FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

A Robust and Explainable Framework for Face Detection and Age Estimation from Videos

Hélder Filipe Pacheco Ferreira

Master in Biomedical Engineering

Supervised by: Prof. Ana Filipa Sequeira

Co-supervised by: Rafael Mamede

Jun 21, 2026

Resumo

A análise automática de faces tem vindo a assumir um papel central em diversas áreas, incluindo segurança, biometria e interação humano–computador. Entre as várias tarefas de *face analytics*, a deteção facial e a estimação de idade destacam-se pela sua relevância prática, mas continuam a apresentar desafios significativos quando aplicadas a dados em vídeo recolhidos em ambientes menos controlados. Variações de pose, iluminação, oclusões, movimento e qualidade de imagem conduzem frequentemente a comportamentos instáveis e a previsões inconsistentes ao longo do tempo. Paralelamente, a natureza opaca dos modelos de aprendizagem profunda levanta preocupações relacionadas com interpretabilidade e confiança nos sistemas.

Esta dissertação estuda a deteção facial e a estimação de idade a partir de vídeo, com particular enfoque na estabilidade das previsões ao longo do tempo e na interpretabilidade dos modelos. O trabalho apresenta uma revisão detalhada do estado da arte em *face analytics*, deteção facial, estimação de idade e inteligência artificial explicável, complementada por uma síntese crítica de abordagens recentes dedicadas à estimação de idade baseada em vídeo.

Com base nessa análise, é definido um *pipeline* modular que integra deteção facial, estimação de idade ao nível da imagem, estratégias de agregação temporal e mecanismos de explicabilidade. A separação clara entre estas componentes permite analisar o impacto individual de cada etapa no desempenho do sistema e facilita a replicação e extensão de metodologias existentes, seguindo um protocolo recente e reprodutível para estimação de idade em vídeo.

Os resultados experimentais demonstram que o YOLOv12m obteve o melhor desempenho de deteção facial entre os modelos avaliados no benchmark WIDER FACE. Para estimação de idade, o modelo MiVOLO obteve um desempenho competitivo quando avaliado no benchmark Casual Conversations, produzindo resultados próximos dos reportados por trabalhos recentes da literatura. Observou-se ainda que o erro de estimação aumenta progressivamente nas faixas etárias mais avançadas, refletindo uma dificuldade comum dos modelos atuais de estimação de idade. A aplicação de estratégias de suavização e agregação temporal permitiu aumentar a estabilidade das previsões ao nível do sujeito, enquanto a análise baseada em Integrated Gradients evidenciou que o modelo utiliza regiões faciais semanticamente relevantes para realizar as suas estimativas. Os testes de validação das explicações confirmaram ainda a consistência e a fiabilidade das atribuições geradas. Como trabalho futuro, poderá ser explorada a adaptação do modelo a diferentes grupos etários, a melhoria da modelação temporal e a validação em conjuntos de dados mais diversificados, de forma a aumentar a robustez e a capacidade de generalização.

Palavras-chave: Deteção facial, estimação de idade, vídeo, análise temporal, inteligência artificial explicável.

Abstract

Automatic face analysis has become a central component of modern computer vision systems, with applications in security, biometrics, and human–computer interaction. Among the various face analytics tasks, face detection and age estimation are particularly relevant, yet they remain challenging when applied to video data acquired in less constrained environments. Variations in pose, illumination, occlusion, motion, and image quality often lead to unstable detections and inconsistent age predictions over time. At the same time, the black-box nature of deep learning models raises concerns related to interpretability and trust.

This dissertation investigates face detection and age estimation from video, with particular emphasis on the temporal stability of predictions and the interpretability of the underlying models. The work presents a comprehensive review of the state of the art in face analytics, face detection, age estimation, and explainable artificial intelligence, complemented by a critical synthesis of recent approaches for video-based age estimation.

Based on this analysis, a modular pipeline is defined, integrating face detection, image-level age estimation, temporal aggregation strategies, and explainability mechanisms. The clear separation between these components enables the impact of each stage on overall system performance to be analysed individually and facilitates the replication and extension of existing methodologies, following a recent and reproducible protocol for video-based age estimation.

Experimental results demonstrate that YOLOv12m achieved the best face detection performance among the evaluated models on the WIDER FACE benchmark. For age estimation, the MiVOLO model achieved competitive performance on the Casual Conversations benchmark, producing results comparable to those reported in recent literature. The age estimation error was found to increase progressively for older age groups, reflecting a well-known challenge in current age estimation systems. Temporal smoothing and aggregation strategies improved the stability of subject-level predictions, while the Integrated Gradients analysis showed that the model relies on semantically meaningful facial regions when estimating age. Furthermore, the explanation validation experiments confirmed the consistency and reliability of the generated attributions. Future work could explore age-specific model adaptation, improved temporal modeling, and validation on more diverse datasets to further improve robustness and generalizability.

Keywords: Face analytics, face detection, age estimation, video analysis, temporal stability, explainable artificial intelligence.

Contents

List of Figures

List of Tables

List of Acronyms

AAM Active Appearance Model

AgeDB Age Database

AI Artificial Intelligence

AP Average Precision

CC Casual Conversations

CCV1 Casual Conversations V1

CCV2 Casual Conversations V2

CNN Convolutional Neural Network

CPU Central Processing Unit

DEX Deep EXpectation

DPM Deformable Part Model

FAN Face Attention Network

FDDB Face Detection Data Set and Benchmark

FG-NET Face and Gesture Recognition Network

GPU Graphics Processing Unit

Grad-CAM Gradient-weighted Class Activation Mapping

GRU Gated Recurrent Unit

HOG Histogram of Oriented Gradients

IG Integrated Gradients

IoU Intersection over Union

LBP Local Binary Pattern

LFW Labeled Faces in the Wild

LSTM Long Short-Term Memory

MAE Mean Absolute Error

MSE Mean Squared Error

MTCNN Multi-task Cascaded Convolutional Networks

PCA Principal Component Analysis

PR Precision–Recall

RMSE Root Mean Squared Error

ResNet Residual Network

S3FD Single Shot Scale-Invariant Face Detector

SDG Sustainable Development Goal

SSD Single Shot MultiBox Detector

SSH Single Stage Headless Face Detector

SVM Support Vector Machine

UTKFace University of Tennessee Knoxville Face Dataset

VGG Visual Geometry Group Network

WIDER FACE WIDER Face Detection Benchmark Dataset

XAI Explainable Artificial Intelligence

YOLO You Only Look Once

Chapter 1

Introduction

The human face is a distinct and easily recognizable biometric feature that is now commonly used in modern identification systems. From high-security environments like airports and border control to everyday consumer devices such as smartphones and laptops, facial recognition and analysis technologies play an increasingly prominent role. However, ensuring that these systems remain reliable under real-world conditions and transparent in their decision-making is an ongoing challenge. This dissertation addresses this challenge by investigating face detection and age estimation from videos captured in less constrained environments, with particular emphasis on the temporal stability of predictions and the interpretability of deep learning models. Through the evaluation and integration of state-of-the-art methods, the work analyses how facial age estimation systems behave under realistic conditions and examines the extent to which their predictions can be explained using modern explainable artificial intelligence techniques.

1.1 Context and Background

Face analysis technologies have advanced significantly over the past decades, becoming central to applications in security, authentication, and consumer electronics. Face detection is often regarded as the foundational step in most face-related tasks. Early face detection systems relied on handcrafted features and cascade-based classifiers, representing important breakthroughs in real-time visual processing [1]. With the advent of deep learning, detection performance improved substantially through data-driven convolutional architectures capable of learning hierarchical representations directly from large-scale datasets [2, 3]. These advances enabled robust operation under less constrained conditions involving pose variation, illumination changes, and background clutter. Comprehensive surveys have documented this methodological evolution and its impact on facial analysis research [4, 5].

Facial age estimation has similarly gained relevance due to its wide application in biometrics, human–computer interaction, and demographic analysis. Deep learning based approaches trained on large-scale datasets such as UTKFace [6] and AgeDB [7] have significantly improved

prediction accuracy. Nonetheless, age estimation remains inherently difficult due to the natural variability in human aging patterns, as highlighted in existing studies [5].

1.2 Motivation

This dissertation is motivated by two major research gaps: the need for robustness and the need for explainability.

First, many face detection and age estimation systems function effectively under controlled environments but degrade significantly in real-world or video-based scenarios. Motion blur, occlusions, drastic illumination changes, and viewpoint variation often result in unstable detections or inconsistent age predictions. Ensuring robust performance across such challenging conditions is critical for real deployments, especially in security or surveillance contexts [8].

Second, modern deep learning systems operate largely as "black boxes" making their internal decision-making opaque. In the context of biometric systems, explainability has become increasingly important, both for regulatory compliance and for user trust. Explainable AI (XAI) methods attempt to reveal which features or regions of the face influence model predictions, providing valuable transparency. Recent work has underscored the need for trustworthy and interpretable facial analytics, such as the studies presented by Rajpal [9].

1.3 Challenges

Several challenges arise in the development of a robust and explainable framework for face detection and age estimation in video:

- **Environmental variability:** Faces in video often appear under inconsistent lighting, occlusions, extreme head poses, or motion blur, making detection more difficult.
- **Diversity of aging patterns:** Individuals of the same age may look significantly different due to genetic, lifestyle, and environmental factors, creating large intra-class variation [5].
- **Temporal inconsistency:** Frame-by-frame predictions in video can fluctuate, requiring temporal smoothing or tracking for stable behavior.
- **Explainability constraints:** Methods such as Grad-CAM or saliency maps must produce faithful, interpretable visual explanations while remaining computationally practical.
- **Dataset limitations:** Image datasets such as UTKFace [6] or AgeDB [7] are widely used, but annotated video datasets remain scarce, limiting research on temporal robustness.

These challenges highlight the need for integrated, well-designed pipelines capable of handling both performance and interpretability requirements.

1.4 Objectives and Scope

The primary objective of this dissertation is to analyse and evaluate face detection and age estimation methods for video data acquired under less constrained conditions. Particular emphasis is placed on the performance of state-of-the-art approaches and on the interpretability of their predictions through explainable artificial intelligence techniques. This objective is pursued through the systematic study and integration of existing computer vision and deep learning methods, enabling the assessment of their accuracy, robustness, and explanation reliability on challenging real-world datasets.

To achieve this objective, the dissertation first presents a review of the state of the art in face analytics, covering face detection, age estimation, and explainable artificial intelligence. This review establishes the theoretical and methodological foundations of the work and provides the context to motivate the experimental choices adopted throughout the dissertation.

The face detection component focuses on the evaluation and comparison of several state-of-the-art detectors using the WIDER FACE benchmark. The objective is to identify the detector that provides the most suitable balance between accuracy, robustness, and computational efficiency for integration into the complete age estimation pipeline.

For age estimation, the selected face detector is combined with the MiVOLO age estimation model, selected based on its competitive performance in recent age estimation literature and its compatibility with the face-only setting adopted in this work, and evaluated on video-based datasets following a subject-level protocol. Particular attention is given to the behaviour of the model under less constrained conditions and to the effectiveness of prediction aggregation strategies for obtaining stable subject-level estimates.

Another central objective is to investigate the interpretability of age estimation predictions through attribution-based explainability analysis, using Integrated Gradients as the main explanation method. The generated attribution maps are analysed to identify the facial regions that contribute more to the model's predictions, while dedicated validation experiments are performed to assess the reliability and faithfulness of the explanations.

Finally, the integrated system is evaluated through a series of quantitative and qualitative analyses covering face detection performance, age estimation accuracy, demographic behaviour, error patterns, and explanation quality. Together, these objectives define the scope of the dissertation and provide a comprehensive assessment of face detection, age estimation, and explainability in less constrained video-based scenarios.

1.5 Contributions

This dissertation contributes to the study of face detection and age estimation in less constrained video scenarios through a combination of experimental evaluation, methodological analysis, and explainability. While substantial progress has been achieved in face analysis from still

images, age estimation from video remains comparatively less explored and presents additional challenges related to motion, image quality variations, and prediction consistency across frames.

The first contribution of this work is the evaluation and replication of the recent video-based age estimation framework proposed by Bešenić et al. [10], chosen as a reference due to its subject-level evaluation protocol and competitive performance on the Casual Conversations benchmark. By reproducing and analysing the evaluation protocol described in that work, including face detection, age prediction, and subject-level aggregation, the dissertation provides an independent assessment of its applicability and performance on challenging video datasets. Such replication contributes to methodological transparency and reproducibility, which remain important aspects of contemporary computer vision research.

The second contribution is the development of an integrated pipeline combining state-of-the-art face detection and age estimation models for video-based analysis. Rather than proposing new architectures, the work focuses on systematically evaluating the interaction between face detection, age prediction, and aggregation strategies in realistic video scenarios. This provides insight into the strengths and limitations of existing approaches when deployed under less constrained conditions.

A third contribution is the incorporation of explainable artificial intelligence techniques into the age estimation inference pipeline through Integrated Gradients. For each sampled frame used in the evaluation, an attribution map is generated from the age estimation model to identify the facial regions contributing most to the corresponding prediction. The generated attribution maps are analysed to identify the facial regions most relevant to age estimation, while dedicated validation experiments based on insertion and deletion tests are used to assess the faithfulness and reliability of the explanations. This contributes to a better understanding of the decision-making process of deep learning models and supports the development of more transparent age estimation systems.

