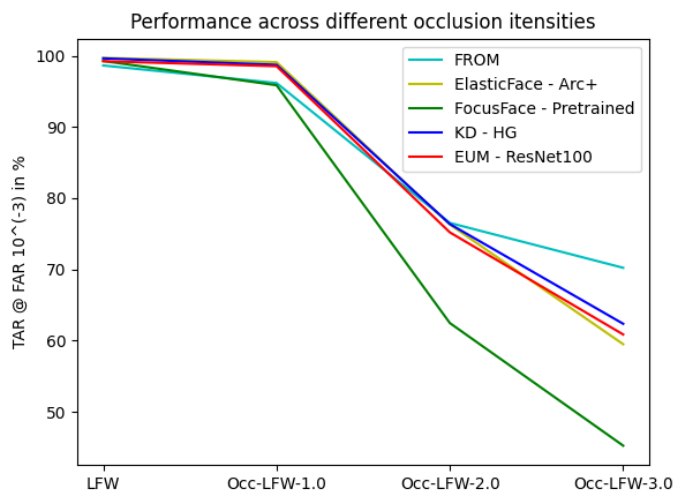


Figure 8.6: Analysis of the performance variation of the algorithms in several versions of the LFW dataset that include different intensity of occlusion.

8.6.4 Results qualitative analysis

An analysis was made focusing on some of the misclassified pairs of images, in order to qualitatively study the performance of the algorithms. For this study, the following models were used: FROM, ElasticFace - Arc+, KD - HG and EUM. Only results from Occ-LFW-1.0 and 2.0 were assessed. Figures 8.7 and 8.8 show some false positive and false negative samples common to all the models considered. Regarding the false positives, there was no single pair that was common to all models on Occ-LFW-1.0. On 2.0, the two given samples present a considerably difficult task. On the other hand, on the false-negative pairs, on 1.0 the examples displayed indicated a difficulty of all the models on pairs that comprise age gaps. The second false-negative pair of the 2.0 dataset presents a nonsensical occlusion, which can propel some discussion regarding the validity of the decision, despite being correct or wrong, in these samples.

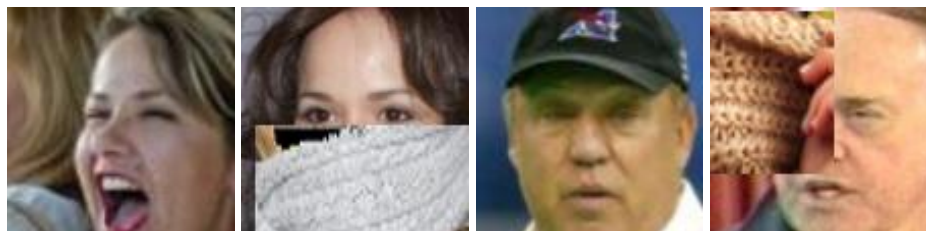


Figure 8.7: Some example pairs from the Occ-LFW-2.0 dataset that were wrongly classified as positive by the following models: FROM, ElasticFace - Arc+, KD - HG and EUM. Each couple of images placed horizontally represents a pair.

Figures 8.9 and 8.10 show the false-positive and false-negative pairs wrongly classified by FROM and correctly classified by the remaining models. It is interesting to note that the majority

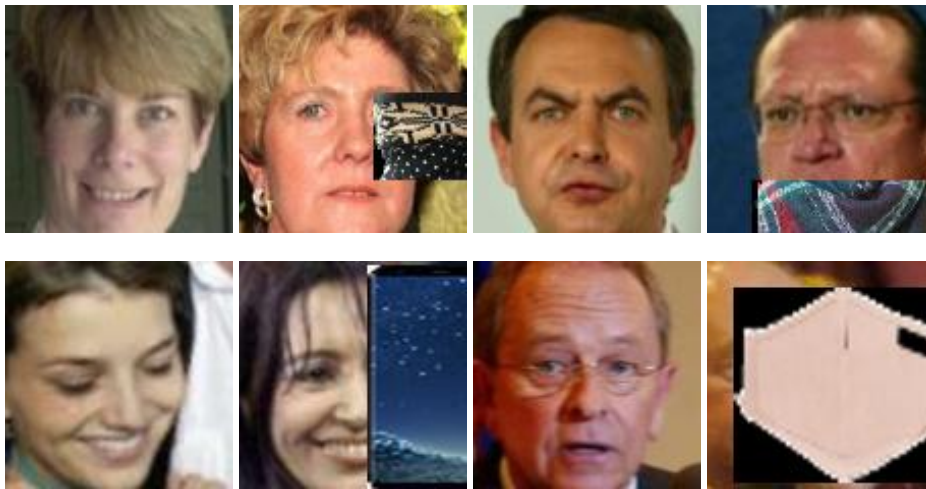


Figure 8.8: Some example pairs from the Occ-LFW dataset that were wrongly classified as negative by the following models: FROM, ElasticFace - Arc+, KD - HG and EUM. First row represents 1.0 dataset and the second the 2.0 dataset. Each couple of images placed horizontally represents a pair.

of the errors committed exclusively by the occluded face recognition model include occlusions on the lower-face area, similar to face masks. It indicates that the performance of masked face recognition models generalizes sufficiently well for these synthetic occlusions. Nonetheless, in Figure 8.9 the occlusions generated are particularly tough and highly unrealistic.

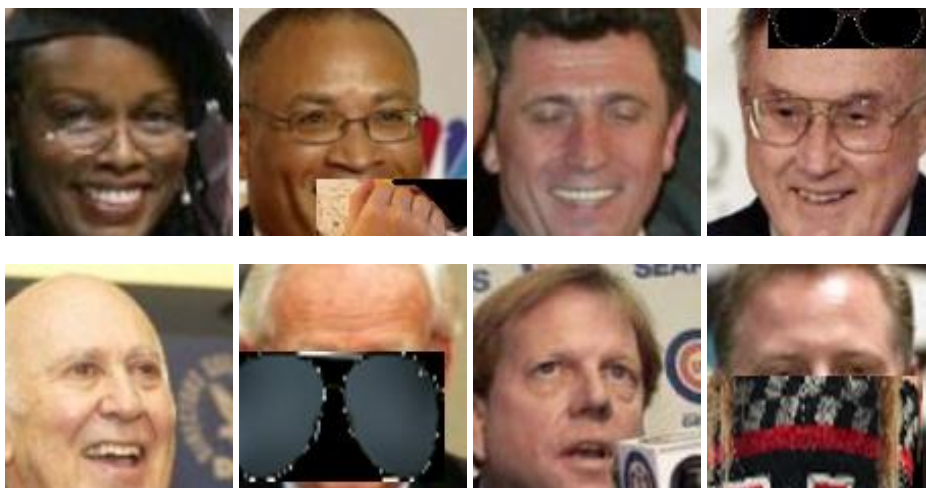


Figure 8.9: Some example pairs from the Occ-LFW dataset that were wrongly classified as positive by the following models: FROM; correctly classified as negative by the following: ElasticFace - Arc+, KD - HG and EUM. First row represents 1.0 dataset and the second the 2.0 dataset. Each couple of images placed horizontally represents a pair.

Finally, we analyze Figures 8.11 and 8.12, which represent the false positives and false negatives correctly predicted by FROM and wrongly predicted by the other models. Differently from

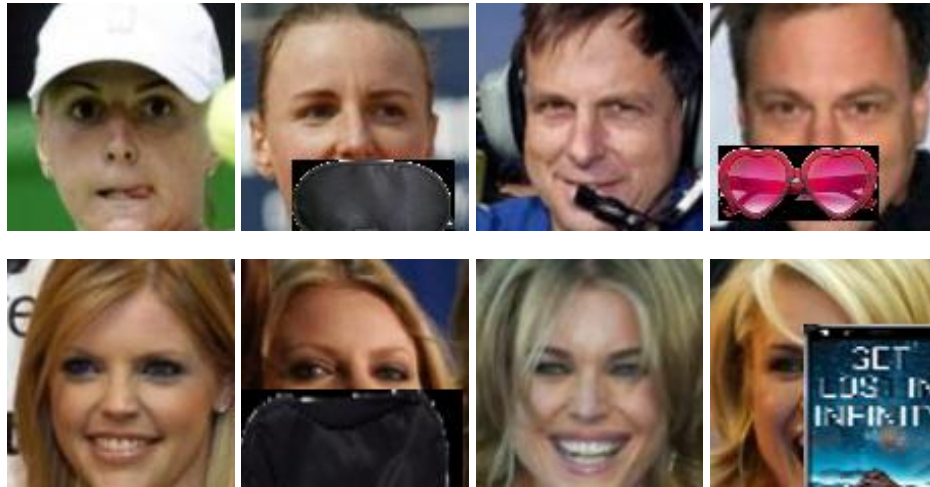


Figure 8.10: Some example pairs from the Occ-LFW dataset that were wrongly classified as negative by the following models: FROM; correctly classified as positive by the following: ElasticFace - Arc+, KD - HG and EUM. First row represents 1.0 dataset and the second the 2.0 dataset. Each couple of images placed horizontally represents a pair.

the previous two figures, these indicate that masked face recognition models are particularly ill against occlusions on the ocular area or forehead. Nevertheless, it must be noted that the number of samples erroneously classified by these models and correctly classified by FROM is considerably small. Due to their training nature, this type of error was already expected from the start. Figure 8.11 shows a false positive between two images from individuals of distinct genders.

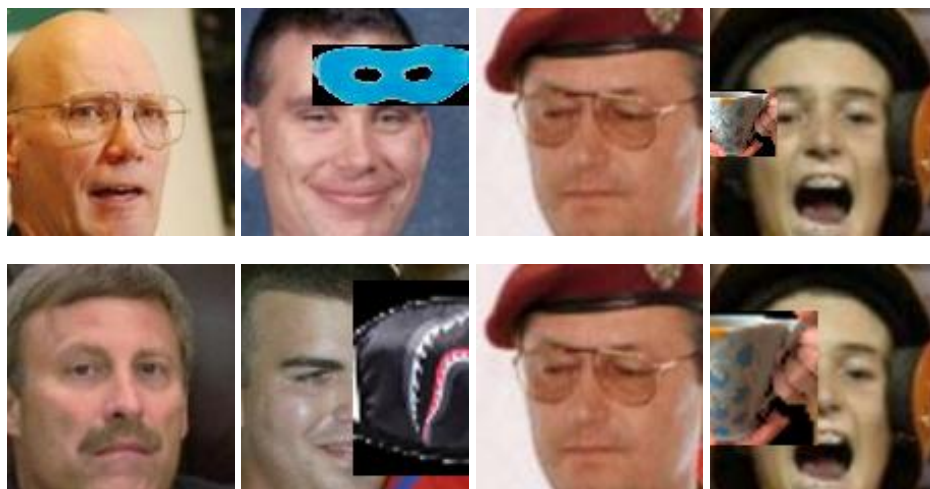


Figure 8.11: Some example pairs from the Occ-LFW dataset that were wrongly classified as positive by the following models: ElasticFace - Arc+, KD - HG and EUM; correctly classified as negative by the following: FROM. First row represents 1.0 dataset and the second the 2.0 dataset. Each couple of images placed horizontally represents a pair.



Figure 8.12: Some example pairs from the Occ-LFW dataset that were wrongly classified as negative by the following models: ElasticFace - Arc+, KD - HG and EUM; correctly classified as positive by the following: FROM. First row represents 1.0 dataset and the second the 2.0 dataset. Each couple of images placed horizontally represents a pair.

8.7 Conclusion

In this Chapter, we studied the interoperable deployability of two similar but still distinct research directions. We carefully crafted a set of experiments, on masked face recognition and occluded face recognition to analyze the performance of these models in the other task. We aimed to learn the differences between these models and the implications, positive and negatives, of deploying them for a task that they were not trained for. From the set of experiments conducted, it became clear that the performance of the current masked face recognition models is more generalizable, robust and uses fewer parameters. This is an important finding since it shows that building on top of the current face recognition models is a viable option that does not increase the complexity of the solution. These experiments were conducted on datasets frequently used to benchmark each of these two research directions.

The capabilities displayed by MFR models foreshadows the future of occluded face recognition. The former, propelled by the urgency of the pandemic, focused on increasing the robustness of current models, which led to simpler, faster to train and more robust solutions when compared to occluded face recognition. However, these findings should not stop researchers from pursuing other research directions within occluded face recognition. We believe that diversity in research and a variety of research directions leads to better breakthroughs. Nonetheless, we still provided sufficient findings to support the investment in an unexplored research direction. Hence, it is possible to conclude that not only masked face recognition models can be used on other occlusions with improved results, but also present a refreshing approach to the task of recognizing under occlusions. In addition we initiated a discussion regarding the validity of decisions on highly occluded and unrealistic samples, which will be addressed in the following Chapter.

Chapter 9

Towards Better Data for Occluded Face Recognition Evaluation

Author Contributions The research work described in this chapter was conducted in collaboration with João R. Pinto, Fadi Boutros and Naser Damer, under the supervision of Ana F. Sequeira and Jaime S. Cardoso. The author of this thesis contributed to this work on the problem conceptualization, data curation, the setup and evaluation of solutions, the results discussion, and publication writing. Additionally, this Chapter originates from hosting a competition, and several participants contributed with different solutions. Some parts of the chapter were originally published in, or adapted from:

- Neto, P. C., Boutros, F., Pinto, J. R., Damer, N., Sequeira, A. F., Cardoso, J. S., ... & Menotti, D. (2022, October). OCFR 2022: Competition on occluded face recognition from synthetically generated structure-aware occlusions. *In 2022 IEEE International Joint Conference on Biometrics (IJCB)* (pp. 1-9). IEEE.

9.1 Context and Motivation

Since 2020, the COVID-19 pandemic has intensified the occurrence of a particular type of face occlusion, the facial mask. Due to sanitary concerns, the majority of countries introduced an obligation to wear masks in closed or crowded environments. As expected, it led to a significant performance drop in face recognition (FR) scenarios [104, 108, 184], as shown on the latest FR vendor test conducted by NIST [375]. Similarly, Damer *et al.* [103] have shown that the performance of humans in recognizing faces is also negatively affected by the presence of a face mask. These findings led to a new direction in FR research. Hence, an additional effort was made to solve masked face recognition (MFR) problem, which benefits from contributions in the form of case-study papers [158, 170, 442, 499], methods [47, 221, 360, 362] and competitions [48, 124]. Moreover, the effect of masks was not exclusively linked to recognition and studies can be found on their effect on face presentation attack detection methods [154, 158]. The volume of new research originated a second part of the NIST FR vendor test, which evaluated the performance of FR algorithms designed after the pandemic [379].

While MFR is a special case of a broader problem known as occluded face recognition (OCFR), the research output on the latter was not significantly affected by the pandemic. Surprisingly, the vast majority of the solutions proposed to solve MFR were not inspired by previous OCFR research as seen in Chapter 8. Thus, there is an opportunity to work on the broader problem with solutions inspired by the more specific one. There are a few works on OCFR post-pandemic. Erakiotan *et al.* [148] proposed a new dataset composed of real faces occluded by sunglasses or face masks. Qiu *et al.* [411] proposed an improved framework over the previous state-of-the-art method by Song *et al.* [463]. Their method, FROM, decreases the training and testing complexity by removing the need for two different inference stages in both training and testing. Both Huang *et al.* [216] and Zeng *et al.* [543] propose architectures that, due to a segmentation module, are occlusion-aware. He *et al.* [207] studied a variation of dropout to maximize the performance under occlusions. Finally, Zeng *et al.* [542] conducted an extensive research on the techniques used for OCFR.

The majority of the works previously mentioned are evaluated on small scale datasets composed of real occlusions or on larger datasets with synthetic occlusions arbitrarily added to face images. These two different approaches for model evaluation produce different results. This finding is stated by Jiang *et al.* [247] after performing the necessary comparative studies. One key highlight from their work is that current models are being tested on strange and abnormal synthetic occlusions, such as blocks. The insertion of a structure to the occlusions added on a synthetic dataset has already been initially approached by Huang *et al.* [217]. However, their approach lacks variety and does not attempt to combine multiple occlusions on different structural areas. Thus, there is a need for more realistic face occlusions data that allow the development of methods that are more robust in real world applications.

One difference between OCFR and MFR research is that the latter usually aims to be an extension of FR, in the sense, that several of the proposed model introduce little to no overhead or complexity. Driven by increasing deployment on consumer and embedded devices, there is an interest in lightweight FR approaches, as shown in 2019 ICCV competition [127]. Some models from that competition and from follow-up research have shown an interesting trade-off between complexity and performance [44, 57, 85]. As stated, research on MFR runs in many cases on top of existing models, making it model-agnostic, such as in [47]. Hence, it can also benefit from these lightweight models, whereas OCFR is mostly application dependent.

Inspired by: (a) the effective research that has been conducted over the past two years on MFR; (b) the larger information availability variation in OCFR in comparison to MFR; (c) the need to revitalize the field with novel research in the direction of improving OCFR performance; and (d) the need for a challenging benchmark truly inspired by real occlusions and a notion of structure on where synthetic occlusions are added, we conducted the IJCB Occluded Face Recognition Competition 2022 (IJCB-OCFR-2022). The final participation toll comprises 3 teams, which have submitted two solutions each, totaling 6 valid solutions. The solutions were evaluated on our novel benchmark including seven databases comprising occlusions placed over four different structural areas, which can also be combined into stronger and broader occlusions. This Chapter summarizes

this competition, the submitted solutions, the novel benchmark and the achieved results in terms of occluded vs occluded face verification accuracy, and occluded vs non-occluded face verification accuracy, as well as the compactness of the recognition models. This comprises an effort towards improving on an issue detected on the previous Chapter. In the following sections, we introduce the new benchmark used for the competition and the associated evaluation criteria, in Section 9.2. In Section 9.3, we describe the six valid solutions submitted by three teams. The results achieved are discussed in Section 9.4. Finally, a brief conclusion is provided in Section 9.5.

9.2 Database and evaluation criteria

9.2.1 OCFR-2022 evaluation benchmark

The evaluation dataset is built from Labeled Faces in the Wild (LFW) dataset [218] by synthetically occluding parts of the face images. This dataset is referred to as occluded face recognition competition data (OCFR-2022) and it is used to evaluate the verification performance of the submitted solutions. The OCFR-2022 evaluation benchmark is publicly available¹ as a batch implementation that would take the LFW as an input and produce the OCFR-2022 evaluation benchmark. In addition, the proposed methodology for this dataset, which is described below, has been used to create a race-balanced occluded face recognition dataset from Racial Faces in the Wild [503]. Hence, further contributing for a bias and robustness analysis under these conditions [323]. Furthermore, one of the top solutions in the Frcsyn challenge at cvpr 2024 [117, 116] leverage these occlusions to create augmentations in the training data.

The synthetic occlusions are added in a realistic manner based on three assumptions: 1) there is an inherent structure to the occlusions that are seen in real datasets; 2) some occlusions are only fit to certain spatial areas in the face; and 3) it is possible to have a combination of occlusions.

Table 9.1: Occlusions

Type	Areas of occlusions
U	From the nose to the forehead
L	From the chin to the nose
T	Parts of the forehead
E	The eyes

For the first assumption, we considered the following structure for occlusions: top of the head (T), eye-based (E), upper (U) and lower (L) face, as seen in Table 9.1. These types of occlusions cover different parts of the face. Thus, as stated in the second assumption, each has specific types of objects (occluders) associated with that area, for instance, sunglasses are not expected to be placed on top of the mouth nor hats are expected to be on the lower part of the face. Some of these occlusions can be seen in Figure 9.1. To comply with this assumption, it was necessary to individually and manually annotate all the occluders with the expected position of at least two of

¹<https://github.com/NetoPedro/OCFR-2022>

the five face landmarks. In other words, for sunglasses, it is necessary to annotate the position of the left eye in the left lens and of the right eye in the right lens. It allows the alignment with the detected landmarks in the original image. We further add noise, drawn from a uniform distribution, to the landmarks for variability purposes.



Figure 9.1: Examples of occluded face images of the OCFR-2022 benchmark comprising top of head (e.g. hats), upper (e.g. carnival masks), eye-based (e.g. sunglasses), and lower (e.g. mouth occlusions such as books, masks) face occlusions. Original non-occluded images are from the Labeled Faces in the Wild (LFW) dataset [218].

Finally, for the third, we propose to combine several occlusions of different types. For instance, an E and an L occlusion can be combined to create a strongly occluded face, as seen in Figure 9.2.

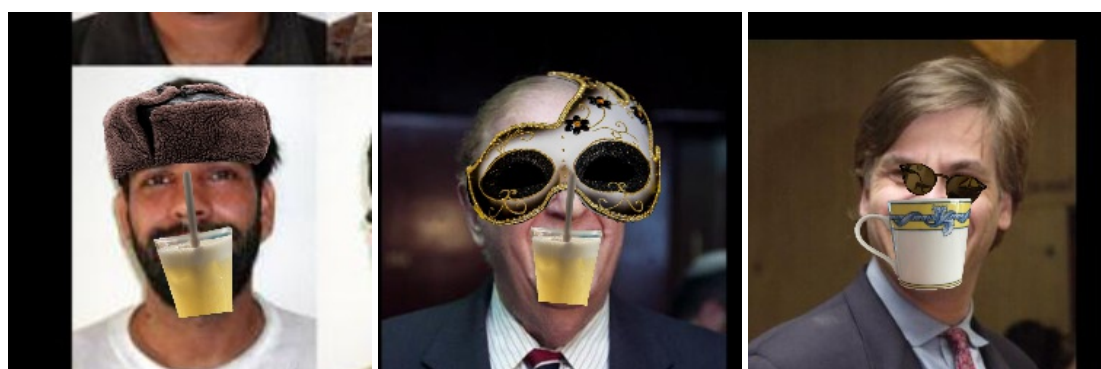


Figure 9.2: Examples of occluded face images comprising, from left to right, a combination of top of head and lower face (T+L); upper and lower face (U+L); and eye-based and lower face occlusions (E+L). Original non-occluded images are from the Labeled Faces in the Wild (LFW) dataset [218].

Despite the liberty to create different configurations through the combination of several occlusions, some combinations are not allowed. For example, adding glasses (occlusion of type E) on top of a carnival mask (occlusion of type U) is not allowed, as they occlude a similar face area. The combinations that were used to create the seven occluded benchmark protocols are in Table 9.2. Our experiments with the baseline have shown that occlusions of type T were not challenging when used alone. Hence, they were never used independently.

Table 9.2: Evaluation protocols

Protocol	Occlusions
#0	No occlusion
#1	U, E or L
#2	L + [UET] or T + [UE]
#3	L + [UET], T + [UE] or #1
#4	L + [ET], T + [UE] or #1
#5	L + T + [UE]
#6	L + T + [UE] or #2
#7	L + T + [UE] or #3

The LFW dataset [218], used to create the evaluation benchmark, was released in 2007, and since, has become a popular benchmark to compare FR models. It is composed of 13,233 face images of 5,749 identities in unconstrained environments. The number of comparison pairs in protocol [218] of LFW is 6000 (3000 genuine and 3000 imposter comparisons). Nonetheless, our approach to generate synthetic occluded images can be applied in any FR dataset without restrictions. Based on the occlusion protocols (Table 9.2) and the 6000 comparisons pairs provided by the LFW official evaluation protocol [218], we constructed 14 different versions of the dataset. For each occlusion protocols (Table 9.2), we created two versions of occluded LFW. The probes are always occluded. The references are only occluded in the second version.

Table 9.3: Participants

Team name	Affiliation
BIOSMC-CDTA	CDTA- Centre de Développement des Technologies Avancées; ESI - Ecole Nationale Supérieure d'Informatique
SiMiT Lab	Smart Interaction and Machine Intelligence Lab, Istanbul Technical University, Turkey
UFPR-UNICO	Federal University of Parana, Brazil

9.2.2 Evaluation criteria

For the evaluation of the models, we compare their performance on seven different sets of the OCFR-2022 benchmark, which comprise different combinations of occlusions. Besides these datasets, we also include an evaluation of the original LFW dataset [218] (without occlusions) and information on the compactness of the models, which are not used to rank the solutions for the competition. For ranking purposes, the models will be evaluated with non-occluded references and occluded probes. We consider this scenario to be the most likely to occur in the real world, and it

will be noted as BLR-OP. We further report the performance of occluded reference and occluded probes on two of the seven datasets. This scenario is noted as OR-OP.

The verification performance is reported as the false non-match rate (FNMR) at two different operation points, FMR100 and FMR1000, which are the lowest FNMR for a false match rate (FMR) $< 1.0\%$ and $< 0.1\%$, respectively. For each evaluation setups, the submitted solutions ranking is based on FMR100. Finally, we further report the Equal Error Rate (EER) for each evaluation setups.

The final team's ranking is based on the average of the ranking attained on the seven protocols. Hence, a smaller value indicates a superior solution in terms of ranking order.

Baseline Methods The state-of-the-art FR performance has grown significantly over time. Hence, to better understand and study the performance of the submitted algorithms, it is necessary to compare them with a state-of-the-art FR baseline approach. The selected baseline is the ElasticFace [46], which achieved state-of-the-art performance on several FR benchmarks. In this work, we use the ElasticFace-Arc+ [46] that avoids the assignment of fixed margin value in margin-penalty softmax loss through a flexible penalty margin. The backbone of the ElasticFace is a ResNet-100 [205] trained on a refined version of the MS-Celeb-1M dataset [198] (MS1MV2). We used a pretrained version of ElasticFace known as ElasticFace-Arc+ provided by the authors [46]². We compare the submitted solutions with two versions of the ElasticFace Arc+: **Baseline UB** and **Baseline**. The first has access to the landmarks detected from the non-occluded version of the data, which are used to align the dataset. This version is considered an upper bound of the baseline performance, hence, it is known as **Baseline UB**. The other version only has access to the bounding boxes, which are the only inputs given in the competition. This version, known as **Baseline**, does not perform alignment nor landmark detection, and it is a less challenging version of the baseline ran on the same conditions of the submitted solutions. One must keep in mind that the baseline was trained to expect an aligned face image.

9.2.3 Submission and evaluation process

Each submission was requested as a Win32 or Linux console application. However, this was later extended to allow python scripts. A template script and the virtual environment requirements were shared with the participants. The solution should accept the evaluation-list (text file), bounding-boxes (text file) and an output path as parameters. The first input contains the pairs of paths for both the reference and the probe image. The bounding-boxes file contains the bounding boxes detected on the original LFW dataset using the MTCNN algorithm [546]. We note that, for realistic evaluation purposes, we did not provide the detected landmarks. The output of the solution is a text file, which contains the comparison scores between each pair in the evaluation list. Additionally, participants who do landmark detection and alignment were asked to provide feedback regarding the success of the detection.

²<https://github.com/fdbtrs/ElasticFace>

9.3 Participants and submitted solutions

The competition aimed to include a diverse set of participants with a significant geographic variation. The call for participation was shared on the International Joint Conference on Biometrics (IJCB 2022) website; on the website dedicated to the competition³; on public computer vision mailing lists (e.g. CVML e-mailing list); and on the organisers private e-mailing lists. The call attracted 8 teams, from which, 3 submitted valid solutions. Each team was authorised to submit two solutions. In total there are 6 valid submissions, all from teams with academic affiliations.

9.3.1 BIOSMC-CDTA Team

The methods submitted by the BIOSM-CDTA team: **AdaFace4M** and **AdaFace12M** are based on the recently released deep models based on Quality Adaptive Margin loss dubbed AdaFace [269]. These methods do not perform face alignment and use the bounding box provided. Considering that AdaFace improves largely the FR performance on low quality academic datasets [269], the aim of the proposed methods is to investigate the performance of AdaFace on the challenging scenario presenting low quality occluded faces. The two proposed solutions (AdaFace4M and AdaFace12M) are trained using million-scale subsets of the cleaned version of the ultra-large-scale face benchmark consisting of 4M identities/260M faces (zhu2021webface260m) [558]. AdaFace12M is trained on a larger subset of 12M of the curated WebFace42M. The training procedure can be found in [269] and the implementation is provided under the MIT license by the authors of [269]. While AdaFace competes with the state of the art angular margin losses on high quality datasets, it outperforms many competing losses on low quality ones, including some quality aware FR systems [269]. Another motivation for proposing AdaFace to tackle the occluded FR challenge is the fact that non-severe occlusions can result in hard to recognize low quality face images, which is the one of the advantages of AdaFace. The submitting team states that to the best of their knowledge, AdaFace has not been investigated for the task of occluded FR. The input size of the presented solution is 112×112 and the output feature embedding size is 512-D.

9.3.2 SiMiT Lab Team

Both submissions of the SiMiT Lab Team: **SMT-OCFR1** and **SMT-OCFR2** are based on a fine-tuned pre-trained model based on the Arcface loss [126]. The pre-trained model used in the competition baseline model (LResNet100E-IR). The LResNet100E-IR model was originally trained on MS-Celeb-1M dataset [198] (MS1MV2), and fine-tuned on real world occlusion datasets. For both upper and lower face occlusions, the authors gathered the Real World Occluded Faces (ROF) [148] dataset that consists of sunglasses, masks and neutral face images. As an addition to the lower face occlusion dataset, the ROF dataset was combined with the MFR2 [23] dataset. The proposed methods perform face detection and alignment. MTCNN [546] model is used to obtain the bounding

³<https://vcml.inesctec.pt/OCFR2022/>

box and five face landmarks. All the face detection and alignment algorithms are obtained from the original Insightface library⁴.

SMT-OCFR1: The objective is to detect if the occlusion is upper or lower face. Depending on the type of occlusion, FR uses models trained using the Arcface loss [126] that are either trained on lower or upper face occlusion datasets. Occlusion detection network is a simple CNN model and trained on ROF sunglasses and ROF masks datasets. Upper face occlusion Arcface model is fine-tuned on ROF sunglasses dataset. Lower face occlusion Arcface model is fine-tuned on ROF masks and MFR2 datasets.

SMT-OCFR2: The objective is fine-tuning a pre-trained model, trained using the Arcface [126] loss, using lower and upper face occlusion datasets together to make a more robust model. The model is fine-tuned on ROF sunglasses, ROF masks, and MFR2 datasets.

Table 9.4: Face verification results on three different protocols: protocol 0, protocol 1 and protocol 2 (in the BLR-OP setting). The values of EER, FMR100, FMR1000 are displayed in percentage (%). The rank orders the solutions from the best to the worst within that protocol.

	Protocol 0				Protocol 1						Protocol 2				
Method	EER	FMR100	FMR1000	Rank	GMean	IMean	AUC	EER	FMR100	FMR1000	Rank	EER	FMR100	FMR1000	Rank
Baseline UB	0.333	0.267	0.333	-	0.522	0.002	0.997	1.633	1.967	4.700	-	13.000	20.567	24.100	-
Baseline	2.500	3.433	8.633	-	0.361	0.029	0.967	8.600	20.533	37.300	-	22.967	46.500	65.467	-
SMT-OCFR1	0.333	0.200	0.400	1	0.394	0.757	0.956	9.667	17.100	27.033	3	25.633	50.033	58.900	4
SMT-OCFR2	0.333	0.300	0.667	2	0.359	0.756	0.953	10.400	20.467	33.500	4	23.167	49.233	61.433	3
AFOIRNet_1	15.866	50.800	67.000	5	0.706	0.638	0.767	31.200	68.800	85.333	5	35.833	84.400	93.300	5
AFOIRNet_2	18.466	56.167	73.400	6	0.673	0.610	0.738	33.400	72.400	84.733	6	37.467	84.500	95.033	6
AdaFace4M	1.333	1.367	2.267	4	0.384	0.014	0.989	4.267	8.433	19.133	2	15.667	29.900	40.967	2
AdaFace12M	1.033	1.067	1.900	3	0.376	0.009	0.991	3.867	6.300	15.400	1	15.700	27.267	38.500	1

9.3.3 UFPR-UNICO Team

The methods submitted by the UFPR-UNICO team: **AFOIRNet_1** and **AFOIRNet_2** are based on fine-tuning the model proposed by [221], with non-occluded and synthetic occluded images as training data, with the objective of making a model that can discriminate occluded faces. These models do not perform face alignment and use the bounding box provided.

To create the two versions of the “Aspirant facial occlusion invariant recognition network” (AFOIRNet), the authors opted to generate two occluded versions of the CASIA-WebFace dataset [534], which originally has 494,414 face images of 10,575 identities. The first version was made by adding facial masks occlusions generated by MaskTheFace method [23] to every image of the dataset, and the second was generated by adding random occlusions in the forehead and periocular regions with one large black box to all images. The original dataset was augmented with the occluded versions, shuffled and split into train, validation and test sets by a ratio of 0.7/0.2/0.1 respectively. The training data was used to fine-tune the mask invariant network proposed by [221], named MaskInvHg, by freezing around 75% of its learnable parameters, with the objective that it learns in a end to end fashion how to be invariant of occlusions in general. The cross entropy loss

⁴<https://github.com/deepinsight/insightface>

was used to train a facial classifier, which is a fully connected layer that receives as input the 512-d feature vector generated by the network and outputs the predicted face of the training dataset. The model was trained for 50 epochs with a batch size of 64, using the SGD optimizer, with a learning rate of 0.001, and finished the training process with around 50% accuracy on the validation dataset. Lastly, the fully connected layer was removed and the feature vector used as its input is the final output of the network. The model AFOIRNet_1 was trained with the full synthetic augmented dataset, and the AFOIRNet_2 was trained on a smaller version of the augmented dataset. The models AFOIRNet_1 and AFOIRNet_2 have 18269004 and 14765727 learnable parameters and size 283MB, 269MB, respectively.