Finally, the dissertation provides a detailed analysis of age estimation performance across different age groups and demographic categories, together with an investigation of common failure cases and current limitations of video-based age estimation systems. These analyses highlight challenges such as increased prediction errors for older subjects and provide practical insights into factors that influence performance in real-world conditions.

Overall, this work contributes to a stronger empirical understanding of face detection, age estimation, and explainability in less constrained video environments, while providing a reproducible evaluation framework for future research.

1.6 Structure of the Dissertation

This dissertation is organized into five chapters.

Chapter 1 introduces the research topic and presents the context, motivation, challenges, objectives, contributions, overall structure of the dissertation, and its relationship to the Sustainable Development Goals (SDGs).

Chapter 2 reviews the state of the art in face analytics, with particular focus on face detection, age estimation, and explainable artificial intelligence. It discusses traditional and deep learning-based approaches, video-based age estimation methods, commonly used datasets, and existing explainability techniques relevant to facial analysis systems.

Chapter 3 describes the methodological framework adopted throughout this work. It presents the reference framework used as the basis for the study, the proposed system pipeline, the datasets and evaluation metrics, the experimental setup, and the protocols used for face detection and age estimation evaluation.

Chapter 4 presents the experimental results. The chapter begins with the evaluation and selection of the face detection model, followed by the assessment of age estimation performance on the Casual Conversations datasets. Performance is analysed across different age groups and demographic categories, together with an investigation of common error patterns. The chapter also includes an explainability analysis based on Integrated Gradients and a validation of the generated explanations through insertion and deletion experiments.

Finally, Chapter 5 summarizes the main findings of the dissertation, discusses the principal limitations of the proposed framework, and outlines potential directions for future work in video-based age estimation and explainable facial analytics.

1.7 Sustainable Development Goals

This dissertation is indirectly related to several Sustainable Development Goals (SDGs), particularly through the study of artificial intelligence methods applied to facial analysis, age estimation, demographic evaluation, and explainability. Table 1.1 summarizes the most relevant SDGs, their associated targets, the contribution of this work, and the indicators used to achieve them.

Table 1.1: Relationship between this dissertation and the Sustainable Development Goals.

SDG	**Target**	**Contribution**	**Performance Indicators and Metrics**
3	3.d	Indirectly contributes to facial analysis technologies with potential applications in healthcare, well-being, and aging-related studies.	Mean Absolute Error (MAE), age-group analysis, bias assessment, and robustness evaluation under less constrained conditions.
9	9.5	Contributes to research in artificial intelligence and computer vision through the evaluation of face detection, age estimation, and explainability methods.	Average Precision (AP), age-group accuracy, MAE, comparison with existing literature, and experimental validation.
10	10.2	Demographic group analysis enables the identification of performance differences and limitations associated with data representativeness.	MAE by demographic group, analyses across age, gender, and country, and identification of error patterns.
16	16.6	The use of explainability methods promotes greater transparency in the interpretation of model predictions.	Integrated Gradients attribution maps, regional attribution analysis, and insertion/deletion tests.

Although the primary contribution of this dissertation is methodological and experimental, the results obtained are indirectly connected to several Sustainable Development Goals. In the context of SDG 3 (Good Health and Well-Being), the study of face detection and age estimation techniques may contribute to the future development of digital health tools, population monitoring systems, and aging-related research. While the work was not designed for a specific clinical application, it enabled the evaluation of artificial intelligence models in estimating age-related characteristics under real-world, less constrained conditions.

Regarding SDG 9 (Industry, Innovation and Infrastructure), this dissertation contributes to research in Artificial Intelligence and Computer Vision through the systematic evaluation of face detection, age estimation, and explainability methods. The comparison of different approaches and the experimental validation of results generated knowledge that may support the development of more effective and reliable systems in future applications.

The connection to SDG 10 (Reduced Inequalities) arises from the analysis of model performance across different demographic groups. Evaluation by age, gender, and country made it possible to identify performance variations among distinct populations and highlight the impact that data representativeness can have on the obtained results. Although bias mitigation was not a direct objective of this work, identifying these differences contributes to a better understanding of

model limitations.

Finally, the relationship with SDG 16 (Peace, Justice and Strong Institutions) is associated with the explainability component developed throughout this dissertation. The use of the Integrated Gradients method enabled the analysis of the facial regions most relevant to the model's predictions, while insertion and deletion tests provided quantitative evidence regarding the faithfulness of the generated explanations. These procedures promote greater transparency and understanding of the decision-making process of artificial intelligence models.

Chapter 2

State of the Art

2.1 Face Analytics

Face analytics is a subdomain of computer vision and pattern recognition focused on the automated analysis of human faces in images and videos. It encompasses a broad range of tasks aimed at extracting identity or semantic information from facial data [4, 11]. The human face is a rich source of visual cues that reveal identity, age, gender, ethnicity, emotion, and other attributes, making facial image analysis a pivotal research area in biometrics and human-centred computing.

The field includes fundamental tasks such as face detection and localisation, geometric alignment to a canonical pose, and the learning of robust facial representations. Building upon these foundational steps, higher-level inference tasks include face recognition (identity verification and identification), age estimation, gender classification, expression recognition, and other attribute predictions (e.g., ethnic background, gaze, or health-related indicators) [5, 12]. These tasks are deeply interrelated: downstream analyses rely on accurate detection and normalisation, and multiple attributes can often be inferred from a shared facial representation. Accordingly, face analytics can be viewed as an umbrella framework covering the entire pipeline from low-level localisation to high-level semantic inference.

2.1.1 Face Analytics tasks

Among the various tasks in face analytics, face detection constitutes the essential first stage upon which all subsequent analyses depend. Automatically locating faces in an image or video frame is a prerequisite for any further processing. As summarized in Table 2.1, once faces are detected they can be aligned, represented through discriminative embeddings, and analysed for higher-level attributes such as identity, age, or expression.

In a typical processing pipeline, a detector outputs bounding boxes that isolate facial regions, which are then cropped and passed to subsequent modules such as alignment, recognition, or attribute estimation [11]. The reliability of this stage is critical, as detection errors propagate to downstream tasks. Even small localisation inaccuracies or misalignments can significantly degrade recognition accuracy or age estimation performance.

Detecting faces in less constrained environments remains challenging due to factors such as pose variation, facial expression changes, illumination differences, partial occlusions, and background clutter. Unlike many generic object detection problems, subtle variations in facial viewpoint or configuration can significantly alter the appearance of discriminative features. These challenges were already highlighted in early face detection research [4], where classical detectors often performed well only under controlled conditions but struggled with extreme pose, occlusion, and illumination variability [4, 13].

Low-resolution facial imagery introduces additional difficulties, as limited pixel information reduces the visibility of age-related facial features and can negatively affect both face detection and age estimation performance [13]. Consequently, modern face analytics systems must operate robustly "in the wild," where variability is the norm rather than the exception.

Table 2.1: Representative tasks in face analytics, their typical outputs, and representative references.

Task	Brief Description	Typical Output	Representative References
Face Detection	Locate faces in images or video.	Bounding boxes	[3, 4, 11]
Face Alignment	Normalize facial geometry using landmarks.	Landmark coordinates	[4]
Face Recognition	Verify or identify subjects from facial features.	Identity / similarity score	[14–16]
Age Estimation	Estimate age from facial appearance.	Age value / age group	[10, 17, 18]
Expression Recognition	Recognize facial expressions or affective state.	Expression label	[19, 20]
Head Pose Estimation	Estimate head orientation relative to the camera.	Yaw / pitch / roll	[21]
Face Parsing	Segment semantic facial regions.	Pixel-wise labels	[4]

2.1.2 Scope of Face Analytics

Once a face is detected, a range of analytic tasks can be applied to the isolated region. A common subsequent step is face alignment, which involves localising key landmarks (e.g., eye corners, nose tip, mouth corners) and geometrically normalising position, scale, and rotation. This process reduces pose-related variability and improves consistency before applying recognition or attribute classifiers [2, 22]. Facial landmark detection itself constitutes both a preprocessing stage and an independent research topic with applications in motion capture and expression analysis.

Beyond alignment, facial representation learning forms the core of most modern face analytics systems. The objective is to transform raw pixel data into a compact feature embedding that preserves discriminative identity or attribute information while remaining invariant to nuisance factors such as pose or illumination. Early approaches relied on hand-crafted descriptors (e.g.,

eigenfaces or local texture features), whereas contemporary systems predominantly learn deep representations from large-scale annotated datasets.

These learned embeddings can then be utilised for diverse tasks. In face recognition, embeddings are compared to determine identity verification or identification [5]. In attribute estimation tasks such as age or gender prediction, the representation is fed into regression or classification models [12]. Expression recognition similarly relies on discriminative features capturing subtle muscle movements or Action Units.

Additional tasks under the face analytics umbrella include head pose estimation, face parsing and segmentation, face reenactment or synthesis, and the inference of softer biometric traits. Collectively, these components enable comprehensive facial analysis extending well beyond identity recognition. Modern benchmarks increasingly evaluate multiple tasks jointly, reflecting the integrated and multi-task nature of contemporary face analytics research [23, 24].

2.1.3 Evolution of Face Analytics

Face analytics methodologies have evolved substantially over the past two decades, transitioning from manually engineered features toward data-driven deep learning frameworks. Early systems relied on carefully designed descriptors combined with traditional machine learning algorithms. In face detection, cascade-based classifiers built upon simple Haar-like features enabled real-time processing on standard CPUs [1]. Similarly, face recognition initially relied on linear subspace methods such as Eigenfaces [25], as well as local texture descriptors including LBP [26] and HOG [27], typically combined with classifiers such as Support Vector Machines.

Age estimation followed a comparable trajectory. Early approaches employed anthropometric measurements, wrinkle analysis, or ageing subspace modeling [28, 29]. While these methods achieved reasonable performance under controlled conditions, they struggled to generalise to less constrained environments due to their limited ability to capture complex variability in pose, illumination, and expression.

The inherent limitations of handcrafted features became increasingly apparent as datasets grew in scale and diversity. Although incremental improvements such as pose normalisation and ensemble strategies were introduced, classical systems remained sensitive to distribution shifts and environmental variability [4, 5]. By the early 2010s, the need for more expressive and scalable learning paradigms became evident, paving the way for deep convolutional architectures that would fundamentally reshape face analytics research.

2.1.4 Deep Learning Impact

Deep convolutional neural networks (CNNs) enabled face analytics systems to learn feature representations directly from data, rather than relying on manually designed descriptors. With the availability of large annotated face databases containing millions of images, CNN-based models rapidly surpassed traditional approaches across major facial analysis tasks.

In face detection, modern deep architectures such as the Multi-Task Cascaded Convolutional Network (MTCNN) [2] and single-shot detectors like RetinaFace [3] leverage hierarchical feature learning to predict face locations with high precision and recall, even under challenging conditions including occlusion, extreme pose, and tiny face scales. These models learn multi-scale filters that capture complex facial patterns and implicitly handle variations that were difficult for handcrafted features. As a result, current detectors reliably operate in scenarios once considered highly challenging, such as partially masked faces or low-light crowd environments.

In face recognition, deep learning produced an unprecedented increase in accuracy. Networks trained on massive identity datasets generate highly discriminative embeddings that approach or surpass human-level performance on benchmark evaluations [5]. Milestone systems such as DeepFace and FaceNet demonstrated that a single CNN architecture could encode identity with remarkable separation power [30]. Subsequent innovations, including margin-based loss functions such as ArcFace [31] and large-scale dataset curation, further strengthened representation compactness and inter-class separability.

Age estimation has undergone a comparable transformation. While early predictors relied on handcrafted features and shallow regressors, modern approaches employ deep networks trained on extensive age-labelled datasets [32, 33]. These models learn complex wrinkle structures, facial morphology variations, and texture-based ageing cues that generalise more effectively across individuals. For instance, the AgeCNN model by Rothe et al., which won the ChaLearn apparent age estimation challenge, demonstrated that pretraining on face recognition tasks followed by fine-tuning for age regression significantly reduces mean absolute error [34]. Numerous subsequent studies confirmed the superiority of CNN-based approaches for age and attribute estimation [33, 35].

Overall, deep learning now constitutes the foundational paradigm for high-performance face analytics systems, primarily due to its capacity to learn robust, hierarchical feature representations from large-scale data.

2.1.5 Current Challenges

Despite substantial progress, several challenges continue to drive research in face analytics. Many of these arise from intrinsic variability in facial appearance across environmental and demographic conditions.

Pose variation remains a persistent difficulty. Non-frontal views introduce self-occlusion and geometric distortion, complicating both detection and recognition. Although deep networks exhibit improved pose tolerance, performance typically decreases for extreme profile views, motivating research into pose-invariant embeddings and 3D-aware modelling strategies.

Occlusion represents a closely related issue. When facial regions are partially hidden by objects, accessories, or other individuals, algorithms must infer identity or attributes from incomplete evidence. Severe occlusions, such as medical masks, significantly degrade performance. Recent methods attempt to mitigate this through occlusion-aware training or image inpainting techniques, yet robust handling of arbitrary occlusion patterns remains unresolved [13].

Illumination variability further complicates facial analysis. Extreme lighting conditions, including backlit or nighttime scenarios, alter texture and shading cues. Although deep models are more illumination-invariant than classical systems, low-light benchmarks such as DarkFace reveal persistent performance degradation under poor visibility [36]. Similarly, low resolution, motion blur, and sensor noise reduce discriminative information, often requiring super-resolution or quality-enhancement modules.