Table 9.5: Face verification results on three different protocols: protocol 3, protocol 4 and protocol 5. All protocols in the BLR-OP setting. The values of EER, FMR100, FMR1000 are displayed in percentage (%). The rank orders the solutions from the best to the worst within that protocol.

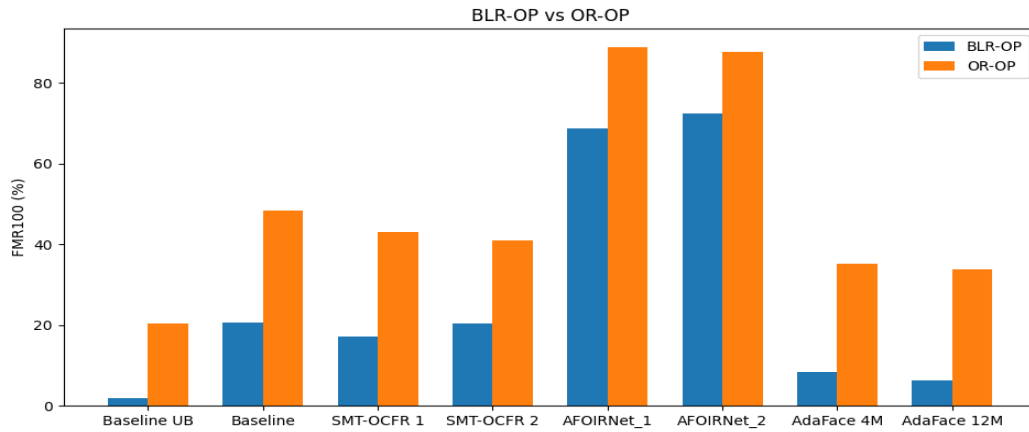
	Protocol 3							Protocol 4				Protocol 5			
Method	GMean	IMean	AUC	EER	FMR100	FMR1000	Rank	EER	FMR100	FMR1000	Rank	EER	FMR100	FMR1000	Rank
Baseline UB	0.417	0.003	0.960	9.233	13.567	16.300	-	1.967	2.800	6.600	-	28.633	59.867	68.533	-
Baseline	0.271	0.023	0.884	18.833	35.867	45.567	-	11.500	30.333	40.467	-	38.200	84.200	93.633	-
SMT-OCFR1	0.488	0.746	0.879	20.800	36.267	45.267	3	14.967	30.767	38.933	3	45.200	94.433	96.933	4
SMT-OCFR2	0.497	0.773	0.880	19.400	38.733	50.600	4	14.633	31.833	49.567	4	41.967	93.467	96.900	3
AFOIRNet_1	0.686	0.631	0.727	34.833	80.700	90.000	5	31.767	78.333	89.367	5	41.600	94.700	98.967	5
AFOIRNet_2	0.650	0.599	0.699	35.800	81.267	91.233	6	32.833	80.233	91.900	6	42.533	95.367	99.400	6
AdaFace4M	0.303	0.015	0.938	12.633	20.833	33.033	2	6.000	12.233	23.766	2	32.133	72.100	88.567	1
AdaFace12M	0.294	0.011	0.940	11.933	20.333	27.800	1	5.933	11.967	20.833	1	33.000	74.133	89.267	2

Table 9.6: Face verification results on two different protocols: protocol 6 and protocol 7. All protocols in the BLR-OP setting. The values of EER, FMR100, FMR1000 are displayed in percentage (%). The rank orders the solutions from the best to the worst within that protocol. Includes information regarding the compactness of the models in terms of the number of learnable parameters and GFLOPs.

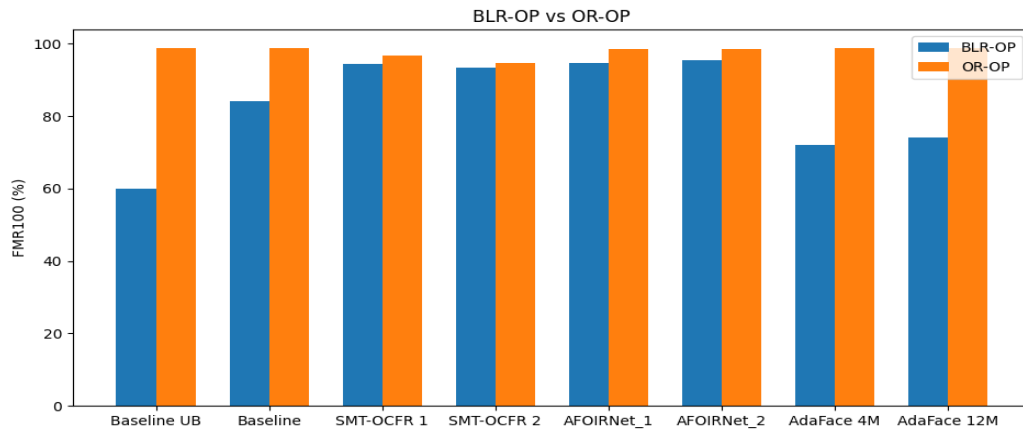
	Protocol 6							Protocol 7				Compactness	
Method	GMean	IMean	AUC	EER	FMR100	FMR1000	Rank	EER	FMR100	FMR1000	Rank	# of learnable parameters	GFLOPs
Baseline UB	0.297	0.002	0.897	18.267	30.400	37.600	-	14.100	22.633	26.600	-	65 155 648	-
Baseline	0.182	0.010	0.793	28.867	55.133	59.667	-	23.533	46.400	61.433	-	65 155 648	-
SMT-OCFR1	0.575	0.762	0.779	30.167	57.567	67.100	3	26.100	48.467	57.800	3	131 206 393	24.35
SMT-OCFR2	0.565	0.739	0.784	29.200	59.467	69.600	4	25.167	49.933	60.300	4	65 131 000	24.20
AFOIRNet_1	0.664	0.624	0.679	37.600	85.833	95.300	5	37.067	81.900	89.933	5	65 150 912	12.09
AFOIRNet_2	0.623	0.585	0.655	39.100	86.533	94.633	6	38.333	83.567	90.800	6	65 150 912	12.09
AdaFace4M	0.213	0.013	0.870	21.300	41.867	54.667	2	17.667	31.633	47.667	2	65 150 912	12.09
AdaFace12M	0.203	0.010	0.869	21.433	41.000	52.367	1	18.067	30.967	44.533	1	65 150 912	12.09

9.4 Results and discussion

In this section we present and discuss the results of the six solutions submitted to the competition and compare them with the results of the baseline approaches. The ranking results are spread across 7 different protocols and the final ranking is reported at the end of the section. AdaFace12M consistently beats the remaining submissions. Hence, it is considered the top ranked solution. Further, we show the results for Occluded-Occluded face verification on two protocols, the least difficult and the hardest one.



(a) Protocol 1



(b) Protocol 5

Figure 9.3: Comparison of the performance of the Baseline algorithms and the submitted solutions on the easiest and the hardest Protocols, 1 and 5 respectively. Side by side analysis of the performance of BLR-OP and OR-OP. The algorithms are displayed in the following order "Baseline UB", "Baseline", "SMT-OCFR 1", "SMT-OCFR 2", "AFOIRNet_1", "AFOIRNet_2", "AdaFace 4M", "AdaFace 12M".

9.4.1 Non-occluded vs Occluded

The major evaluation of this competition relies on a face verification setting where the reference is not occluded and the probe is occluded. To understand the implication of small and single occlusions, it is first necessary to understand how the solutions perform on the non-occluded dataset (Protocol 0). Table 9.4 presents the results for the first three protocols (0,1,2). It is interesting to see that four of the six solutions outperform the Baseline (not UB). Moreover, SMT-OCFR1 achieves better FMR100 results than the Baseline UB, which is not the case with EER and FMR1000. It is not surprising that SMT-OCFR1 and SMT-OCFR2 are the only two solutions to achieve comparable results to the Baseline UB since they are the only solutions to align the input images. Surprisingly, in Protocol 1, a drop in the alignment success rate - measured from the alignment success flag output of the solutions - of just 3.1 percentage points together with simple occlu-

sions is enough to create a significant performance gap between SMT-OCFR1 and the Baseline UB. Nonetheless, the number of solutions that remain superior to the Baseline (not UB) does not change. AdaFace4M and AdaFace12M have shown impressive robustness to occlusions, which is visible in the FMR100 value being at least 2.4x smaller. The performance degrades rather quickly when we evaluate in a protocol composed exclusively of a combination of two occlusions as seen in the performance of Protocol 2. While Baseline UB shows that having knowledge of the landmarks strongly benefits the performance, SMT-OCFR1 and SMT-OCFR2 show that failing to align the images might be more harmful to the performance than not using any alignment at all like AdaFace4M and AdaFace12M. These latter models show impressive results that are significantly better than the Baseline (without alignment), although they use alignment.

The performance on Protocol 2 is mitigated when there is no constraint to only use two joint occlusions. For instance, Protocol 3, shown in Table 9.5 shows a reduction in all the errors for all the methods in comparison to Protocol 2 in Table 9.4. Nonetheless, the landscape remains intact as only AdaFace4M and AdaFace12M surpass the Baseline, but not the Baseline UB. If we create a restriction that Protocol 3 cannot have simultaneous upper (U) and lower (L) face occlusions, which hide the majority of the face, then we obtain Protocol 4 also seen in Table 9.5. The results are still far from the ones achieved with single occlusions in Protocol 1, which shows that despite being a factor for a large portion of the errors, the combination of U+L occlusions is not a single-handedly responsible for the performance degradation, since E+L, T+L and E+T consist of a significant source of error to the models. Protocol 5 is virtually the most difficult since it only allows triple occlusions. The gap between the Baseline, SMT-OCFR1, and SMT-OCFR2 increases as these submitted solutions fall short to have an FMR100 smaller than 90%, which might be caused by attempts to align the input images with a success rate of 61.2%. On the other hand, the performance of AdaFace4M and AdaFace12M, which work with unaligned samples shines and keeps surpassing the Baseline.

Table 9.6 shows the results on Protocols 6 and 7, which are relaxations of Protocol 5 to include double and single occlusions, respectively. The overall performance remains similar to previous results. It is also possible to observe the number of parameters and GFLOPs of the models. Moreover, while it is not stated explicitly, SMT-OCFR1 and SMT-OCFR2 require two extraordinary steps of preprocessing, which also increase their computational complexity.

Table 9.7: Final ranking of the solutions submitted. Solutions in bold surpassed the performance of the baseline (without access to the landmarks) in at least one protocol.

Method	Avg. rank (7 protocols)	Rank
AdaFace12M	1.14	1
AdaFace4M	1.85	2
SMT-OCFR1	3.29	3
SMT-OCFR2	3.71	4
AFOIRNet_1	5.00	5
AFOIRNet_2	6.00	6

The solutions are ranked accordingly to the average of performance on the seven protocols. The final ranking is shown in Table 9.7, as expected from the individual analysis of each protocol AdaFace12M takes the lead closely followed by AdaFace4M. This highlights the lower performance of algorithms that strongly rely on alignment for occluded FR. In a sense, this result creates new grounds for further research on which situations are beneficial to discard the alignment step.

9.4.2 Occluded vs Occluded

We further extended our analysis to occluded reference versus occluded probe recognition. The main goal of this analysis is to understand how these algorithms behave when exposed to stronger occlusions on both sides of the comparison. Figure 9.3 shows the performance of the algorithm in Protocol 1 (Figure 9.3a) and Protocol 5 (Figure 9.3b), which are the less and the more occluded, respectively. It is also interesting to note that the SMT-OCFR alignment success rate decreases to 22.2% in Protocol 5 when both the probe and the reference are occluded. And while in Protocol 1, the best performing algorithms retain their advantage, the strong occlusions present in Protocol 5 redefine the order of the best performing solutions. All the solutions achieve results above the 90% FMR100, which is considered inadequate. Nonetheless, it is worth noting that this is a demanding scenario, and a significant portion of the pairs might not have sufficient information available to support a valid verification.

9.5 Conclusion

Holding IJCB-OCFR-2022 allowed us to develop a novel variation of LFW with synthetic occlusions that benefit from domain information regarding their position in the face. In addition, this technique has been adapted for different datasets allowing a deeper study on the different levels of robustness across demographic groups. In this particular competition, we created seven different evaluation protocols with progressive levels of difficulty. To this, we provided two baseline solutions, one which represents an approximation to an upper bound, and a second which represents a true baseline.

Three different teams of researchers explored their interest in the topic by submitting six different solutions. We carefully evaluated each solution in our dataset and performed a comparative analysis between the different strategies, the adoption of alignment and the success of the alignment. Our contributions can be further extended to create training datasets with robust occlusions, as demonstrated by one of the top solutions in the Fracsyn challenge at cvpr 2024.

Chapter 10

On the Domain Gap Between Synthetic and Real Datasets

Author Contributions The research work described in this chapter was conducted in collaboration with Rafael M. Mamede, Tiago Gonçalves and Carolina Albuquerque, under the supervision of Ana F. Sequeira. The author of this thesis contributed to this work on the problem conceptualization, data curation, the preparation and conduction of experiments, the results discussion, and publication writing. Some parts of the chapter were originally published in, or adapted from:

- Neto, P. C., Mamede, R. M., Albuquerque, C., Gonçalves, T., & Sequeira, A. F. (2024). Massively Annotated Datasets for Assessment of Synthetic and Real Data in Face Recognition. *In 18th IEEE International Conference on Automatic Face and Gesture Recognition 2024*

10.1 Context and Motivation

Complex Face Recognition systems have matched and surpassed human-level performance [472]. Recent advances led to deep learning-based neural networks that can learn to distinguish between the identities from a single image. The resourcefulness of such models has led to a continuous focus on improving the best-performing models. Over the years, this improvement was supported by three strong pillars. 1) Exponential increase in computing power; 2) Novel architectures and more expressive deep learning models; 3) Very large datasets.

As mentioned, one of the approaches to further enhance these models relied on the collection and curation of large datasets [558]. These datasets vary in the number of identities, from 10k to 672k, and in the number of images, from 500k to 17M [20, 358, 534]. However, the collection of these datasets has raised privacy and ethical concerns regarding the consent of the individuals present in the data [53]. This led to the retraction of several of these previously publicly available datasets [71]. Moreover, a dataset that is composed of real images with proper curation and consent is not static, since according to the European Union (EU) General Data Protection Regulation (GDPR) [150], consent can be removed at any time. Additionally, individual anonymization of the sample faces is not feasible and further removes the utility of the face to train a face recognition

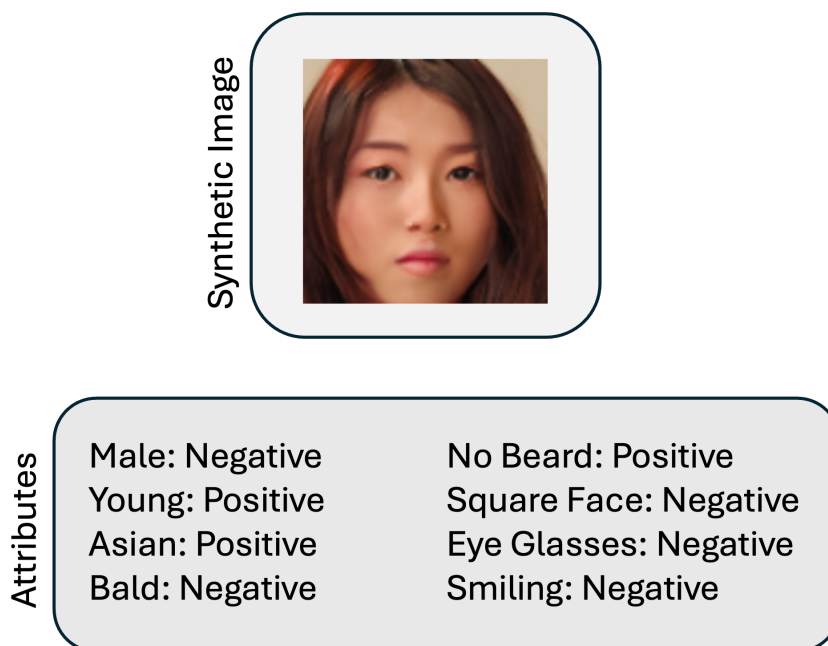


Figure 10.1: An example of an annotated synthetic image. It is possible to observe some of the well defined attributes.

system [333]. This poses a problem for current face recognition research, which requires the use of large-scale datasets that are being limited and removed from public access.

Recently, there has been significant growth in generative artificial intelligence approaches, leading to state-of-the-art methods that can synthesize images that closely resemble real images [12, 270]. Since the initial generative adversarial neural network (GAN) [191] and their improved versions [262], there have been several advances that led to the development of diffusion models. These models are easier to train and lead to better-quality images. Recently, some generative models have been proposed for face data, allowing researchers to condition the identity or other attributes [53, 54]. Following the improvements in generative artificial intelligence, researchers have redirected their efforts into how to synthesize new datasets for face recognition that could remove the dependency on the previously used real datasets.

Despite efforts to develop synthetic data that faithfully represents real data, the performance of models trained on these new synthetic datasets is yet to achieve a performance similar to models trained on real data [53], even when the data is constructed through different synthesizing approaches [30]. These former models seem to perform considerably worse on certain ethnicities and other variations of the traditional face verification setting, such as cross-pose or cross-age. This behaviour can be sustained by several factors, and Huber *et al.* [224] has already explored the diversity of synthetic datasets with regards to gender, ethnicity, age and head position. We firmly believe that the diversity of face datasets can be further described by other attributes. The work of Terhörst *et al.* [476] aims to create datasets that are annotated for 47 distinct attributes, which can be leveraged to highlight differences between synthetic and real datasets. Previous research has noted a potential domain gap between real and synthetic data [292, 436, 520].

As seen in Figure 10.1, the performance of an FR system trained on synthetic samples might be restricted by the fact that those samples do not capture the complete variation and full spectrum of real samples. In this Chapter, we aim to understand how closely the synthetic data mimics the distribution of the real data. For this, we have leveraged two real datasets BUPT-BalancedFace and BUPT-GlobalFace [504] in addition to two synthetic datasets generated, one generated through diffusion (IDiff-Face [53]) and the other with a GAN [49]. Using Terhörst *et al.* [476] Massive Attribute Classifier (MAC) method, we have computed annotations for all the samples in the four datasets. Following this, we have conducted several studies on the distribution of these annotations in order to extract information regarding the diversity of each dataset.

This Chapter is organized into four main sections. The first of these, presents related work regarding automatic annotation strategies and synthetic data generation. Afterwards, in the Methods section, the fundamental details of the experiment's design and setup are discussed in detail. This latter section is followed by a Results section, which aims to present the main findings. Finally, we conclude with a discussion of future work and a summary of the most important elements of this research. The contributions of this work are the following:

- Created annotations, which will be publicly available, for two real datasets. One of the datasets is balanced for ethnicity, whereas the other follows the world ethnicity distribution;
- Replicated the annotation process on two synthetic datasets, enabling future research on the soft-biometrics of these datasets using the released annotations;
- Performed a statistical analysis of the diversity of these datasets through the study of their annotations.

10.2 Related Work

Face recognition from synthetic data has gained popularity in recent years. It presented some challenges to researchers as it was not possible to generate several samples from the same identity, nor generate sufficiently realistic samples. Over the years, these problems have been mitigated [117]. In 2022, Boutros *et al.* [55] proposed SFace, a generative adversarial network-based approach to create new samples, and in 2024 proposed SFace2 [54] achieving state-of-the-art results compared to other GAN-based methods [277, 412]. To mitigate the existence of a single image per identity, some works have also explored unsupervised approaches to train face recognition systems [56]. IDiff-Face [53] comprised a novel diffusion-based technique that conditions the model on a desired identity, leading to a dataset that achieves, in some datasets, a performance comparable to a model trained on real datasets. Kim *et al.* [270] leveraged the conditioning of diffusion models to generate samples from a specific identity and with a specific style. For example, it is possible to generate an image of a certain person using glasses.

Soft-biometric annotations for face images provide contextual information not dependent on a specific identity (such as gender, age, or ethnicity) and are essential for exploring the variability of the data. Low variability of some characteristics can make models trained on that dataset less

robust to applications on real-world data, where inference on examples with such characteristics is necessary. With this information, it is also possible to explore, disclose, and correct demographic biases, addressing fairness concerns. The process of annotating manually is labor-intensive, which can be unfeasible for large datasets. Some works have proposed classification-based estimation of soft-biometric characteristics which can aid the annotation of large datasets: Karkkainen *et al.* [261] (ethnicity, gender, age), Gonzalez-Sosa *et al.* [189] (gender, age, craniofacial features, skin color, subjective annotation), Mirabet *et al.* [339] (hidden face attributes), Merler *et al.* [336] (gender, age, glasses, beard, and mustache), Terh r st *et al.* [476] (47 attribute annotations covering gender, ethnicity, age, accessories, facial hair, hairstyles, subjective annotation, and other facial characteristics). Following the completeness of the 47 attributes generated by the model proposed by Terh r st *et al.*, it presents itself as the ideal approach to studying the broad diversity of synthetic datasets. A similar annotation approach has been followed for DeepFake datasets, and through the analysis of such annotations it was possible to understand and detect certain biases on DeepFake detection systems [523]. This further highlights the importance of having this information available.

10.3 Methodology

10.3.1 Datasets

In this section we present the four datasets used in our experiments. We provide details on their composition and, in the case of synthetic datasets, the generative approach.

10.3.1.1 Real datasets

BUPT-Balanced and BUPT-GlobalFace have been proposed by Wang *et al.* [504] and were intended to create a framework to study the biases of face recognition models. Each identity on the dataset has been labeled according to its skin tone into one of the following ethnicities: African, Asian, Caucasian and Indian. BUPT-Balanced balances the number of identities that belong to each of these four categories and is composed of 1.3 million images with 28k identities, which means that there are 7k identities per ethnicity. On the other hand, BUPT-Globalface contains two million images from 38k identities, and the ethnicity distribution of the identities follows the same distribution seen in the world’s population.

Comprising two distinct ethnicity distributions these datasets allow for a comparison of the Real vs. Real to be used as a baseline of the differences that can be expected from datasets that are known to have different distributions on certain attributes.

10.3.1.2 Synthetic datasets

For the study of synthetic data, we have selected two fundamentally different datasets. The first dataset, referred to throughout the rest of this Chapter as Syn-GAN, was introduced as a tool to be used in quantization scenarios [49]. It is comprised of 500k images that have no information

regarding their identity. They were generated with a generative adversarial network [191]. Using noise sampled from a Gaussian distribution, a pre-trained generator¹ of StyleGAN2-AD is used to create novel face samples. As shown in Chapter 11 [363] the quantization of deep neural networks with this synthetic dataset has an impact on the bias of the final face recognition system. The higher robustness shown highlights a possibility that this data comprises samples that slightly deviate from the real data distribution.

The second dataset exploited diffusion models to generate identity-conditioned samples. IDiff-Face [53] was used to create a dataset called CPD-25 (Two-Stage), which comprises 10k identities and 50 images for each identity. This dataset is significantly more realistic and has been shown to have a performance that reduces the gap between models trained on real and synthetic datasets. Besides the identity, no other attribute is conditioned.

10.3.1.3 Annotations

None of the datasets includes annotations beyond identity, with an exception for the skin tone-based labels on the two real datasets. Hence, studies on the diversity of these datasets from the point of view of soft-biometrics was limited. Additionally, fairness assessments were not trivial. Knowing the impact that these annotations might have on future research, we released, for each image in each dataset, 45 different attributes². For ethical reasons, we have decided to exclude annotations for the fields “Chubby” and “Attractive” seen in the original paper. In total we provide roughly 189M annotations, which is slightly larger than the annotations given by Terh rst *et al.* [476] at 124M.

10.3.2 Experimental Design

10.3.2.1 Annotation process

The generation of annotations for each of the aforementioned datasets used the methodology proposed by Terh rst *et al.* [475, 476]. Given an image x aligned with MTCNN [546], we map the image to face template space given by FaceNet [439], which is further mapped to 47 different soft-biometric attributes via the Massive Attribute Classifier (MAC), a multi-objective NN-based classifier. Each attribute is considered to be an individual classification task, and the majority of the attributes can be “Positive”, “Negative” or “Undefined”.

The proposed implementation strategy leveraged the idea of reliability as a metric for the confidence of a prediction of each attribute [478]. Here, we consider an ensemble of m MAC classifiers where dropout is applied individually with a given probability ($p_{drop} = 0.5$) during test. This results in m estimators with slightly different architectures due to the zeroed connections. This dropout process mimics the behaviour of a Gaussian process and the final prediction for each of the A attributes is computed through a majority vote of the classifiers in the ensemble. Considering each element of the ensemble as a classifier $f_i : \mathbb{R}^N \rightarrow [0, 1]^A$, where $i \in \{1, 2, \dots, m\}$,

¹<https://github.com/NVlabs/stylegan2-ada>

²<https://github.com/NetoPedro/Synth-MAAD-Face>

the reliability of the prediction for the attribute a , $rel(f(x)^{(a)}) = rel(y^{(a)})$, for $a \in \{1, 2, \dots, A\}$, is given by:

$$rel(y^{(a)}) = \frac{(1 - \alpha)}{m} \sum_{i=1}^m y_i^{(a)} - \frac{\alpha}{m^2} \sum_{i=1}^m \sum_{j=1}^m |y_i^{(a)} - y_j^{(a)}| \quad (10.1)$$

The parameter α balances the impact of the centrality measure versus the impact of the dispersion metric and, for equal impact of these factors, we chose this value to be 0.5. Similarly to the original paper, the number of estimators was set as $m = 100$.

The original implementation of MAC creates a new ensemble of models for each image, and that ensemble is then used to estimate the soft-biometric attributes. Hence, each image is passed, one by one, 100 times through the network. Our implementation creates a batch of images during the inference step. Each image in the batch will be evaluated by the same ensemble of models. However, since the same image is still seen by 100 distinct architectures, we retain the advantages of the Gaussian Process while being able to run experiments hundreds of times faster. This allows the usage of the MAAD methodology to be more easily applicable to label large datasets. We used a labeling batch size of 1024.

10.3.2.2 Comparison of the different datasets

Comparing two datasets at the image level is a non-trivial task. Hence, dataset comparison is easier if done at the level of the latent space of the deep neural network. However, with regards to faces, it might not encapsulate the variety of attributes that might be present in an image, since two individuals with a very similar set of attributes will have some distance between them. Additionally, it is not direct to understand the “differences” in this space. However, in the MAAD attribute space, not only comparing different samples with clarity is easier, but individuals with similar characteristics share the space.

Taking into consideration the advantages of this attribute-space, we have devised several strategies to measure the discrepancy between real and synthetic data. One of the first approaches was to measure the relative frequency of “Undefined” predictions on each of the four datasets, which indicated very similar results on average. Afterwards, we focused on the comparison of individual attributes and how much different was the prevalence of positives and negatives for datasets of different sources. This brought some interesting perspectives on the distribution of each attribute.

Additionally, and considering that learned models might be useful to detect the distinct patterns of data sources, we attempted to measure the relative classifiability of “Real vs. Synthetic”. This was done with two strategies: using a classifier; creating two clusters with K-Means [320] and validating how many samples of each source fall within each of the learned clusters.

Considering that the individual analysis of the distribution of a single dataset is a poor metric, we propose to model the prediction of each dataset with a Kernel Density Estimation (KDE) [114, 389] approach on the attribute space. This information takes into consideration the several configurations that each individual might take.

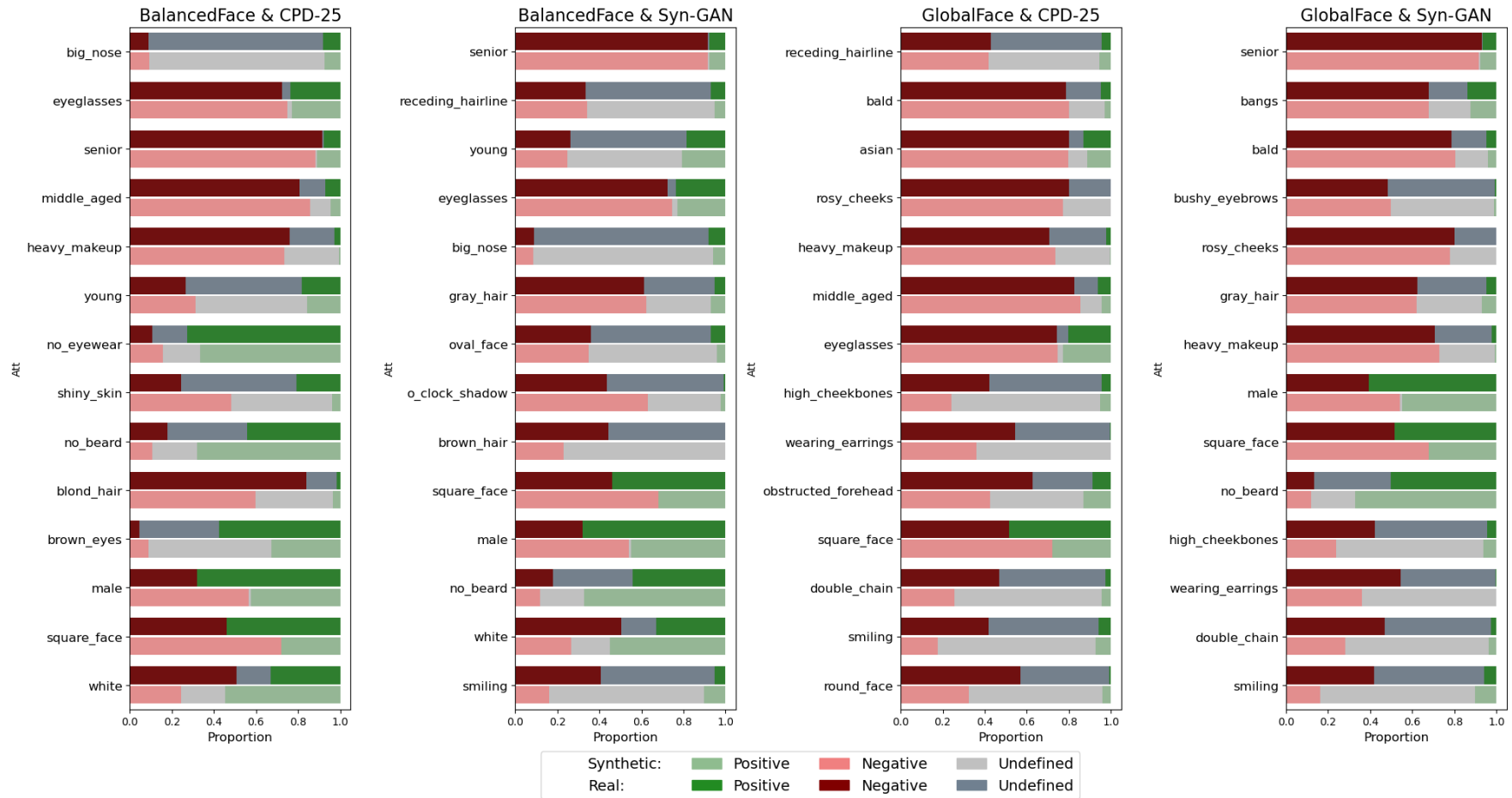


Figure 10.2: Comparison of the real and the synthetic datasets on individual attribute distribution. From the left to the right we have a comparison between BalancedFace and CPD-25, BalancedFace and Syn-GAN, GlobalFace and CPD-25, and GlobalFace and Syn-GAN. The first seven entries of each plot represent the most similar attribute distributions, whereas the seven bottom attributes represent the less similar distributions.

Before computing the distribution we take the mode of all the attribute-sets of an identity. Finally, having learnt the distribution of each dataset, we can compute the Kullback–Leibler (KL) divergence (Eq. 10.2) [289], on both sides, between real and synthetic datasets.

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \quad (10.2)$$

This distance function has the particularity of not being a metric, since it does not respect the symmetry property. Hence results differ if we swap P and Q . A possible interpretation of this distance is the quantification of the information lost when using Q to approximate the distribution P . This is particularly useful in our scenario as it can tell us the information lost when we approximate the distribution, at the attribute space, of real data using synthetic data. Ideally, this distance should be close to zero as the information contained in one distribution would be reflected in the other distribution.

10.3.3 Experimental Setup

All the annotation experiments were conducted in a GPU cluster, leveraging a NVIDIA A100 GPU with 80GB of VRAM. The batch size for the inference was set at 1024, and we have conducted several tests to find the optimal batch size for our configuration. Additionally, we have measured the impact of having different batch sizes on the predictions and did not find statistically significant differences between different runs. We separated the face template extraction and the attribute computation steps so that it could be possible to create batches on the latter stage without affecting the approximation to a Gaussian Process.

The remaining experiments were conducted in a consumer grade laptop without GPU. Annotations were saved for later use and release.

10.4 Results

10.4.0.1 A study on the attributes

Following the previously described methodology, we aimed to understand how the different datasets behaved with respect to their attributes. In the initial stage, we aimed to understand the attributes of each dataset individually, hence, for the different combinations of real datasets with synthetic datasets we calculated the seven most similar attribute distributions and the seven most non-similar.

Looking at Figure 10.2 it is possible to observe the four different combinations. For each plot, the top seven attributes are the ones that are more in line when the two datasets are compared, whereas the bottom seven are the most dissimilar. Some attributes are frequently displayed in the dissimilar set: Square Face, Male, No Beard and Smiling. In the synthetic data, samples are generally more skewed towards being female, whereas in real data the opposite happens. Probably related to this prevalence, there are significantly fewer samples with beards on synthetic datasets.

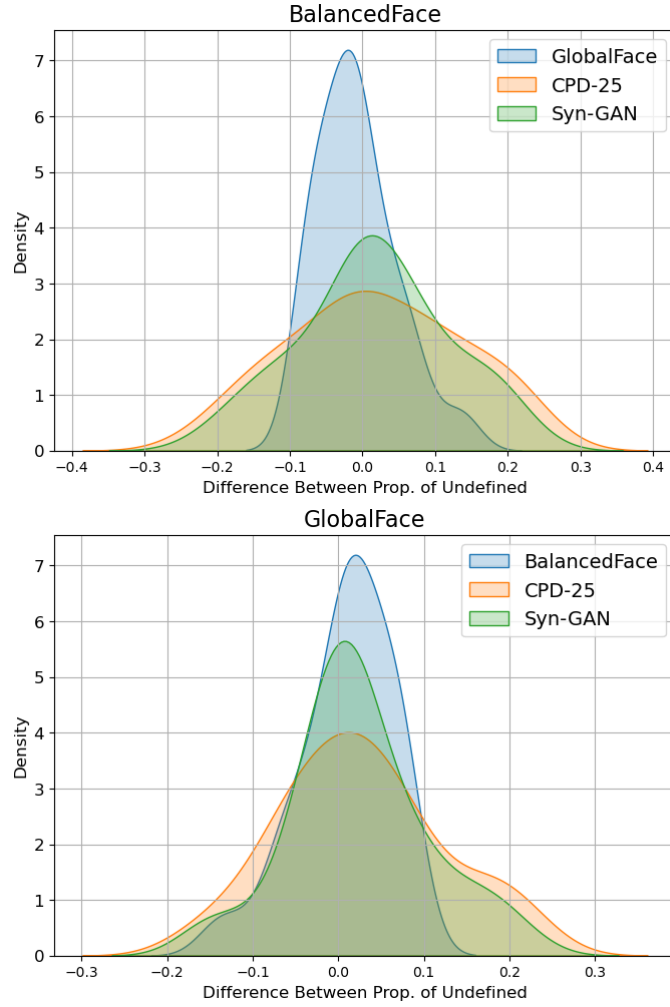


Figure 10.3: Distributions $Prop0_{d1}^{(a)} - Prop0_{d2}^{(a)}$, where $Prop0_{d'}^{(a)}$ represents the proportion of the Undefined prediction of MAC for attribute a' in dataset d' . The title of each plot represents the $d2$ set, and the label of each KDE plot represents the $d1$ set.

Although it seems a minor issue, the beard represents one of the most natural and common forms of face occlusions. One other relevant aspect is the inability to properly detect smiles on synthetic data, as the majority of the annotations are “Undefined”. This might impact the variability of the samples as the emotions/face reactions are one of the elements that most affect the perception of a face. One additional relevant element is the presence of White as one of the most dissimilar attribute distributions when the real dataset is BalancedFace. While this dataset focuses on balancing the different skin tones, synthetic datasets contain mostly white individuals.

On the other hand, a few attributes are consistently displayed as ones with the most similar distribution. Age-related attributes such as Senior, Young and Middle-aged are considerably similar in both types of datasets. Although this seems to be a good indicator, we have not measured if the age distribution within each identity is similar. For instance, we have 100 images uniformly spread between 18 and 70 years old, but having only one age group within each identity. Or we

could have images of different ages within each identity. It is also visible that on synthetic datasets it is extremely rare to have a sample face wearing heavy makeup.

10.4.0.2 Dispersion of Undefined Samples

Another relevant question we can raise when discussing the labeling of the synthetic data with MAC is whether we get an abnormal amount of Undefined predictions. Abnormal behavior of this outcome could indicate underlying problems of the labeling method, and its adequacy when applied to our synthetic datasets. If we noticed that in general there is a higher amount of this label across all attributes in the synthetic datasets, we could infer that the MAC is having trouble defining meaningful predictions for each class, either because the synthetic data has a poor representation of that attribute, or it is not represented.

To verify underlying trends in the Undefined prediction, we analyse the differences between the proportions of this label, given two datasets, for each attribute. We take notice of changes in the behavior of these distributions when analysing Real-Real or Synthetic-Real differences. The results, Figure 10.3, show no significant difference in the mean of these differences, meaning we do not observe generalised trends for Undefined labeling across all attributes. What we can verify is the difference in the dispersion of these differences, meaning that some attributes have more extreme fluctuations. We also noted that the attributes that have the highest decrease in the proportion of Undefined when compared to a Real dataset were consistently attributes related to facial hair (such as 'o_clock_shadow', 'no_beard', 'sideburns', 'goatee'). This is consistent with a balancing of the gender attribute in the synthetic dataset, since prediction for these attributes tends to be easier in the 'female' class. On the other hand attributes that consistently have an increase in unpredictability include 'smiling' and accessory use such as 'wearing lipstick' and 'wearing earrings'. Emotions and their expressions might be particularly difficult to model with a generative system, as well as artifacts such as accessories. It might be the case that since these share no link to the identity information, they might trigger the generative model to ignore or smooth them.

10.4.0.3 Clustering Synthetic vs. Real

In addition to the comparison of the different attributes with respect to their real and synthetic distribution, we have attempted to fit a K-means ($k = 3$) model to three out of the four datasets. Figure 10.4 shows the distribution of the samples of each dataset within one of the three clusters. As expected, BalancedFace and GlobalFace share very similar distributions across clusters. Surprisingly, the presence of CPD-25 in the third cluster is rather small, and the dataset is heavily inserted in the second cluster. This evidence already highlights some potential differences between the attribute space of synthetic and real samples. Especially if we consider that even on the most prevalent cluster for synthetic data, real data is significantly present, whereas the opposite is not verifiable.

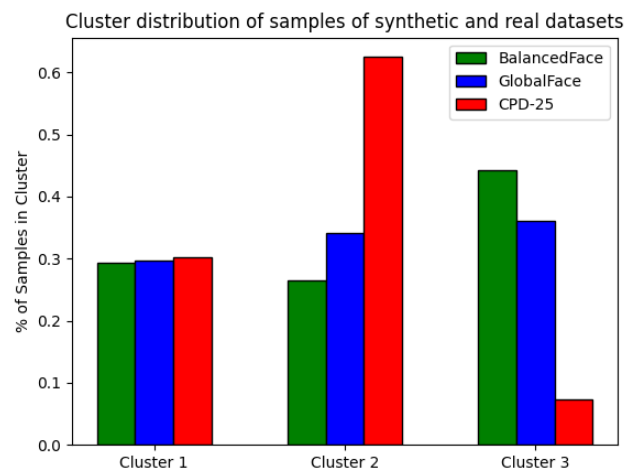


Figure 10.4: Cluster aggregation of the different identities on each of three datasets (BalancedFace, GlobalFace and CPD-25). Clusters calculated with K-Means. Syn-GAN was removed from the comparison, since it consists of 500k images of distinct identities. Synthetic data has higher presence in cluster 2 ($> 60\%$) and lower in cluster 3 ($< 10\%$).

Table 10.1: KL divergence between all the distribution learnt by KDE for all four datasets. Interesting to denote the lower information loss when real data is used to approximate synthetic data in comparison with the higher information loss when we swap P and Q.

P \ Q	GlobalFace	BalancedFace	Syn-GAN	CPD-25
GlobalFace	-	0.708	3.007	2.223
BalancedFace	0.409	-	1.984	1.223
Syn-GAN	1.010	0.613	-	0.249
CPD-25	0.507	0.247	0.371	-

10.4.0.4 Synthetic vs. Real Divergence

Following the potential difference in the distribution of the attribute set of real and synthetic datasets, we have used a KDE model to estimate and approximate the distribution for each one of the four datasets. Afterwards, as seen in Table 10.1, we have measured the Kullback–Leibler divergence for all combinations of datasets. When P is set to the distribution of GlobalFace, we noticed that both Syn-GAN and CPD-25 as Q lead to significant information loss. On the other hand, BalancedFace as Q allows for diminished information loss. Setting BalancedFace as P leads to very similar results with a synthetic Q, but considerably better than the previous comparison. On the other hand, GlobalFace is quite accurate at approximating the other real datasets. Both synthetic datasets can be easily explained by the other datasets, leading to very small values of information loss. This discrepancy in the KL values highlights the lack of diversity shown in synthetic datasets for face recognition and how they must improve to replace real data.

10.5 Conclusion

Throughout this Chapter we have covered several strategies to uncover the reason behind the inferior performance of models trained on synthetic data when compared to models trained on real data. Acknowledging the difficulties of comparing these datasets at image level, we have proposed to use MAC to create massive annotations for each of four datasets. This process allowed for a few studies on the diversity and gap between real and synthetic datasets.

Leveraging the annotations, it was possible to immediately inspect some attributes and their difference in distribution across all datasets. It was also possible to measure the undefined dispersion on synthetic datasets and uncovering that attributes such as smiling are difficult to measure on synthetic data, as shown by the quantity of undefined samples for this attribute.

Considering the attribute set as a whole combination of attributes, it was possible to place the samples of each dataset on one of two cluster and extract hints regarding the lower variability of synthetic data. Additionally, after modelling this distributions, we manage to use the Kullback–Leibler divergence to measure the information difference between the four datasets. As expected, synthetic datasets shown a poor capability to approximate real data.

In summary, we have not yet find a clear answer to the reason behind the performance difference of models trained on these datasets. Yet, we have made contributions on the gaps between both types of datasets and we have released the annotations. Future researchers can leverage these annotations to further condition diffusion models, to find correlations between a set of attributes and the performance, or to build better automatic annotation tools. There are several directions in which this research might lead. In further Chapters we extend the discussion for the relevance of using some of these attributes in a discrete or continuous space.

Part IV

Biases in Face Recognition

Chapter 11

Inspecting the Biases Through Model Compression

Author Contributions The research work described in this chapter was conducted in collaboration with Eduarda Caldeira under the supervision of Ana F. Sequeira and Jaime S. Cardoso. The author of this thesis contributed to this work on the problem conceptualisation, data curation, the preparation and conduction of experiments, the results discussion, and publication writing. Some parts of the chapter were originally published in, or adapted from:

- Neto, P. C., Caldeira, E., Cardoso, J. S., & Sequeira, A. F. (2023, September). Compressed models decompress race biases: What quantized models forget for fair face recognition. *In 2023 International Conference of the Biometrics Special Interest Group (BIOSIG) (pp. 1-5). IEEE.*
- Neto, P. C., Caldeira, E., Cardoso, J.S., Sequeira, A.F. (2023). Compressed Models Decompress Race Biases on Face Recognition. *In 29th Portuguese Conference on Pattern Recognition (RECPAD 2023), Portugal.*

11.1 Context and Motivation

Face recognition methods have made significant progress over the previous years [46]. Current systems are capable of rivaling with humans under certain conditions and are quickly reducing the gap on the remaining test scenarios [400]. The urge to keep the current rate of improvement on these deep learning-based approaches led to an era of complex and obscure models. As such, despite their extraordinary performance, there are two pressing concerns. First, there are hardware limitations that affect the complexity of the models that can be deployed and used in real scenarios. These limitations affect both storage, memory and processing time. The second concern is that the behavior of a deep neural network is not easily understood [367]. As such, besides the valuable information, also irrelevant or even harmful correlations can be learned by these models, and hidden within their obscure nature.

Addressing these two concerns is of utter importance. To tackle them individually, one must be careful to avoid a potential trade-off between their mitigation. For instance, considering the

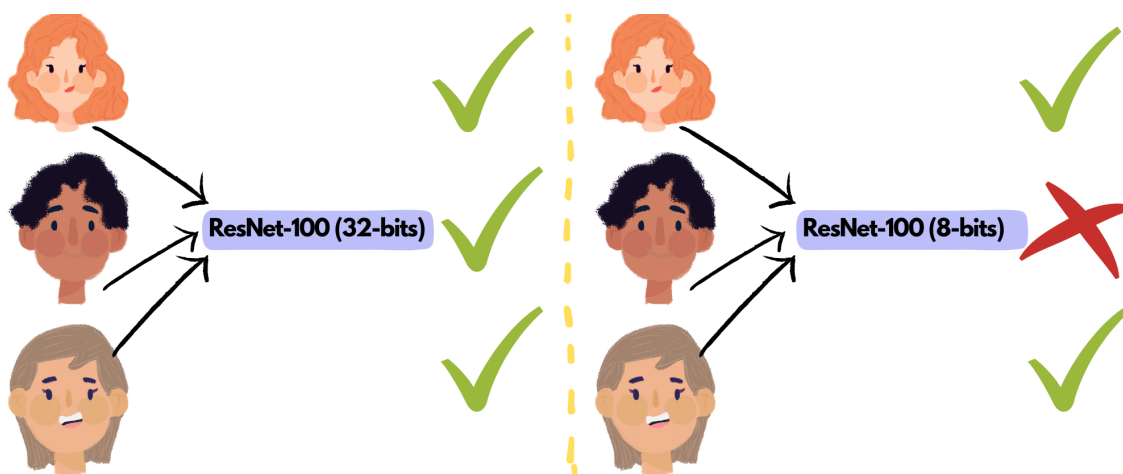


Figure 11.1: Representation on a potential source of biases created after quantizing a deep neural network. Less represented classes or classes prone to suffer from discrimination might be easier to forget, leading to errors focused on those classes.

possibility of an existing bias on the models, further reducing the model size can lead to an increased bias. Moreover, unless that the model is reduced to an interpretable version of itself, these growing biases remain hidden within the black-box model.

Instead of using a smaller model, current work is investigating different model compression approaches. Quantization, knowledge distillation and pruning are the most common [68]. In this work, we aim to study the impact of quantization on the mitigation or amplification of existing biases. Quantization focuses on re-defining how the weights of a neural network are represented. For instance, usually, these weights are represented with 32-bits, also known as full-precision. With quantization we aim to represent each weight with N -bits, where $N < 32$. It can also comprise a fine-tuning step on the new weight representation.

Hooker *et al.* [214] presented a set of experiments that indicates a potential increase of the previous biases and tried to identify the elements forgotten by the deep neural network [213]. To further extend this research, we framed our problem within the context of racial biases in face recognition systems. Stoychev *et al.* [467] presented mixed results on a face-related task, and the effects on the biases were dependent on the training dataset. For this reason, our work, starting from Boutros *et al.* [49] quantization approach, further includes the usage of real and synthetic data and the usage of distinct datasets for training. This study also aims to understand the current trade-offs between small models and hidden biases.

Within the context of this Chapter, we aim to answer three research questions: 1) *Are smaller models more biased?* 2) *Are quantized models more biased? (As represented in Figure 11.1)* 3) *What is the impact of using synthetic data to quantize these models?* We individually address each of these questions and provide a potential explanation for the behavior displayed by the models. Furthermore, the usage of synthetic data is motivated by the possibility of further growing our current datasets without compromising ethical concerns or privacy. Besides, with synthetic data, it is possible to create a higher degree of variability. Given these research questions, we present

the following contributions:

- A study on racial bias on four differently sized models trained on MS1MV2 [126] and of two models trained on BUPT-Balancedface and BUPT-Globalface [504];
- Using QuantFace [49] and real data we study the racial bias of the quantized version of all the models previously evaluated;
- We repeat the previous analysis with synthetic data;
- While confirming some of the Hooker *et al.* [214] findings, we have also discovered that quantization with synthetic data mitigates the racial bias of the resulting model.

While we conducted our studies on a face recognition task, we theorize that the results seen are plausible for similar verification or classification problems, and other biometric modalities. In other words, smaller models and quantization might impact the class with higher risk of discrimination in the same way that we describe in this study. Further analysis on different compression techniques have been studied by the author, but are out of the scope of this Chapter [66, 364].

The following sections are divided into four major topics and a conclusion. Section 11.2 describes the five distinct datasets utilized in this study. Afterwards, in Section 11.3 and 11.4 the methodology and experimental setup are described in detail. Finally, the results are shown and discussed in Section 11.5.

11.2 Datasets

In this study we leveraged five datasets, two for training and quantization, one just for quantization and one for testing. The first three were MS1MV2 [126], BUPT-Balancedface [504] and BUPT-Globalface [504]. And the other two were a synthetic dataset [49] and Racial Faces in-the-wild (RFW) [503], respectively.

11.2.1 MS1MV2

MS1MV2 [126] is widely used in the literature to train and explore several deep face recognition models [46, 360]. It is a refined version of the original MS-Celeb-1M dataset [198], which further improved the training of these systems. The dataset contains 85k different identities and almost six million images and it is not balanced with respect to the race.

11.2.2 BUPT-Balancedface and BUPT-Globalface

Wang *et al.* [504] introduced two distinct datasets to train deep face recognition systems. These datasets have been created to study and mitigate race bias on face recognition through skin tone labeling as African, Asian, Caucasian and Indian. BUPT-Globalface contains two million images from 38k different identities, and the distribution of identities per race follows their distribution in the world. On the other hand, BUPT-Balancedface contains 1.3 million images from 28k identities

which are divided into 7k identities per race. As such, this second dataset is race balanced. The notion of what is a balanced dataset will be further questioned and challenged in this document.

11.2.3 Synthetic data

This dataset, introduced in [49] contains approximately 500k unlabeled synthetic images. These images have been generated by a generative adversarial network [191, 264]. The noise used as input to generate the images was sampled from a Gaussian distribution and fed to a pre-trained generator (official open source implementation ¹ of StyleGAN2-AD). The use of synthetic data is often seen to result in sub-optimal performances [412] which might be caused by a domain gap between real and synthetic data [292, 436, 520]. In this work, the goal is not to use the synthetic data to learn the representations from scratch, and we further argue that there might exist advantages of this domain gap.

11.2.4 RFW

Proposed by the same authors of BUPT-Balancedface, RFW [503] was designed as a benchmarking dataset for fair face verification. Similarly, it includes labels for ethnicity, which allows for a fair assessment of potential biases. For each ethnicity, it contains 3000 individuals with 6000 image pairs for face verification. Each pair is a composition of two images, reference and probe, that share same ethnicity label.

11.3 Methods

The methodology was designed in two different steps. First, it is necessary to understand if there is a bias problem on quantized models, and for this we have used the publicly available QuantFace models. If we can identify a bias-related problem with quantization, it is necessary to understand if it replicates on models trained on other datasets (balanced and non-balanced).

QuantFace has four different architectures available: MobileFaceNet [86], ResNet-18 [206], ResNet-50 and ResNet-100. Each of these architectures is available in five distinct shapes: the original full-precision model, the 8-bit model quantized with real data, the 8-bit model quantized with synthetic data, the 6-bit model quantized with real data, and finally the 6-bit model quantized with synthetic data. This variability allows us to study and understand if the behavior of the quantized model changes with its precision, the network architecture and the data used in the quantization process. Hence, the training dataset for this is fixed (MS1MV2), so we can ignore the training data as a factor of variability.

For the first step, one must keep in mind that MS1MV2 and RFW have some overlapping samples. Hence, there is some identity leakage for the test data. This means that models trained with MS1MV2 are expected to have better performance than models on the remaining steps, and this might be even more noticeable when the quantization resorts to this dataset too. Despite being

¹<https://github.com/NVlabs/stylegan2-ada>

aware of this leakage, we have decided to still conduct this first step as all different architecture are in the same conditions, and leverage the information of the second step to corroborate the information on datasets without identity leakage.

For the second step of the study, we inspect the impact of the training data on the behavior of the model. For this, we fix the architecture as a ResNet-34. Two of these models have been trained, the first was trained on BUPT-Balancedface and the second was trained on BUPT-Globalface. The first is available in the four different quantized versions described above, whereas the second was only studied in its 8-bit version with real data quantization from two different sources and synthetic quantization. Again, the network architecture is fixed so that the variability factors are limited to the data used for training and quantization. As described, this step explores some options where the model is trained with one real dataset and quantized with another real dataset. This is to infer if performance changes are due to using different quantization and training datasets, or due to the synthetic dataset that is used for quantization in the remaining configurations. If both have comparable changes, one might theorize that different identities on quantization are having an impact, either positive or negative. On the other hand, if they differ one might consider the difference in distribution between real and synthetic samples as the culprit of such change.

As no ethnicity information was available for the synthetic dataset, we used the an ethnicity classifier, described in Chapter 12, to infer these labels. This allows for a study on the distribution of synthetic samples with respect to the race of the different images.

11.4 Experimental Setup

For the quantization process we have utilized the open-source implementation of QuantFace² with a batch size of 128 on a Nvidia Tesla v100 32GB. We have utilized the same configuration as proposed by Boutros *et al.* [49]. For training the face recognition models on BUPT-Balancedface and BUPT-Globalface, we have utilized the same protocol of Deng *et al.* [126] to pre-process the images, reduce the learning rate and stop the training.

11.4.1 Evaluation Metrics

For each ethnicity and each model, we computed its performance as the percentage of correct predictions (accuracy). From this, we can compare different models regarding their overall performance, but we can also compute fairness-related metrics. With this in mind, we explored the standard deviation (STD) between the accuracies on the different ethnic groups for a model, and the skewed error ratio (SER), seen in Equation 11.1, which measures the ratio between the largest and smallest error.

$$SER = \frac{100 - \min(acc)}{100 - \max(acc)} \quad (11.1)$$

²<https://github.com/fdbtrs/QuantFace/>

With STD one evaluates the variance between the different accuracies, where values approaching 0 are ideal, and large values are to be avoided. On the other hand, SER measures or much larger is the worst error when compared with the better error. This is important to understand the relative proportion between the different accuracy values. As a relative evaluation metric, SER is highly sensitive when the accuracy is above 99%. This happens because as the errors get below 1% their relative difference also change accordingly. For instance, a SER computed for a maximum accuracy of 90% and a minimum accuracy of 80% is the same if these accuracy values were 99.9% and 99.8%. STD is highly sensitive to absolute differences, and grows large on sets with lower accuracy values. Although the validity of such metrics can be questioned, these have been widely adopted in the literature [504]. Hence, for the sake of reproducible research and compliance with the literature, we have chosen to retain the metrics previously used.

11.5 Results

Table 11.1: Table comprising the results, evaluated on RFW, from the different models trained on MS1MV2 and their respective quantized versions for different bits and quantization strategies (real or synthetic data). The versions of the models quantized with synthetic data seem to display better fairness metrics at a comparable average performance.

Model	Bits	Quant.	Caucasian	Indian	Asian	African	Avg.	STD	SER
MobileFaceNets	32	-	95.18%	92.00%	89.93%	90.22%	91.83%	2.41	2.09
	8	Real	95.32%	91.60%	89.27%	90.08%	91.57%	2.68	2.29
	8	Synth.	94.18%	91.83%	88.85%	89.72%	91.15%	2.38	1.92
	6	Real	90.05%	86.52%	82.88%	83.18%	85.66%	3.36	1.72
	6	Synth.	89.97%	86.95%	83.13%	84.40%	86.11%	3.02	1.68
ResNet-18	32	-	97.48%	95.38%	93.72%	94.27%	95.21%	1.66	2.49
	8	Real	97.42%	95.33%	93.55%	94.20%	95.13%	1.70	2.50
	8	Synth.	96.95%	95.07%	93.30%	93.87%	94.80%	1.61	2.20
	6	Real	96.93%	94.65%	92.52%	93.22%	94.33%	1.95	2.44
	6	Synth.	96.80%	94.78%	92.35%	93.28%	94.30%	1.94	2.39
ResNet-50	32	-	99.00%	98.15%	97.62%	98.32%	98.27%	0.57	2.38
	8	Real	99.07%	98.07%	97.65%	98.40%	98.30%	0.60	2.53
	8	Synth.	99.02%	97.72%	97.33%	97.88%	97.99%	0.73	2.72
	6	Real	98.32%	96.27%	94.55%	95.87%	96.25%	1.56	3.24
	6	Synth.	97.95%	96.63%	94.97%	96.20%	96.44%	1.23	2.45
ResNet-100	32	-	99.65%	98.88%	98.50%	99.00%	99.01%	0.48	4.29
	8	Real	99.57%	98.87%	98.15%	98.77%	98.84%	0.58	4.30
	8	Synth.	99.37%	98.72%	98.13%	98.78%	98.75%	0.51	2.97
	6	Real	95.27%	93.15%	90.32%	91.70%	92.61%	2.12	2.05
	6	Synth.	95.93%	93.40%	91.92%	92.60%	93.46%	1.75	1.99

11.5.1 Bias on different architectures

A careful analysis of the performance of the different sized models at full precision (Table 11.1) shows that smaller models tend to have higher biases and lower performance in terms of average accuracy. ResNet-100 is an exception and this difference might be connected to the fact that SER becomes highly sensitive when the errors are below 1%. Within the same family of networks, ResNet-18 shows a considerable performance drop when compared to ResNet-50. What is most interesting is that on Asians and Africans the performance drop is significantly larger than on Caucasians. This is, of course, illustrated by and higher SER and STD on ResNet-18.

The quantized version of these models seems to retain the performance and bias advantaged when compared to simpler models. As theorized, the quantization has a negative impact on the bias, and in most cases on the performance too. A higher compression ratio, meaning less bits to represent each weight, leads to higher bias too. Surprisingly, we noticed that fine-tuning the quantized model with synthetic data, regardless of the number of bits of the final model, it generally has lower bias than its counterpart trained with real data. This is rather interesting due to the already discussed overlap between the real dataset in this experiment (MS1MV2) and the test dataset (RFW). However, it is not clear if the improvement is due to a specific characteristic of the synthetic data, or some other factor.

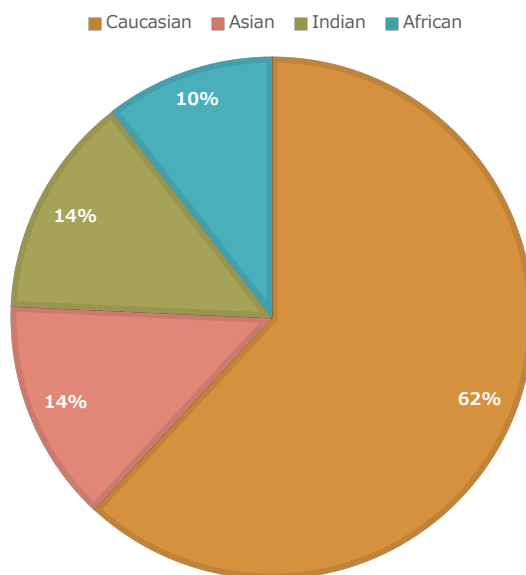


Figure 11.2: The estimated distribution of the different ethnicity groups on the synthetic dataset used for quantization. Caucasians represent around 62% of all samples. Thus, demonstrating the unbalanced nature of these generated images.

This results sparked some interest regarding the nature of the synthetic data and how it distributes, ethnicity-wise. Hence, we leveraged the model trained to classify an image in one of four ethnicity groups. Our estimations, shows in percentage in Figure 11.2, indicate 365889 Caucasians, 81568 Asians, 81568 Indians and 61966 Africans. This highly unbalance nature discards any relation between the bias reduction and the ethnicity distribution of the synthetic data.

11.5.2 Bias on different datasets

Table 11.2: Table comprising the results, evaluated on RFW, from two ResNet-34 models trained on BUPT-Balancedface (BL) and BUPT-Globalface (GL) and their respective quantized versions for different bits and quantization strategies (BL, GL or synthetic data). The versions of the models quantized with synthetic data seem to perform outstandingly well.

Train Data	Bits	Quant.	Caucasian	Indian	Asian	African	Avg.	STD	SER
BL	32	-	96.60%	94.50%	94.03%	93.37%	94.63%	1.40	1.95
	8	BL	94.98%	93.60%	92.77%	90.95%	93.08%	1.68	1.80
	8	Synth.	96.03%	94.40%	93.97%	92.50%	94.23%	1.45	1.89
	6	BL	89.22%	87.87%	86.25%	82.80%	86.54%	2.77	1.60
	6	Synth.	94.58%	92.88%	91.45%	91.13%	92.51%	1.58	1.64
GL	32	-	97.67%	95.52%	94.15%	93.87%	95.30%	1.74	2.63
	8	BL	95.42%	92.75%	91.83%	89.88%	92.47%	2.30	2.21
	8	GL.	94.70%	92.15%	90.23%	88.75%	91.46%	2.57	2.12
	8	Synth.	97.33%	95.15%	94.17%	93.55%	95.05%	1.66	2.42

Having studied the impact of changing the system architecture, and after noticing the impact of synthetic data, we have decided to extend these experiments on two additional real datasets. Furthermore, one of these datasets is race-balanced, and we expected it to improve bias across different experiments. As seen in Table 11.2, training two ResNet-34 on BUPT-Balancedface and BUPT-Globalface shows, at full precision, that despite a higher performance of the latter, the balance of the former is essential to ensure better bias metrics. On the model trained with the BUPT-Balancedface the versions quantized with synthetic data has not only kept the same tendency of the previous experiments, but it has also surpassed, by a large margin, the version of the method quantized with the balanced data.

The ResNet-34 trained on the BUPT-Globalface performs better if quantized with the data from the BUPT-Balancedface instead of using the data from training. Once again, it might be possible that introducing variability and unseen data for the quantization increases the capability of the model to be robust for all ethnicities. This is further validated by the version quantized with the synthetic data, which leads to a performance similar to the full precision model.

After capturing the behavior of quantization across different datasets, architectures and compression strength, one can easily understand that a different dataset for training and fine-tuning yields greater benefits in both average accuracy and bias metrics.

11.6 Conclusion

In this Chapter, we tackled three research questions, and we have provided answers to all of them. In 1) and 2) it was possible to infer that the quantized models and the smaller models are indeed more biased; 3) it was also verifiable that the use of synthetic data for quantization positively impacts the fairness metrics. We have extended previous literature on the assessment

of information loss on quantized models and further introduced a novel topic regarding the use of synthetic data for bias mitigation.

Despite the results shown by our experiments, there are several gaps in the literature that should be tackled in future work. For instance, it is not known whether this behavior is the same for gender biases or if synthetic data harms gender biases while helping to mitigate race biases. A more comprehensive study is required. The usage of a combination of real and synthetic data should be also analyzed to understand how, just by changing the training data, one can mitigate these biases without performance drop. Furthermore, we still do not know if these findings hold for different traits and tasks, and further studies are required to confirm the generalization of these findings.

While the results shown are still preliminary, they introduce a few research directions that might be relevant for the future of biometrics in an era of increasing concern with these biases.

Chapter 12

A Truly Balanced Dataset for Fair Face Recognition

Author Contributions The research work described in this chapter was conducted in collaboration with Naser Damer, under supervision of Ana F. Sequeira and Jaime S. Cardoso. The author of this thesis contributed to this work on the problem conceptualization, data curation, the preparation and conduction of experiments, the results discussion, and publication writing. Some parts of the chapter were originally published in, or adapted from:

- Neto, P. C., Damer, N., Cardoso, J. S. & Sequeira, A. F., (2025). Balancing Beyond Discrete Categories: Continuous Demographic Labels for Fair Face Recognition. Submitted to Transactions on Information Forensics and Security

12.1 Context and Motivation

Face recognition (FR) has grown over the recent years. Deep learning (DL)-based FR systems are currently available for diverse purposes, ranging from border control to mobile authentication devices. While the performance of the FR systems has been advancing over time, it was deep learning that enabled outstanding improvements in the realm of recognition with face biometrics. These improvements and the apparent nearly flawless performance led to the implementation and deployment of face recognition systems in critical domains. Many of these deployments have led to negative consequences due to wrong predictions, lack of explainability and, most importantly, notable biases regarding certain demographic groups. This raised concerns on both researchers and users, and it is a problem that is still occurring [28, 285]. Hence, these flawed systems keep on been negatively mentioned by news outlets¹, often being associated with wrong predictions in court-related tasks² [251].