Beyond imaging factors, demographic variability and algorithmic bias have emerged as critical concerns. Performance disparities across gender, ethnicity, and age groups can arise from unbalanced training data or biased representations. For example, commercial gender classification systems exhibited significantly higher error rates for darker-skinned females compared to lighter-skinned males [37]. In age estimation, models trained predominantly on adult faces may underperform for children or elderly subjects. Recent research therefore emphasises fairness auditing, balanced dataset curation, and bias-aware optimisation strategies [35].

Age-related variability introduces an additional challenge. Ageing is a complex and nonlinear biological process influenced by genetics, health, and lifestyle [33]. Individuals of the same chronological age may exhibit markedly different facial ageing patterns, complicating regression accuracy and cross-age recognition. Recognising individuals across decades remains difficult due to progressive morphological changes [29]. Although deep embeddings exhibit partial age invariance, substantial age gaps can still lead to failure cases.

2.1.6 Practical Relevance and Applications

Face analytics underpins a wide range of real-world applications across security, human–computer interaction, healthcare, and demographic analysis.

In biometrics and security, face recognition has become mainstream for identity verification, from consumer devices to automated border control systems. Law enforcement agencies employ detection and recognition systems in surveillance and forensic investigations [5]. Attribute estimation, including age and gender prediction, assists in narrowing search spaces during identification processes.

In human–computer interaction, facial analysis enables adaptive and personalised interfaces. Video conferencing platforms rely on detection and alignment for frame centring, while interactive displays use demographic estimation to tailor content. Expression recognition supports emotionally aware systems capable of detecting engagement or distress.

Emerging biomedical applications further highlight the societal impact of face analytics. Deep learning systems trained on clinical facial datasets can identify syndromic patterns associated with genetic disorders. For example, DeepGestalt demonstrated expert-level performance in recognising craniofacial phenotypes across numerous syndromes [38]. Additional applications include automated pain assessment and neurological disorder screening, where high reliability is essential.

In demographic research, large-scale age and gender estimation enables population-level analytics, such as analysing social media demographics or estimating foot traffic distributions in

urban environments [24]. While these uses raise important ethical and privacy considerations, they illustrate the broad practical reach of face analytics technologies.

In summary, face analytics provides a comprehensive framework for automated facial understanding. The following sections examine two core components, Face Detection and Age Estimation, within this broader context, highlighting both shared challenges and domain-specific considerations for real-world deployment.

2.2 Face Detection

Face detection is the process of determining whether human faces appear within an image or video and estimating their spatial location. It is the first and most essential stage of any facial analysis framework, as all subsequent tasks such as face recognition, age estimation, or facial expression analysis depend on reliable localization [11].

Over roughly three decades, face detection has evolved from rule-based systems and handcrafted feature extractors to deep learning-based approaches capable of operating under challenging real-world conditions. Figure 1 summarises the main stages of this evolution, highlighting representative methods discussed throughout this section. The progression ranges from classical approaches such as Viola–Jones and HOG+SVM to modern deep learning-based detectors including MTCNN, RetinaFace, and recent YOLO-based architectures.

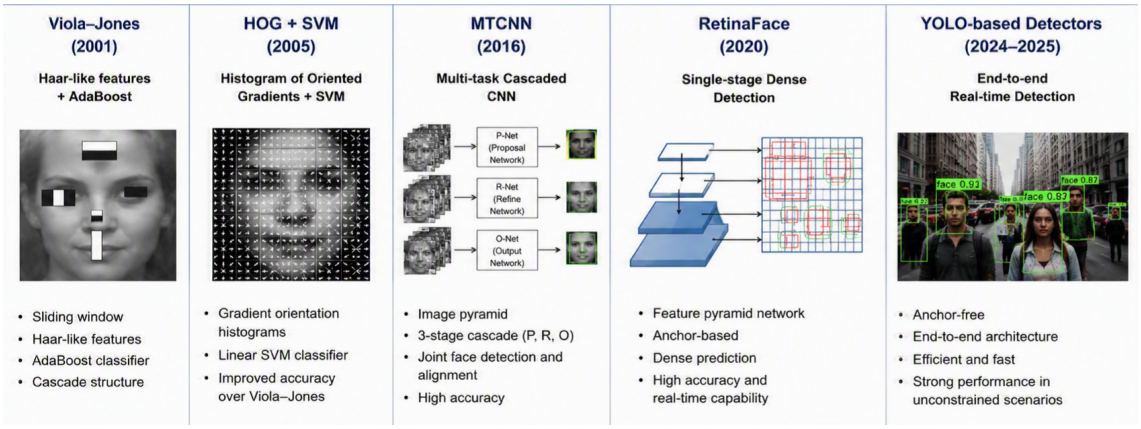


Figure 1: Evolution of representative face detection methods, from traditional handcrafted-feature approaches to modern deep learning-based detectors.

In the following sections, a structured state-of-the-art review is presented following this chronological progression. The discussion begins with classical pre-deep learning methods, followed by convolutional neural network-based detectors, and concludes with approaches designed for less constrained and video-based scenarios, as discussed in the surveys by Zafeiriou et al. [4] and Hassaballah et al. [5].

2.2.1 Traditional Face Detection Methods

Before deep learning, face detection was dominated by handcrafted systems based on prior knowledge, geometric constraints, and low-level image features. Classical taxonomies categorise early approaches into knowledge-based, template-matching, feature-based, and appearance-based methods [11]. Knowledge-based models attempted to codify the expected spatial relationships between facial components such as relative eye symmetry, the vertical ordering of nose and mouth, or the rough proportions of a frontal face. Although computationally lightweight, these models lacked robustness to real-world variability and often generated false positives when non-face objects happened to match coarse geometric templates [39]. Template-matching approaches applied static or deformable face prototypes across sliding windows and compared them to local image patches. While effective for well-aligned, frontal faces, template-based detectors typically failed under scale variation, rotations, and non-frontal poses [40].

Feature-based methods attempted to identify local structures characteristic of facial appearance, such as edge patterns, gradient distributions, or skin-color regions. These local cues were then combined using heuristics or shallow classifiers. Such systems performed reasonably well under controlled conditions but were highly sensitive to illumination, background clutter, and occlusion [41]. Appearance-based models, such as PCA-based eigenfaces, used global intensity patterns to characterize faces, but these models were computationally demanding and sensitive to image variation [25].

A major breakthrough came with the Viola–Jones framework [1], which marked a significant departure from earlier heuristics by introducing a scalable and fast detection pipeline based on three key innovations: (i) Haar-like rectangular features computed efficiently using integral images, (ii) AdaBoost for selecting the most discriminative features and combining weak learners into strong ones, and (iii) a cascade architecture that rejected non-face regions early. This structure is illustrated in Figure 2.

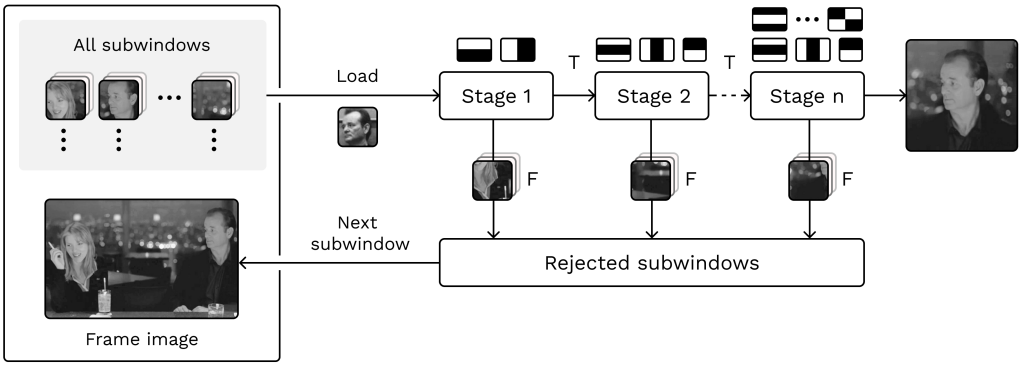


Figure 2: Overview of the Viola–Jones face detection framework, illustrating Haar-like features with AdaBoost feature selection, and the cascade structure that enables real-time face detection. Source: Paralect Blog [42].

This combination enabled real-time detection of around 15 frames per second on low-end CPUs, making Viola–Jones the foundational detector of the 2000s [11]. Its success stemmed from

high detection rates on frontal faces and computational efficiency. However, the method's heavy reliance on upright frontal geometry made it vulnerable to non-frontal poses, rotations, and partial occlusions. Although extensions employing rotated Haar features [43] or alternative descriptors like Local Binary Patterns [26] provided modest improvements but did not fundamentally overcome these limitations.

In the late 2000s, feature-descriptor approaches gained prominence. Dalal and Triggs' Histograms of Oriented Gradients (HOG) [27], initially proposed for pedestrian detection, effectively captured local gradient orientation distributions and proved highly suitable for frontal face detection. Combined with a linear SVM, HOG-based detectors offered superior robustness to changes in illumination and moderate pose variation compared to Haar cascades [27]. The Dlib library subsequently popularized a HOG+SVM face detector that remains widely used for lightweight applications [44].

Another influential classical technique was the deformable part model (DPM), which represented objects as collections of parts connected by flexible spatial relationships. Xiangxin Zhu's mixture-of-trees model [22] applied a part-based formulation to faces, enabling the detection of both frontal and profile views by using multiple part configurations. This model demonstrated significantly improved robustness to pose variation and moderate occlusion relative to earlier global-template approaches. However, DPMs were computationally expensive and sensitive to background clutter, and by the early 2010s they approached their practical performance limits [22, 45]. On benchmarks such as FDDB, classical detectors plateaued: they were effective for controlled datasets but degraded sharply in unconstrained "in-the-wild" environments [46]. This stagnation created the conditions for deep learning based approaches to dominate the field.

2.2.2 Deep Learning Based Face Detection

Deep convolutional neural networks (CNNs) transformed face detection by learning hierarchical visual representations directly from data. Although early neural-network-based detectors appeared in the late 1990s [47], the success of deeper architectures, particularly after AlexNet [48], accelerated the transition toward fully data-driven approaches.

One of the first influential deep face detectors was the Multi-Task Cascaded Convolutional Network (MTCNN) [2]. MTCNN uses a three-stage CNN cascade that progressively refines candidate regions while jointly predicting face bounding boxes and facial landmarks. This formulation improved robustness to variations in pose, illumination, and scale, achieving strong results on benchmarks such as FDDB [46] and WIDER FACE [49].

Modern detectors are commonly divided into two-stage and single-stage architectures. Two-stage methods, such as Faster R-CNN [50], first generate region proposals and subsequently classify and refine them, generally providing strong localization at increased computational cost. In contrast, single-stage approaches such as SSD [51] and YOLO [52] predict bounding boxes directly from dense feature maps, favouring efficient inference and real-time applications.

The WIDER FACE dataset [49] played an important role in the development of these methods due to its large variations in face scale, pose, and occlusion. Its "Hard" subset contains particularly small and partially occluded faces, exposing weaknesses in earlier detectors and encouraging architectures designed to preserve information at multiple spatial scales [53].

SSH [54] introduced context modules and scale-aware anchors to improve small-face detection while maintaining single-stage efficiency. DSFD [55] later incorporated dual-shot supervision to strengthen feature representations, while RetinaFace [3] combined multi-scale detection with joint prediction of bounding boxes, facial landmarks, and coarse 3D facial structure.

Subsequent architectures continued to refine scale handling and feature representation. TinaFace [56] improved anchor assignment, ASFD [57] introduced adaptive receptive fields, and HP-FLDet [58] jointly addresses face detection and landmark localisation. The relatively small performance differences among these recent methods suggest that progress on WIDER FACE increasingly depends not only on accuracy, but also on computational efficiency and robustness.

SCRFD [59] explicitly targets this accuracy–efficiency trade-off by redistributing computation and training samples across network stages. Its strong performance across different computational budgets, particularly on difficult WIDER FACE cases, makes it a relevant reference for efficient face detection.

YOLOv12 [60] represents a more recent generation of real-time object detectors, combining attention-based mechanisms with the efficiency of single-stage YOLO architectures. Although not originally designed specifically for face detection, its favourable accuracy–latency trade-off makes it a relevant candidate for face-analysis pipelines requiring both robustness and efficient inference.

In the context of this dissertation, face detection serves as the first stage of the age estimation pipeline, and the resulting face crops are subsequently processed by MiVOLO [61] for frame-level age prediction. MiVOLO was pretrained on a combination of large-scale age and gender datasets, including UTKFace, rather than being trained specifically on the Casual Conversations data used in this work. Consequently, detector robustness and localisation quality are particularly important, as detection errors may directly affect the input provided to the downstream age estimation model.

Table 2.2 summarizes representative deep learning face detectors evaluated on the WIDER FACE test set (Hard subset) [49]. Modern methods consistently achieve high AP values, while recent developments increasingly emphasize efficiency, scalability, and robustness to challenging real-world conditions [36, 37].

Table 2.2: Comparison of representative face detection models on the WIDER FACE test set (Hard subset). AP = Average Precision.

Method	Backbone	Architecture Type	AP (Hard)
SSH (2017) [54]	VGG-16	Anchor-based (single-stage)	0.844
DSFD (2019) [55]	ResNet-152	Dual-shot detector	0.900
RetinaFace (2020) [3]	ResNet-50	Multi-task anchor-based	0.914
TinaFace (2020) [56]	ResNet-50	Anchor refinement	0.924
ASFD (2021) [57]	ResNet-152	Adaptive receptive fields	0.921
HP-FLDet (2024) [58]	ResNet-50	Joint face and landmark detection	0.921

2.2.3 Face Detection in Video and Unconstrained Settings

Although deep learning has largely solved face detection for static, high-quality images, the task remains substantially more challenging in video. Real-world video streams often include rapid head movements, motion blur, abrupt changes in illumination, occlusions from objects or other people, and frames where the face is too small or partially outside the field of view [62]. Additionally, video-based systems must ensure temporal stability: a face should not spontaneously disappear and reappear across frames, nor should random false positives flicker in and out. These temporal inconsistencies can significantly degrade downstream tasks such as tracking, recognition, and age estimation [63].

The most straightforward video pipeline applies a still-image face detector such as MTCNN, RetinaFace, or a lightweight single-stage CNN to every frame independently. With modern GPUs or efficient mobile architectures, per-frame face detection at video frame rates can often be computationally feasible. However, when detection is combined with subsequent processing stages, such as tracking, age estimation, or explainability analysis, the overall computational requirements may become substantially higher. [64]. A detector may confidently detect a face in stable frames but fail momentarily when confronted with blur, occlusion, or non-frontal pose. Similarly, false positives may appear for only a single frame due to reflections or sudden lighting shifts. These behaviours motivate temporal smoothing mechanisms [65].

The dominant paradigm for achieving temporal consistency is tracking-by-detection. In this approach, a detector provides candidate bounding boxes for each frame, and a tracking module links detections over time into continuous trajectories. Popular tracking algorithms include Kalman filtering coupled with the Hungarian algorithm for data association [63], as well as more advanced trackers such as SORT [63] and DeepSORT [66]. These systems estimate short-term future face positions, allowing the tracker to bridge occasional missed detections. They also filter out spurious detections that fail to persist across frames and assign consistent identities over time.

Tracking-by-detection is computationally efficient and modular, making it widely used in surveillance, human–computer interaction, and autonomous driving.

In addition to tracking, motion cues can improve robustness under difficult video conditions. Faces exhibit characteristic micro-motions such as blinking, subtle head movements, or lip motion. Optical flow techniques like Farnebäck flow [67] can detect regions moving coherently with human head dynamics, allowing the system to prioritize candidate regions even before appearance-based confidence returns. Although motion-based signals are secondary to CNN predictions, they remain valuable in cases involving low light, heavy motion blur, or compression artifacts [36].

Researchers have also explored deep architectures that explicitly incorporate temporal information. Some models use 3D convolutional layers to process short frame sequences jointly [68, 69], while others add recurrent modules (LSTMs or GRUs) on top of frame-level CNN features to propagate face-relevant information over time [70]. Despite their appeal, these temporally aware networks remain less common due to increased computational cost. For many applications, a strong still-image detector combined with a robust tracker provides a sufficient solution.

Real-time requirements further shape system design. Applications such as driver monitoring, access control, and smart-camera analytics often operate with limited computational resources. In these settings, lightweight face detectors such as MobileNet-based backbones [71], binarized neural networks, or architectures like FaceBoxes [72] are attractive due to their ability to run efficiently on CPUs. Additional strategies include down-sampling frames, performing detection every n-th frame (with tracking interpolating positions), or restricting detection to foreground-motion regions. Although such strategies may reduce accuracy, they preserve responsiveness under tight computational budgets.

Video introduces further challenges not present in still images. Long-term partial occlusions such as a hand covering part of the face or masks can cause detectors to alternate between success and failure. Very small faces, common in wide-angle surveillance footage, pose a fundamental challenge: below a certain resolution, even advanced CNNs lack sufficient visual information. Super-resolution and deblurring modules have been proposed to mitigate these limitations [73], though they add computational overhead.

Finally, fairness and demographic robustness are increasingly important considerations in video settings. Several studies report that face detectors may perform unevenly across demographic groups or under varied lighting and pose conditions [37]. Ensuring reliable detection for all populations requires balanced training data, appropriate augmentation strategies, and in some cases, explicit fairness-aware optimisation.

In summary, modern deep detectors combined with temporal smoothing and tracking provide strong performance for video-based applications. Persistent challenges include motion blur, extended occlusions, extremely small faces, and demographic fairness. These aspects are central to the framework developed in this thesis, which aims for reliable and explainable face analysis in realistic, unconstrained environments.

2.2.4 Face Detection Datasets

Over the years, several large-scale benchmarks have shaped the evolution of modern face detectors.

One of the first widely adopted unconstrained benchmark was the FDDB (Face Detection Data Set and Benchmark) [46] containing 2,845 images with 5,171 annotated faces with moderate pose and illumination variations. However, FDDB is relatively small by modern standards and contains limited extreme-scale or heavily occluded faces.

WIDER FACE [49] significantly expanded evaluation complexity, comprising over 32,000 images and nearly 400,000 annotated faces. It includes substantial scale variation, heavy occlusions, crowd scenes, and extreme pose diversity. The dataset defines "easy", "medium", and "hard" subsets, making it particularly suitable for benchmarking robustness to tiny faces and difficult environmental conditions. Nevertheless, WIDER FACE is limited to static images and does not provide temporal annotations.

Other datasets such as AFW [22] and PASCAL Face [74] contributed to early benchmarking efforts by introducing unconstrained, in-the-wild imagery with multiple faces and moderate pose variation. However, their limited scale and relatively low diversity in occlusion, tiny faces, and adverse imaging conditions restrict their relevance for evaluating modern deep detectors.

Subsequent benchmarks introduced more challenging scenarios. MALF [75], for example, incorporated attribute annotations such as occlusion, enabling finer-grained robustness analysis. The IJB series [76] included images and video frames with extreme pose and quality variations, while UFDD [77] explicitly evaluated performance under degradations such as blur, rain, and low resolution. These datasets reflect a shift from basic detection accuracy toward robustness under realistic and adverse conditions.

2.3 Age Estimation

Age estimation is the task of automatically predicting a person's age or age group from their face image [78, 79]. It is a key component in face analytics with applications ranging from security control to personalized human–computer interaction. Given a detected face, an age estimation system may output either a precise age or an age range (category). Unlike attributes such as identity or gender, age is a slowly varying attribute that progresses irreversibly and differently for each individual. This makes automatic age estimation challenging: the facial signs of aging are subtle, personalized, and influenced by many factors such as health, lifestyle, and ethnicity [80]. Moreover, a distinction exists between chronological age (actual age) and apparent age (how old one looks), which has been addressed by several benchmarks such as ChaLearn LAP [32]. In general, age estimation can be formulated as a regression problem (predicting a continuous age) or a classification problem (predicting an age group), both exhibiting an inherent ordinal structure [81, 82].

Over the past two decades, automatic age estimation from faces has evolved significantly [29]. Early approaches relied on hand-crafted facial features and conventional machine learning, evaluated on relatively constrained image datasets. In the last decade, deep learning on large-scale face data has dramatically improved performance, with convolutional neural networks now dominating the field [79, 81]. In this section, we review the state-of-the-art in age estimation chronologically:

starting from classical feature-based methods, then modern CNN-based approaches, and finally methods for less constrained and video-based scenarios.

2.3.1 Early Age Estimation Methods

Research on automatic age estimation began in the late 1990s. One of the earliest approaches was proposed by Kwon and Lobo [28], who introduced a simple age classification scheme based on anthropometric facial measurements and wrinkle analysis. Their method used ratios between key facial landmarks to distinguish infants from adults and analysed wrinkle patterns (e.g., forehead and eye corners) to further separate younger and older adults. Although evaluated on a small dataset, the system demonstrated the feasibility of automated age classification.

Subsequent studies focused on designing features that capture age-related facial characteristics. Lanitis *et al.* [83] employed Active Appearance Models (AAM) to jointly represent facial shape and texture, learning a mapping from AAM parameters to age. Their experiments showed that nonlinear regression combined with hierarchical estimation (coarse age group followed by refined prediction) improved accuracy. Other work attempted to explicitly model aging dynamics. For instance, Geng *et al.* [84] proposed an "aging pattern" framework that represents sequences of an individual's face images across time as subspaces describing personal aging trajectories. However, such approaches required longitudinal datasets and were less practical for general age estimation.

Throughout the 2000s, numerous hand-crafted features and traditional machine learning algorithms were explored [83, 85]. Global facial descriptors (e.g., PCA or AAM parameters) were commonly combined with local texture features designed to capture wrinkles and skin smoothness. Methods based on Local Binary Patterns (LBP) or Gabor wavelets encoded fine-grained skin details and were typically coupled with classifiers or regressors such as Support Vector Machines or shallow neural networks. Under controlled conditions, these techniques achieved age-group classification accuracies of approximately 80–85% [84, 86], but their performance often degraded in less constrained environments due to variations in illumination, pose, and image quality [29].

A notable development in the pre-deep learning era was the Bio-Inspired Features (BIF) framework proposed by Guo *et al.* [86]. BIF descriptors are derived from multi-scale and multi-orientation Gabor filters designed to emulate aspects of human visual perception. When combined with Support Vector Regression, BIF achieved state-of-the-art results for the time, reaching a mean absolute error (MAE) of approximately 4.4 years on the MORPH dataset. Subsequent refinements, including gender-specific models to account for different aging patterns, reduced the error to about 2.6 years on certain benchmarks [87].

Another important insight was that age prediction inherently involves ordered labels. To exploit this property, several works adopted ordinal regression formulations. Chang *et al.* [88], for example, proposed the Ordinal Hyperplane Ranking (OHRank) approach, which trained multiple SVM classifiers to determine whether a face is older than predefined age thresholds. By incorporating ordinal constraints, these models achieved improved accuracy, with errors around 4.5 years MAE on datasets such as FG-NET and MORPH. Despite these advances, progress with

handcrafted features began to plateau by the early 2010s, largely due to limitations in feature representation and the relatively small size of available aging datasets [29, 78, 80].

2.3.2 Deep Learning-based Age Estimation

The mid-2010s marked a major shift in age estimation with the widespread adoption of deep learning. Convolutional neural networks (CNNs) began to outperform traditional handcrafted-feature approaches by learning age-related representations directly from large facial image datasets [32, 81]. One of the earliest successful CNN-based approaches was proposed by Levi and Hassner [12], who applied deep learning to age-group classification on the Adience dataset, demonstrating that CNNs could achieve competitive performance under less constrained conditions.

A significant advance was later introduced by Rothe *et al.* [34] through the Deep EXpectation (DEX) model. Using the large-scale IMDb-WIKI dataset and a VGG-16 architecture, DEX formulated age estimation as a classification problem and converted class probabilities into continuous age predictions. This approach substantially improved performance and established deep learning as the dominant paradigm for age estimation.

Subsequent research explored architectures and learning strategies specifically designed for age prediction. Niu *et al.* [89] proposed an ordinal CNN that explicitly models the ordered nature of age labels, while other works investigated ranking-based methods, label distribution learning, and age-aware loss functions [79, 84]. Multi-task learning approaches were also introduced to jointly estimate age and related attributes, such as gender, allowing networks to learn richer facial representations.

Overall, deep learning has significantly improved the accuracy and robustness of automated age estimation systems. Nevertheless, reliable age prediction remains challenging in less constrained environments, where variations in pose, illumination, occlusion, image quality, and demographic characteristics can substantially affect performance. In addition, age estimation accuracy is often lower for very young and elderly individuals due to limited training data and greater variability in facial ageing patterns [80]. These challenges continue to motivate the development of more robust and generalizable age estimation methods [90].

2.3.3 Unconstrained and Video-Based Age Estimation

The shift to deep learning has significantly improved the robustness of age estimation systems in unconstrained environments, where variations in pose, illumination, facial expression, and image quality are common. Modern convolutional neural network (CNN) based approaches can generalize well to in-the-wild conditions when trained on large and diverse datasets. In practical pipelines, face preprocessing steps such as detection and alignment are typically employed to normalize facial regions, which helps mitigate variability and improve prediction accuracy [32]. However, extreme head poses, motion blur, and occlusions continue to degrade performance, as they obscure critical facial features required for reliable age inference [80].

Extending age estimation from images to videos introduces additional challenges, but also opportunities. Early works on video-based age estimation explored the use of temporal information through handcrafted features. For instance, Hadid [19] distinguished between frame-based approaches, which aggregate independent predictions, and spatio-temporal methods that explicitly model facial dynamics. Similarly, Dibeklioğlu *et al.* [18, 91] demonstrated that incorporating facial motion patterns, such as smile dynamics, can improve age estimation accuracy compared to static image-based methods. These approaches, however, were limited by the use of controlled datasets and handcrafted feature representations.

With the advent of deep learning, more advanced video-based models have been proposed. A common baseline strategy consists of applying an image-based age estimator to individual frames and aggregating the predictions using statistical measures such as the mean or median [20, 79]. While simple and effective, this approach ignores temporal dependencies and treats each frame independently. More sophisticated methods integrate temporal modeling components, such as recurrent neural networks (RNNs), temporal attention mechanisms, or convolutional temporal models, to capture relationships across frames. For example, Pei *et al.* [20] proposed an end-to-end architecture combining CNNs with spatial and temporal attention modules, while Ji *et al.* [92] introduced stability-aware aggregation to reduce frame-level prediction noise. These approaches improve temporal consistency by emphasizing stable facial characteristics and suppressing transient artifacts.