Deep learning success has been driven by three important elements. First, the improvements in computational power propelled the ability to run computationally expensive methods that were

¹<https://www.scientificamerican.com/article/police-facial-recognition-technology-cant-tell-black-white/>

²<https://capitalnews.org/facial-recognition-wrongful-arrests/>

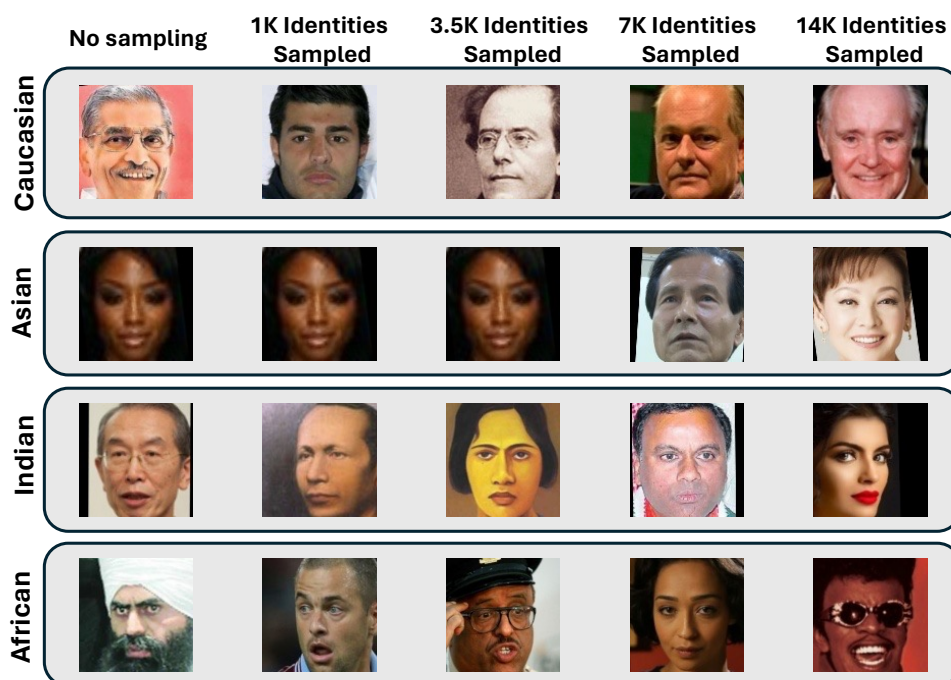


Figure 12.1: **Demonstration of the Sampling Effect:** For each sampling intensity (using sampling strategy A) the sample with the lowest confidence of each label is shown. For instance, on No Sampling, it is possible to see an Asian image (according to the dataset label) that the ethnicity classifier model has a low confidence in the Asian class. It is also noticeable that as the sampling intensifies, the ethnicity traits become more prominent.

previously disregarded due to their computational cost [287]. Additionally, new techniques allowed for an increased depth in neural networks' architectures [205]. Thus, the complexity of DL models has grown significantly. Finally, the ability to scrap large scale datasets from the web increased the quantity of data available to train these deep systems [199, ?]. The first two elements came at the cost of transparency, while the non-rigorous collection of data led to unbalanced datasets that fed the demographic biases found in face recognition models. Together, they create an environment that favors the presence of undetected biases.

The relation of a biased dataset and a biased model is well-studied, and evidence indicates this relationship can also be found in face recognition tasks. In 2021, Wang *et al.* [504] has proposed two face recognition datasets. The first of these datasets follows an ethnicity distribution similar to the real-world distribution, whereas the other balanced the number of identities per each ethnicity. The same authors have also proposed RFW, a balanced dataset for benchmark and bias detection [503]. Similarly, Wu and Bowyer [517] investigated which elements should be balanced on a balanced dataset and proposed a test dataset, sampled from VGGFace2 [71], that is balanced for head-pose, quality and brightness. BFW [427] dataset was also sampled from VGGFace2, and balanced its data on both gender and ethnicity. In general, these datasets focused on balancing the discrete labels that identify each gender/ethnicity or any other attribute. However, in practice, one can pose the question "How Asian is my Asian sample?", as the prevalence of certain ethnicity

traits is not equal in all individuals. Hence, we argue that not all training samples from a certain demographic trait contribute equally to the discriminative power of a model on test samples from that trait.

Having proper representativeness in the dataset is essential to achieve a sufficiently good discriminative function. Thus, based on the previously described balancing strategy, we defined a set of novel strategies that provide grounds for a stronger selection of identities with respect to each ethnicity. These strategies aim to share, across research fields, that discrete labels-based balancing might yield a sub-optimal result with respect to the fairness of the model across demographic groups. With this in mind, we propose a set of steps to balance data with respect to a demographic attribute while using continuous labels. For this, we start by proposing a method based on a state-of-the-art FR system to create these continuous labels. Additionally, we leverage these labels to assess the true balance of the dataset, showcasing the unbalanced nature of the data. Finally, we propose three different strategies to leverage these labels to create a new set of samples with a higher balance. We validate our strategies by removing 3.5%, 12.5%, 25% and 50% of the identities in the training dataset. Figure 12.1 shows the least representative identity of each group at a given point of one of the sampling strategies. We compare the performance of our strategy versus models trained on the original dataset and also on a set constructed by randomly removing the same number of identities that retain the balance of the discrete labels. Our results highlight the ability of the three approaches in the reduction of two different bias metrics. Additionally, in terms of accuracy, models trained on some identity-reduced sets created with our approach surpass or are on par with the models trained on the entire dataset. These results were acquired for multiple backbones and losses.

We can summarize our contributions as follows:

- Through the usage of continuous labels, we analyze discretely balanced datasets, showcasing their unbalance on the continuous space;
- We propose three sampling strategies to remove identities and balance the data in the continuous space;
- We release all the “datasets” created with these strategies and recommend their use;
- We showcase, through experiments with different backbones and losses, that models trained on these smaller but balanced sets can perform on par with the whole dataset while improving fairness metrics.

We organize this Chapter in four main sections. It starts by discussing the current literature on the topics of face recognition and fairness in this context. Afterwards, we present the main contributions with a discussion regarding the ethnicity classifier, the notation, and the formulation of the sampling strategies. Before presenting and discussing the results in the last section, we introduce the experimental setup with all the necessary details to reproduce this work.

12.2 Related work

Face Recognition: Face Recognition tasks are normally designated as Verification and Identification. The former handles 1-1 comparisons where a reference image is given and the objective is to classify a probe image as belonging to the same person, or not. The latter task requires 1-N comparisons to match the input face to an identity within a given gallery. In order to be able to conduct these matching processes, the models utilized require a high discriminative power to mitigate both false positives and false negatives. With the rise of deep learning, contrastive losses, such as Triplet Loss, gained momentum and achieved state-of-the-art results [71, 439]. These works learn to distinguish different faces by ensuring that the distance between an anchor and an image from the same identity is smaller than the distance of the anchor to a sample from a different identity. Recently, Softmax-based losses have gained traction as they do not suffer from the network collapse problems associated with contrastive methods [294]. In 2018, Wang *et al.* [500] proposed CosFace, where they proposed an additive margin to the resulting cosine of the angle between weights and features. Furthermore, CosFace normalizes the weights and features to have norm 1, further scaling it to a constant s . Deng *et al.* [126] extended CosFace by adding the margin to the angle between features and weights, and not to the resulting cosine. Boutros *et al.* [46] studied variations of both ArcFace and CosFace, where the margin is modeled by a random variable sampled, for each sample-epoch combination, from a Gaussian distribution. Huang *et al.* [220], inspired by the principles of curriculum learning, proposed CurricularFace, a novel loss that tackles samples from the easiest to the hardest as training progresses.

Bias in Face Recognition: The differences in the performance of Face Recognition with respect to certain demographic groups have been discussed in commercial systems by Boulamwini and Gebru [62], highlighting superior performance on males with lighter skin tones. Additionally, this work proposed one of the first balanced evaluation datasets. The repercussions of such findings led to updates on commercial systems (e.g. IBM and Microsoft) [417]. Over time, the problem of biases in commercial systems has been widely discussed [137, 140, 242]. Huber *et al.* [223] noted that explanation maps reflect the hidden biases of a model that they aim to explain.

Over the years, researchers have approached the mitigation of this problem from two different perspectives. First, mitigating the bias that results from the dataset [200]. Secondly, understanding and reducing the bias resulting from the deep architecture chosen [136]. Despite claims of favoring one research direction, research efforts have been allocated to both directions.

Gong *et al.* [188] proposed a learnable adaptive set of masks for convolutional filters. Different demographic images are split into different groups, and for each group, a set of masks is learnt in order to enhance the prevalence of ethnicity-specific features in the final embedding. Similar to the previous work, Wang *et al.* [504] method leveraged the differences of the different demographic groups to learn independent margins for each of these groups. Huang *et al.* [219] explores the gradient activation maps of the different ethnicity groups. Their proposed method leverages adversarial learning to create race-invariant gradient activation maps. Although their reported bias decreases, it has lower overall performance in all four ethnicities when compared to [504]. Serna *et*

al. [447] modified the triplet generation process of the triplet loss. Their restricted semi-hard selection selects triplets from the same demographic group in order to increase the discriminative capabilities intra-group.

After hypothesizing that demographic bias could originate from the architecture, Dooley *et al.* [136] developed a Neural Architecture Search pipeline to find an architecture and a set of hyperparameters that minimize the bias of the final model. In Chapter 11 we investigated the behavior of the model's bias before and after quantitation on several architectures and datasets. Our findings have shown that quantitation with synthetic data has the potential to mitigate bias to a certain degree.

SensitiveNets [350] propose a bias mitigation strategy that arises from an increase in the privacy and protection of sensitive attributes. Hence, their approach maps the feature vector to a space where these attributes are suppressed. Leveraging a different perspective, Gong *et al.* [187] interprets demographic information as an important piece of the debiasing process. Hence, their work proposes a novel architecture that does disentangled prediction of identity, gender, race and age. To ensure no mutual information between the different predictions, an adversarial strategy was adopted. This latter strategy has a negative impact on the overall performance of the model.

To validate their proposed unsupervised approach to bias mitigation, Wang *et al.* [503] introduced the validation dataset Racial Faces in-the-wild (RFW). This dataset has been used over the past six years to evaluate both commercial and research systems. Two years later, Robinson *et al.* [427] introduced Balanced Faces in-the-wild (BFW), a benchmark dataset that balances for ethnicity and gender. Despite balancing for more demographics, this dataset has roughly 4x fewer faces and 15x fewer identities.

Gwilliam *et al.* [200] challenged prior assumptions that a balanced dataset leads to higher bias mitigation. We believe that the previous definition is outdated, and we further show that models trained on datasets balanced with continuous labels (which are imbalanced if we consider discrete labels), perform better than models trained on datasets balanced for discrete labels. Hence, we also reformulate the assumption of Gwilliam *et al.* [200] as our current datasets are not truly balanced.

DeAndres-Tame *et al.* [116, 118] hosted a competition aimed at evaluating face recognition system trained on synthetic data and evaluated on the different ethnicity groups. Yucer *et al.* [537] introduced a solution that consisted of two steps. First, resampling from VGGFace2 [71] a set of 1200 identities balanced for race. Afterwards, their approach would generate versions of the same identity with its ethnicity flipped. Improvements came at the cost of overall performance. Similar experiments have been conducted by Chandaliya *et al.* [78] to construct a more challenging and balanced dataset.

Wang *et al.* [508] studied a sampling strategy in a study regarding different bias mitigation strategies. Their sampling approach focuses on increasing the likelihood that the model sees a sample from a rare class through oversampling. Our proposed approach differs significantly as we propose to downsample the data to approach a more balanced configuration.

12.3 Methodology

In this Chapter we propose a novel theoretical framework to interpret ethnicity labels as values belonging to a continuous spectrum. This further defies current notions of data balancing and equilibrium with respect to ethnicity. As such, we show that not all samples from a demographic group are equally strong in their representativeness of its group. Hence, we theorize that equal number of samples/identities per group does not suffice nor contributes meaningfully to a balanced dataset. Rather, we believe that the level of belonging to a demographic group of each sample directly influences the balancing act. To validate our approach we propose three sampling approaches to balance face recognition datasets for ethnicity labels in the continuous space. These sampling approaches, which would not be possible in a discrete formulation, leverage the flexibility of the novel interpretation of ethnicity labels.

Figure 12.2 highlights, in each row, the inter-ethnicity variability for each of the four distinct demographic groups. On each row, from left to the right one can observe the images with lowest to the highest prevalence of ethnicity specific traits. As shown by the four left images of each row, one can verify not only the existence of possibly noisy labels, but also non-representative samples of each ethnicity and samples that could belong to more than one of the four demographic groups, a clear representation of the continuous nature of demographic labels.

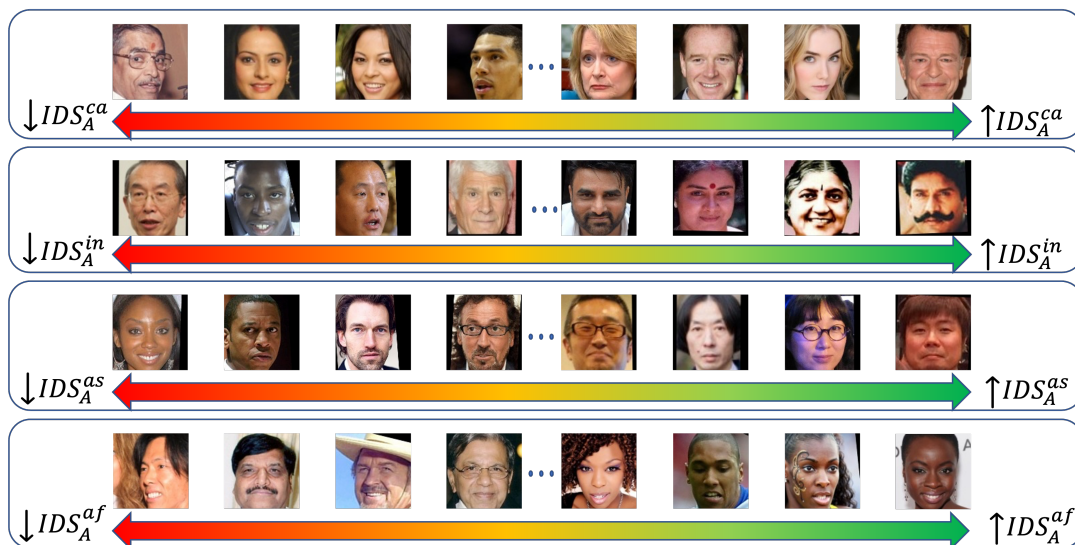


Figure 12.2: **Ethnicity Shift in Demographic Groups:** given the labels of the dataset, we show one sample of the identities that are the least and the most aligned with their label (introduced later in Section 12.3.1.1). It is visible the presence of noise in this label, as well as the high variability even within samples with high scores and the non-descriptive nature of discrete labels.

In the following paragraphs we discuss the methodology used to verify our hypothesis on the continuous nature of demographic labels. Given a function that maps from the image space to a four-dimensional continuous space, which indicates the degree of belonging of an image to four different ethnicities. We show, individually, for each ethnicity, that removing strongly-scored samples degrades the performance much faster than removing weakly-scored ones. Then, considering

the label distribution in the continuous space, we show that by evening these distributions one can achieve similar performance with an unbalanced number of identities per ethnicity group. In fact, we show that this balancing improves fairness metrics even if we do not consider the number of images per identity, nor the number of identities per ethnicity group, which further increases the unbalance of the data when seen from the perspective of discrete labels.

12.3.1 Balancing Training Data

Given a face recognition dataset where each identity is labeled with an ethnicity from the set $\mathcal{D}$, where $|\mathcal{D}| = d$, and an ethnicity score agent $f : \mathbb{R}^{h \times w} \rightarrow \mathbb{R}^d$ that infers the degree of belonging of an image to each of the d ethnicity labels in $\mathcal{D}$, one can produce image-specific continuous ethnicity labels. In this direction, identity-specific scores can be computed from these image score and a final representativeness score can be given to each ethnicity from the identity scores. To assess the overall representativeness of an ethnicity and the identity score, we designed a set of possible methods with different focuses and trade-offs. The first of these scores, can be viewed as comparative proxy between ethnicities to infer the balancing of the dataset, whereas the identity score highlights the contributions of each identity to their ethnicity score.

To validate our theory, we conducted several experiments. Following different sampling strategies, we show that having the same number of identities per ethnicity group does not mean that the dataset is balanced.

12.3.1.1 Notation

Before presenting our approach to balance a face recognition dataset, it is necessary to introduce some information regarding the variables and the notation to be used. For the remaining of this Chapter, let us consider $d = 4$. As such, we present the ethnicity label y , where $y \in \mathcal{D}$ and $\mathcal{D} = \{Af, As, In, Ca\}$. In addition, we introduce three variables that are associated with the ethnicity score at three different granularity levels, ies for the image score, IDS for the identity score and ES for the ethnicity score. The image score, $ies_{j,i} \in \mathbb{R}^4$, can be defined as follows:

$$ies_{j,i} = \begin{pmatrix} ies_{j,i}^{Af} \\ ies_{j,i}^{Ca} \\ ies_{j,i}^{As} \\ ies_{j,i}^{In} \end{pmatrix}, \quad (12.1)$$

where j and i , both scalars, represent the identity of the image and the image itself, respectively. Within the vector, it is possible to observe four entries which correspond to the image ethnicity score with respect to each ethnicity group. The identity score $IDS_j \in \mathbb{R}^4$ is defined as:

$$IDS_j = \begin{pmatrix} IDS_j^{Af} \\ IDS_j^{Ca} \\ IDS_j^{As} \\ IDS_j^{In} \end{pmatrix} \quad (12.2)$$

In a similar fashion, each entry of the vector corresponds to the identity ethnicity score with respect to each ethnicity. Finally, $ES \in \mathbb{R}^{4 \times 4}$ is a matrix defined as:

$$ES = \begin{pmatrix} ES_{Af}^{Af} & ES_{Af}^{Ca} & ES_{Af}^{As} & ES_{Af}^{In} \\ ES_{Ca}^{Af} & ES_{Ca}^{Ca} & ES_{Ca}^{As} & ES_{Ca}^{In} \\ ES_{As}^{Af} & ES_{As}^{Ca} & ES_{As}^{As} & ES_{As}^{In} \\ ES_{In}^{Af} & ES_{In}^{Ca} & ES_{In}^{As} & ES_{In}^{In} \end{pmatrix} \quad (12.3)$$

ES_{Ca}^{Af} refers to the ethnicity score for all the identities labeled as “Caucasians” with respect to how representative they are of the “African” group. In the context of this research, we focus our attention on the ethnicity score calculated with respect to the ground truth ethnicity. Hence, ES can be simplified as $diag(ES)$, where we refer to each element as $diag(ES)^{(y)}$.

12.3.1.2 Protocol A

Two elements of strong importance for the performance of face recognition systems are the number of identities and of images per identity in its training data. We hypothesized that our balancing approach should work regardless of our focus on the number of identities, and images per identity. As such, we have designed a sampling protocol that completely ignores both these elements, only focusing on the ethnicity score. This behaviour implies that an identity with 100 images could be removed in favour of an identity with ten images, if the second had a higher score. On this Protocol, A, for an image i from identity j with an image ethnicity score vector $ies_{j,i}$, the identity ethnicity score vector, IDS_{Aj} , can be calculated as

$$IDS_{Aj} = \sum_i^{M_j} \frac{ies_{j,i}}{M_j}, \quad (12.4)$$

where M_j represents the number of images from the identity j . From this score, we get the average ethnicity score of this identity. Its formulation, as an average, ignores the impact of the number of images associated with an identity. On this protocol, we can further calculate the Ethnicity Score ES_{Ay} , for a demographic group y with N_y identities as

$$ES_{Ay} = \sum_j^{N_y} \frac{IDS_{Aj}}{N_y}. \quad (12.5)$$

This score ignores the number of identities of a given ethnicity group due to its averaging over all identities. Given this ethnicity score, the ethnicity imbalance is quite significant, as shown in subsequent sections. $IDS_A^{(y)}$ can be used to rank all identities within a demographic group y . Following this, one can remove the identity j with the lowest $IDS_{Aj}^{(y)}$ from the ethnicity y with the smaller value for $diag(ES_A)^{(y)}$. This will lead to an increase in the latter score. This is shown in Algorithm 1.

Algorithm 1 Sampling Process for Protocol A

Input: $Z \geq 0$
 $\mathcal{E} \leftarrow \{Af, As, Ca, In\}$
 $\triangleright$ Ethnic Groups: African, Asian, Caucasian, Indian

 $N \leftarrow 0$
while $N \neq Z$ **do**

 for $y \in \mathcal{E}$ **do**

 $ES_A(y) \leftarrow 0$
 $\triangleright$ Initialize ethnicity-level score

 $K_y \leftarrow 0$
 $\triangleright$ Ethnicity count for sample j

 for $j \in y$ **do**

 $IDS_A(j) \leftarrow 0$
 $\triangleright$ Initialize ID-level score

 $M_j \leftarrow 0$
 $\triangleright$ Image count for sample j

 for $img \in j$ **do**

 $IDS_A(j) \leftarrow IDS_A(j) + ies(img)$

 $M_j \leftarrow M_j + 1$

 end for

 $IDS_A(j) \leftarrow \frac{IDS_A(j)}{M_j}$
 $\triangleright$ Mean for ID j

 $ES_A(y) \leftarrow ES_A(y) + IDS_A(j)$

 $K_y \leftarrow K_y + 1$

 end for

 $ES_A(y) \leftarrow \frac{ES_A(y)}{K_y}$
 $\triangleright$ Mean for Ethnicity y

 end for

 $y^* \leftarrow \arg \min_{y \in \mathcal{E}} ES_A(y)$
 $\triangleright$ Ethnicity with $\downarrow$ score

 $j^* \leftarrow \arg \min_{j \in y^*} IDS_A(j)$
 $\triangleright$ ID with $\downarrow$ score

 Remove j^* from y^*

 $N \leftarrow N + 1$
end while

12.3.1.3 Protocol B

Protocol A relies on two strong assumptions: 1) The number of identities in an ethnicity is not relevant for balancing; 2) The number of images per identity is not relevant too. Although these assumptions seem to hold sufficiently well in practice, it can be noticed that retaining the maximum performance requires a relaxation of the second assumption. For this, we theorize that a sampling approach that considers, not the average score of all images, but the sum of the score of all images leads to more robust solutions. With this change, the choice between two identical identities,

regarding their ethnicity score, relies on their number of images. For this, Protocol B, IDS_{Bj} is given by

$$IDS_{Bj} = \sum_i^{M_j} ies_{j,i} \quad (12.6)$$

This formulation incorporates information regarding the number of images, preferring images that maximize both the image score and the number of images. So, between two ethnically similar identities, the one with the lowest amount of images is more likely to be discarded. This allows for a greater performance retention, while still providing a strong bias reduction scheme. The final score is calculated as

$$ES_{By} = \sum_j^{N_y} \frac{IDS_{Bj}}{N_y} \quad (12.7)$$

Despite considering the number of images within an identity, the ethnicity score ignores the number of identities contained in the ethnicity group. The elimination of identities follows a similar process to Protocol A as shown in Algorithm 2.

Algorithm 2 Sampling Process for Protocol B

Input: $Z \geq 0$

$\mathcal{E} \leftarrow \{Af, As, Ca, In\}$

▷ Ethnic Groups: African, Asian, Caucasian, Indian

$N \leftarrow 0$

while $N \neq Z$ **do**

for $y \in \mathcal{E}$ **do**

$ES_B(y) \leftarrow 0$

▷ Initialize ethnicity-level score

$K_y \leftarrow 0$

▷ Ethnicity count for sample j

for $j \in y$ **do**

$IDS_B(j) \leftarrow 0$

▷ Initialize ID-level score

for $img \in j$ **do**

$IDS_B(j) \leftarrow IDS_B(j) + ies(img)$

end for

$ES_B(y) \leftarrow ES_B(y) + IDS_B(j)$

$K_y \leftarrow K_y + 1$

end for

$ES_B(y) \leftarrow \frac{ES_B(y)}{K_y}$

▷ Mean for Ethnicity y

end for

$y^* \leftarrow \arg \min_{y \in \mathcal{E}} ES_B(y)$

▷ Ethnicity with ↓ score

$j^* \leftarrow \arg \min_{j \in y^*} IDS_B(j)$

▷ ID with ↓ score

Remove j^* from y^*

$N \leftarrow N + 1$

end while

12.3.1.4 Protocol C

Similarly to the relaxation on the assumptions of Protocol A to construct Protocol B, one can further relax the first assumption regarding the independence between fairness-performance trade-offs and the number of identities per ethnicity. Hence, we perform a similar relaxation step on the computation of the ethnicity score. Besides this relaxation, one strong benefit of this hypothesis is the different sampling strategy, which supports extreme identity removal strategies. In theory, such method could function until there is only a single identity per ethnicity, creating a robust approach for highly imbalanced datasets. In Protocol C, IDS_{Cj} is computed as

$$IDS_{Cj} = \sum_i^{M_j} ies_{j,i}, \quad (12.8)$$

in a similar approach to IDS_B , hence, considering the number of images within an identity. ES_{Cy} is given by

$$ES_{Cy} = \sum_j^{N_y} IDS_{Cj} \quad (12.9)$$

This score considers both the number of images in an identity, propagated from IDS_{Cj} , and the number of different identities belonging to a demographic group y . This formulation as a sum over the different identities implies a different strategy to select which identity should be removed. On Protocols A and B, the final score is computed as a mean over all identities, so if we remove the one with the lowest ethnicity score, the average score increases, thus approximating that group to other high-scoring ethnicity groups. In this case, an identity is removed from the group y with the highest $diag(ES_C)^{(y)}$, hence, pushing it down to the level of the other groups. This is highlighted in Algorithm 3.

12.3.1.5 Relabeling

Low-scored samples might be an indicative of the presence of mislabeled samples. In fact, their presence might pose the question regarding the potential performance of face recognition models on properly labeled datasets. Hence, given a dataset we propose the creation of a second version of that dataset with adjusted labels, determined by an ethnicity score agent. This relabeled dataset provides an alternative starting point for our balancing protocols, enabling us to assess the impact of mislabeling on dataset balancing. In our relabeling process, each identity's ethnicity label y is reassigned based on the maximum score inferred by the ethnicity score agent: $y = \operatorname{argmax}_{\hat{y} \in \mathcal{Y}} IDS_{Aj}^{(\hat{y})}$. Identities that were considered low-score in one dataset can become high-score following their ethnicity label re-assignment, thus reducing their likelihood of being removed. Our theory shall hold even when these noisy signals are removed through the relabeling

Algorithm 3 Sampling Process for Protocol C

Input: $Z \geq 0$
 $\mathcal{E} \leftarrow \{Af, As, Ca, In\}$ ▷ Ethnic Groups: African, Asian, Caucasian, Indian
 $N \leftarrow 0$
while $N \neq Z$ **do**
 for $y \in \mathcal{E}$ **do**
 $ES_C(y) \leftarrow 0$ ▷ Initialize ethnicity-level score
 for $j \in y$ **do**
 $IDS_C(j) \leftarrow 0$ ▷ Initialize ID-level score
 for $img \in j$ **do**
 $IDS_C(j) \leftarrow IDS_C(j) + ies(img)$
 end for
 $ES_C(y) \leftarrow ES_C(y) + IDS_C(j)$
 end for
 end for
 $y^* \leftarrow \arg \max_{y \in \mathcal{E}} ES_C(y)$ ▷ Ethnicity with $\uparrow$ score
 $j^* \leftarrow \arg \min_{j \in y^*} IDS_C(j)$ ▷ ID with $\downarrow$ score
 Remove j^* from y^*
 $N \leftarrow N + 1$
end while

process. Thus, by replicating protocols on both datasets, we show that balancing goes beyond a simple strategy to remove mislabeled samples.

12.4 Experimental setup

In this section, we expand on the details regarding the experiments and reproducibility of the methods previously described. All sampled versions, the code for sampling and other elements are publicly available on GitHub³.

12.4.1 Datasets

Provided our hypothesis, and our methodology to test it, one can leverage datasets that contain, or not, ethnicity information. However, for a comparison between a discretely balanced dataset and a continuously balanced one, it was essential to select a dataset that was discretely balanced with regards to ethnicity. For this, BUPT-BalancedFace [504] was chosen as it is naturally balanced for a same number of identities per ethnicity. BUPT-GlobalFace [504] was used for extended experiments which started from an unbalanced starting point, and evaluation was conducted on a ethnicity-balanced dataset, known as RFW [503], to assess fairness metrics. These datasets are described in the following subsections.

³<https://github.com/NetoPedro/Equilibrium-Face-Recognition>

12.4.1.1 RFW

Racial Faces in-the-wild (RFW) [503] has been designed to benchmark face recognition algorithms on the different ethnicity groups. Hence, it is divided, by the original authors, into four groups, each corresponding to an ethnicity. Each of these groups has roughly 6000 images, and in total the dataset is composed of approximately 9000 different individuals. It should be noted that in face recognition scenarios, the presence of bias is measured intra-ethnicity. In other words, if pairs of ethnicity A are easier to classify than pairs of ethnicity B, then our model is biased against ethnicity B.

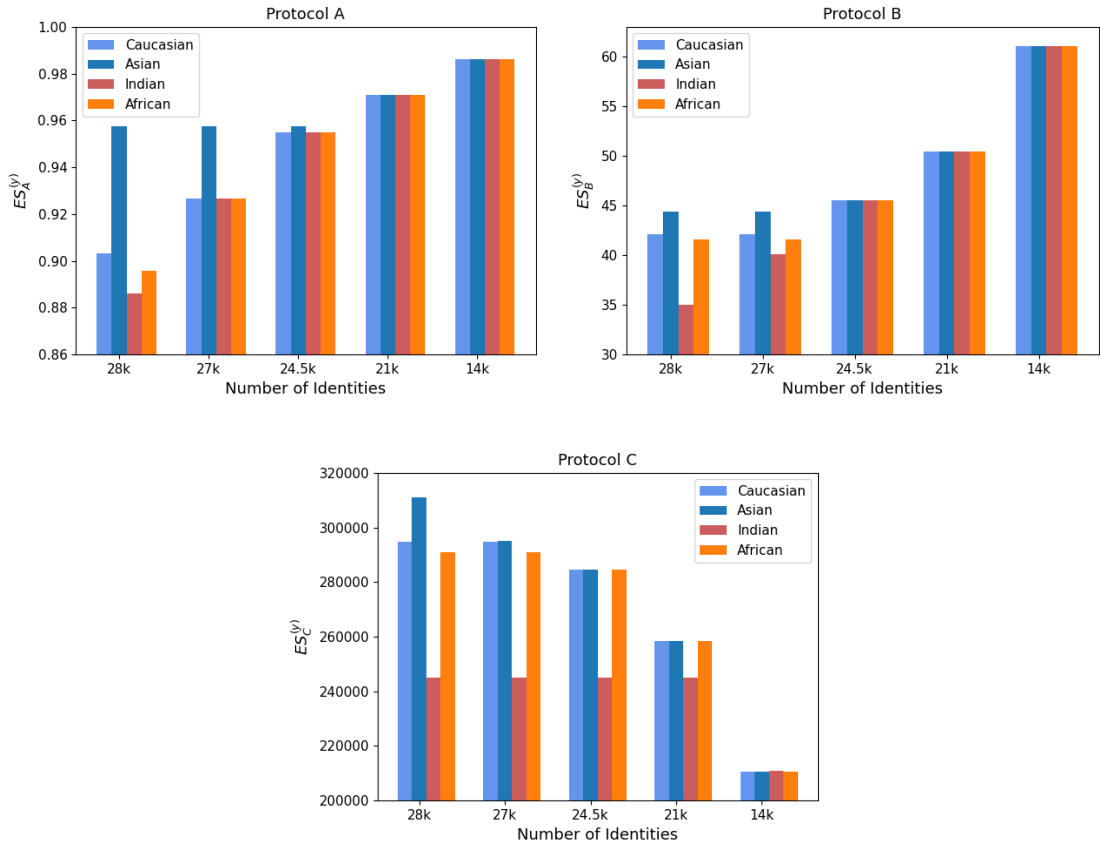


Figure 12.3: The evolution of the different scores $diag(ES_A)^{(y)}$, $diag(ES_B)^{(y)}$, and $diag(ES_C)^{(y)}$ (y-axis) as the number of identities in the dataset decreases (x-axis). It can be seen that for the first two protocols, balance is quickly achieved, whereas for the other only the last iteration is balanced.