Despite these advances, the development of fully video-based age estimation models remains limited by the scarcity of large-scale annotated video datasets. As highlighted in recent work by Bešenić *et al.* [10], most existing methods rely on image-based training and only apply temporal aggregation at inference time. Their study emphasizes that the lack of standardized benchmarks and labeled video data has hindered progress in this area. To address this, they proposed a dedicated video evaluation protocol and demonstrated that incorporating temporal modeling can improve both accuracy and prediction stability compared to frame-based baselines.

In practice, due to the limited availability of suitable video datasets and the increased complexity of temporal models, many real-world systems still adopt the simpler frame-by-frame estimation strategy followed by temporal smoothing or aggregation. This approach offers a strong trade-off between performance, computational cost, and implementation simplicity, and serves as a reliable baseline for comparison with more advanced video-based methods.

2.4 Age Estimation Datasets

The performance and generalization capability of age estimation systems are strongly influenced by the characteristics of the datasets used for training and evaluation. Key factors include dataset scale, annotation quality, demographic diversity, and the modality of the data (static images versus videos). While significant progress has been achieved using large-scale image datasets, the extension to video-based age estimation remains limited due to the scarcity of suitable annotated

benchmarks. Table 2.3 summarises the principal datasets commonly used in age estimation research and highlights the differences in scale, modality, and demographic coverage. This section provides a comprehensive overview of these datasets, with a particular focus on their properties, advantages, and limitations.

Table 2.3: Summary of representative image-based and video-based age estimation datasets.

Dataset	Type	Samples	Subjects	Age Range	Main Characteristics
FG-NET	Image	1,002	82	0–69	Longitudinal ageing
MORPH	Image	55,134	13,618	16–77	Large-scale; demographic imbalance
UTKFace	Image	23,708	–	0–116	Age, gender and ethnicity labels
AgeDB	Image	16,488	568	1–101	Cross-age benchmark
IMDb-WIKI	Image	523,051	–	0–100+	Largest dataset; noisy labels
UvAge	Video	6,898	516	16–83	Unconstrained videos
Casual Conversations	Video	45,186	3,011	18–80+	Diverse demographics
Casual Conversations V2	Video	56,847	5,559	18–81	Expanded demographic coverage

2.4.1 Image-Based Age Estimation Datasets

Most advances in age estimation have been driven by large-scale image datasets, which provide labeled facial images across different age groups. Early datasets such as FG-NET [93] contain 1,002 images of 82 subjects and are particularly useful for longitudinal analysis, as they include multiple images of the same individuals across different ages. However, their limited size makes them unsuitable for training deep learning models.

The introduction of larger datasets such as MORPH [94], which contains over 55,000 images, enabled more robust statistical modeling and facilitated the transition to data-driven approaches. Nevertheless, MORPH exhibits demographic imbalances, particularly with respect to gender and ethnicity, which can bias learned models. Similarly, UTKFace [82] provides over 20,000 images annotated with age, gender, and ethnicity, and is widely used for regression-based evaluation. However, due to its semi-automatic collection process, it contains labeling inaccuracies and noise.

Other commonly used datasets include AgeDB [7], which contains 16,488 images of 568 individuals, and CACD [15], which provides a large number of celebrity images across age variations. While these datasets support cross-age evaluation, they often suffer from limited diversity or weak labeling. A major milestone was the introduction of the IMDb-WIKI dataset [32], which contains over 500,000 images and enabled large-scale deep learning approaches such as DEX. Despite its scale, the dataset relies on metadata-derived labels, leading to significant annotation noise.

Overall, while image-based datasets have been instrumental in advancing age estimation, they lack temporal continuity and cannot capture the dynamic facial cues present in videos. This limitation restricts their suitability for studying temporal stability and real-world deployment scenarios involving video data.

2.4.2 Video-Based Age Estimation Datasets

In contrast to image datasets, video-based datasets for age estimation are significantly less developed. One of the earliest datasets used in this context is the UvA-NEMO Smile dataset [91], which contains 1,240 videos from 400 subjects aged between 8 and 76 years. A related dataset, UvA-NEMO Disgust [18], includes 518 videos from 324 subjects recorded under similar conditions. Although both datasets provide age metadata, they were originally designed for facial expression analysis and were recorded in controlled environments with constrained illumination and mostly frontal head poses. As a result, their applicability to less constrained age estimation scenarios is limited.

A more dedicated effort is the UvAge dataset [17], which contains 6,898 videos from 516 subjects with ages ranging from 16 to 83 years. The dataset was constructed by collecting celebrity videos from public events and associating them with known birth dates. While UvAge provides real age annotations in unconstrained conditions, it is not publicly accessible in practice, limiting its use for reproducible research.

More recently, the Casual Conversations (CC) dataset [95] has emerged as the most relevant large-scale video dataset for age estimation. It consists of 45,186 videos from 3,011 subjects, with self-reported age labels and additional annotations such as gender, skin tone, and lighting conditions. The dataset was collected in diverse, real-world environments, making it particularly suitable for fairness-aware evaluation. A balanced subset, known as CCMini, contains 6,022 videos selected to ensure a more uniform distribution of lighting conditions.

Building upon this dataset, Bešenić *et al.* [10] proposed a refined benchmark known as CCMiniVID, which introduces continuous face-tracked sequences for video-based evaluation. Unlike previous image-based protocols that rely on randomly sampled frames (e.g., CCMiniIMG), CCMiniVID provides temporally consistent sequences of approximately 20 seconds (around 600 frames) per video, along with face tracking metadata such as landmarks, head pose, and tracking quality. This enables the evaluation of both frame-based aggregation methods and true temporal models under reproducible conditions. The authors further define multiple dataset variants (e.g., CCMiniVID-O, CCMiniVID-A, and CCMiniVID-R) to address issues such as missing data, sequence consistency, and dataset size.

In addition to these datasets, recent work has introduced new video-based benchmarks such as the P-Age dataset [96], which leverages real-world videos to model apparent age classification under challenging conditions including motion blur, occlusions, and varying viewpoints. Unlike traditional face-only approaches, this dataset also explores the use of body cues and spatio-temporal transformers, reflecting a shift toward more holistic and robust age estimation systems.

2.4.3 Limitations of Existing Datasets

Despite recent progress, the development of video-based age estimation remains constrained by several fundamental limitations in available datasets. First, there is a lack of large-scale, publicly accessible video datasets with reliable age annotations. While datasets such as Casual Con-

versations provide high-quality annotations, their usage is often restricted to evaluation purposes, limiting their applicability for supervised training [10]. Second, many existing datasets either contain a limited number of subjects or are collected in controlled environments, reducing their generalizability to real-world scenarios.

Another important limitation is the lack of temporal annotations, such as identity tracking and frame-level correspondence, which are necessary for training and evaluating spatio-temporal models. In many cases, researchers must rely on frame-level processing or construct pseudo-video datasets from static images, which does not fully capture the temporal dynamics of facial aging. Additionally, demographic imbalances across age, gender, and ethnicity persist in many datasets, potentially introducing bias into trained models.

As a consequence of these limitations, most existing approaches to video-based age estimation rely on applying image-based models independently to each frame and aggregating the predictions using statistical measures such as the mean or median. While this strategy provides a practical baseline, it does not exploit temporal information explicitly, highlighting a key research gap in the development of robust and explainable video-based age estimation systems.

2.5 Explainable Artificial Intelligence

The increasing adoption of deep learning models in high-stakes applications has raised significant concerns regarding transparency, interpretability, and trustworthiness. In facial analytics tasks such as age estimation, where predictions may influence biomedical decisions, understanding the reasoning behind model outputs is particularly important. Explainable Artificial Intelligence (XAI) seeks to address this challenge by providing methods that make the internal behavior of complex models more interpretable to humans [97].

2.5.1 Categories of Explainability Methods

Explainability techniques are commonly categorized as either *intrinsic* or *post-hoc*. Intrinsic interpretability refers to models that are inherently transparent by design, such as decision trees or linear models. However, such models typically lack the representational capacity required for high-dimensional tasks like face detection or age estimation. Consequently, most research in facial analytics relies on post-hoc explanation methods applied to trained deep networks [98].

Post-hoc methods can be further divided into gradient-based, perturbation-based, and surrogate-model approaches. Each of these categories provides different perspectives on model reasoning [97].

2.5.2 Gradient-Based Visual Explanations

Gradient-based methods are among the most widely used techniques in computer vision. They rely on backpropagation to estimate the contribution of each pixel or feature map to the final prediction.

One influential method is Grad-CAM (Gradient-weighted Class Activation Mapping) [99], which produces coarse localization heatmaps highlighting the spatial regions most responsible for a specific class prediction. Grad-CAM has been extensively applied to face-related tasks, including age and gender estimation, due to its simplicity and compatibility with standard CNN architectures.

Variants such as Grad-CAM++ [100] improve localization quality by better handling multiple object instances or distributed evidence. Other gradient-based approaches include saliency maps [101] and Integrated Gradients [102].

Integrated Gradients estimates feature importance by accumulating gradients along a path between a reference baseline and the actual input. In contrast to standard saliency maps, which rely on gradients evaluated only at the input, Integrated Gradients considers how the model response evolves across this path, providing a more theoretically grounded attribution. The method was specifically designed to satisfy properties such as sensitivity and implementation invariance, which support more consistent interpretation of feature contributions [102].

These characteristics make Integrated Gradients particularly relevant for face-based age estimation, where the objective is to analyse fine-grained spatial contributions from facial regions rather than only obtain coarse localization maps. For this reason, Integrated Gradients was selected in this work as the primary attribution method, while its generated explanations were further assessed through dedicated faithfulness validation experiments.

While computationally efficient, gradient-based methods may produce noisy or visually unstable explanations, particularly under small input perturbations. Their reliability in reflecting true causal importance remains an active area of research.

2.5.3 Perturbation and Model-Agnostic Methods

Perturbation-based approaches evaluate feature importance by systematically modifying parts of the input and observing changes in the output.

LIME (Local Interpretable Model-agnostic Explanations) [103] approximates the model locally with an interpretable surrogate model, while SHAP (SHapley Additive exPlanations) [104] computes feature attributions based on cooperative game theory. These methods are model-agnostic and theoretically grounded, but can be computationally expensive when applied to high-resolution images.

In computer vision, perturbation-based heatmaps are often generated by masking image regions and measuring the resulting changes in the model prediction. Because these methods require multiple modified versions of the same image to be evaluated, they are generally more computationally demanding than gradient-based approaches. Although they provide a more direct measure of the influence of individual image regions on the prediction, their computational cost often limits their applicability in real-time or large-scale systems.

2.5.4 Explainability in Face Analytics

Explainability has gained increasing attention in face recognition and age estimation research. Studies have used saliency-based methods to identify whether models rely on semantically meaningful facial regions, such as wrinkles, eye contours, or facial shape, when predicting age. These analyses help determine whether networks capture physiologically plausible aging cues or rely on spurious correlations, such as background artifacts.

In biometric contexts, explainability also plays a role in auditing demographic bias and detecting failure modes. Visualization of activation maps can reveal whether models disproportionately focus on specific facial regions across demographic groups [37]. This is particularly relevant for age estimation, where subtle texture cues may vary across populations.

2.5.5 Explainability in Video-Based Systems

Most explainability techniques were originally developed for image-based tasks and are therefore typically applied to individual video frames rather than to video sequences as a whole. When used in video-based applications, attribution maps can vary across frames due to changes in pose, illumination, facial expression, and image quality, making the interpretation of model behaviour more challenging than in static-image settings.

Recent research has explored the extension of explainability methods to video data, including temporally aggregated explanations and approaches that account for temporal dependencies between consecutive frames [105]. Nevertheless, explainability in video-based age estimation remains relatively underexplored when compared to image-based age estimation.

In this dissertation, explainability is incorporated through the application of Integrated Gradients to the age estimation model. Rather than analysing temporal consistency of explanations across video sequences, the focus is placed on understanding which facial regions contribute most strongly to individual age predictions and on validating the reliability of the generated attribution maps through insertion and deletion experiments. This analysis provides insights into the decision-making process of the model while complementing the quantitative evaluation of age estimation performance.

2.6 Summary of Relevant Works

To position this dissertation within the current landscape of video-based age estimation, we summarize a set of representative and recent works that explicitly address temporal information, pose and expression variability, and reproducible evaluation protocols. The goal of this comparison is twofold: (i) to clarify which modeling strategies have been explored to exploit video dynamics (e.g., temporal attention, temporal smoothing, pose-aware frame selection), and (ii) to identify a strong and recent methodological baseline to replicate and extend in the experimental plan. In particular, the benchmark protocol and semi-supervised video framework proposed by

Bešenić *et al.* are especially aligned with the objectives of this thesis, as they formalize reproducible offline and online evaluation for video-based age estimation and explicitly target temporal stability, which is a key practical challenge when moving from images to videos [10].

Table 2.4 consolidates the main characteristics of the selected works, including the datasets employed, the core ideas used to handle temporal information, headline performance results as reported by the authors, and the main limitations identified in each approach. While the reviewed methods differ in supervision level, data type, and evaluation protocol, a consistent conclusion emerges: exploiting temporal structure can improve either accuracy or, more importantly, the *stability* of age predictions over time. However, the magnitude of these gains depends strongly on dataset scale, variability, and the availability of suitable video-oriented training labels [10, 20, 21, 106].