12.4.1.2 BUPT-BalancedFace and BUPT-GlobalFace

Both BUPT-BalancedFace and BUPT-GlobalFace *et al.* [504] have been collected to be used as training datasets for fair face recognition models. Proposed by the same authors of RFW, these two datasets can be used to study the bias inherent to the dataset and to the architecture itself. Similarly to RFW, each identity has a label based on its skin tone and is further categorized into

one of the following ethnicities: African, Asian, Indian or Caucasian. BUPT-GlobalFace is the largest of the two datasets, with two million images from 38k identities. The ethnicity distribution of these identities follows a similar distribution to the world’s population. BUPT-BalancedFace equalizes the number of identities that belong to each ethnicity, and is composed of 1.3 million images from 28k distinct identities. This dataset is balanced with respect to its discrete labels.

12.4.2 Balancing Criteria

Different protocols reach their equilibrium at different points, for instance, some might originate a balanced dataset (in the continuous space) after removing three thousand identities, others might require a higher number of removals. Figure 12.6 shows the equilibrium of the different ethnicities for different numbers of identities for all the three protocols. As expected, Protocol A and B reach even scores around a similar number of identities, while Protocol C requires significantly more steps to balance all four ethnicities.

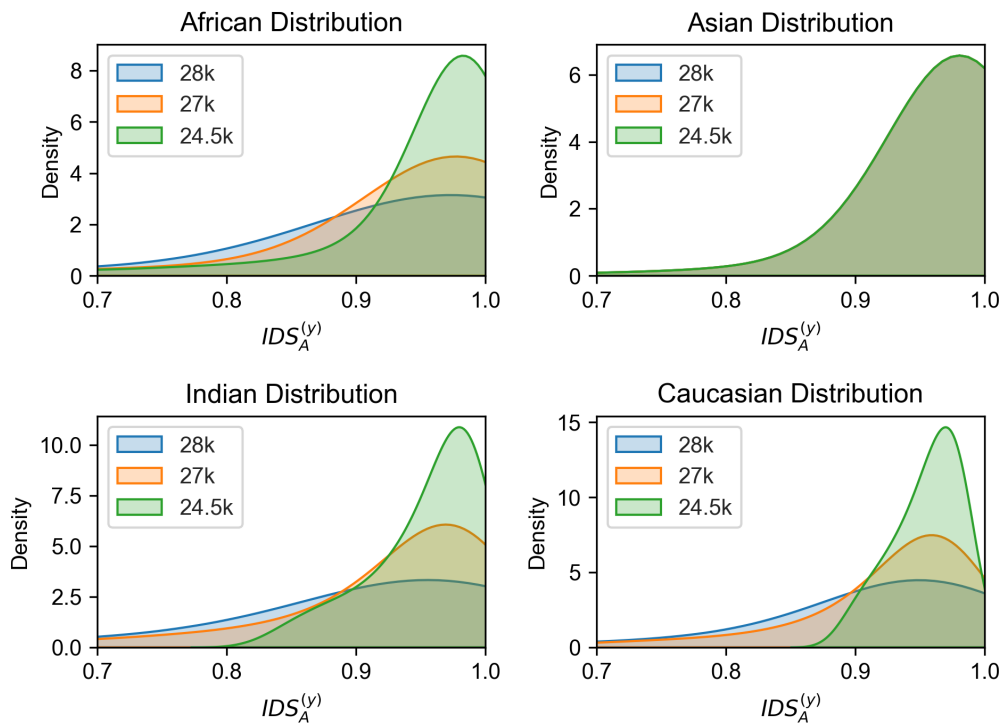


Figure 12.4: Distribution shift of the ethnicity score on the different ethnicity groups when sampling identities with Protocol A. It can be seen that the Asian distribution is not affected, as no sample is removed. It would only be affected when the sampling removes up to 25% of the identities. Additionally, this shows that distributions shift from a flatter look into one that peaks near 1.

A fair comparison of all these protocols requires them to have a similar number of identities. As such, within the context of this work, we have decided to compare the different approaches through the removal of a fixed number of identities. We compare the original dataset with 28k identities with sampled versions with 27k, 24.5k, 21k and 14k, thus removing from 3.5% to 50%

of the identities. One should note that this are reference values for comparison, and might not necessarily reflect the optimal value for identity removal. In Figure 12.4 we show a density approximation of Protocol A identity scores computed from the available identities. Once we remove 1k and 3.5k identities, the mean of these densities, for three ethnicities, starts to converge towards to 1, and the variance starts to decrease. In addition, we can see that $IDS_A^{(y)}$ shows no change as we remove up to 3.5k identities from the dataset, reflecting the behavior of this protocol, which only affects this demographic group for higher values of identities removed. This results from the higher baseline $IDS_A^{(y)}$ for this group.

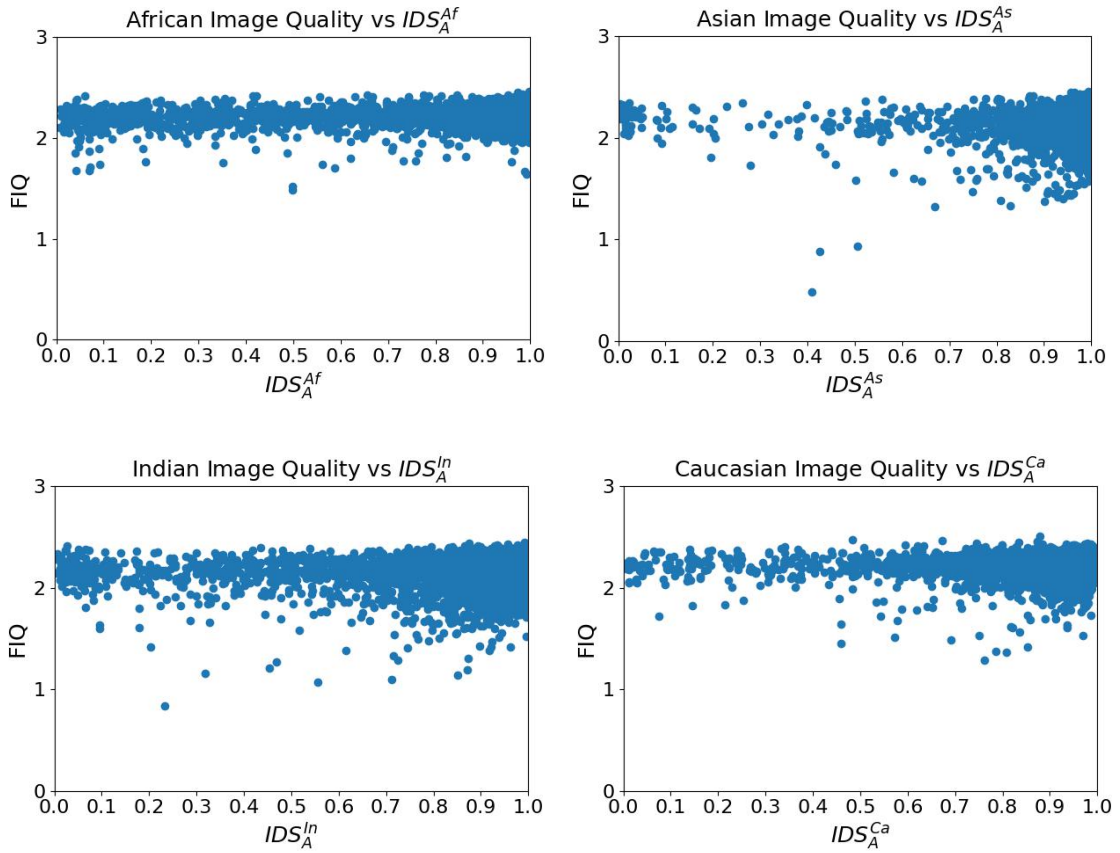


Figure 12.5: Utility metrics and Ethnicity Score relation. It seems that, despite difference across demographic groups, does not seem to exist a correlation between the Ethnicity Score and CR-FIQA utility score.

12.4.3 Ethnicity Score as an Utility Measure

The nature of this score and its link to performance improvements raises the question of whether it acts as a utility measure. In face recognition, utility-like scores have been quite popular due to their ability to give higher or lower importance to images given their utility for training a face recognition system. Some of these utility measures are linked to specific models, others are more

agnostic. In practice, our ethnicity score behaves like an agnostic utility system, providing insights on the most impactful samples per ethnicity.

Due to the similarities with utility, one might pose the question regarding how our score differs from utility scores, and if they can be combined. For this assessment, we performed a comparison between $ies_{j,i}^{(y)}$ and a quality score. For this, we compared our score with the quality score produced by CR-FIQA [52].

Figure 12.5 highlights that high-scores for the quality metric do not imply high scores on our ethnicity score. Furthermore, it can be seen that samples with maximum representativeness of their ethnicity group might still have subpar quality scores in comparison with less representative samples. Hence, despite the visible differences in the quality distribution of different ethnicities (e.g. IDS_A^{In} shows higher dispersion on the y-axis), samples removed by our balancing strategy are not necessarily low-utility ones. As such, the ethnicity score can be seen as an uncorrelated utility metric when compared to quality-based metrics, indicating the potential of combining both metrics for the selection of fair and high-performing datasets.

12.4.4 Ethnicity Classifier

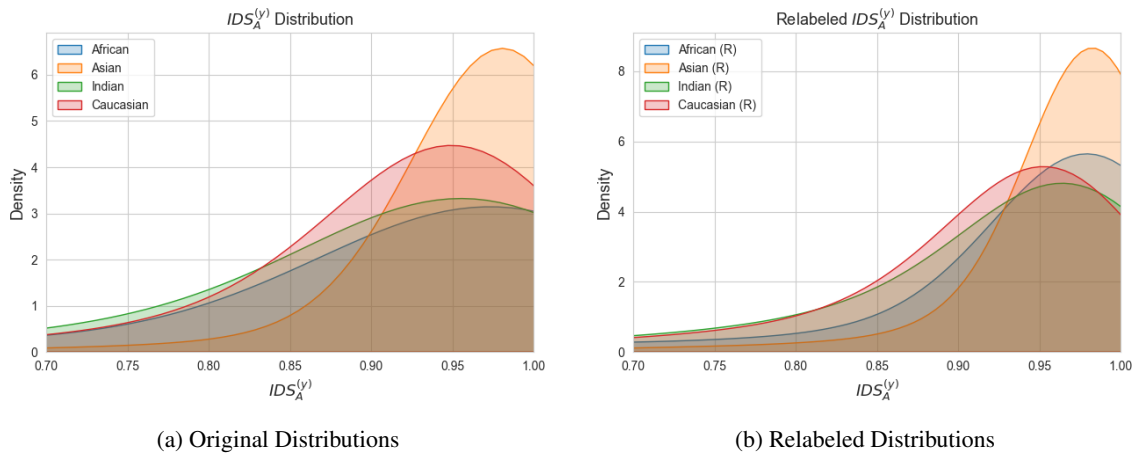


Figure 12.6: Identity Score ($IDS_A^{(y)}$) Distribution for the different demographic groups for both the original ethnicity labels and after a relabelling process.

Both training datasets have ethnicity information in the form of discrete labels. To study the impact of continuous labels, there is no dataset available in the literature that contains such information. Hence, we have designed a pseudo-labelling strategy to infer such information. Starting from a pretrained face recognition model (ElasticArc [46] with a ResNet100 [205]), we add a classification layer with four output classes, one for each ethnicity. Despite believing that the distance between African and Caucasian samples, and between Caucasian and Indian samples is not equal, an ordinal formulation of this classification problem would be an under-fitted formulation of those distances. Hence, we train our classification model on BUPT-BalancedFace with each ethnicity hot-encoded in a vector. During training, only the weights of the classification layer are updated.

We verified that the latent space of FR systems encodes ethnicity information, as this approach achieves far stronger results than training a deep neural network from scratch, in addition of being more robust to overfitting.

After training our classifier, we use it to infer a probability vector for each image in both BUPT-BalancedFace and BUPT-GlobalFace datasets. One important point is to take into account that an image with a high probability on the Caucasian group is not necessarily representative of the definition of Caucasian, it means that that particular image is easily classified as belonging to that group of images in the dataset. Given this, we can also verify that images with the lowest probability are likely to not contribute significantly to the performance of that demographic group. Figure 12.6a shows, for each group, a density function of our proposed score, further described in the previous sections. In the image, it is visible that the distribution of these scores in different groups varies significantly. This can be mitigated, if we relabel the ethnicity label of each identity with the new ES, as seen in Figure 12.6b.

12.4.5 Implementation Details

In our experiments, we utilise three different architectures: iResNet-34, -50 and -100 [205], which output an embedding of 512-D. This embedding size is standard across the literature, and we utilize iResNet-34 for the majority of the experiments, as it is faster to train and validate the effects of different training datasets. All the networks have been trained from scratch without resorting to any pre-trained weights. We optimize the networks with margin-based losses, and for a more robust evaluation, we conduct experiments with two different losses: ElasticArcFace [46] and CurricularFace [220]. Both losses use the same value for the margin, $m = 0.50$, and for the scaling factor, $s = 64$. In addition, ElasticArcFace has a standard deviation parameter set to 0.05 as described in the original paper.

Each network is optimized in batches of 256 images for 26 epochs, with an initial learning rate of 0.1. We schedule the learning rate of the stochastic gradient descent (SGD) optimizer to decay at epochs 8, 14, 20 and 25. Random horizontal flips are the only augmentation strategy used.

12.4.6 Performance Evaluation

To properly evaluate the different sampling approaches and to compare models trained with them with other models, we followed a strategy of comparing datasets with the same number of identities. So, having three proposed sampling approaches, we wanted to compare them with a fourth model trained on an approach that retained the balance of the dataset with respect to the original discrete labels. Hence, we perform random sampling, which will retain the discretely balanced nature of the original set. So, for each sampling strategy, we compare a model trained on its respective dataset with models trained on the original dataset and a randomly sampled dataset. The first comparison highlights if there is any performance and fairness change, whereas the second shows the advantages of continuous labels against discrete labels on equally sized datasets.

To evaluate both performance and fairness, we use RFW and individually compute the accuracy for each demographic group. This accuracy across demographic groups can be used to compute a global accuracy, which showcases the discriminative power of the final model. Additionally, the standard-deviation (STD) across the accuracy of different groups can be measured as a guide to its fairness. The skewed error rate (SER) can further enhance our understanding of the relation between the higher and lower error rates within the four ethnicities. SER can be described, with respect to the accuracy, as follows:

$$SER = \frac{1 - \min([Acc_{af}, Acc_{as}, Acc_{ca}, Acc_{in}])}{1 - \max([Acc_{af}, Acc_{as}, Acc_{ca}, Acc_{in}])} \quad (12.10)$$

The combination of performance and fairness metrics ensures that we aim for the best-performing model at a specific bias level. In other words, reducing the bias at a high performance cost can increase the errors across all ethnicity. For this, we further analyse this trade-off through the study of the Pareto-Front of all models with respect to prediction error and bias.

12.4.7 Hardware

All face recognition models and the ethnicity classifier have been trained on a server with a single NVIDIA A100 80GB GPU. The sampled versions of the datasets have been generated on a commercial Apple's MacBook Pro with a M1 Pro chip.

12.5 Results

Our hypothesis that a continuous label is essential to ensure and measure the balance nature of a dataset can be further validated through experimental results. In this following section, we focus on validating the previously discussed assumptions that shall hold for our hypothesis to be verified. Hence, the following section discusses the validation of the two main assumptions: 1) If continuous labels have an impact removing most representative samples has a different effect from removing the less representative, or randomly; 2) Regardless of the Protocol we follow for balancing the data in the continuous space (through sample removal), it shall attain fairer results than a discretely balanced dataset with the same number of identities.

All the findings presented in this section originate from experiments using an iResNet-34 with ElasticArcFace. To showcase the robustness of our hypothesis beyond network architecture or loss, experiments with other architectures, losses, and training datasets are also shown.

12.5.1 Single Ethnicity Analysis

In our hypothesis, the first assumption is that not all samples contribute equally to the performance of the model on a given ethnicity. Hence, some might be deemed more representative, or less representative of a given ethnicity. Given this assumption, one can estimate that retaining less representative samples of one ethnicity would lead to lower performance on that specific ethnicity

than retaining the most representative ones. This shall be more noticeable on demographic groups frequently associated with unfairness.

To validate such assumption, we propose to remove exactly half of the identities in a given ethnicity. To ensure that less representative samples lead to lower performance on that ethnicity when compared to more representative ones, we propose **Min.** and **Max.**, the retention of the less and the most representative half, respectively. To add an additional layer of granularity, we also state that randomly (**Rand.**) removing half of the identities shall lead to lower performance too.

As expected, Table 12.1 validates our assumption that training with the most representative samples of an ethnicity (**Max.**) leads to higher performance on that given ethnicity. This is strongly noticeable on the African group with nearly three percent point difference. Max. for both African and Indian groups retains performances comparable to a model trained on the entire dataset. An additional observation is that despite training on significantly less data, some versions seem to improve, on certain ethnicities, when compared to a model trained on the entire dataset. Fairness is not considered in this specific analysis, as the focus was not to balance the data, but to validate the first assumption that not all identities of a given ethnicity contribute equally to that given ethnicity.

Table 12.1: Results for an iResNet-34 (ElasticArcFace) when half of the samples from a single identity are removed. Three strategies are followed: Retaining 3500 random identities (Rand), the 3500 identities that minimise the score for that ethnicity (Min), and the 3500 identities that maximise this score (Max).

#ID	Strategy	Accuracy per skin tone (%) $\uparrow$			
		Caucasian	Indian	Asian	African
28k	None	96.67%	94.88%	94.22%	93.38%
24.5k	Rand. (Af)	96.08%	94.28%	93.82%	91.92%
24.5k	Min. (Af)	96.30%	94.52%	94.37%	89.95%
24.5k	Max. (Af)	96.12%	94.77%	94.18%	92.78%
24.5k	Rand. (In)	96.30%	94.02%	93.85%	93.38%
24.5k	Min. (In)	96.35%	93.78%	94.23%	93.15%
24.5k	Max. (In)	96.38%	94.68%	93.98%	93.13%
24.5k	Rand. (As)	96.43%	94.67%	92.95%	93.45%
24.5k	Min. (As)	96.28%	94.42%	93.00%	93.30%
24.5k	Max. (As)	96.46%	94.52%	93.00%	93.85%
24.5k	Rand. (Ca)	95.60%	94.63%	93.80%	93.48%
24.5k	Min. (Ca)	95.70%	94.72%	94.07%	93.43%
24.5k	Max. (Ca)	95.77%	94.97%	93.97%	93.52%

12.5.2 Balancing in the Continuous Space

Following the verification of the first assumption, we now are assured that the representativeness of the identity is, to some extent, represented by the continuous label. Hence, we now have an

additional assumption that must be held if one wants to validate our hypothesis. Balancing the data in this new continuous space must lead to superior performance vs. fairness trade-offs.

The concept of balancing is, in itself, a vague concept, which might refer to balanced with regards to the identities, to the density of each identity, or even just to the raw number of images. And thus, to show the clearly superior performance of our hypothesis, leaving no question regarding the continuous nature of the problem, we show independently that our assumption holds for these three distinct perspectives of a balanced dataset.

Table 12.2: Results for iResNet-34 (ElasticArcFace) trained on several versions of the training set. Ranging from no sampling to three different sampling strategies, where two are based on Protocol A - **A** and **A(R)**. These latter strategies seem to hold significant advantages with respect to random sampling. Results are shown for several variations of the number of identities kept in the training set.

#ID	Protocol	Accuracy per skin tone (%) $\uparrow$				Average	Fairness $\downarrow$	
		Caucasian	Indian	Asian	African		STD	SER
28k	None	96.67%	94.88%	94.22%	93.38%	94.79%	1.39	1.99
27k	Random	96.65%	94.82%	94.20%	93.77%	94.86%	1.27	1.86
24.5k	Random	96.28%	94.47%	93.55%	93.15%	94.36%	1.39	1.84
21k	Random	95.78%	93.97%	93.20%	92.43%	93.85%	1.44	1.79
14k	Random	93.82%	92.10%	91.08%	90.00%	91.75%	1.62	1.62
27k	A	96.48%	94.90%	94.33%	93.77%	94.87%	1.17	1.76
24.5k	A	96.11%	94.82%	94.23%	93.42%	94.65%	1.13	1.69
21k	A	95.50%	93.92%	93.78%	92.82%	94.01%	1.11	1.59
14k	A	92.01%	91.83%	92.20%	90.95%	91.75%	0.55	1.16
27k	A(R)	96.35%	94.80%	94.30%	93.58%	94.76%	1.17	1.76
24.5k	A(R)	96.00%	94.37%	94.13%	93.52%	94.51%	1.06	1.62
21k	A(R)	95.13%	93.87%	93.63%	93.03%	93.92%	0.88	1.43
14k	A(R)	92.22%	92.42%	92.12%	90.68%	91.86%	0.80	1.23

12.5.3 Protocol A

In the first scenario, one does not account for the density of each identity nor to the number of identities left in an ethnicity group. Rather, one uses Protocol A, which only considers our proposed label.

On Table 12.2 three distinct approaches for sampling are used. 1) A fixed amount of identities are sampled randomly, hence, retaining the balancing of the data in the discrete space - **Random**; 2) using the original ethnicity labels and the proposed score, a fixed amount of identities are removed - **A**; 3) similar to the latter, it utilises the revised discrete labels and the proposed score - **A(R)**. Given our initial assumption that fairness metrics shall improve under our strategies, we found that beyond fairness metrics, balancing the data in the continuous space consistently led to improved performance metrics too. Hence, not only was the assumption verified, but it also showed that continuous labels have such an impact on data balancing that even without accounting

for performance-related factors (#IDs and #images), performance can be improved with correct balancing.

Table 12.3: Results for iResNet-34 (ElasticArcFace) trained on several versions of the training set. Ranging from no sampling to three different sampling strategies, where two are based on Protocol B - **B** and **B(R)**. These latter strategies seem to hold significant advantages with respect to random sampling. Results shown for several variations of the number of identities kept in the training set.

#ID	Strategy	Accuracy per skin tone (%) $\uparrow$				Average	Fairness $\downarrow$	
		Caucasian	Indian	Asian	African		STD	SER
28k	None	96.67%	94.88%	94.22%	93.38%	94.79%	1.39	1.99
27k	Random	96.65%	94.82%	94.20%	93.77%	94.86%	1.27	1.86
24.5k	Random	96.28%	94.47%	93.55%	93.15%	94.36%	1.39	1.84
21k	Random	95.78%	93.97%	93.20%	92.43%	93.85%	1.44	1.79
14k	Random	93.82%	92.10%	91.08%	90.00%	91.75%	1.62	1.62
27k	B	96.60%	95.20%	94.45%	93.77%	95.02%	1.24	1.87
24.5k	B	96.18%	94.73%	94.32%	93.98%	94.80%	0.97	1.58
21k	B	96.27%	94.70%	94.25%	93.68%	94.73%	1.11	1.69
14k	B	95.03%	93.18%	92.42%	91.22%	92.96%	1.60	1.77
27k	B(R)	96.37%	95.08%	94.42%	93.73%	94.90%	1.12	1.73
24.5k	B(R)	96.25%	94.78%	94.33%	93.65%	94.75%	1.10	1.69
21k	B(R)	96.10%	94.38%	94.07%	93.13%	94.42%	1.24	1.76
14k	B(R)	95.10%	93.70%	92.38%	91.83%	93.25%	1.46	1.67

12.5.4 Protocol B

Differently from the strategy previously discussed, Protocol B leverages continuous labels, but also accounts for the number of images within an identity. Through this, we add an additional focus on the exploitation of the trade-off between fairness and performance. To validate our assumption regarding balancing in the continuous space, this protocol would have to show reduced fairness, and improved performance. We consider two different approaches for this protocol. One with a revision of the original labels, **B(R)**, and one with the original ethnicity labels, **B**.

Table 12.3 highlights some of the relevant aspects of exploring continuous labels in line with the number of images per identity. First, although having slightly higher bias than Protocol A in some iterations, it is always better than the bias present in all versions of random sampling. Additionally, the average accuracy is greatly improved compared to both protocol A and Random. When we remove 1k of the identities, the performance increases significantly with improvements on both STD and SER. But the most surprising result is that after removing one quarter of the dataset (7k identities), the performance degradation is nearly non-existent (0.06 percent points) with a high improvement on the bias. This means that we managed to train a comparable model with respect to performance, which is more fair and takes roughly 25% less time to train. When we remove half of the identities, the proposed B(R) achieves a performance 1.5 percent points above

the random equivalent, while also being at 1.5 percent points distance from the version trained on the full dataset.

These results further reinforce our assumption that continuous labels provide a improved strategy to balance datasets while accounting for both performance and fairness.

Table 12.4: Results for iResNet-34 (ElasticArcFace) trained on several versions of the training set. Ranging from no sampling to three different sampling strategies, where two are based on Protocol C - C and C(R). These latter strategies seem to hold significant advantages with respect to random sampling. Results shown for several variations of the number of identities kept in the training set.

#ID	Strategy	Accuracy per skin tone (%) $\uparrow$				Average	Fairness $\downarrow$	
		Caucasian	Indian	Asian	African		STD	SER
28k	None	96.67%	94.88%	94.22%	93.38%	94.79%	1.39	1.99
27k	Random	96.65%	94.82%	94.20%	93.77%	94.86%	1.27	1.86
24.5k	Random	96.28%	94.47%	93.55%	93.15%	94.36%	1.39	1.84
21k	Random	95.78%	93.97%	93.20%	92.43%	93.85%	1.44	1.79
14k	Random	93.82%	92.10%	91.08%	90.00%	91.75%	1.62	1.62
27k	C	96.68%	94.73%	94.07%	93.93%	94.85%	1.27	1.83
24.5k	C	96.43%	94.80%	93.72%	93.53%	94.62%	1.33	1.81
21k	C	96.12%	94.55%	93.73%	92.28%	94.17%	1.60	1.99
14k	C	94.60%	93.13%	92.32%	90.85%	92.73%	1.57	1.69
27k	C(R)	96.52%	94.98%	94.33%	93.32%	94.79%	1.34	1.92
24.5k	C(R)	96.18%	94.78%	94.22%	93.33%	94.63%	1.19	1.75
21k	C(R)	96.28%	94.28%	93.58%	92.87%	94.25%	1.47	1.92
14k	C(R)	94.70%	93.38%	92.12%	90.73%	92.73%	1.70	1.75

12.5.5 Protocol C

Following the previous analysis, we further explored the impact of the accounting for both the number of images and identities while balancing the dataset with our proposed score. In the other side of the spectrum, when compared to Protocol A, this Protocol C demonstrated a clear dominance of performance-related metrics when compared to bias-related ones.

C achieves competitive results, performance-wise, with protocol B, and manages to surpass the model trained on the full dataset under some scenarios. However, it diminish the gap between random sampling and ethnicity oriented sampling with regards to bias metrics. This is seen in Table 12.4, where, despite differences in overall performance of the models, bias metrics are similar between Random, and both C and C(R). This results reinforce that even if we consider the number of identities and images per ethnicity group, the ethnicity score of each identity is far more relevant and meaningful.

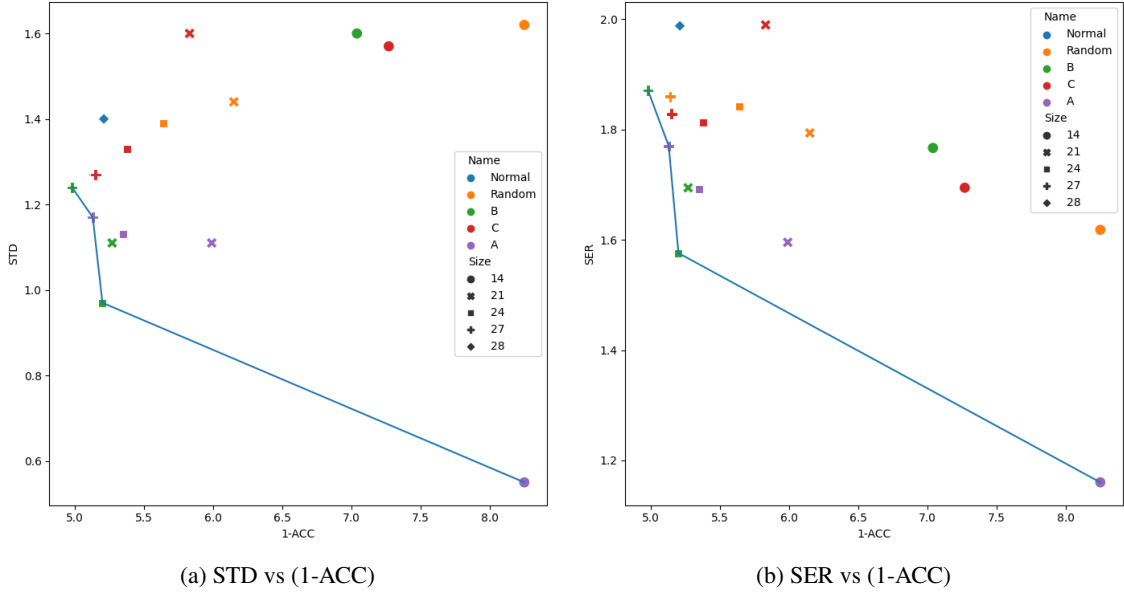


Figure 12.7: Trade-off between bias and performance visually highlighted through the computation of a pareto frontier with all the dominant solutions. Different shapes on the points represent different dataset sizes, while the colour reflects the sampling strategy. Pareto frontier elements are connected through a blue line.

12.5.6 Comparison between Protocols

In order to compare the different protocols, their trade-offs and the ideal number of identities, one can focus on visualising the differences. So, for an easier and more visual comparison, we propose the computation of a Pareto Front between all trained models.

Figure 12.7 shows the trade-off between the error (the complement of the accuracy) and STD, and also between the error and SER. On Figure 12.7a it is visible the superior performance of sampling strategies A and B, with the both having half of their models on the pareto frontier. These four models dominate all other instances, even the one without sampling. Random sampling shows inferior performance being the worse for all sampling magnitudes. Similar findings are displayed in Figure 12.7b, where B and A outperform the remaining strategies.

If one conducts a similar analysis per each dataset size, the results reflect the same conclusions with both A and B leading with regards to performance and bias. While B's performance is the best in all scenarios, A's dominance is only seen in three out of four dataset sizes.

12.5.7 Different Architectures

In order to validate previous findings on larger architectures, we trained a ResNet-50, and a ResNet-100 on the data given by each protocol in addition to the random data. Due to the large number of potential experiments, we decided to not consider relabeling, and to only remove 3.5k samples. Nonetheless we firmly believe that previous experiments in addition to the ones presented here, constitute sufficient evidence of the effectiveness of continuous labels.

Table 12.5: Results for iResNet-50 (ElasticArcFace) trained on several versions of the training set. Ranging from no sampling to four different sampling strategies, A, B, C and Random.

#ID	Strategy	Accuracy per skin tone (%) $\uparrow$				Average	Fairness $\downarrow$	
		Caucasian	Indian	Asian	African		STD	SER
28k	None	96.83%	95.33%	95.00%	94.16%	95.33%	1.11	1.84
24.5k	Random	96.32%	94.73%	94.27%	93.67%	94.75%	1.13	1.72
24.5	A	96.35%	95.02%	94.47%	93.88%	94.93%	1.05	1.68
24.5k	B	97.02%	95.00%	94.93%	94.05%	95.25%	1.26	2.00
24.5k	C	96.87%	95.32%	94.33%	94.25%	95.19%	1.22	1.84

ResNet-50 results are presented in Table 12.5. From these results, one can quickly infer a behavior similar to ResNet-34. When compared to our sampling approaches, random sampling is only close, yet with some difference, in performance to strategy A. However, this strategy achieves better fairness metrics than Random sampling. Strategies B and C seem to have lower fairness, but when carefully analysing them, their average accuracy is far higher than random sampling, and each individual demographic has improved performance in comparison.

Table 12.6: Results for iResNet-100 (ElasticArcFace) trained on several versions of the training set. Ranging from no sampling to four different sampling strategies, A, B, C and Random.

#ID	Strategy	Accuracy per skin tone (%) $\uparrow$				Average	Fairness $\downarrow$	
		Caucasian	Indian	Asian	African		STD	SER
28k	None	97.12%	95.78%	94.93%	95.36%	95.80%	0.95	1.76
24.5k	Random	97.08%	95.72%	94.73%	94.87%	95.60%	1.08	1.80
24.5	A	96.93%	95.65%	95.13%	95.22%	95.73%	0.83	1.59
24.5k	B	97.27%	95.65%	95.15%	95.28%	95.84%	0.98	1.78
24.5k	C	97.15%	96.02%	95.05%	95.38%	95.90%	0.93	1.74

An analysis of the results of ResNet-100, shown in Table 12.6, leads to similar conclusions. However, in this case, strategies B and C not only surpass random in average accuracy, but achieve better values for the fairness metrics too. Moreover, they even surpass, with respect to average accuracy, the model trained on the full dataset, with C also achieving lower values of bias. These results further motivate the use of continuous labels to balance face recognition datasets.

12.5.8 Different Loss

In addition to validate on different architectures, we explored ResNet-34 with a different face recognition loss. This is motivated by the fact that, in this case, our Ethnicity classifier comes from a model pre-trained with ElasticArcFace. Hence, we repeat a large portion of the experiments with CurricularFace, highlighting that our sampled sets are effective, less biased and originate better models than sets balanced in the discrete space.

Table 12.7: Results for iResNet-34 (CurricularFace) trained on several versions of the training set. Ranging from no sampling to four different sampling strategies, A, B, C and Random. These latter strategies seem to hold significant advantages with respect to random sampling. Results shown for several variations of the number of identities kept in the training set.

#ID	Strategy	Accuracy per skin tone (%) $\uparrow$				Average	Fairness $\downarrow$	
		Caucasian	Indian	Asian	African		STD	SER
28k	None	96.55%	94.90%	94.37%	93.97%	94.95%	1.13	1.75
27k	Random	96.55%	94.88%	94.37%	93.75%	94.89%	1.20	1.81
24.5k	Random	96.27%	94.72%	93.97%	93.30%	94.57%	1.28	1.80
21k	Random	95.50%	94.15%	93.28%	92.47%	93.85%	1.30	1.67
14k	Random	93.95%	92.72%	90.98%	89.87%	91.88%	1.81	1.67
27k	A	96.83%	94.93%	94.30%	94.03%	95.02%	1.26	1.88
24.5	A	96.07%	94.87%	94.10%	93.38%	94.61%	1.15	1.68
21k	A	95.22%	94.10%	93.23%	93.05%	93.90%	0.99	1.45
14k	A	92.17%	92.33%	91.98%	91.25%	91.93%	0.48	1.14
27k	B	96.43%	95.15%	94.25%	93.67%	94.88%	1.20	1.77
24.5k	B	96.58%	94.95%	94.23%	93.75%	94.88%	1.24	1.83
21k	B	96.10%	94.47%	94.15%	93.10%	94.46%	1.24	1.77
14k	B	95.10%	93.72%	92.25%	91.72%	93.20%	1.52	1.69
27k	C	96.42%	95.22%	94.37%	93.68%	94.92%	1.18	1.77
24.5k	C	96.22%	94.85%	94.07%	93.45%	94.65%	1.19	1.73
21k	C	96.20%	94.60%	93.53%	92.67%	94.25%	1.52	1.93
14k	C	94.62%	93.72%	92.07%	91.12%	92.88%	1.58	1.65

In Table 12.7, it is visible that Random sampling was unable to surpass the remaining models on every dataset size and for all metrics, either performance or fairness ones. Furthermore, it is possible to note the large impact that strategy A has on the bias reduction, and that B has on the performance maintenance as the size of the training set reduces. More impressive is the fact that such performance maintenance does not make this model less fair than the ones trained on Random sampling. Having our hypothesis validated on a second face recognition loss further motivates its adoption in future research.

12.6 Conclusion

During this work we have introduced the concept of an ethnicity flow, as ethnicity labels should be seen in a continuous space instead of the traditional discrete categorization. We propose a strategy to calculate new labels for each of the images in the dataset, and we show that these are aligned with misclassified samples too. This continuous space reflected a clear presence of strong ethnicity features on higher scores, and less evident features on lower score samples.

Our proposed formulation is straightforward and can be validated by removing samples that unbalance the dataset. For this, we propose three strategies to infer the balancing of the dataset on a continuous space, and to remove samples that break this balance. The three strategies showcased

advantages and disadvantages, and the ideal option should be decided per task. We have shown that there is more to balanced datasets, and that one can aim to further understand it.

We further release the set of samples produced by our sampling approach, and the code that can be used to replicate, reproduce and enhance the proposed methodology. In addition, algorithmic bias mitigation strategies can be applied on top of these set, hopefully boosting the performance and bias mitigation even further.

The strategies explored in this document have been making an impact in the literature. In 2024, these ideas were employed in one of the top solutions of the fairness in face recognition with synthetic data competition, with the sampling strategies being a crucial point to select the final balanced synthetic dataset. Similarly, it also contributed for a knowledge distillation problem with synthetic and mixed data.

Part V

Epilogue

Chapter 13

Conclusion

The current Chapter presents itself as the concluding Chapter of this Thesis. It serves as a quick summary of the findings, results and topics of this document, but also as reminder that small contributions might have a growing, and not initially observed impact. As such, on the topics discussed in this document, and further detailed below, there are dozens, hundreds, or even thousand of small contributions to be made. Some will not work, some will improve on the work presented in this document, and some will completely follow a different direction. Together, these efforts support the sustained growth of our research field.

Following the structure of the document, we have accomplished several contributions that span across a myriad of different biometric problems. In explainability, we contributed with a major revision, which follows our view of the field, of categorization of xAI methods through the lens of Causality. This categorization allowed us to study around 80 different works on explainable AI for biometrics, some of which were fully introduced in the following chapters. In those Chapters, we begin by proposing the use of the simplest and less effective strategy for xAI, visualization maps. However, instead of using them to explain the result of the model, we leveraged their strengths to study the model shift in the presence or absence of background information. For morphing attack detection and masked face recognition, we explored more effective interpretability methods that focus on inserting domain information directly in the training process. Through this, we became aware of certain behaviors promoted by us, and by the constraints we have inserted into the training loop. Finally, in this topic we further studied interventions on the prediction, which means that although the model remains a black-box, we aimed at creating an interpretable framework for the predicted score. We achieved outstanding results with our proposed score. Yet, there is a long way to go, especially if one aims at getting to the third rung of the xAI ladder.

As privacy concerns arise, synthetic data increases in popularity. However, how we use this data must be also a concern. Not only synthetic data comprised faces generated from scratch, but also artificial artifacts added to the image to simulate a certain event. For instance, the addition of occlusions. We dedicated two Chapters to the analysis of the performance of several models on masked and occluded face recognition datasets, and to the creation of a more robust and realistic evaluation dataset containing synthetic occlusions. Although not included in this document, such

strategy to add occlusions was part of the solution of one of the top competitors at the CVPR synthetic face recognition competition. Finally, on the topic of completely generated samples, we identified some elements that might pose as the source of the performance gap between models trained on real and synthetic datasets. Our attribute-based analysis of both datasets provided indications that certain occurrences are less likely on synthetic dataset, even small occlusions such as beards. Several other challenges arise, either from the point of view of generating new samples, or the interpretation and usefulness of such samples.

Lastly, fairness has been a hot topic in biometrics, and in face recognition. Although this was largely ignored when deploying highly biased systems, these event created a general awareness regarding the needs of these systems. Researchers have been debating if bias comes from the architecture/training process or from the data itself. We argue that it comes from both, in different shapes, but from both. We start by studying the bias in the models, and in the choice of the architecture. This is rather important as face recognition deployment scenarios are largely hardware-constraint. We identified a growing bias on both smaller and quantized models, which might be partially mitigated with quantization with out-of-the-distribution data. We finalize by studying the balance of balanced face recognition datasets. Recognizing that not all Africans, Caucasians look like the remaining individuals of their demographic group, it is fair to assume that not all hold a similar intensity of demographic-specific traits. Hence, we proposed the notion of continuous demographic labels, which allowed us to balance a balanced dataset, achieving far superior performance and better fairness metrics with less data.

The research presented in this document does not end on itself, in the same manner it was supported by many other small contributions to biometrics and similar domains, it is expected that it can be a contributing factor for future research and groundbreaking discoveries.

References

- [1] Ban dangerous facial recognition technology that amplifies racist policing | Amnesty International. [shorturl.at/zJZ09](https://www.amnesty.org/en/documents/AMR/2021/08/07/443001/). *Online*; accessed on 2021-08-07.
- [2] ISO - ISO/IEC 30107-1:2016 - Information technology — Biometric presentation attack detection — Part 1: Framework. <https://www.iso.org/standard/53227.html>. *Online*; accessed on 2021-07-30.
- [3] Wrongfully Accused by an Algorithm - The New York Times. <https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html>. *Online*; accessed on 2021-08-07.
- [4] Salem Hamed Abdurrahim, Salina Abdul Samad, and Aqilah Baseri Huddin. Review on the effects of age, gender, and race demographics on automatic face recognition. *The Visual Computer*, 34(11):1617–1630, 2018.
- [5] M Abo-Zahhad, Sabah M Ahmed, and S N Abbas. Biometric authentication based on PCG and ECG signals: present status and future directions. *Signal, Image and Video Processing*, 8(4):739–751, 2014.
- [6] Mohamed Abomhara, Sule Yildirim Yayilgan, Anne Hilde Nymoen, Marina Shalaginova, Zoltán Székely, and Ogerta Elezaj. How to Do It Right: A Framework for Biometrics Supported Border Control. In Sokratis Katsikas and Vasilios Zorkadis, editors, *E-Democracy – Safeguarding Democracy and Human Rights in the Digital Age*, pages 94–109, Cham, 2020. Springer International Publishing.
- [7] Andy Adler and Stephanie Schuckers. *Biometric Vulnerabilities, Overview*, pages 160–168. Springer US, Boston, MA, 2009.
- [8] Akshay Agarwal, Daksha Yadav, Naman Kohli, Richa Singh, Mayank Vatsa, and Afzel Noore. Face presentation attack with latex masks in multispectral videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 275–283, 2017.
- [9] Poorya Aghdaie, Baaria Chaudhary, Sobhan Soleymani, Jeremy Dawson, and Nasser M Nasrabadi. Attention aware wavelet-based detection of morphed face images. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.
- [10] F. Agraftoti, Jiexin Gao, and D. Hatzinakos. *0 heart biometrics : Theory , methods and applications*. 2012.
- [11] Mazida A Ahmed, Ridip Dev Choudhury, and Kishore Kashyap. Race estimation with deep networks. *Journal of King Saud University-Computer and Information Sciences*, 2020.

- [12] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- [13] Vitor Albiero, Krishnapriya KS, Kushal Vangara, Kai Zhang, Michael C King, and Kevin W Bowyer. Analysis of gender inequality in face recognition accuracy. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 81–89, 2020.
- [14] Tomé Albuquerque, Ana Moreira, Beatriz Barros, Diana Montezuma, Sara Pires de Oliveira, Pedro C. Neto, João C. Monteiro, Liliana Ribeiro, Sofia Gonçalves, Ana Monteiro, Isabel M. Pinto, and Jaime S. Cardoso. Quality control in digital pathology: Automatic fragment detection and counting. In *EMBC*, pages 588–593. IEEE, 2022.
- [15] Stephan Algermissen and Max Hörnlein. Person identification by footstep sound using convolutional neural networks. *Applied Mechanics*, 2(2):257–273, 2021.
- [16] Raed Alharbi, Minh N Vu, and My T Thai. Learning interpretation with explainable knowledge distillation. In *IEEE International Conference on Big Data (Big Data)*, pages 705–714. IEEE, 2021.
- [17] Miguel Almeida, Paulo Lobato Correia, and Peter Kastmand Larsen. Biofov – an open platform for forensic video analysis and biometric data extraction. In *4th International Conference on Biometrics and Forensics (IWBF)*, pages 1–6, 2016.
- [18] Fernando Alonso-Fernandez, Julian Fierrez, Javier Ortega-Garcia, Joaquin Gonzalez-Rodriguez, Hartwig Fronthaler, Klaus Kollreider, and Josef Bigun. A comparative study of fingerprint image-quality estimation methods. *IEEE Transactions on Information Forensics and Security*, 2(4):734–743, 2007.
- [19] Hammam Alshazly, Christoph Linse, Erhardt Barth, Sahar Ahmed Idris, and Thomas Martinetz. Towards explainable ear recognition systems using deep residual networks. *IEEE Access*, 9:122254–122273, 2021.
- [20] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021.
- [21] David Anghelone, Cunjian Chen, Philippe Faure, Arun Ross, and Antitza Dantcheva. Explainable thermal to visible face recognition using latent-guided generative adversarial network. In *16th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–8. IEEE, 2021.
- [22] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [23] Aqeel Anwar and Arijit Raychowdhury. Masked face recognition for secure authentication. *arXiv preprint arXiv:2008.11104*, 2020.

- [24] Gustavo Aquino, Marly GF Costa, and Cicero FF Costa Filho. Explaining one-dimensional convolutional models in human activity recognition and biometric identification tasks. *Sensors*, 22(15):5644, 2022.
- [25] Tashreef Abdullah Araf, Ayesha Siddika, Sadullah Karimi, and Md Golam Rabiul Alam. Real-time face emotion recognition and visualization using grad-cam. In *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–5. IEEE, 2022.
- [26] Shwetank Arya, Neeraj Pratap, and Karamjit Bhatia. Future of Face Recognition: A Review. *Procedia Computer Science*, 58:578–585, jan 2015.
- [27] Lei Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, page 2654–2662, Cambridge, MA, USA, 2014. MIT Press.
- [28] Fabio Bacchini and Ludovica Lorusso. Race, again: how face recognition technology reinforces racial discrimination. *Journal of information, communication and ethics in society*, 17(3):321–335, 2019.
- [29] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.
- [30] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3526–3535, 2023.
- [31] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.
- [32] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.
- [33] Sudipta Banerjee, Prateek Jaiswal, and Arun Ross. Facial de-morphing: Extracting component faces from a single morph. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2022.
- [34] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.
- [35] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017.
- [36] Yashasvi Baweja, Poojan Oza, Pramuditha Perera, and Vishal M Patel. Anomaly detection-based unknown face presentation attack detection. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2020.

- [37] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.
- [38] Ying Bian, Peng Zhang, Jingjing Wang, Chunmao Wang, and Shiliang Pu. Learning multiple explainable and generalizable cues for face anti-spoofing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2310–2314. IEEE, 2022.
- [39] Rubel Biswas, Víctor González-Castro, Eduardo Fidalgo, and Enrique Alegre. A new perceptual hashing method for verification and identity classification of occluded faces. *Image and Vision Computing*, 113:104245, 2021.
- [40] Ruud M Bolle, Jonathan H Connell, Sharath Pankanti, Nalini K Ratha, and Andrew W Senior. *Guide to biometrics*. Springer Science & Business Media, 2013.
- [41] Serena Booth, Yilun Zhou, Ankit Shah, and Julie Shah. Bayes-trex: a bayesian sampling approach to model transparency by example. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11423–11432, 2021.
- [42] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830, 2016.
- [43] Naima Bousnina, João Ascenso, Paulo Lobato Correia, and Fernando Pereira. A rise-based explainability method for genuine and impostor face verification. In *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6. IEEE, 2023.
- [44] Fadi Boutros, Naser Damer, Meiling Fang, Florian Kirchbuchner, and Arjan Kuijper. Mix-facenets: Extremely efficient face recognition networks. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.
- [45] Fadi Boutros, Naser Damer, Meiling Fang, Florian Kirchbuchner, and Arjan Kuijper. Mix-facenets: Extremely efficient face recognition networks. In *International IEEE Joint Conference on Biometrics, IJCB 2021, Shenzhen, China, August 4-7, 2021*, pages 1–8. IEEE, 2021.
- [46] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. pages 1578–1587, June 2022.
- [47] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Self-restrained triplet loss for accurate masked face recognition. *Pattern Recognition*, 124:108473, 2022.
- [48] Fadi Boutros, Naser Damer, Jan Niklas Kolf, Kiran Raja, Florian Kirchbuchner, Raghavendra Ramachandra, Arjan Kuijper, Pengcheng Fang, Chao Zhang, Fei Wang, et al. Mfr 2021: Masked face recognition competition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2021.
- [49] Fadi Boutros, Naser Damer, and Arjan Kuijper. Quantface: Towards lightweight face recognition by synthetic data low-bit quantization, 2022.
- [50] Fadi Boutros, Naser Damer, Kiran Raja, Florian Kirchbuchner, and Arjan Kuijper. Template-driven knowledge distillation for compact and accurate periocular biometrics deep-learning models. *Sensors*, 22(5):1921, 2022.

- [51] Fadi Boutros, Meiling Fang, Marcel Klemmt, Biying Fu, and Naser Damer. Cr-fiqa: face image quality assessment by learning sample relative classifiability. *arXiv preprint arXiv:2112.06592*, 2021.
- [52] Fadi Boutros, Meiling Fang, Marcel Klemmt, Biying Fu, and Naser Damer. CR-FIQA: face image quality assessment by learning sample relative classifiability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 5836–5845. IEEE, 2023.
- [53] Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fuzzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19650–19661, 2023.
- [54] Fadi Boutros, Marco Huber, Anh Thi Luu, Patrick Siebke, and Naser Damer. Sface2: Synthetic-based face recognition with w-space identity-driven sampling. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.
- [55] Fadi Boutros, Marco Huber, Patrick Siebke, Tim Rieber, and Naser Damer. Sface: Privacy-friendly and accurate face recognition using synthetic data, 2022.
- [56] Fadi Boutros, Marcel Klemmt, Meiling Fang, Arjan Kuijper, and Naser Damer. Unsupervised face recognition using unlabeled synthetic data. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023.
- [57] Fadi Boutros, Patrick Siebke, Marcel Klemmt, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Pocketnet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation. *IEEE Access*, 10:46823–46833, 2022.
- [58] Aidan Boyd, Daniel Moreira, Andrey Kuehlkamp, Kevin Bowyer, and Adam Czajka. Human saliency-driven patch-based matching for interpretable post-mortem iris recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 701–710, 2023.
- [59] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.
- [60] Ioan Buciuc and Alexandru Gacsadi. Biometrics systems and technologies: A survey. *International Journal of Computers Communications & Control*, 11(3):315–330, 2016.
- [61] Mario Bunge. *Causality and modern science*. Routledge, 2017.
- [62] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 2018.
- [63] Ruth M. J. Byrne. Précis of the rational imagination: How people create alternatives to reality. *Behavioral and Brain Sciences*, 30(5-6):439–453, 2007.
- [64] Ruth M J Byrne. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, {IJCAI-19}*, pages 6276–6282. International Joint Conferences on Artificial Intelligence Organization, 2019.

- [65] Rizhao Cai, Haoliang Li, Shiqi Wang, Changsheng Chen, and Alex C Kot. Drl-fas: A novel framework based on deep reinforcement learning for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 16:937–951, 2020.
- [66] Eduarda Caldeira, Jaime S Cardoso, Ana F Sequeira, and Pedro C Neto. Mst-kd: Multiple specialized teachers knowledge distillation for fair face recognition. pages 211–228, 2024.
- [67] Eduarda Caldeira, Pedro C Neto, Tiago Gonçalves, Naser Damer, Ana F Sequeira, and Jaime S Cardoso. Unveiling the two-faced truth: Disentangling morphed identities for face morphing detection. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 955–959. IEEE, 2023.
- [68] Eduarda Caldeira, Pedro C Neto, Marco Huber, Naser Damer, and Ana F Sequeira. Model compression techniques in biometrics applications: A survey. *Information Fusion*, 114:102657, 2025.
- [69] Eduarda Caldeira, Pedro C Neto, Marco Huber, Naser Damer, and Ana F Sequeira. Model compression techniques in biometrics applications: A survey. *Information Fusion*, 114:102657, 2025.
- [70] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pages 13–18. IEEE, 2009.
- [71] Q. Cao, L. Shen, W. Xie, O. Parkhi, and A. Zisserman. VGGFace2: A Dataset for Recognising Faces across Pose and Age. pages 67–74, 2018.
- [72] R. Cappelli, D. Maio, A. Lumini, and D. Maltoni. Fingerprint image reconstruction from standard templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1489–1503, 2007.
- [73] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, et al. Black-box access is insufficient for rigorous ai audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2254–2272, 2024.
- [74] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.
- [75] Eduardo Castro, Jose Costa Pereira, and Jaime S Cardoso. Symmetry-based regularization in deep breast cancer screening. *Medical Image Analysis*, 83:102690, 2023.
- [76] Mario Cesarelli, Fabio Martinelli, Francesco Mercaldo, and Antonella Santone. Emotion recognition from facial expression using explainable deep learning. In *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCOM/CyberSciTech)*, pages 1–6. IEEE, 2022.
- [77] Praveen Kumar Chandaliya and Neeta Nain. Child face age progression and regression using self-attention multi-scale patch gan. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.

- [78] Praveen Kumar Chandaliya, Kiran Raja, Raghavendra Ramachandra, Zahid Akhtar, and Christoph Busch. Towards inclusive face recognition through synthetic ethnicity alteration. *arXiv preprint arXiv:2405.01273*, 2024.
- [79] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018.
- [80] Angelos Chatzimparmpas, Rafael M Martins, Ilir Jusufi, and Andreas Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, 19(3):207–233, 2020.
- [81] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.
- [82] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [83] Cunjian Chen and Arun Ross. An explainable attention-guided iris presentation attack detector. In *2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 97–106, 2021.
- [84] Shen Chen, Taiping Yao, Keyue Zhang, Yang Chen, Ke Sun, Shouhong Ding, Jilin Li, Feiyue Huang, and Rongrong Ji. A dual-stream framework for 3d mask face presentation attack detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 834–841, 2021.
- [85] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018.
- [86] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings*, volume 10996 of *Lecture Notes in Computer Science*, pages 428–438. Springer, 2018.
- [87] Weiping Chen and Yongsheng Gao. Recognizing partially occluded faces from a single sample per class using string-based matching. In *European Conference on Computer Vision*, pages 496–509. Springer, 2010.
- [88] Yi Chen, Sarat C Dass, and Anil K Jain. Fingerprint quality indices for predicting authentication performance. In *International conference on audio-and video-based biometric person authentication*, pages 160–170. Springer, 2005.
- [89] Yi Chen, Sarat C Dass, and Anil K Jain. Localized iris image quality using 2-d wavelets. In *International conference on biometrics*, pages 373–381. Springer, 2006.
- [90] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

- [91] I. Chingovska and A. R. dos Anjos. On the use of client identity information for face antispoofing. *IEEE Transactions on Information Forensics and Security*, 10(4):787–796, 2015.
- [92] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012.
- [93] Jonghyun Choi, Shuowen Hu, S Susan Young, and Larry S Davis. Thermal to visible face recognition. In *Sensing Technologies for Global Health, Military Medicine, Disaster Response, and Environmental Monitoring II; and Biometric Technology for Human Identification IX*, volume 8371, page 83711L. International Society for Optics and Photonics, 2012.
- [94] Lee Ying Chong, Thian Song Ong, and Andrew Beng Jin Teoh. Feature fusions for 2.5D face recognition in Random Maxout Extreme Learning Machine. *Applied Soft Computing*, 75:358–372, feb 2019.
- [95] Anurag Chowdhury, Simon Kirchgasser, Andreas Uhl, and Arun Ross. Can a cnn automatically learn the significance of minutiae points for fingerprint matching? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 351–359, 2020.
- [96] P. Coli, G. L. Marcialis, and F. Roli. Fingerprint silicon replicas: Static and dynamic features for vitality detection using an optical capture device. *Int. J. Image Graph.*, 8:495–512, 2008.
- [97] Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.
- [98] Cynthia M. Cook, John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):32–41, 2019.
- [99] Ricardo Correia, Paulo Correia, and Fernando Pereira. Face verification explainability heatmap generation. In *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2023.
- [100] Lynne Coventry, Antonella De Angeli, and Graham Johnson. Honest it’s me! self service verification. In *The ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–4, 2003.
- [101] Antonio R Damasio. Descartes’ error: Emotion, rationality and the human brain. *New York: Putnam*, 352, 1994.
- [102] Naser Damer, Fadi Boutros, Marius Süßmilch, Meiling Fang, Florian Kirchbuchner, and Arjan Kuijper. Masked face recognition: Human vs. machine. *arXiv preprint arXiv:2103.01924*, 2021.

- [103] Naser Damer, Fadi Boutros, Marius Süßmilch, Meiling Fang, Florian Kirchbuchner, and Arjan Kuijper. Masked face recognition: Human versus machine. *IET Biometrics*, 2022.
- [104] Naser Damer, Fadi Boutros, Marius Süßmilch, Florian Kirchbuchner, and Arjan Kuijper. Extended evaluation of the effect of real and simulated masks on face recognition performance. *IET Biometrics*, 10(5):548–561, 2021.
- [105] Naser Damer, Meiling Fang, Patrick Siebke, Jan Niklas Kolf, Marco Huber, and Fadi Boutros. Mordiff: Recognition vulnerability and attack detectability of face morphing attacks created by diffusion autoencoders. *arXiv preprint arXiv:2302.01843*, 2023.
- [106] Naser Damer, Jonas Henry Grebe, Cong Chen, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. The effect of wearing a mask on face recognition performance: an exploratory study. In *BIOSIG 2020 - Proceedings of the 19th International Conference of the Biometrics Special Interest Group, online, 16.-18. September 2020*, volume P-306 of *LNI*, pages 1–10. Gesellschaft für Informatik e.V., 2020.
- [107] Naser Damer, Jonas Henry Grebe, Cong Chen, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. The effect of wearing a mask on face recognition performance: an exploratory study. In *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6. IEEE, 2020.
- [108] Naser Damer, Jonas Henry Grebe, Cong Chen, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. The effect of wearing a mask on face recognition performance: an exploratory study. In *BIOSIG 2020 - Proceedings of the 19th International Conference of the Biometrics Special Interest Group, online, 16.-18. September 2020*, volume P-306 of *LNI*, pages 1–10. Gesellschaft für Informatik e.V., 2020.
- [109] Naser Damer, César Augusto Fontanillo López, Meiling Fang, Noémie Spiller, Minh Vu Pham, and Fadi Boutros. Privacy-friendly synthetic data for the development of face morphing attack detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1606–1617, 2022.
- [110] Naser Damer, César Augusto Fontanillo López, Meiling Fang, Noémie Spiller, Minh Vu Pham, and Fadi Boutros. Privacy-friendly synthetic data for the development of face morphing attack detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1606–1617, June 2022.
- [111] Naser Damer, Alexandra Mosegui Saladie, Andreas Braun, and Arjan Kuijper. Morgan: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network. In *2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–10. IEEE, 2018.
- [112] Naser Damer, Noémie Spiller, Meiling Fang, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. Pw-mad: Pixel-wise supervision for generalized face morphing attack detection. In *International Symposium on Visual Computing*, pages 291–304. Springer, 2021.
- [113] Laurine Dargaud, Mathias Ibsen, Juan Tapia, and Christoph Busch. A principal component analysis-based approach for single morphing attack detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 683–692, 2023.
- [114] Richard A Davis, Keh-Shin Lii, and Dimitris N Politis. Remarks on some nonparametric estimates of a density function. *Selected Works of Murray Rosenblatt*, pages 95–100, 2011.

- [115] Petter Eilif De Lange, Borger Melsom, Christian Bakke Vennerød, and Sjur Westgaard. Explainable ai for credit assessment in banks. *Journal of Risk and Financial Management*, 15(12):556, 2022.
- [116] Ivan DeAndres-Tame, Ruben Tolosana, Pietro Melzi, Ruben Vera-Rodriguez, Minchul Kim, Christian Rathgeb, Xiaoming Liu, Luis F Gomez, Aythami Morales, Julian Fierrez, et al. Second frcsyn-ongoing: Winning solutions and post-challenge analysis to improve face recognition with synthetic data. *Information Fusion*, 120:103099, 2025.
- [117] Ivan DeAndres-Tame, Ruben Tolosana, Pietro Melzi, Ruben Vera-Rodriguez, Minchul Kim, Christian Rathgeb, Xiaoming Liu, Aythami Morales, Julian Fierrez, Javier Ortega-Garcia, et al. Fracsyn challenge at cvpr 2024: Face recognition challenge in the era of synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3173–3183, 2024.
- [118] Ivan DeAndres-Tame, Ruben Tolosana, Pietro Melzi, Rubén Vera-Rodríguez, Minchul Kim, Christian Rathgeb, Xiaoming Liu, Aythami Morales, Julian Fierrez, Javier Ortega-Garcia, Zhizhou Zhong, Yuge Huang, Yuxi Mi, Shouhong Ding, Shuigeng Zhou, Shuai He, Lingzhi Fu, Heng Cong, Rongyu Zhang, Zhihong Xiao, Evgeny Smirnov, Anton Pimenov, Aleksei Grigorev, Denis Timoshenko, Kaleb Mesfin Asfaw, Cheng-Yaw Low, Hao Liu, Chuyi Wang, Qing Zuo, Zhixiang He, Hatef Otroschi-Shahreza, Anjith George, Alexander Unnervik, Parsa Rahimi, Sébastien Marcel, Pedro C. Neto, Marco Huber, Jan Niklas Kolf, Naser Damer, Fadi Boutros, Jaime S. Cardoso, Ana Filipa Sequeira, Andrea Atzori, Gianni Fenu, Mirko Marras, Vitomir Struc, Jiang Yu, Zhangjie Li, Jichun Li, Weisong Zhao, Zhen Lei, Xiangyu Zhu, Xiao-Yu Zhang, Bernardo Biesseck, Pedro Vidal, Luiz Coelho, Roger Granada, and David Menotti. Second edition fracsyn challenge at CVPR 2024: Face recognition challenge in the era of synthetic data. *CoRR*, abs/2404.10378, 2024.
- [119] Frank Deane, Kate Barrelle, Ron Henderson, and Doug Mahar. Perceived acceptability of biometric security systems. *Computers & Security*, 14(3):225–231, 1995.
- [120] Lisa DeBruine. debruine/webmorph: Beta release 2. *Zenodo* [https://doi.org/10, 5281](https://doi.org/10.5281/2018), 2018.
- [121] Lisa DeBruine and Benedict Jones. Face research lab london set, Apr 2021.
- [122] Ashley Deeks. The judicial demand for explainable artificial intelligence. *Columbia Law Review*, 119(7):1829–1850, 2019.
- [123] J Deng, W Dong, R Socher, L.-J. Li, K Li, and L Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- [124] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. Masked face recognition challenge: The insightface track report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1437–1444, 2021.
- [125] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [126] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern*

- Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE, 2019.
- [127] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.
 - [128] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019.
 - [129] Dorothy E Denning. Why i love biometrics: It is ‘liveness,’ not secrecy, that counts.
 - [130] T. I. Dhamecha, A. Nigam, R. Singh, and M. Vatsa. Disguise detection and face recognition in visible and thermal spectrums. In *2013 International Conference on Biometrics (ICB)*, pages 1–8, 2013.
 - [131] K. Dharavath, F. A. Talukdar, and R. H. Laskar. Study on biometric authentication systems, challenges and future trends: A review. In *2013 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–7, 2013.
 - [132] Moises Diaz, Miguel A Ferrer, and Gennaro Vessio. Explainable offline automatic signature verifier to support forensic handwriting examiners. *Available at SSRN 4395568*.
 - [133] Changxing Ding and Dacheng Tao. Robust face recognition via multimodal deep face representation. *IEEE Transactions on Multimedia*, 17(11):2049–2058, 2015.
 - [134] Changxing Ding and Dacheng Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):1002–1014, 2017.
 - [135] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. Deformable protopnet: An interpretable image classifier using deformable prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10265–10275, 2022.
 - [136] Samuel Dooley, Rhea Sukthanker, John P. Dickerson, Colin White, Frank Hutter, and Micah Goldblum. Rethinking bias mitigation: Fairer architectures make for fairer face recognition. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
 - [137] Samuel Dooley, George Z. Wei, Tom Goldstein, and John Dickerson. Robustness disparities in face detection. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.
 - [138] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
 - [139] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4829–4837, 2016.