Table 2.4: Representative works related to video-based age estimation.

Year	**Dataset**	**Main Contribution**	**Key Finding**	**Ref.**
2020	UvA–NEMO	CNN–RNN architecture combining spatial and temporal attention for age estimation from facial expression videos.	Improved performance compared with appearance-only approaches.	[20]
2022	AFAD, CACD + videos	Head-pose-aware frame selection to improve video-based age estimation.	Reduced prediction variability across video sequences.	[21]
2022	Idiap Video Age	Mobile-oriented age analysis framework incorporating temporal information.	Strong performance for minor/adult classification.	[106]
2024	Casual Conversations	Video age estimation benchmark and temporal framework based on pseudo-labeling.	Reduced MAE from 5.25 to 4.95 years and improved stability by over 50%.	[10]

Chapter 3

Methods

3.1 Reference Framework

Rather than strictly reproducing a single existing methodology, this dissertation draws on recent work in video-based age estimation. Current approaches broadly range from end-to-end spatiotemporal models, which learn relationships across video frames directly, to modular pipelines in which age is estimated at the frame level and subsequently combined through temporal smoothing or aggregation.

Within this context, the framework proposed by Bešenić et al. in *Let Me Take a Better Look: Towards Video-Based Age Estimation* [10] was selected as the main conceptual reference because of its close alignment with the scope of this dissertation. It considers age estimation in less constrained video conditions, relies on a modular pipeline that can be reproduced using pretrained components, and explicitly evaluates subject-level aggregation strategies. These characteristics make it a suitable basis for studying the behaviour of individual pipeline stages without requiring the development of a dedicated end-to-end video model.

The reference framework detects faces independently in each frame, performs frame-level age estimation, and subsequently aggregates the resulting predictions to obtain a more stable subject-level estimate. It also distinguishes between offline and online evaluation settings, enabling both full-sequence analysis and incremental prediction, while explicitly considering the variability of frame-level estimates over time.

This dissertation adopts the general decomposition and evaluation philosophy of the framework, as well as the use of the Casual Conversations dataset [95], while redefining several methodological components. In particular, greater emphasis is placed on the selection and evaluation of modern face detection models and on the integration of recent age estimation approaches. This allows the influence of face detection quality on downstream age estimation performance to be examined more directly, while preserving a modular and reproducible experimental structure.

The framework therefore serves as a methodological baseline rather than a template to be reproduced unchanged. The proposed pipeline extends this foundation with more recent components and additional analyses aimed at improving robustness and addressing the specific research

questions of this dissertation.

3.2 Proposed System Pipeline

The proposed system processes less constrained video data through sequential stages of face detection, age estimation, temporal smoothing, and subject-level aggregation, as illustrated in Figure 3. The following sections describe the implementation and evaluation of each stage.

Before constructing the final age estimation pipeline, a face detector selection stage is performed. Several state-of-the-art face detection models are evaluated on the WIDER FACE [49] Hard subset because this split concentrates the most challenging detection conditions, including small faces, partial occlusions, and large pose variations, making it particularly suitable for differentiating the robustness of modern detectors under conditions closer to those encountered in unconstrained video. The comparison considers quantitative detection performance under challenging conditions, including small faces, occlusions, and large pose variations. The detector achieving the best overall performance is then selected and integrated into the proposed pipeline.

The pipeline begins with face detection on the input video frames using the selected detector. The detected face regions are subsequently cropped using the predicted bounding boxes with a 20% spatial expansion around the detected face, preserving additional surrounding facial context while reducing the risk of excluding relevant facial regions near the bounding-box boundaries. No additional landmark-based alignment is applied before age estimation. This preprocessing step provides a more consistent facial crop across frames while retaining the original facial geometry used by the downstream age estimation model.

Following preprocessing, frame-level age estimation is performed using the MiVOLO model [61]. MiVOLO was selected due to its competitive performance in recent age estimation literature and its compatibility with the face-only inference setting adopted in this work, allowing its integration as a pretrained estimator without requiring additional model training. Each detected face is processed independently, producing an age prediction for every valid frame. At this stage, temporal information is not yet incorporated, and predictions are generated solely from the visual content of individual frames.

The next stage consists of temporal aggregation and smoothing. Since frame-level predictions may fluctuate due to image quality variations, facial expressions, motion blur, or detection inconsistencies, temporal smoothing is applied to improve prediction stability across time. The resulting frame-level estimates are then aggregated to obtain a final subject-level age prediction representative of the entire video.

In parallel with the prediction pipeline, several evaluation and analysis procedures are performed. Face detection performance is assessed using the WIDER FACE [49] benchmark during detector selection. The resulting age estimation pipeline is subsequently evaluated on the Casual Conversations V1 and V2 datasets [95, 107], enabling comparison with previous literature and a detailed analysis of performance across age groups and demographic categories. Explainability is investigated primarily through Integrated Gradients [108], selected because it provides pixel-level

attributions with a theoretically grounded formulation and is well suited to identifying fine-grained facial regions contributing to age predictions. Rather than relying solely on visual inspection of the attribution maps, their faithfulness is further assessed using insertion and deletion validation experiments, which quantify how model predictions respond when highly attributed regions are progressively introduced or removed. Although additional attribution methods could provide complementary perspectives, the analysis is deliberately centred on a single explanation technique so that its behaviour can be examined consistently and validated in greater depth across the evaluation set.

A key characteristic of the proposed framework is the independent evaluation of its individual components. Face detection is assessed separately before integration with age estimation, while age prediction and explainability analyses are conducted using a common downstream pipeline. This modular design facilitates reproducibility and enables a clearer interpretation of the contribution of each stage to the overall system.

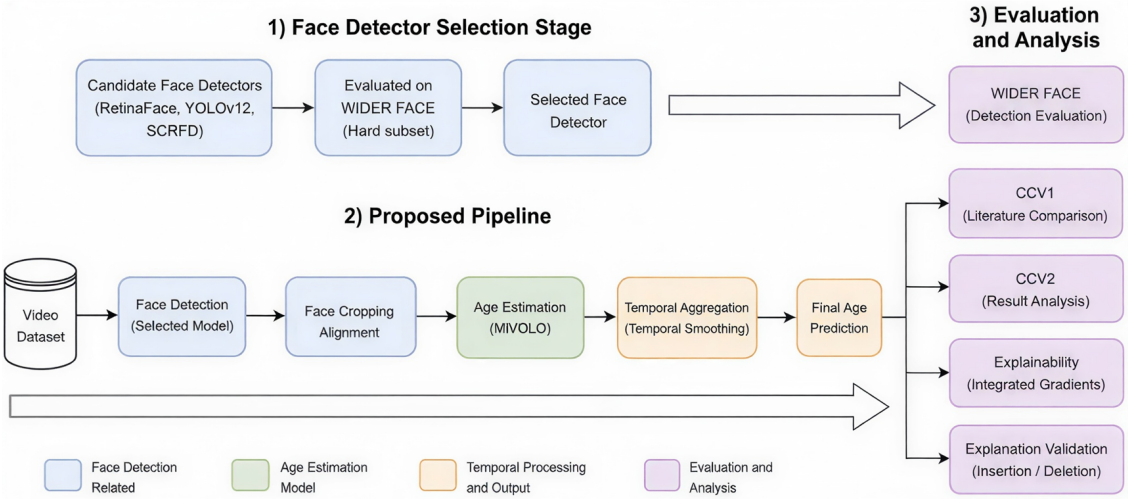


Figure 3: Overview of the proposed modular pipeline for face detection and age estimation from video, illustrating the main processing stages.

3.3 Datasets

This dissertation addresses face detection and age estimation under less constrained conditions, requiring the use of multiple complementary datasets. Since no single dataset provides annotations for face detection, age estimation, and temporal analysis simultaneously, the experimental design combines image-based datasets with video-based benchmarks, each serving a specific role within the proposed pipeline.

3.3.1 WIDER FACE

For face detection evaluation, the WIDER FACE dataset [49] is used as the primary benchmark, with particular focus on the Hard subset. The intended application scenario of this work is

age estimation from less constrained video, where subjects may present variations in pose, facial visibility, scale, and image quality across frames. Although the WIDER FACE Hard subset also contains more extreme cases than those expected in the final age estimation setting, its high concentration of small faces, occlusions, and large pose variations provides a demanding stress test for comparing detector robustness under challenging visual conditions.

The subset is therefore not intended to reproduce the exact deployment scenario of the proposed pipeline, but rather to distinguish candidate detectors under difficult conditions that may partially occur in unconstrained video. In this work, WIDER FACE is used exclusively to compare candidate detection models and support the selection of the most appropriate detector for the proposed pipeline.

Figure 4: Examples of challenging images from the WIDER FACE Hard subset [49].

3.3.2 Labeled Faces in the Wild (LFW)

The Labeled Faces in the Wild (LFW) dataset [14] was used exclusively in the preliminary face detection study presented in Section 4.1. LFW contains more than 13,000 face images collected from the web under unconstrained conditions, exhibiting substantial variations in pose, illumination, facial expression, background, image quality, and partial occlusions. These characteristics make the dataset suitable for evaluating the robustness of face detection algorithms in realistic scenarios.

Figure 5: Example images from the Labeled Faces in the Wild (LFW) dataset, illustrating variations in pose, illumination, facial expression, background, and image quality.

In this dissertation, LFW was not used for model selection or age estimation. Instead, it served as the evaluation benchmark for the preliminary comparison of Haar Cascade, MTCNN, and RetinaFace, providing initial experience with face detector assessment and performance analysis under challenging real-world conditions.

3.3.3 Casual Conversations

For video-based evaluation, this work follows the general protocol introduced by Bešenić *et al.* [10], which utilises subsets of the Casual Conversations dataset [95]. The dataset contains less constrained video recordings collected from participants with diverse demographic backgrounds and recording conditions, making it particularly suitable for evaluating age estimation systems in realistic scenarios.

Two subsets of Casual Conversations are used throughout this dissertation. Casual Conversations V1 is employed to enable direct comparison with previously published work, while Casual Conversations V2, being a more recent release that was not available to the earlier studies used for comparison, is used for the large-scale evaluation of the proposed framework, including demographic analysis and explainability experiments.

For the Casual Conversations V2 evaluation, a total of 5,559 subjects were retained after the data filtering procedure. The resulting dataset covers an age range from 18 to 81 years, with a mean age of 30.3 years and a median age of 28 years. Figure 6 presents the distribution of subjects across 5-year age groups. Most participants are concentrated between 20 and 35 years of age, while the number of subjects decreases progressively for older age ranges.

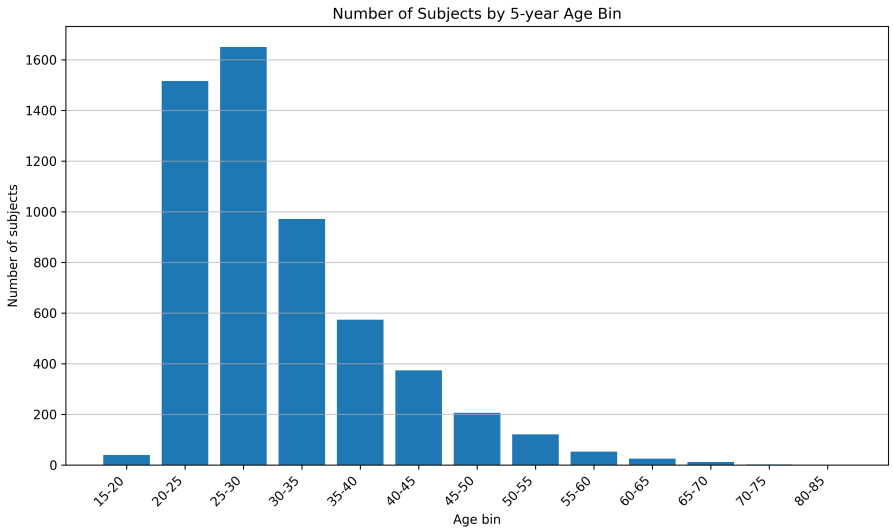


Figure 6: Distribution of subjects across 5-year age groups in the Casual Conversations V2 dataset.

Table 3.1 summarises the main demographic characteristics of the dataset. The participant population is relatively balanced with respect to gender and includes subjects from multiple countries and skin-tone categories, supporting the demographic analyses presented later in this dissertation.

Table 3.1: Summary of the demographic characteristics of the Casual Conversations V2 dataset.

Characteristic	**Value**
Eligible subjects	5,559
Age range	18–81 years
Mean age	30.3 years
Median age	28 years
Cisgender women	2,951
Cisgender men	2,380
Other gender identities	228
India	1,756
Brazil	1,632
Philippines	995
Indonesia	399
United States of America	374
Mexico	367
Vietnam	10
Fitzpatrick types represented	I–VI

The demographic diversity of Casual Conversations V2, combined with its less constrained recording conditions, provides a challenging evaluation environment that more closely reflects

real-world deployment scenarios than controlled benchmark datasets. However, the age distribution is imbalanced, with substantially fewer subjects represented in the oldest age groups. This characteristic should be considered when interpreting age-specific performance results presented later in this dissertation.

The combination of WIDER FACE and Casual Conversations enables a multi-stage evaluation of the proposed framework. WIDER FACE provides a challenging benchmark for detector selection, and Casual Conversations enables the assessment of temporal behaviour and subject-level age estimation in realistic video scenarios.

3.4 Evaluation Metrics

The evaluation of the proposed system is designed to assess both the performance of individual components and the overall behaviour of the pipeline in image-based and video-based scenarios. Given the modular structure adopted in this dissertation, different metrics are employed to analyse face detection accuracy, age estimation performance, and temporal consistency.

For face detection, evaluation is conducted on the WIDER FACE dataset using standard metrics based on precision-recall analysis. In particular, Average Precision (AP) is used to quantify the ability of each model to correctly localize faces under challenging conditions, including small scales, occlusions, and complex backgrounds. These metrics enable a comparative assessment of candidate detectors and support the selection of the most suitable model for integration into the pipeline.

For age estimation, the primary quantitative metric is the Mean Absolute Error (MAE), defined as the average absolute difference between predicted and ground-truth ages. MAE is widely adopted in the literature due to its interpretability and direct comparability across different methods, although it can still be influenced by large prediction errors. In addition, complementary measures such as exact accuracy and one-off accuracy are considered by grouping predictions into age intervals, providing further insight into model performance under both strict and relaxed evaluation criteria.

For video-based evaluation, temporal behaviour is analysed through stability-oriented metrics that capture the variability of predictions across frames. In particular, the standard deviation of frame-level predictions is used to quantify temporal consistency, with lower values indicating more stable and reliable behaviour. This analysis allows the assessment of the impact of prediction smoothing and temporal aggregation strategies on the overall robustness of the system.

Finally, qualitative evaluation is conducted through the visual inspection of prediction trajectories and explainability maps. This complementary analysis supports the interpretation of quantitative results and enables the identification of consistent patterns, failure cases, and potential sources of error within the proposed pipeline.

3.5 Experimental Setup

This section describes the experimental configuration adopted throughout this dissertation. The objective is to ensure reproducibility while providing a clear description of how the different components of the proposed framework are integrated and evaluated.

Face detection models were first evaluated independently on the WIDER FACE dataset to identify the most suitable detector for the proposed pipeline. Following the detector selection stage, age estimation experiments were conducted using the MiVOLO model operating in face-only mode without additional fine-tuning.

The final age estimation pipeline consisted of face detection followed by age prediction. Face localisation was performed using the YOLOv12m face detector selected during the face detection evaluation stage. Detected faces were filtered using confidence and size constraints, and when multiple faces were present in a frame, the most relevant detection was selected before being cropped and forwarded to the age estimation model.

Video-based experiments were conducted using the Casual Conversations dataset. For the large-scale evaluation on Casual Conversations V2, each subject was represented by 20 frames sampled at two-second intervals throughout the video duration. Since the shortest recording lasted 40 seconds, all subjects contributed the same number of sampled frames.Consecutive sampled frames were therefore separated by approximately two seconds, reducing temporal redundancy while maintaining coverage across the recording. In total, 111,340 frames from 5,559 subjects were processed.

Frame-level age predictions were generated independently for each valid face crop. To improve temporal consistency, a rolling median filter with a window size of three frames was applied to the resulting predictions. Subject-level age estimates were then obtained by aggregating predictions across all valid frames belonging to the same individual.

Several aggregation strategies were evaluated, including the mean, median, trimmed mean, confidence-weighted averaging, and confidence-based frame selection. For the confidence-based strategies, the face detector confidence score associated with each detected face was used as a heuristic indicator of detection quality. These scores represent the detector's confidence in the presence of a face and should not be interpreted as calibrated measures of age-prediction reliability. Moreover, because detector scores tend to be concentrated toward high values for successfully detected faces, their ability to distinguish between reliable and unreliable age predictions may be limited. Among these approaches, the raw mean strategy produced the most reliable subject-level estimates and was therefore adopted throughout the subsequent experiments.

Only subjects with at least eight valid frame-level predictions were considered eligible for evaluation. All 5,559 subjects satisfied this requirement and were therefore retained for analysis. Ground-truth ages and demographic metadata provided by the dataset were preserved throughout the evaluation process, enabling the age-group, demographic, error, and explainability analyses presented in Chapter 4.

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 5090 GPU. Unless otherwise specified, default model parameters were used, as the primary objective of this work was the evaluation and integration of state-of-the-art face detection and age estimation methods rather than model retraining or hyperparameter optimisation.

3.6 Face Detection Evaluation Protocol

All face detection models are evaluated on the WIDER FACE dataset, with particular focus on the Hard subset. For each detector, inference is performed independently on all images, and predictions are stored following the standard WIDER FACE evaluation format.

To improve robustness on challenging cases, multi-scale inference is performed by resizing each image to three target scales of 800, 1200, and 1600 pixels and combining the resulting detections. Horizontal flipping is also considered to increase detection coverage. A low confidence threshold of 0.005 is used to retain a broad set of candidate detections, followed by non-maximum suppression (NMS) with an Intersection over Union (IoU) threshold of 0.3 to remove redundant bounding boxes.

For each image, detected bounding boxes are represented by their top-left coordinates, width, height, and confidence score, and are sorted in descending order of confidence. This ensures compatibility with the official WIDER FACE evaluation protocol and enables the generation of precision–recall curves.

Performance is summarized using Average Precision (AP), following the metric definition introduced previously. A detection is considered a valid match when its bounding box reaches the IoU criterion used by the evaluation protocol.

In addition to the quantitative evaluation, qualitative inspection of detection outputs is conducted to analyse failure cases, particularly in scenarios involving small faces, occlusions, and challenging imaging conditions. This complementary analysis helps identify limitations that may not be fully captured by aggregate performance metrics alone.

Chapter 4

Experimental Results

This chapter presents the experimental results of the proposed framework. It begins with a preliminary face detection study conducted as part of a curricular project related to the dissertation topic. The chapter then presents the face detector selection process, age estimation model selection, and a comprehensive evaluation on the Casual Conversations datasets. Finally, explainability analyses and validation experiments are presented to assess the reliability of the generated explanations.

4.1 Preliminary Face Detection Study

Prior to the main work presented in this dissertation, a preliminary study was conducted as part of a curricular project to familiarise with the topic of facial analysis and the evaluation of face detection systems. The study focused on comparing classical and deep learning-based face detectors under less constrained conditions, providing practical experience with dataset preparation, performance evaluation, and quantitative analysis methodologies that were later applied throughout this dissertation.

The study evaluated three widely used face detection approaches: Haar Cascade, MTCNN, and RetinaFace [1–3]. Experiments were performed on the Labelled Faces in the Wild (LFW) dataset [14], which contains images collected under real-world conditions with substantial variations in pose, illumination, facial expression, occlusion, and image quality. Unlike many previous comparisons that focus primarily on detection accuracy, the evaluation considered multiple complementary metrics, including detection accuracy, small-face detection rate, average detection time, edge localisation failures, and detection count variability.

Table 4.1 summarises the main results obtained in this preliminary study.

Table 4.1: Summary of the preliminary face detection study on the LFW dataset [14]. Adapted from the curricular project developed prior to this dissertation.

Metric	Haar Cascade	MTCNN	RetinaFace
Average Precision (%)	99.37	**99.99**	99.71
Small Face Detection (%)	0.46	**2.83**	2.74
Detection Time (ms)	**25.93**	68.10	173.87
Detection Count StdDev	0.32	0.49	**0.56**

Among them, MTCNN [2] achieved the highest detection accuracy and small-face detection rate, with small faces defined as faces occupying a relatively small bounding-box area within the image, while RetinaFace [3] provided comparable performance at the expense of a higher computational cost.

Although this preliminary study does not constitute a direct contribution of the present dissertation, it provided valuable practical experience in face detector evaluation and reinforced the importance of detector selection in facial analysis pipelines. The insights obtained motivated the more extensive face detector assessment presented in the following sections, where detector performance was evaluated on the WIDER FACE [49] benchmark before integration into the proposed age estimation framework.

4.2 Face Detection Model Selection and Evaluation

Before performing age estimation, it was necessary to select a face detection model capable of accurately locating faces under a wide range of real-world conditions. Since the age estimation stage relies directly on the quality of the detected facial regions, errors at this stage can propagate through the pipeline and negatively affect subsequent age predictions. For this reason, several recent face detection models were evaluated and compared using the WIDER FACE dataset [49].

The objective of this evaluation was to identify the detector that provided the best balance between accuracy and robustness, particularly under challenging conditions involving small faces, occlusions, and large variations in pose and image quality. Both quantitative and qualitative analyses were conducted to assess detector performance and to better understand the strengths and limitations of each approach.

The results presented in the following subsections guided the selection of the face detector adopted throughout the remainder of this dissertation, ensuring that the subsequent age estimation experiments were based on reliable facial crops.

4.2.1 Quantitative Performance

The performance of the evaluated face detection models is summarized in Table 4.2. Results are reported in terms of Average Precision (AP) across the Easy, Medium, and Hard subsets of the WIDER FACE dataset [49].

It is important to note that the AP values reported in Table 4.2 correspond to the specific implementations, pretrained weights, and evaluation settings adopted in this dissertation. Therefore, these results should not be directly compared with the values presented in Chapter 2, which were obtained from the original publications under their respective experimental protocols. Differences in model variants, training procedures, inference strategies, and benchmark evaluation settings can lead to noticeable variations in reported performance.

Table 4.2: Face detection performance on the WIDER FACE dataset (AP at IoU = 0.5).

Model	**Easy**	**Medium**	**Hard**
RetinaFace	0.9286	0.8940	0.7301
YOLOv12s	0.9462	0.9191	0.7592
YOLOv12m (best)	**0.9498**	**0.9432**	**0.8877**
SCRFD	0.8599	0.8597	0.7540

Among the evaluated models, YOLOv12m [60] achieves the best overall performance, obtaining AP values of 0.9498, 0.9432, and 0.8877 on the Easy, Medium, and Hard subsets, respectively. The improvement is particularly evident on the Hard subset, which contains a large number of small, heavily occluded, and low-resolution faces. This suggests that YOLOv12m [60] is better able to maintain detection accuracy under challenging real-world conditions.

RetinaFace achieves competitive performance on the Easy and Medium subsets but exhibits a larger drop on the Hard subset. Although RetinaFace [3] is widely recognised as a strong face detector, the performance observed here reflects the specific implementation and evaluation configuration used in this work rather than the best results reported in the literature.

YOLOv12s [60] provides stronger performance than RetinaFace [3] on the Hard subset while maintaining high accuracy on the easier subsets. However, it remains consistently below YOLOv12m, indicating that the larger model benefits from increased representational capacity when dealing with difficult detection scenarios.

SCRFD [59] obtains performance comparable to YOLOv12s [60] on the Hard subset but underperforms on the Easy and Medium subsets. This behaviour suggests a trade-off between robustness to difficult cases and overall detection accuracy across the full dataset.

Overall, the results indicate that YOLOv12m achieves the highest detection performance among the evaluated models. Its superiority is consistent across all evaluation subsets and is particularly evident on the Hard subset, which contains a large number of small, occluded, and low-resolution faces. Based on these results, YOLOv12m was selected as the face detector integrated into the age estimation pipeline used throughout the remainder of this dissertation.

4.2.2 Precision–Recall Analysis

To further analyse detector performance, precision–recall (PR) curves are presented in Figure 7. These curves provide a more detailed view of the trade-off between detection precision and

recall across different confidence thresholds, using an IoU threshold of 0.5 for matching predicted and ground-truth bounding boxes.

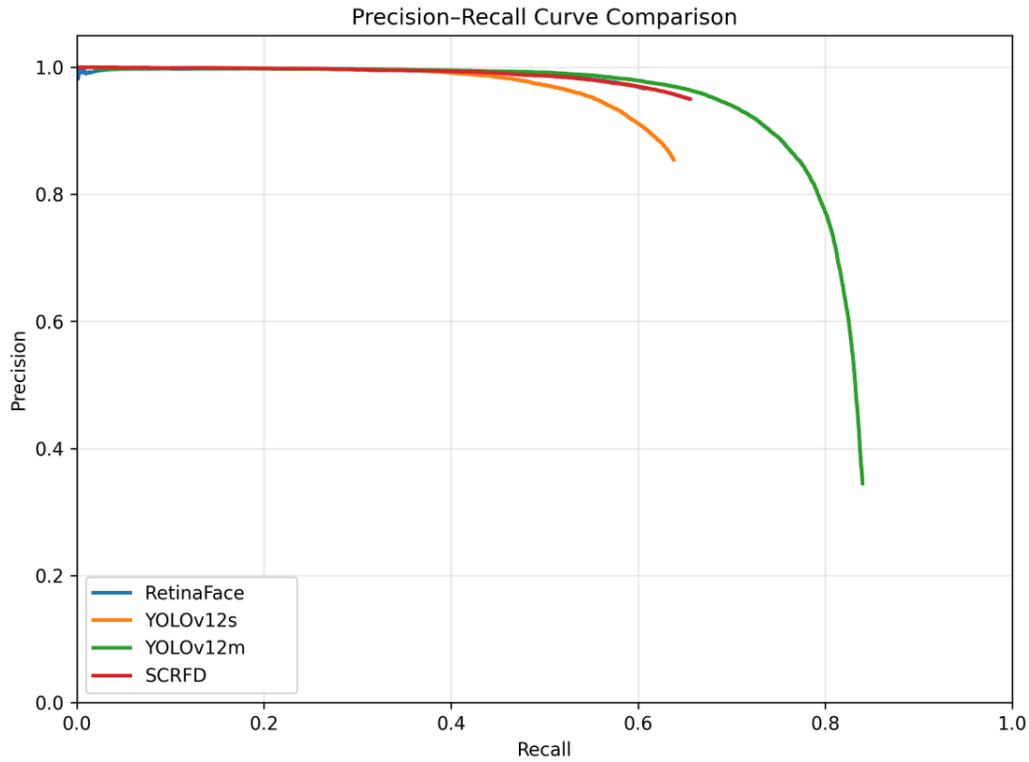


Figure 7: Precision–recall curves for the evaluated face detection models on the WIDER FACE dataset, illustrating the trade-off between precision and recall across different confidence thresholds.