- [140] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020.
- [141] Yuting Du, Tong Qiao, Ming Xu, and Ning Zheng. Towards Face Presentation Attack Detection Based on Residual Color Texture Representation. *Security and Communication Networks*, 2021:6652727, 2021.
- [142] Yueqi Duan, Jiwen Lu, and Jie Zhou. Uniformface: Learning deep equidistributed representation for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2019.
- [143] Rudresh Dwivedi, Ritesh Kumar, Deepak Chopra, Pranay Kothari, and Manjot Singh. An efficient ensemble explainable ai (xai) approach for morphed face detection. *arXiv preprint arXiv:2304.14509*, 2023.
- [144] Eran Eiding, Roe Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Trans. Information Forensics and Security*, 9(12):2170–2179, 2014.
- [145] Mohamad El-Abed, Romain Giot, Baptiste Hemery, and Christophe Rosenberger. A study of users’ acceptance and satisfaction of biometric systems. In *44th Annual 2010 IEEE International Carnahan Conference on Security Technology*, pages 170–178. IEEE, 2010.
- [146] Mohamad El-Abed, Romain Giot, Baptiste Hemery, and Christophe Rosenberger. Evaluation of biometric systems: A study of users’ acceptance and satisfaction. *International Journal of Biometrics*, 4(3):265–290, 2012.
- [147] Stephen J Elliott, Sarah A Massie, and Mathias J Sutton. The perception of biometric technology: A survey. In *2007 IEEE Workshop on Automatic Identification Advanced Technologies*, pages 259–264. IEEE, 2007.
- [148] Mustafa Ekrem Erakın, Uğur Demir, and Hazım Kemal Ekenel. On recognizing occluded faces in the wild. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2021.
- [149] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.
- [150] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council.
- [151] European Parliament, Council of the European Union. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS COM/2021/206 final. <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206>. *Online*; accessed on 2021-08-07.
- [152] European Parliament, Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and

- repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). <http://data.europa.eu/eli/reg/2016/679/oj>, 2016. *Online*; accessed on 2021-08-07.
- [153] Meiling Fang, Fadi Boutros, and Naser Damer. Unsupervised face morphing attack detection via self-paced anomaly detection. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2022.
- [154] Meiling Fang, Fadi Boutros, Arjan Kuijper, and Naser Damer. Partial attack supervision and regional weighted inference for masked face presentation attack detection. In *16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021, Jodhpur, India, December 15-18, 2021*, pages 1–8. IEEE, 2021.
- [155] Meiling Fang, Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. Iris presentation attack detection by attention-based and deep pixel-wise binary supervision network. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.
- [156] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Real masks and spoof faces: On the masked face presentation attack detection. *Pattern Recognition*, page 108398, 2021.
- [157] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Learnable multi-level frequency decomposition and hierarchical attention mechanism for generalized face presentation attack detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3722–3731, 2022.
- [158] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Real masks and spoof faces: On the masked face presentation attack detection. *Pattern Recognit.*, 123:108398, 2022.
- [159] Sajad Farokhi, Jan Flusser, and Usman Ullah Sheikh. Near infrared face recognition: A literature survey. *Computer Science Review*, 21:1–17, 2016.
- [160] Kelwin Fernandes, Jaime S Cardoso, and Birgitte Schmidt Astrup. A deep learning approach for the forensic evaluation of sexual assault. *Pattern Analysis and Applications*, 21(3):629–640, 2018.
- [161] FICO. Introduction to scorecard for fico model builder. Technical report, 2006.
- [162] Danilo Franco, Nicolo Navarin, Michele Donini, Davide Anguita, and Luca Oneto. Deep fair models for complex data: Graphs labeling and explainable face recognition. *Neuro-computing*, 470:318–334, 2022.
- [163] Danilo Franco, Luca Oneto, Nicolò Navarin, and Davide Anguita. Learn and visually explain deep fair models: an application to face recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2021.
- [164] Gil Friedrich and Yehezkel Yeshurun. Seeing People in the Dark: Face Recognition in Infrared Images. In Heinrich H Bülthoff, Christian Wallraven, Seong-Whan Lee, and Tomaso A Poggio, editors, *Biologically Motivated Computer Vision*, pages 348–359, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

- [165] Frontex. Best practice technical guidelines for automated border control (abc) systems. 2017.
- [166] Biying Fu and Naser Damer. Explainability of the implications of supervised and unsupervised face image quality estimations through activation map variation analyses in face recognition models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 349–358, 2022.
- [167] Biying Fu and Naser Damer. Towards explaining demographic bias through the eyes of face recognition models. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2022.
- [168] Biying Fu, Florian Kirchbuchner, and Naser Damer. The effect of wearing a face mask on face image quality. *CoRR*, abs/2110.11283, 2021.
- [169] Biying Fu, Florian Kirchbuchner, and Naser Damer. The effect of wearing a face mask on face image quality. In *16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021, Jodhpur, India, December 15-18, 2021*, pages 1–8. IEEE, 2021.
- [170] Biying Fu, Florian Kirchbuchner, and Naser Damer. The effect of wearing a face mask on face image quality. In *16th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.
- [171] Robert D Furberg, Travis Taniguchi, Brian Aagaard, Alexa M Ortiz, Meghan Hegarty-Craver, Kristin H Gilchrist, and Ty A Ridenour. Biometrics and Policing: A Protocol for Multichannel Sensor Data Collection and Exploratory Analysis of Contextualized Psychophysiological Response During Law Enforcement Operations. *JMIR Research Protocols*, 6(3):e44, mar 2017.
- [172] J. Galbally, S. Marcel, and J. Fierrez. Biometric antispooofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552, 2014.
- [173] Javier Galbally, R. Plamondon, Julian Fierrez, and J. Ortega-Garcia. Synthetic on-line signature generation. part i: Methodology and algorithms. *Pattern Recognit.*, 45:2610–2621, 2012.
- [174] Javier Galbally Herrero. Vulnerabilities and attack protection in security systems based on biometric recognition. 2009.
- [175] Xiufeng Gao, Stan Z Li, Rong Liu, and Peiren Zhang. Standardization of face image sample quality. In *International Conference on Biometrics*, pages 242–251. Springer, 2007.
- [176] Raul Vicente Garcia, Lukasz Wandzik, Louisa Grabner, and Joerg Krueger. The harms of demographic bias in deep face recognition research. In *2019 International Conference on Biometrics (ICB)*, pages 1–6, 2019.
- [177] C. Garvie, Georgetown University. Center on Privacy, Technology, and Georgetown University. Law Center. Center on Privacy & Technology. *The Perpetual Line-up: Unregulated Police Face Recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016.
- [178] Mengyue Geng, Peixi Peng, Yangru Huang, and Yonghong Tian. Masked face recognition with generative data augmentation and domain constrained ranking. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2246–2254, 2020.

- [179] Angelo Genovese, Vincenzo Piuri, and Fabio Scotti. Towards explainable face aging with generative adversarial networks. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3806–3810, 2019.
- [180] Markos Georgopoulos, James Oldfield, Mihalis A Nicolaou, Yannis Panagakis, and Maja Pantic. Mitigating Demographic Bias in Facial Datasets with Style-Based Multi-attribute Transfer. *International Journal of Computer Vision*, 129(7):2288–2307, 2021.
- [181] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.
- [182] Marta Gomez-Barrero, Pawel Drozdowski, Christian Rathgeb, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Naser Damer, Jannis Priesnitz, Nicholas W. D. Evans, and Christoph Busch. Biometrics in the era of COVID-19: challenges and opportunities. *CoRR*, abs/2102.09258, 2021.
- [183] Marta Gomez-Barrero, Pawel Drozdowski, Christian Rathgeb, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Naser Damer, Jannis Priesnitz, Nicholas W. D. Evans, and Christoph Busch. Biometrics in the era of COVID-19: challenges and opportunities. *CoRR*, abs/2102.09258, 2021.
- [184] Marta Gomez-Barrero, Pawel Drozdowski, Christian Rathgeb, Jose Patino, Massimiliano Todisco, Andras Nautsch, Naser Damer, Jannis Priesnitz, Nicholas Evans, and Christoph Busch. Biometrics in the era of covid-19: challenges and opportunities. *arXiv preprint arXiv:2102.09258*, 2021.
- [185] Tiago Gonçalves, Isabel Rio-Torto, Luís F Teixeira, and Jaime S Cardoso. A survey on attention mechanisms for medical applications: are we moving towards better algorithms? 2022.
- [186] Sixue Gong, Xiaoming Liu, and Anil K Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *European Conference on Computer Vision*, pages 330–347. Springer, 2020.
- [187] Sixue Gong, Xiaoming Liu, and Anil K. Jain. Jointly de-biasing face recognition and demographic attribute estimation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, pages 330–347. Springer, 2020.
- [188] Sixue Gong, Xiaoming Liu, and Anil K. Jain. Mitigating face recognition bias via group adaptive classifier. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3414–3424. Computer Vision Foundation / IEEE, 2021.
- [189] Ester Gonzalez-Sosa, Julian Fierrez, Ruben Vera-Rodriguez, and Fernando Alonso-Fernandez. Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation, and COTS Evaluation. *IEEE Transactions on Information Forensics and Security*, 13(8):2001–2014, 2018.

- [190] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.
- [191] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.
- [192] Dmitry Gorodnichy, Svetlana Yanushkevich, and Vlad Shmerko. Automated border control: Problem formalization. In *2014 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, pages 118–125. IEEE, 2014.
- [193] Margarida Gouveia, Eduardo Meca, Ana Rebelo, and Jaime S. Cardoso. Deep minutiae fingerprint extraction using equivariance priors. In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS)*, 2023.
- [194] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*. National Institute of Standards and Technology, 2019.
- [195] Patrick Grother and Elham Tabassi. Performance of biometric quality measures. *IEEE transactions on pattern analysis and machine intelligence*, 29(4):531–543, 2007.
- [196] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.
- [197] Han Guo, Nazneen Fatema Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. Fastif: Scalable influence functions for efficient model interpretation and debugging. *arXiv preprint arXiv:2012.15781*, 2020.
- [198] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.
- [199] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world. In Jan P. Allebach, Zhigang Fan, and Qian Lin, editors, *Imaging and Multimedia Analytics in a Web and Mobile World 2016, San Francisco, CA, USA, February 14-18, 2016*, pages 1–6. Society for Imaging Science and Technology, 2016.
- [200] Matthew Gwilliam, Srinidhi Hegde, Lade Tinubu, and Alex Hanson. Rethinking common assumptions to mitigate racial bias in face recognition datasets. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021*, pages 4106–4115. IEEE, 2021.
- [201] Chunrui Han, Shiguang Shan, Meina Kan, Shuzhe Wu, and Xilin Chen. Personalized convolution for face recognition. *International Journal of Computer Vision*, pages 1–19, 2022.

- [202] Md. Mehedi Hasan, Md. Salah Uddin Yusuf, Tanbin Islam Rohan, and Shidhartho Roy. Efficient two stage approach to detect face liveness : Motion based and deep learning based. In *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, pages 1–6, 2019.
- [203] Peter Hase, Chaofan Chen, Oscar Li, and Cynthia Rudin. Interpretable image recognition with hierarchical prototypes. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 32–40, 2019.
- [204] Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297 – 310, 1986.
- [205] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [206] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.
- [207] Mingjie He, Jie Zhang, Shiguang Shan, Xiao Liu, Zhongqin Wu, and Xilin Chen. Locality-aware channel-wise dropout for occluded face recognition. *IEEE Trans. on Image Proc.*, 31:788–798, 2021.
- [208] Ran He, Wei-Shi Zheng, Bao-Gang Hu, and Xiang-Wei Kong. A regularized correntropy framework for robust pattern recognition. *Neural computation*, 23(8):2074–2100, 2011.
- [209] S. Hemalatha. A systematic review on fingerprint based biometric authentication system. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–4, 2020.
- [210] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016.
- [211] Lisa Anne Marie Hendricks. *Visual Understanding through Natural Language*. University of California, Berkeley, 2019.
- [212] Perry Hinton. Implicit stereotypes and the predictive brain: cognition and culture in “biased” person perception. *Palgrave Communications*, 3(1):17086, 2017.
- [213] Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.
- [214] Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.
- [215] Qingqiao Hu, Siyang Yin, Huiyang Ni, and Yisiyuan Huang. An end to end deep neural network for iris recognition. *Procedia Computer Science*, 174:505–517, 2020.
- [216] Baojin Huang, Zhongyuan Wang, Kui Jiang, Qin Zou, Xin Tian, Tao Lu, and Zhen Han. Joint segmentation and identification feature learning for occlusion face recognition. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2022.

- [217] Baojin Huang, Zhongyuan Wang, Guangcheng Wang, Kui Jiang, Kangli Zeng, Zhen Han, Xin Tian, and Yuhong Yang. When face recognition meets occlusion: A new benchmark. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4240–4244. IEEE, 2021.
- [218] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.
- [219] Linzhi Huang, Mei Wang, Jiahao Liang, Weihong Deng, Hongzhi Shi, Dongchao Wen, Yingjie Zhang, and Jian Zhao. Gradient attention balance network: Mitigating face recognition racial bias via gradient attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 38–47. IEEE, 2023.
- [220] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020.
- [221] Marco Huber, Fadi Boutros, Florian Kirchbuchner, and Naser Damer. Mask-invariant face recognition through template-level knowledge distillation. In *16th IEEE Int. Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.
- [222] Marco Huber, Fadi Boutros, Anh Thi Luu, Kiran Raja, Raghavendra Ramachandra, Naser Damer, Pedro C Neto, Tiago Gonçalves, Ana F Sequeira, Jaime S Cardoso, et al. Syn-mad 2022: Competition on face morphing attack detection based on privacy-aware synthetic training data. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2022.
- [223] Marco Huber, Meiling Fang, Fadi Boutros, and Naser Damer. Are explainability tools gender biased? a case study on face presentation attack detection. *arXiv preprint arXiv:2304.13419*, 2023.
- [224] Marco Huber, Anh Thi Luu, Fadi Boutros, Arjan Kuijper, and Naser Damer. Bias and diversity in synthetic-based face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6215–6226, 2024.
- [225] Marco Huber, Anh Thi Luu, Philipp Terhörst, and Naser Damer. Efficient explainable face verification based on similarity score argument backpropagation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4736–4745, 2024.
- [226] Marco Huber, Philipp Terhörst, Florian Kirchbuchner, Naser Damer, and Arjan Kuijper. Stating comparison score uncertainty and verification decision confidence towards transparent face recognition. In *33rd British Machine Vision Conference 2022, BMVC 2022, November 21-24, 2022*. BMVA Press, 2022.
- [227] Marco Huber, Philipp Terhörst, Anh Thi Luu, Florian Kirchbuchner, and Naser Damer. Verification of sitter identity across historical portrait paintings by confidence-aware face recognition. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022.

- [228] Jamal Hussain Shah, Muhammad Sharif, Mudassar Raza, Marryam Murtaza, and Saeed-Ur-Rehman. Robust Face Recognition Technique under Varying Illumination. *Journal of Applied Research and Technology*, 13(1):97–105, 2015.
- [229] Frederik Hvilshøj, Alexandros Iosifidis, and Ira Assent. Ecinn: Efficient counterfactuals from invertible neural networks. *arXiv preprint arXiv:2103.13701*, 2021.
- [230] ISO/IEC 19795-1:2006 Information technology — Biometric performance testing and reporting. Standard, International Organization for Standardization, 2016.
- [231] ISO/IEC JTC1 SC37. Information Technology - Biometrics - Presentation attack detection Part 3: Testing and Reporting. *ISO Int. Organization for Standardization*, 2017.
- [232] Anil K. Jain, Ruud Bolle, and Sharath Pankanti. *Biometrics: Personal Identification in Networked Society*. Kluwer Academic Publishers, Norwell, Massachusetts, 1999.
- [233] Anil K Jain, Sarat C Dass, and Karthik Nandakumar. Soft Biometric Traits for Personal Recognition Systems. In David Zhang and Anil K Jain, editors, *Biometric Authentication*, pages 731–738, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.
- [234] Anil K Jain, Debayan Deb, and Joshua J Engelsma. Biometrics: Trust, but verify. *arXiv preprint arXiv:2105.06625*, 2021.
- [235] Anil K. Jain, Patrick Flynn, and Arun A. Ross. *Handbook of Biometrics*. Springer Publishing Company, Incorporated, 1st edition, 2010.
- [236] Anil K. Jain and Karthik Nandakumar. Biometric System Design, Overview. *Encyclopedia of Biometrics*, pages 135–140, 2009.
- [237] Anil K Jain, Karthik Nandakumar, and Arun Ross. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters*, 79:80–105, 2016.
- [238] Anil K. Jain, Arun A. Ross, and Karthik Nandakumar. *Introduction to Biometrics*. Springer US, 2011.
- [239] Rubal Jain and Chander Kant. Attacks on biometric systems: an overview. *International Journal of Advances in Scientific Research*, 1(07):283–288, 2015.
- [240] Ayush Jaiswal, Rex Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Unsupervised Adversarial Invariance. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [241] Ayush Jaiswal, Shuai Xia, Iacopo Masi, and Wael AbdAlmageed. Ropad: Robust presentation attack detection through unsupervised adversarial invariance. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.
- [242] Siddharth D. Jaiswal, Karthikeya Duggirala, Abhisek Dash, and Animesh Mukherjee. Two-face: Adversarial audit of commercial face recognition systems. In Ceren Budak, Meeyoung Cha, and Daniele Quercia, editors, *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022*, pages 381–392. AAAI Press, 2022.

- [243] Govind Jeevan, Geevar C Zacharias, Madhu S Nair, and Jeny Rajan. An empirical study of the impact of masks on face recognition. *Pattern Recognition*, 122:108308, 2022.
- [244] Govind Jeevan, Geevar C Zacharias, Madhu S Nair, and Jeny Rajan. An empirical study of the impact of masks on face recognition. *Pattern Recognition*, 122:108308, 2022.
- [245] Ziwei Ji, YU Tiezheng, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating llm hallucination via self reflection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- [246] Haoran Jiang and Dan Zeng. Explainable face recognition based on accurate facial compositions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1503–1512, 2021.
- [247] Wang Jiang, Lin Ye, Zhang Yi, and Cheng Peng. A new occluded face recognition framework with combination of both deocclusion and feature filtering methods. *Multimedia Tools and Applications*, pages 1–30, 2022.
- [248] Weina Jin, Xiaoxiao Li, Mostafa Fatehi, and Ghassan Hamarneh. Guidelines and evaluation of clinical explainable ai in medical image analysis. *Medical image analysis*, 84:102684, 2023.
- [249] P. A. Johnson, B. Tan, and S. Schuckers. Multimodal fusion vulnerability to non-zero effort (spoof) imposters. In *2010 IEEE International Workshop on Information Forensics and Security*, pages 1–5, 2010.
- [250] Peter Johnson, Richard Lazarick, Emanuela Marasco, Elaine Newton, Arun Ross, and Stephanie Schuckers. Biometric Liveness Detection: Framework and Metrics. In *International Biometric Performance Conference (IBPC)*, 2012.
- [251] Thaddeus L Johnson, Natasha N Johnson, Denise McCurdy, and Michael S Olajide. Facial recognition systems in policing and racial disparities in arrests. *Government Information Quarterly*, 39(4):101753, 2022.
- [252] Laurie A Jones, Annie I Antón, and Julia B Earp. Towards understanding user perceptions of authentication technologies. In *Proceedings of the 2007 ACM workshop on Privacy in electronic society*, pages 91–98, 2007.
- [253] Indu Joshi, Riya Kothari, Ayush Utkarsh, Vinod K Kurmi, Antitza Dantcheva, Sumantra Dutta Roy, and Prem Kumar Kalra. Explainable fingerprint roi segmentation using monte carlo dropout. In *IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 60–69, 2021.
- [254] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 290–306, 2018.
- [255] Karl Ricanek Jr. and Tamirat Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2006)*, 10-12 April 2006, Southampton, UK, pages 341–345. IEEE Computer Society, 2006.

- [256] Hong-Gyu Jung, Sin-Han Kang, Hee-Dong Kim, Dong-Ok Won, and Seong-Wan Lee. Counterfactual explanation based on gradual construction for deep networks. *arXiv preprint arXiv:2008.01897*, 2020.
- [257] Amba Kak. Regulating Biometrics: Global Approaches and Urgent Questions. Technical report, AI Now Institute, sep 2020.
- [258] Nathan D Kalka, Jinyu Zuo, Natalia A Schmid, and Bojan Cukic. Image quality assessment for iris biometric. In *Biometric technology for human identification III*, volume 6202, page 62020D. International Society for Optics and Photonics, 2006.
- [259] Margot E Kaminski. The right to explanation, explained. *Berkeley Tech. LJ*, 34:189, 2019.
- [260] Alperen Kantarcı, Hasan Dertli, and Hazım Kemal Ekenel. Shuffled patch-wise supervision for presentation attack detection. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2021.
- [261] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 1547–1557. IEEE, 2021.
- [262] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.
- [263] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.
- [264] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
- [265] G. Kaur, Gurpreet Singh, and V. Kumar. A review on biometric recognition. In *BSBT 2014*, 2014.
- [266] Frank Keil, Leonid Rozenblit, and Candice Mills. What lies beneath? understanding the limits of understanding. *Thinking and seeing: Visual metacognition in adults and children*, pages 227–49, 2004.
- [267] Saeed Khorram and Li Fuxin. Cycle-consistent counterfactuals by latent transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10203–10212, 2022.
- [268] Been Kim and Finale Doshi-Velez. Interpretable machine learning: the fuss, the concrete and the questions. *ICML Tutorial on interpretable machine learning*, 2017.
- [269] Minchul Kim, Anil K. Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition, 2022.

- [270] Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Dcfac: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12715–12725, 2023.
- [271] D. E. King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.
- [272] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [273] Dakshina Ranjan Kisku and Rinku Datta Rakshit. Face spoofing and counter-spoofing: a survey of state-of-the-art algorithms. *Transactions on Machine Learning and Artificial Intelligence*, 5(2):31, 2017.
- [274] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.
- [275] Sanneke Kloppenburg and Irma van der Ploeg. Securing Identities: Biometric Technologies and the Enactment of Human Bodily Differences. *Science as Culture*, 29(1):57–76, 2020.
- [276] Martin Knoche, Torben Teepe, Stefan Hörmann, and Gerhard Rigoll. Explainable model-agnostic similarity and confidence in face verification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 711–718, 2023.
- [277] Jan Niklas Kolf, Tim Rieber, Jurek Elliesen, Fadi Boutros, Arjan Kuijper, and Naser Damer. Identity-driven three-player generative adversarial network for synthetic-based face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 806–816, 2023.
- [278] Jukka Komulainen, Abdenour Hadid, and Matti Pietikäinen. Context based face anti-spoofing. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2013.
- [279] Pavel Korshunov, Anubhav Jain, and Sébastien Marcel. Custom attribution loss for improving generalization and interpretability of deepfake detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8972–8976. IEEE, 2022.
- [280] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Empirically analyzing the effect of dataset biases on deep face recognition systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
- [281] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2261–2268, 2019.
- [282] N. Kose and J. Dugelay. Reflectance analysis based countermeasure technique to detect face mask attacks. In *2013 18th International Conference on Digital Signal Processing (DSP)*, pages 1–6, 2013.

- [283] Ranjana Koshy and Ausif Mahmood. Enhanced Deep Learning Architectures for Face Liveness Detection for Static and Video Sequences. *Entropy*, 22(10), 2020.
- [284] Emine Krichen, Sonia Garcia-Salicetti, and Bernadette Dorizzi. A new probabilistic iris quality measure for comprehensive noise detection. In *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6. IEEE, 2007.
- [285] KS Krishnapriya, Vítor Albiero, Kushal Vangara, Michael C King, and Kevin W Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020.
- [286] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.
- [287] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012.
- [288] Andrey Kuehlkamp, Aidan Boyd, Adam Czajka, Kevin Bowyer, Patrick Flynn, Dennis Chute, and Eric Benjamin. Interpretable deep learning-based forensic iris segmentation and recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 359–368, 2022.
- [289] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.
- [290] Kenneth Lai, Helder C. R. Oliveira, Ming Hou, Svetlana N. Yanushkevich, and Vlad P. Shmerko. Risk, trust, and bias: Causal regulators of biometric-enabled decision support. *IEEE Access*, 8:148779–148792, 2020.
- [291] Sanghoon Lee, Chulhan Lee, and Jaihie Kim. Model-based quality estimation of fingerprint images. In *International Conference on Biometrics*, pages 229–235. Springer, 2006.
- [292] Sangrok Lee, Eunsoo Park, Hongsuk Yi, and Sang Hun Lee. Strdan: Synthetic-to-real domain adaptation network for vehicle re-identification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 2590–2597. Computer Vision Foundation / IEEE, 2020.
- [293] Pinja Lehtonen and Pami Aalto. Smart and secure borders through automated border control systems in the EU? The views of political stakeholders in the Member States. *European Security*, 26(2):207–225, 2017.
- [294] Elad Levi, Tete Xiao, Xiaolong Wang, and Trevor Darrell. Rethinking preventing class-collapsing in metric learning with margin-based losses. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10296–10305. IEEE, 2021.
- [295] Chenyu Li, Shiming Ge, Daichi Zhang, and Jia Li. Look through masks: Towards masked face recognition with de-occlusion distillation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3016–3024, 2020.

- [296] Haoliang Li, Peisong He, Shiqi Wang, Anderson Rocha, Xinghao Jiang, and Alex C. Kot. Learning generalized deep feature representation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(10):2639–2652, 2018.
- [297] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C. Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7):1794–1809, 2018.
- [298] Pengyu Li. Bionet: A biologically-inspired network for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10344–10354, 2023.
- [299] Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. Sharp multiple instance learning for deepfake video detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1864–1872, 2020.
- [300] Yande Li, Kun Guo, Yonggang Lu, and Li Liu. Cropping and attention based approach for masked face recognition. *Applied Intelligence*, 51(5):3012–3025, 2021.
- [301] Suk-Young Lim, Dong-Kyu Chae, and Sang-Chul Lee. Detecting deepfake voice using explainable deep learning techniques. *Applied Sciences*, 12(8):3926, 2022.
- [302] Xiaofeng Lin, Seungbae Kim, and Jungseock Joo. Fairgrape: Fairness-aware gradient pruning method for face attribute classification. In *European Conference on Computer Vision*, pages 414–432. Springer, 2022.
- [303] Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247–266, 1990.
- [304] Decheng Liu, Xinbo Gao, Chunlei Peng, Nannan Wang, and Jie Li. Heterogeneous face interpretable disentangled representation for joint face recognition and synthesis. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [305] Shanhong Liu. • Global biometrics market revenue 2018-2027 | Statista. <https://www.statista.com/statistics/1048705/worldwide-biometrics-market-revenue/>, 2020. *Online*; accessed on 2021-08-07.
- [306] Shuhua Liu, Yu Song, Mengyu Zhang, Jianwei Zhao, Shihao Yang, and Kun Hou. An Identity Authentication Method Combining Liveness Detection and Face Recognition. *Sensors*, 19(21), 2019.
- [307] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphreface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.
- [308] Xuan Liu, Xiaoguang Wang, and Stan Matwin. Improving the interpretability of deep neural networks with knowledge distillation. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 905–912. IEEE, 2018.
- [309] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 389–398, 2018.

- [310] Yaojie Liu, Joel Stehouwer, and Xiaoming Liu. On disentangling spoof trace for generic face anti-spoofing. In *European Conference on Computer Vision*, pages 406–422. Springer, 2020.
- [311] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of Int. Conf. on Computer Vision (ICCV)*, December 2015.
- [312] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.
- [313] Mohamed Loey, Gunasekaran Manogaran, Mohamed Hamed N Taha, and Nour Eldeen M Khalifa. A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the covid-19 pandemic. *Measurement*, 167:108288, 2021.
- [314] Max Maria Losch, Mario Fritz, and Bernt Schiele. Semantic bottlenecks: Quantifying and improving inspectability of deep representations. In *Pattern Recognition: 42nd DAGM German Conference, DAGM GCPR, Tübingen, Germany, Proceedings 42*, pages 15–29. Springer, 2020.
- [315] Giulio Lovisotto, Raghav Malik, Ivo Služanovic, Marc Roeschlin, Paul Trueman, and Ivan Martinovic. Mobile biometrics in financial services: A five factor framework. *University of Oxford, Oxford, UK*, 2017.
- [316] Boyu Lu, Jun-Cheng Chen, Carlos D Castillo, and Rama Chellappa. An experimental evaluation of covariates effects on unconstrained face verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):42–55, 2019.
- [317] Yuhang Lu and Touradj Ebrahimi. Explanation of face recognition via saliency maps. *arXiv preprint arXiv:2304.06118*, 2023.
- [318] Michael J. Lyons. "Excavating AI" Re-excavated: Debunking a Fallacious Account of the JAFFE Dataset. n.b. All JAFFE images in this article are subject to specific terms of use and may not be reused without permission, regardless of the license applied to the document as a whole., July 2021.
- [319] Michael J Lyons, Miyuki Kamachi, and Jiro Gyoba. Coding Facial Expressions with Gabor Wavelets (IVC Special Issue), September 2020. This manuscript is a modified version of a conference article, that was invited for publication in a special issue of Image and Vision Computing dedicated to a selection of articles from the IEEE Face & Gesture 1998 conference. The special issue never materialized.
- [320] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.
- [321] Ahmed Mahfouz, Tarek M Mahmoud, and Ahmed Sharaf Eldin. A survey on behavioral biometric authentication on smartphones. *Journal of Information Security and Applications*, 37:28–37, 2017.
- [322] Satya Mallick. Face morph using opencv — c++ / python | learnopencv, Mar 2016.
- [323] Rafael M Mamede, Pedro C Neto, and Ana F Sequeira. Fairness under cover: Evaluating the impact of occlusions on demographic bias in facial recognition. pages 84–98, 2024.

- [324] Bishwas Mandal, Adaeze Okeukwu, and Yihong Theis. Masked face recognition using resnet-50. *arXiv preprint arXiv:2104.08997*, 2021.
- [325] Anthony Mansfield. Information technology–biometric performance testing and reporting–part 1: Principles and framework. *ISO/IEC*, pages 19795–1, 2006.
- [326] Aleix Martinez and Robert Benavente. The ar face database: Cvc technical report, 24. 1998.
- [327] Rogério Martins. Why Are We Not Able to See Beyond Three Dimensions? *The Mathematical Intelligencer*, 38(4):46–51, 2016.
- [328] Sherin Mathews, Shivangee Trivedi, Amanda House, Steve Povolny, and Celeste Fralick. An explainable deepfake detection framework on a novel unconstrained dataset. *Complex & Intelligent Systems*, pages 1–13, 2023.
- [329] Rita Matulionyte. Reconciling trade secrets and explainable ai: face recognition technology as a case study. *European Intellectual Property Review*, 44(1):36–42, 2022.
- [330] Nick Maynard. Mobile Payment Authentication Market Report. In *Juniper Research*, 2021.
- [331] Ghazal Mazaheri and Amit K Roy-Chowdhury. Detection and localization of facial expression manipulations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1035–1045, 2022.
- [332] Niall McLaughlin, Ji Ming, and Danny Crookes. Largest matching areas for illumination and occlusion robust face recognition. *IEEE transactions on cybernetics*, 47(3):796–808, 2016.
- [333] Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 16:4147–4183, 2021.
- [334] Iurii Medvedev, Farhad Shadmand, and Nuno Gonçalves. Mordeephy: Face morphing detection via fused classification. *arXiv preprint arXiv:2208.03110*, 2022.
- [335] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14225–14234. Computer Vision Foundation / IEEE, 2021.
- [336] Michele Merler, Nalini K. Ratha, Rogério Schmidt Feris, and John R. Smith. Diversity in Faces. *ArXiv*, abs/1901.10436, 2019.
- [337] Domingo Mery and Bernardita Morris. On black-box explanation for face verification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3418–3427, 2022.
- [338] Yao Ming, Panpan Xu, Huamin Qu, and Liu Ren. Interpretable and steerable sequence learning via prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 903–913, 2019.

- [339] Nelida Mirabet-Herranz and Jean-Luc Dugelay. Lvt face database: A benchmark database for visible and hidden face biometrics. In *BIOSIG 2023, 22nd International Conference of the Biometrics Special Interest Group*, 2023.
- [340] Christoph Molnar. *Interpretable Machine Learning*. 2019. <https://christophm.github.io/interpretable-ml-book/>.
- [341] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.
- [342] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1 – 15, 2018.
- [343] Helena Montenegro and Jaime S Cardoso. Anonymizing medical case-based explanations through disentanglement. *arXiv preprint arXiv:2311.04833*, 2023.
- [344] Helena Montenegro, Pedro C. Neto, Cristiano Patrício, Isabel Rio-Torto, Tiago Gonçalves, and Luís F. Teixeira. Evaluating privacy on synthetic images generated using gans: Contributions of the VCMi team to imageclefmedical gans 2023. In *CLEF (Working Notes)*, volume 3497 of *CEUR Workshop Proceedings*, pages 1596–1610. CEUR-WS.org, 2023.
- [345] Helena Montenegro, Wilson Silva, and Jaime S Cardoso. Privacy-preserving generative adversarial network for case-based explainability in medical image analysis. *IEEE Access*, 9:148037–148047, 2021.
- [346] Helena Montenegro, Wilson Silva, Alex Gaudio, Matt Fredrikson, Asim Smailagic, and Jaime S Cardoso. Privacy-preserving case-based explanations: Enabling visual interpretability by protecting privacy. *IEEE Access*, 2022.
- [347] David Montero, Marcos Nieto, Peter Leskovsky, and Naiara Aginako. Boosting masked face recognition with multi-task arcface. *arXiv preprint arXiv:2104.09874*, 2021.
- [348] Diana Montezuma, Sara P Oliveira, Pedro C Neto, Domingos Oliveira, Ana Monteiro, Jaime S Cardoso, and Isabel Macedo-Pinto. Annotating for artificial intelligence applications in digital pathology: a practical guide for pathologists and researchers. *Modern Pathology*, 36(4):100086, 2023.
- [349] Janette Moody. Public perceptions of biometric devices: The effect of misinformation on acceptance and use. *Issues in informing science & information technology*, 1, 2004.
- [350] Aythami Morales, Julian Fierrez, Rubén Vera-Rodríguez, and Ruben Tolosana. Sensitivenets: Learning agnostic representations with application to face images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(6):2158–2164, 2021.
- [351] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017.
- [352] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R Varshney. Understanding unequal gender classification accuracy from face images. *arXiv preprint arXiv:1812.00099*, 2018.

- [353] Henning Myhrvold, Haoyu Zhang, Juan Tapia, Raghavendra Ramachandra, and Christoph Günther Busch. Explainable visualization for morphing attack detection. 2022.
- [354] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2901–2907. AAAI Press, 2015.
- [355] Borum Nam, Joo Young Kim, Yeongmyeong Kim, Jiyeon Kim, Soonwon So, Hyung Youn Choi, and In Young Kim. Facialcuenet: An interpretable deception detection model for criminal interrogation using facial expression. *Available at SSRN 4208671*.
- [356] Anoop Namboodiri, Shailesh Saini, Xiaoguang Lu, and Anil Jain. Skilled Forgery Detection in On-Line Signatures: A Multimodal Approach. In *Lecture Notes in Computer Science*, volume 3072, pages 505–511, 2004.
- [357] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14933–14943, 2021.
- [358] Aaron Nech and Ira Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3406–3415. IEEE Computer Society, 2017.
- [359] P. C. Neto, A. F. Sequeira, and J. S. Cardoso. Myope models - are face presentation attack detection models short-sighted? In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 390–399, Los Alamitos, CA, USA, jan 2022. IEEE Computer Society.
- [360] Pedro C Neto, Fadi Boutros, João Ribeiro Pinto, Naser Damer, Ana F Sequeira, and Jaime S Cardoso. Focusface: Multi-task contrastive learning for masked face recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021.
- [361] Pedro C Neto, Fadi Boutros, Joao Ribeiro Pinto, Naser Damer, Ana F Sequeira, Jaime S Cardoso, Messaoud Bengherabi, Abderaouf Bousnat, Sana Boucheta, Nesrine Hebbadj, et al. Ocfr 2022: Competition on occluded face recognition from synthetically generated structure-aware occlusions. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2022.
- [362] Pedro C Neto, Fadi Boutros, João Ribeiro Pinto, Mohsen Saffari, Naser Damer, Ana F Sequeira, and Jaime S Cardoso. My eyes are up here: Promoting focus on uncovered regions in masked face recognition. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2021.
- [363] Pedro C Neto, Eduarda Caldeira, Jaime S Cardoso, and Ana F Sequeira. Compressed models decompress race biases: What quantized models forget for fair face recognition. In *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2023.
- [364] Pedro C Neto, Ivona Colakovic, Sašo Karakatič, and Ana F Sequeira. How knowledge distillation mitigates the synthetic gap in fair face recognition. pages 351–367, 2024.

- [365] Pedro C Neto, Naser Damer, Jaime S Cardoso, and Ana F Sequeira. Balancing beyond discrete categories: Continuous demographic labels for fair face recognition. *arXiv preprint arXiv:2506.01532*, 2025.
- [366] Pedro C Neto, Tiago Gonçalves, Marco Huber, Naser Damer, Ana F Sequeira, and Jaime S Cardoso. Orthomad: Morphing attack detection through orthogonal identity disentanglement. In *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2022.
- [367] Pedro C. Neto, Tiago Gonçalves, João Ribeiro Pinto, Wilson Silva, Ana Filipa Sequeira, Arun Ross, and Jaime S. Cardoso. Causality-inspired taxonomy for explainable artificial intelligence. *CoRR*, abs/2208.09500, 2022.
- [368] Pedro C. Neto, Rafael M. Mamede, Carolina Albuquerque, Tiago Gonçalves, and Ana Filipa Sequeira. Massively annotated datasets for assessment of synthetic and real data in face recognition. In *FG*, pages 1–7. IEEE, 2024.
- [369] Pedro C Neto, Diana Montezuma, Sara P Oliveira, Domingos Oliveira, João Fraga, Ana Monteiro, João Monteiro, Liliana Ribeiro, Sofia Gonçalves, Stefan Reinhard, et al. An interpretable machine learning system for colorectal cancer diagnosis from pathology slides. *NPJ precision oncology*, 8(1):56, 2024.
- [370] Pedro C Neto, Sara P Oliveira, Diana Montezuma, João Fraga, Ana Monteiro, Liliana Ribeiro, Sofia Gonçalves, Isabel M Pinto, and Jaime S Cardoso. imil4path: A semi-supervised interpretable approach for colorectal whole-slide images. *Cancers*, 14(10):2489, 2022.
- [371] Pedro C Neto, Joao Ribeiro Pinto, Fadi Boutros, Naser Damer, Ana F Sequeira, and Jaime S Cardoso. Beyond masks: On the generalization of masked face recognition models to occluded face recognition. *IEEE Access*, 10:86222–86233, 2022.
- [372] Pedro C Neto, Ana F Sequeira, Jaime S Cardoso, and Philipp Terhörst. Pic-score: Probabilistic interpretable comparison score for optimal matching confidence in single-and multi-biometric face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1021–1029, 2023.
- [373] Tom Neubert, Andrey Makrushin, Mario Hildebrandt, Christian Kraetzer, and Jana Dittmann. Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images. *IET Biometrics*, 7(4):325–332, 2018.
- [374] M. L. Ngan, P. J. Grother, and K. K. Hanaoka. Ongoing face recognition vendor test (frvt) part 6b: Face recognition accuracy with face masks using post-covid-19 algorithms. 2020.
- [375] Mei Ngan, Patrick Grother, and Kayee Hanaoka. Ongoing Face Recognition Vendor Test (FRVT) Part 6A: Face recognition accuracy with face masks using pre-COVID-19 algorithms, 2020.
- [376] Mei Ngan, Patrick J Grother, and Mei Ngan. *Face recognition vendor test (FRVT) performance of automated gender classification algorithms*. US Department of Commerce, National Institute of Standards and Technology, 2015.
- [377] Mei L Ngan, Patrick J Grother, Kayee K Hanaoka, et al. Ongoing face recognition vendor test (frvt) part 6a: Face recognition accuracy with masks using pre-covid-19 algorithms. 2020.

- [378] Mei L Ngan, Patrick J Grother, Kayee K Hanaoka, et al. Ongoing face recognition vendor test (frvt) part 6b: Face recognition accuracy with face masks using post-covid-19 algorithms. 2020.
- [379] Mei L Ngan, Patrick J Grother, Kayee K Hanaoka, et al. Ongoing face recognition vendor test (frvt) part 6b: Face recognition accuracy with face masks using post-covid-19 algorithms. 2020.
- [380] Ewa Magdalena Nowara, Ashutosh Sabharwal, and Ashok Veeraraghavan. Ppgsecure: Biometric presentation attack detection using photoplethysmograms. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 56–62. IEEE, 2017.
- [381] S. O’Dea. • Fingerprint recognition technology share 2018-2022 | Statista. <https://www.statista.com/statistics/1003600/smartphone-fingerprint-recognition-technology-share/>, 2021. *Online*; accessed on 2021-08-07.
- [382] Sara P Oliveira, Diana Montezuma, Ana Moreira, Domingos Oliveira, Pedro C Neto, Ana Monteiro, João Monteiro, Liliana Ribeiro, Sofia Gonçalves, Isabel M Pinto, et al. A cad system for automatic dysplasia grading on h&e cervical whole-slide images. *Scientific Reports*, 13(1):3970, 2023.
- [383] Sara P Oliveira, Pedro C Neto, João Fraga, Diana Montezuma, Ana Monteiro, João Monteiro, Liliana Ribeiro, Sofia Gonçalves, Isabel M Pinto, and Jaime S Cardoso. Cad systems for colorectal cancer from wsi are still not ready for clinical acceptance. *Scientific Reports*, 11(1):14358, 2021.
- [384] Muhtahir O. Oloyede, Gerhard P. Hancke, and Hermanus C. Myburgh. A review on face recognition systems: recent approaches and challenges. *Multimedia Tools and Applications* 2020 79:37, 79(37):27891–27922, jul 2020.
- [385] Luca Oneto, Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer, and Massimiliano Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. *Advances in Neural Information Processing Systems*, 33:15360–15370, 2020.
- [386] Michael Opitz, Georg Waltner, Georg Poier, Horst Possegger, and Horst Bischof. Grid loss: Detecting occluded faces. In *European conference on computer vision*, pages 386–402. Springer, 2016.
- [387] Shi Pan, Sanaul Hoque, and Farzin Deravi. An attention-guided framework for explainable biometric presentation attack detection. *Sensors*, 22(9):3365, 2022.
- [388] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018.
- [389] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

- [390] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.
- [391] Vatsa S Patel, Zhongliang Nie, Trung-Nghia Le, and Tam V Nguyen. Masked face analysis via multi-task deep learning. *Journal of Imaging*, 7(10):204, 2021.
- [392] Joseph Pato and Lynnette Millett. *Biometric Recognition: Challenges and Opportunities*. dec 2010.
- [393] Judea Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. 1985.
- [394] Judea Pearl. *Causality*. Cambridge university press, 2009.
- [395] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.
- [396] Martin Pernuš, Mansi Bhatnagar, Badr Samad, Divyanshu Singh, Peter Peer, Vitomir štruc, and Simon Dobrišek. Childnet: Structural kinship face synthesis model with appearance control mechanisms. *IEEE Access*, 11:49971–49991, 2023.
- [397] Dane K Peterson and Gordon F Pitz. Confidence, uncertainty, and the use of information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1):85, 1988.
- [398] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.
- [399] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1090–1104, 2000.
- [400] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.
- [401] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, Jin Chang, K. Hoffman, J. Marques, Jaesik Min, and W. Worek. Overview of the face recognition grand challenge. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 947–954 vol. 1, 2005.
- [402] Allan Pinto, Siome Goldenstein, Alexandre Ferreira, Tiago Carvalho, Helio Pedrini, and Anderson Rocha. Leveraging shape, reflectance and albedo from shading for face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 15:3347–3358, 2020.
- [403] J. R. Pinto and J. S. Cardoso. Explaining ECG Biometrics: Is It All In The QRS? In *Proceedings of the 19th International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, Sep. 2020.

- [404] Norman Poh, Thirimachos Bourlai, Josef Kittler, Lorene Allano, Fernando Alonso-Fernandez, Onkar Ambekar, John Baker, Bernadette Dorizzi, Omolara Fatukasi, Julian Fierrez, Harald Ganster, Javier Ortega-Garcia, Donald Maurer, Albert Ali Salah, Tobias Scheidat, and Claus Vielhauer. Benchmarking quality-dependent and cost-sensitive score-level multimodal biometric fusion algorithms. *IEEE Transactions on Information Forensics and Security*, 4(4):849–866, 2009.
- [405] Alexander P Pons and Peter Polak. Understanding user perspectives on biometric technology. *Communications of the ACM*, 51(9):115–118, 2008.
- [406] Salil Prabhakar, Sharath Pankanti, and Anil K. Jain. Biometric recognition: Security and privacy concerns, mar 2003.
- [407] Shyam Sunder Prasad, Naval Kishore Mehta, Ankit Shukla, Pranav Mahajan, Arshdeep Singh, Sumeet Saurav, and Sanjay Singh. A self-explainable face anti-spoofing solution based on depth estimation. In *International Conference on Frontiers of Intelligent Computing: Theory and Applications*, pages 103–113. Springer, 2022.
- [408] S. Prema, Mohamed Riyas V. S. Deen, Murali V. P. Krishna, and S. Praveen. Vehicle and license authentication using finger print. In *5th International Conference on Advanced Computing Communication Systems (ICACCS)*, pages 737–740, 2019.
- [409] Zhongang Qi, Saeed Khorram, and Li Fuxin. Embedding deep networks into visual explanations. *Artificial Intelligence*, 292:103435, 2021.
- [410] Bosheng Qin and Dongxiao Li. Identifying facemask-wearing condition using image super-resolution with classification network to prevent covid-19. *Sensors*, 20(18):5236, 2020.
- [411] Haibo Qiu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. End2end occluded face recognition by masking corrupted features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [412] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. Syn-face: Face recognition with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10880–10890, 2021.
- [413] Ying Qiu, Vitor Albiero, Michael C. King, and Kevin W. Bowyer. Does face recognition error echo gender classification error? In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2021.
- [414] Alyssa Quek. Facemorpher, 2019.
- [415] Ramachandra Raghavendra, KiranB Raja, Sushma Venkatesh, and Christoph Busch. Face morphing versus face averaging: Vulnerability and detection. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 555–563. IEEE, 2017.
- [416] Anuj Rai, Somnath Dey, Pradeep Patidar, and Prakhar Rai. Expressnet: An explainable residual slim network for fingerprint presentation attack detection. *arXiv preprint arXiv:2305.09397*, 2023.
- [417] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing revisited: Investigating the impact of publicly naming biased performance results of commercial AI products. *Commun. ACM*, 66(1):101–108, 2023.

- [418] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.
- [419] Ankit Rajpal, Khushwant Sehra, Rashika Bagri, and Pooja Sikka. Xai-fr: Explainable ai-based face recognition using deep neural networks. *Wireless Personal Communications*, 129(1):663–680, 2023.
- [420] Raghavendra Ramachandra and Hailin Li. Finger-nestnet: Interpretable fingerphoto verification on smartphone using deep nested residual network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 693–700, 2023.
- [421] Manish Rathod, Chirag Dalvi, Kulveen Kaur, Shruti Patil, Shilpa Gite, Pooja Kamat, Ketan Kotecha, Ajith Abraham, and Lubna Abdelkareim Gabralla. Kids’ emotion recognition using various deep-learning models with explainable ai. *Sensors*, 22(20):8066, 2022.
- [422] Jennifer Raymond. Most of us are biased. *Nature*, 495(7439):33–34, 2013.
- [423] Joao Ribeiro Pinto, Jaime S. Cardoso, and Andre Lourenco. Evolution, current challenges, and future possibilities in ECG Biometrics, jun 2018.
- [424] Isabel Rio-Torto, Jaime S Cardoso, and Luis Filipe Teixeira. Parameter-efficient generation of natural language explanations for chest x-ray classification. In *Submitted to Medical Imaging with Deep Learning*, 2024. under review.
- [425] Ronald L Rivest. Learning decision lists. *Machine learning*, 2(3):229–246, 1987.
- [426] Dan Robinson. Biometric payment authentication is coming - but there’s hurdles in the way. <https://www.nsbanking.com/analysis/biometric-payment-authentication/>, 2020. *Online*; accessed on 2021-08-02.
- [427] Joseph P. Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: Too bias, or not too bias? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 1–10. Computer Vision Foundation / IEEE, 2020.
- [428] Andrea Macarulla Rodriguez, Luis Unzueta, Zeno Geradts, Marcel Worring, and Unai Elordi. Multi-task explainable quality networks for large-scale forensic facial recognition. *IEEE Journal of Selected Topics in Signal Processing*, 2023.
- [429] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
- [430] Hiranmoy Roy, Debotosh Bhattacharjee, and Ondrej Krejcar. Interpretable local frequency binary pattern (lfrbp) based joint continual learning network for heterogeneous face recognition. *IEEE Transactions on Information Forensics and Security*, 17:2125–2136, 2022.
- [431] F.S. Samaria and A.C. Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, pages 138–142, 1994.

- [432] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.
- [433] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.
- [434] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [435] Jitao Sang, Zhen Lei, and Stan Z Li. Face image quality evaluation for iso/iec standards 19794-5 and 29794-5. In *International Conference on Biometrics*, pages 229–238. Springer, 2009.
- [436] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser-Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3752–3761. Computer Vision Foundation / IEEE Computer Society, 2018.
- [437] Sascha Saralajew, Lars Holdijk, Maïke Rees, Ebubekir Asan, and Thomas Villmann. Classification-by-components: probabilistic modeling of reasoning over a set of components. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 2792–2803, 2019.
- [438] Eklavya Sarkar, Pavel Korshunov, Laurent Colbois, and Sébastien Marcel. Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks. *arXiv preprint arXiv:2012.05344*, 2020.
- [439] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [440] Clemens Seibold, Anna Hilsmann, and Peter Eisert. Focused lrp: Explainable ai for face morphing attack detection. In *WACV (Workshops)*, pages 88–96, 2021.
- [441] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [442] Sachith Seneviratne, Nuran Kasthuriarachchi, Sanka Rasnayaka, Danula Hettiachchi, and Ridwan Shariffdeen. Does a face mask protect my privacy?: Deep learning to predict protected attributes from masked face images. In *Australasian Joint Conference on Artificial Intelligence*, pages 91–102. Springer, 2022.
- [443] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016.

- [444] A. F. Sequeira, W. Silva, J. R. Pinto, T. Gonçalves, and J. S. Cardoso. Interpretable biometrics: Should we rethink how presentation attack detection is evaluated? In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2020.
- [445] Ana F Sequeira, Tiago Gonçalves, Wilson Silva, João Ribeiro Pinto, and Jaime S Cardoso. An exploratory study of interpretability for face presentation attack detection. *IET Biometrics*, 2021.
- [446] Ana F. Sequeira, Tiago Gonçalves, Wilson Silva, João Ribeiro Pinto, and Jaime S. Cardoso. An exploratory study of interpretability for face presentation attack detection. *IET Biometrics*, n/a(n/a).
- [447] Ignacio Serna, Aythami Morales, Julian Fierrez, and Nick Obradovich. Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence*, 305:103682, 2022.
- [448] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Regularized fine-grained meta face anti-spoofing. In *AAAI*, pages 11974–11981, 2020.
- [449] Renu Sharma and Arun Ross. D-netpad: An explainable and interpretable iris presentation attack detector. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020.
- [450] Renu Sharma and Arun Ross. Viability of optical coherence tomography for iris presentation attack detection. In *25th International Conference on Pattern Recognition (ICPR)*, pages 6165–6172. IEEE, 2021.
- [451] Sahil Sharma and Vijay Kumar. Voxel-based 3D face reconstruction and its application to face recognition using sequential deep learning. *Multimedia Tools and Applications* 2020 79:25, 79(25):17303–17330, feb 2020.
- [452] LinLin Shen, Alex Kot, and WaiMun Koo. Quality measures of fingerprint images. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 266–271. Springer, 2001.
- [453] Yichun Shi and Anil K. Jain. Probabilistic face embeddings. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6901–6910. IEEE, 2019.
- [454] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.
- [455] Joseph Sill. Monotonic networks. 1998.
- [456] Samuel Henrique Silva, Mazal Bethany, Alexis Megan Votto, Ian Henry Scarff, Nicole Beebe, and Peyman Najafirad. Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Science International: Synergy*, 4:100217, 2022.
- [457] Wilson Silva, Alexander Poellinger, Jaime S Cardoso, and Mauricio Reyes. Interpretability-guided content-based medical image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–314. Springer, 2020.

- [458] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer, 2014.
- [459] Gurmail Singh and Kin-Choong Yow. These do not look like those: An interpretable deep learning model for image recognition. *IEEE Access*, 9:41482–41493, 2021.
- [460] Maneet Singh, Richa Singh, and Arun Ross. A comprehensive overview of biometric fusion. *Inf. Fusion*, 52:187–205, 2019.
- [461] Andrea C Skelly, Joseph R Dettori, and Erika D Brodt. Assessing bias: the importance of considering confounding. *Evidence-based spine-care journal*, 3(01):9–12, 2012.
- [462] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.
- [463] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 773–782, 2019.
- [464] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.
- [465] Nisha Srinivas, Karl Ricanek, Dana Michalski, David S. Bolme, and Michael King. Face recognition algorithm bias: Performance differences on images of children and adults. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2269–2277, 2019.
- [466] Dave Steinkrau, Patrice Y. Simard, and Ian Buck. Using gpus for machine learning algorithms. In *Eighth International Conference on Document Analysis and Recognition (ICDAR 2005)*, 29 August - 1 September 2005, Seoul, Korea, pages 1115–1119. IEEE Computer Society, 2005.
- [467] Samuil Stoychev and Hatice Gunes. The effect of model compression on fairness in facial expression recognition. *arXiv preprint arXiv:2201.01709*, 2022.
- [468] Yifan Sun, Changmao Cheng, Yuhao Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020.
- [469] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.
- [470] (SYRIS Technology Corp). About FAR, FRR and EER. *Technical document*, pages 0–4, 2004. *Online*; accessed on 2021-08-02.
- [471] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

- [472] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.
- [473] Xiaoyang Tan, Yi Li, Jun Liu, and Lin Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *European Conference on Computer Vision*, pages 504–517. Springer, 2010.
- [474] Murat Taskiran, Nihan Kahraman, and Cigdem Eroglu Erdem. Face recognition: Past, present and future (a review). *Digital Signal Processing*, 106:102809, 2020.
- [475] Philipp Terhörst, Daniel Fährmann, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Beyond Identity: What Information Is Stored in Biometric Face Templates? In *International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2020.
- [476] Philipp Terhorst, Daniel Fahrman, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Maad-Face: A Massively Annotated Attribute Dataset for Face Images. *IEEE Transactions on Information Forensics and Security*, 16:3942–3957, 2021.
- [477] Philipp Terhörst, Marco Huber, Naser Damer, Florian Kirchbuchner, Kiran Raja, and Arjan Kuijper. Pixel-level face image quality assessment for explainable face recognition. *arXiv preprint arXiv:2110.11001*, 2021.
- [478] Philipp Terhörst, Marco Huber, Jan Niklas Kolf, Ines Zelch, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Reliable age and gender estimation from face images: Stating the confidence of model predictions. In *10th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, Tampa, FL, USA, September 23-26, 2019*, pages 1–8. IEEE, 2019.
- [479] Philipp Terhörst, Malte Ihlefeld, Marco Huber, Naser Damer, Florian Kirchbuchner, Kiran B. Raja, and Arjan Kuijper. Qmagface: Simple and accurate quality-aware face recognition. *CoRR*, abs/2111.13475, 2021.
- [480] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. SER-FIQ: unsupervised estimation of face image quality based on stochastic embedding robustness. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5650–5659. Computer Vision Foundation / IEEE, 2020.
- [481] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, pages 1–1, 2021.
- [482] Mary Theofanos, Brian Stanton, Charles Sheppard, Ross Micheals, N Zhang, W Wydler, Larry Nadel, and R Rubin. Usability testing of height and angles of ten-print fingerprint capture. *NISTIR*, June, 20(42):210, 2008.
- [483] Mary F Theofanos, Brian C Stanton, Shahram Orandi, Ross J Micheals, Nien Fan Zhang, et al. *Usability testing of ten-print fingerprint capture*. US Department of Commerce, National Institute of Standards and Technology, 2007.

- [484] Mateusz Trokielewicz, Adam Czajka, and Piotr Maciejewicz. Presentation attack detection for cadaver iris. In *IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10. IEEE, 2018.
- [485] Mateusz Trokielewicz, Adam Czajka, and Piotr Maciejewicz. Perception of image features in post-mortem iris recognition: Humans vs machines. In *IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2019.
- [486] M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991.
- [487] Shahadat Uddin, Haohui Lu, Ashfaqur Rahman, and Junbin Gao. A novel approach for assessing fairness in deployed machine learning algorithms. *Scientific Reports*, 14(1):17753, 2024.
- [488] Lev Utkin, Pavel Drobintsev, Maxim Kovalev, and Andrei Konstantinov. Combining an autoencoder and a variational autoencoder for explaining the machine learning model predictions. In *28th Conference of Open Innovations Association (FRUCT)*, pages 489–494. IEEE, 2021.
- [489] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.
- [490] Nicole Van Hoeck, Patrick D Watson, and Aron K Barbey. Cognitive neuroscience of human counterfactual reasoning. *Frontiers in Human Neuroscience*, 9:420, 2015.
- [491] Kushal Vangara, Michael C King, Vitor Albiero, Kevin Bowyer, et al. Characterizing the variability in face recognition accuracy relative to race. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- [492] Rafael Henrique Vareto and William Robson Schwartz. Face spoofing detection via ensemble of classifiers toward low-power devices. *Pattern Analysis and Applications*, 24(2):511–521, 2021.
- [493] Kush R Varshney, Jamie C Rasmussen, Aleksandra Mojsilović, Moninder Singh, and Joan M DiMicco. Interactive visual salesforce analytics. 2012.
- [494] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [495] Sushma Venkatesh, Haoyu Zhang, Raghavendra Ramachandra, Kiran Raja, Naser Damer, and Christoph Busch. Can gan generated morphs threaten face recognition systems equally as landmark based morphs?-vulnerability and detection. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2020.
- [496] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.
- [497] Michael Wagner and Girija Chetty. *Liveness Assurance in Face Authentication*, pages 908–916. Springer US, Boston, MA, 2009.

- [498] Weitao Wan and Jiansheng Chen. Occlusion robust face recognition based on mask learning. In *2017 IEEE international conference on image processing (ICIP)*, pages 3795–3799. IEEE, 2017.
- [499] Bingshu Wang, Jiangbin Zheng, and CL Philip Chen. A survey on masked facial detection methods and datasets for fighting against covid-19. *IEEE Trans. on AI*, 3(3):323–343, 2021.
- [500] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.
- [501] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.
- [502] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.
- [503] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 692–702. IEEE, 2019.
- [504] Mei Wang, Yaobin Zhang, and Weihong Deng. Meta balanced network for fair face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):8433–8448, 2021.
- [505] Qiangchang Wang and Guodong Guo. Dsa-face: Diverse and sparse attentions for face recognition robust to pose variation and occlusion. *IEEE Transactions on Information Forensics and Security*, 16:4534–4543, 2021.
- [506] Tong Wang, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry Mac-Neille. A bayesian framework for learning rule sets for interpretable classification. *The Journal of Machine Learning Research*, 18(1):2357–2393, 2017.
- [507] Yu-Chun Wang, Chien-Yi Wang, and Shang-Hong Lai. Disentangled representation with dual-stage feature learning for face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1955–1964, 2022.
- [508] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8916–8925. Computer Vision Foundation / IEEE, 2020.
- [509] Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, and Zhen Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5042–5051, 2020.

- [510] Renliang Weng, Jiwen Lu, Junlin Hu, Gao Yang, and Yap-Peng Tan. Robust feature set matching for partial face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 601–608, 2013.
- [511] Renliang Weng, Jiwen Lu, and Yap-Peng Tan. Robust point set matching for partial face recognition. *IEEE transactions on image processing*, 25(3):1163–1176, 2016.
- [512] Jonathan R Williford, Brandon B May, and Jeffrey Byrne. Explainable face recognition. In *European Conference on Computer Vision*, pages 248–263. Springer, 2020.
- [513] Martin Winter, Werner Bailer, and Georg Thallinger. Demystifying face-recognition with locally interpretable boosted features (libf). In *2022 10th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE, 2022.
- [514] J B Woodward. Making Things Happen: A Theory of Causal Explanation. *Making Things Happen: A Theory of Causal Explanation*, 2003.
- [515] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2008.
- [516] Haiyu Wu and Kevin W Bowyer. A real balanced dataset for understanding bias? factors that impact accuracy, not numbers of identities and images. *arXiv preprint arXiv:2304.09818*, 2023.
- [517] Haiyu Wu and Kevin W. Bowyer. What should be balanced in a "balanced" face recognition dataset? In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*, page 235. BMVA Press, 2023.
- [518] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Automated, general-purpose counterfactual generation. *arXiv preprint arXiv:2101.00288*, 2021.
- [519] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo I. Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. In *NeurIPS*, 2022.
- [520] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 6502–6509. AAAI Press, 2020.
- [521] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. In *Computer Vision—ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 506–523. Springer, 2020.
- [522] Ying Xu, Kiran Raja, and Marius Pedersen. Supervised contrastive learning for generalizable and explainable deepfakes detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 379–389, 2022.
- [523] Ying Xu, Philipp Terhörst, Kiran Raja, and Marius Pedersen. A comprehensive analysis of ai biases in deepfake detection with massively annotated databases. *arXiv preprint arXiv:2208.05845*, 2022.

- [524] Yuanyuan Xu, Wan Yan, Genke Yang, Jiliang Luo, Tao Li, and Jianan He. Centerface: joint face detection and alignment using face as point. *Scientific Programming*, 2020, 2020.
- [525] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.
- [526] Jian Yang, Lei Luo, Jianjun Qian, Ying Tai, Fanlong Zhang, and Yong Xu. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):156–171, 2016.
- [527] Jianwei Yang, Zhen Lei, and Stan Z Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*, 2014.
- [528] Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li. Face liveness detection with component dependent descriptor. In *2013 International Conference on Biometrics (ICB)*, pages 1–6. IEEE, 2013.
- [529] Meng Yang, Lei Zhang, Simon Chi-Keung Shiu, and David Zhang. Robust kernel representation with statistical local features for face recognition. *IEEE transactions on neural networks and learning systems*, 24(6):900–912, 2013.
- [530] Meng Yang, Lei Zhang, Jian Yang, and David Zhang. Robust sparse coding for face recognition. In *CVPR 2011*, pages 625–632. IEEE, 2011.
- [531] Xiao Yang, Wenhan Luo, Linchao Bao, Yuan Gao, Dihong Gong, Shibao Zheng, Zhifeng Li, and Wei Liu. Face anti-spoofing: Model matters, so does data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3507–3516, 2019.
- [532] Zhanjia Yang, Xiangping Zhu, Changyuan Jiang, Wenshuang Liu, and Linlin Shen. Ram-face: Race adaptive margin based face recognition for racial bias mitigation. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2021.
- [533] Nick Yeung and Christopher Summerfield. Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1310–1321, 2012.
- [534] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.
- [535] Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu. Towards interpretable face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9348–9357, 2019.
- [536] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5295–5305, 2020.
- [537] Seyma Yucer, Samet Akçay, Noura Al Moubayed, and Toby P. Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 83–92. Computer Vision Foundation / IEEE, 2020.

- [538] Timothy Zee, Geeta Gali, and Ifeoma Nwogu. Enhancing human face recognition with an interpretable neural network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.
- [539] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.
- [540] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 International Conference on Computer Vision*, pages 2018–2025. IEEE, 2011.
- [541] Chris G. Zeinstra, Didier Meuwly, Raymond N. J. Veldhuis, and Luuk J. Spreeuwers. Mind the gap: A practical framework for classifiers in a forensic context. In *9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, Redondo Beach, CA, USA, October 22-25, 2018*, pages 1–9. IEEE, 2018.
- [542] Dan Zeng, Raymond Veldhuis, and Luuk Spreeuwers. A survey of face recognition techniques under occlusion. *IET biometrics*, 10(6):581–606, 2021.
- [543] Dan Zeng, Raymond Veldhuis, Luuk Spreeuwers, and Richard Arendsen. Occlusion-invariant face recognition using simultaneous segmentation. *IET biometrics*, 10(6):679–691, 2021.
- [544] Dan Zeng, Raymond N. J. Veldhuis, and Luuk J. Spreeuwers. A survey of face recognition techniques under occlusion. *IET Biom.*, 10(6):581–606, 2021.
- [545] Guangpeng Zhang and Yunhong Wang. Asymmetry-based quality assessment of face images. In *International Symposium on Visual Computing*, pages 499–508. Springer, 2009.
- [546] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.
- [547] Qinglong Zhang, Lu Rao, and Yubin Yang. Group-cam: Group score-weighted visual explanations for deep convolutional networks. *arXiv preprint arXiv:2103.13859*, 2021.
- [548] Zihui Zhang, Cuican Yu, Shuang Xu, and Huibin Li. Learning Flexibly Distributional Representation for Low-quality 3D Face Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3465–3473, 2021.
- [549] Fang Zhao, Jiashi Feng, Jian Zhao, Wenhan Yang, and Shuicheng Yan. Robust lstm-autoencoders for face de-occlusion in the wild. *IEEE Transactions on Image Processing*, 27(2):778–790, 2017.
- [550] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diff-swap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8568–8577, 2023.
- [551] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5:7, 2018.

- [552] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017.
- [553] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.
- [554] Song Zhou and Sheng Xiao. 3D face recognition: a survey. *Human-centric Computing and Information Sciences* 8:1, 8(1):1–27, nov 2018.
- [555] Zihan Zhou, Andrew Wagner, Hossein Mobahi, John Wright, and Yi Ma. Face recognition with contiguous occlusion using markov random fields. In *2009 IEEE 12th international conference on computer vision*, pages 1050–1057. IEEE, 2009.
- [556] Hongbo Zhu, Chuang Yu, and Angelo Cangelosi. Explainable emotion recognition for trustworthy human–robot interaction. In *Proc. Workshop Context-Awareness Hum.-Robot Interact. Approaches Challenges ACM/IEEE HRI, Sapporo, Japan*, 2022.
- [557] Pengkai Zhu, Ruizhao Zhu, Samarth Mishra, and Venkatesh Saligrama. Low dimensional visual attributes: An interpretable image encoding. In *International Conference on Pattern Recognition*, pages 90–102. Springer, 2021.
- [558] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021.