YOLOv12m [60] demonstrates the most favourable behaviour, maintaining high precision across a wide range of recall values. This indicates that the model is capable of detecting a large proportion of faces while keeping the number of false positives relatively low. In contrast, YOLOv12s [60] exhibits a sharper decline in precision at higher recall levels, suggesting reduced robustness when attempting to detect more challenging faces.

RetinaFace [3] and SCRFD [59] present more stable but overall lower performance curves, reflecting their comparatively lower detection capacity in complex scenarios.

It is worth noting that none of the curves reaches a recall value of 1.0. This behaviour is expected on the WIDER FACE Hard subset, which contains numerous extremely small, heavily occluded, and low-quality faces. Some of these faces are not detected by the evaluated models under any confidence threshold, causing recall to saturate before reaching its theoretical maximum. Consequently, the curves terminate at the highest achievable recall obtained by each detector rather than extending to the right boundary of the plot.

These results reinforce the quantitative findings reported in Table 4.2 and further highlight the superior performance of YOLOv12m [60] in balancing detection accuracy and robustness under challenging conditions.

4.2.3 Detection Behaviour Analysis

To further analyse the behaviour of the selected detector, an additional evaluation was conducted to investigate detection performance across different face-size categories. This analysis complements the quantitative results presented previously by examining how detection accuracy varies with the scale of the target face, which is a particularly relevant factor in the WIDER FACE Hard subset [49].

For this purpose, ground-truth faces were grouped according to the area of their annotated bounding boxes. Faces with an area smaller than 32×32 pixels were classified as *small*, faces between 32×32 and 96×96 pixels as *medium*, and faces larger than 96×96 pixels as *large*. These thresholds follow the scale categories commonly adopted in face detection literature.

Table 4.3 summarizes the number of ground-truth faces, successfully detected faces (true positives), missed faces, and the corresponding recall values for each category.

Table 4.3: Detection performance across face-size categories on the WIDER FACE Hard subset.

Face Size	**Ground Truth**	**True Positives**	**Missed Faces**	**Recall (%)**
Small	28775	22729	6046	78.99
Medium	8606	8334	272	96.84
Large	2327	2276	51	97.81

The results show a clear dependence between detection performance and face size. While YOLOv12m achieves recall values above 96% for medium and large faces, performance decreases to 78.99% for small faces. This behaviour is expected given the characteristics of the WIDER FACE Hard subset, where many faces occupy only a limited number of pixels and contain less visual information. Small faces are therefore inherently more difficult to localise accurately than larger faces.

Despite this reduction in recall, the detector is still able to correctly identify the majority of small faces present in the dataset. Considering that small faces represent the largest category within the Hard subset, this result helps explain the strong overall performance achieved by YOLOv12m in the previous evaluations.

Further insight is provided by the confidence score distribution shown in Figure 8. A clear separation between true positives and false positives can be observed. True detections are predominantly associated with higher confidence values, whereas false positives are concentrated at lower confidence ranges.

This behaviour indicates that the detector is able to effectively discriminate between correct and incorrect detections using its confidence scores. The concentration of false positives at low confidence levels suggests that many incorrect detections could be filtered out by increasing the confidence threshold. However, such filtering would also remove a fraction of true detections, illustrating the trade-off between precision and recall.

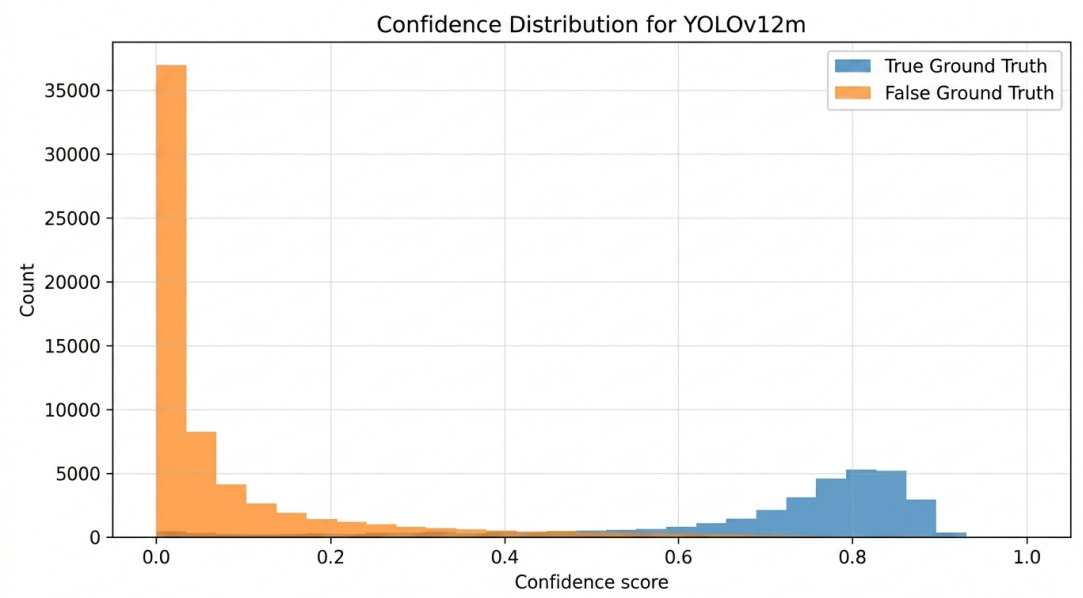
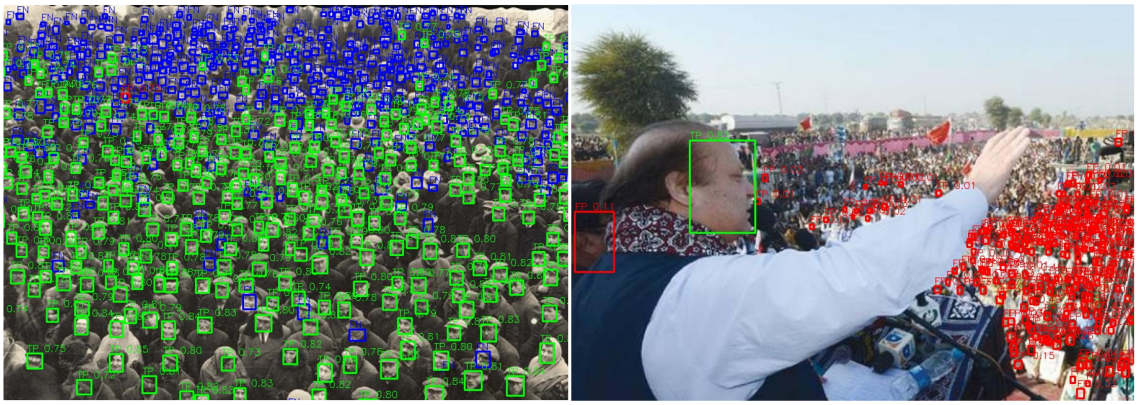


Figure 8: Confidence score distribution for YOLOv12m detections. True positives correspond to detections matched to a ground-truth face according to the evaluation protocol, while false positives correspond to unmatched detections.

4.2.4 Qualitative Analysis of Failure Cases

To further analyse the limitations of the selected detector, qualitative examples are presented in Figure 9. These examples highlight challenging scenarios in the WIDER FACE dataset and provide insight into the model's behaviour beyond quantitative metrics.

In Figure 9 (A), a dense crowd scenario is shown, where a large number of faces are present at very small scales. While many faces are correctly detected, a significant number of ground-truth faces (highlighted in blue) are not detected. This behaviour is primarily due to the extreme density and low resolution of the faces, which makes them inherently difficult to detect. In such scenarios, even state-of-the-art models struggle to maintain high recall, as distinguishing facial features becomes increasingly challenging. Despite these missed detections, the model successfully identifies the majority of visible faces, demonstrating strong robustness under highly congested conditions.



A) B)

Figure 9: Examples of detection behaviour under challenging conditions. Green boxes represent correct detections, blue boxes indicate missed ground-truth faces, and red boxes correspond to false positives. (A) Dense crowd scenario with numerous small faces. (B) Scene with a large number of low-confidence false positives, from [49]

Figure 9 (B) illustrates a different type of behaviour, where a large number of detections in the crowd are classified as false positives (red boxes). At first glance, this may appear as a significant failure; however, a closer inspection reveals that many of these detections correspond to regions that visually resemble faces or contain partial facial features. Furthermore, these detections are generally associated with low confidence scores, reflecting uncertainty in the model's predictions.

This behaviour is consistent with the use of a low detection threshold, which is intentionally adopted to maximise recall and ensure that difficult faces are not missed. As a result, the model produces a higher number of candidate detections, including uncertain ones. In practical applications, these false positives can be effectively reduced by applying a higher confidence threshold, without significantly affecting true positive detections.

Overall, these examples highlight the trade-off between recall and precision in challenging scenarios. While some missed detections and false positives are observed, they are largely a consequence of the dataset difficulty and the chosen inference configuration, rather than fundamental limitations of the model. The detector demonstrates strong robustness in complex scenes and provides reliable confidence estimates that can be leveraged for further filtering.

4.3 Age Estimation Model Selection

Following the selection of the face detection model, the next step was to choose an age estimation model for the proposed framework. Several age estimation approaches have been introduced in recent years, ranging from traditional convolutional neural networks to more recent transformer-based architectures. Among these, MiVOLO was selected due to its strong performance reported in the literature and its ability to jointly exploit facial and contextual information when available.

Before applying the model to the main evaluation dataset, an initial assessment was conducted on the Casual Conversations V1 dataset. This preliminary experiment served two purposes. First,

it allowed the behaviour of MiVOLO to be evaluated on a publicly available benchmark under controlled conditions. Second, it enabled a direct comparison with the work of Bešenić et al., who reported some of the strongest results available for video-based age estimation on the Casual Conversations benchmark, providing a valuable reference point for assessing the performance of the selected model.

4.3.1 MiVOLO Evaluation on Casual Conversations V1

Before applying the age estimation pipeline to the Casual Conversations V2 dataset [107], an initial evaluation was performed on Casual Conversations V1 [95]. This experiment was designed to verify the suitability of MiVOLO [61] for the proposed framework and to establish a direct comparison with previously reported results on the same benchmark.

The evaluation followed a frame-based protocol similar to the CCMiniIMG setting described by Bešenić et al [10]. Each video was represented by a collection of extracted frames, which were processed independently. Faces were first detected using the selected YOLOv12m detector [60] and subsequently cropped before being provided as input to MiVOLO [61]. Age predictions obtained from individual frames belonging to the same subject were then aggregated to produce a final subject-level prediction.

A total of 3 011 subjects were processed, corresponding to more than 600 000 extracted frames. Face detection was successful for 601 649 frames, resulting in an effective detection rate above 99%, ensuring that the age estimation stage was not significantly affected by missed detections. The final evaluation was performed at subject level using the aggregated predictions.

Table 4.4 summarises the obtained results.

Table 4.4: Subject-level age estimation performance of MiVOLO on Casual Conversations V1.

Metric	Value
MAE (Mean Absolute Error)	6.24 years
RMSE (Root Mean Squared Error)	7.99 years
Median Absolute Error	5.10 years
Accuracy within $\pm$5 years	49.1%
Accuracy within $\pm$10 years	80.2%

The obtained results indicate that MiVOLO is capable of capturing the general age distribution of the dataset, achieving a strong correlation between predicted and ground-truth ages. In particular, approximately half of all predictions fall within five years of the true age, while more than 80% remain within a ten-year margin.

Although the model demonstrates good overall performance, the error distribution is not uniform across age groups. To better understand this behaviour, additional analyses were conducted using age-bin confusion matrices and age-specific error measurements.

4.3.2 Comparison with Previous Literature

To place the obtained results into context, the performance of MiVOLO [61] was compared with the CCMiniIMG [95] results reported by Bešenić et al [10]. Following the same evaluation protocol, age predictions were grouped into three age categories (18–30, 31–45, and 46–85 years), and classification accuracy was computed at subject level. This metric therefore reflects the ability of the model to assign subjects to the correct age group rather than predict the exact chronological age.

Table 4.5 presents the comparison between both approaches.

Table 4.5: Comparison with the CCMiniIMG results reported by Bešenić et al.

Metric	Bešenić et al.	MiVOLO Results
Accuracy	73.06%	72.0%
MAE	5.25 years	6.24 years

The results obtained with MiVOLO [61] are broadly consistent with those reported by Bešenić et al [10]. In terms of age-group classification, the difference between the two approaches is relatively small, with MiVOLO achieving an accuracy of 72.0% compared with 73.06% reported in the literature. This indicates that both methods are similarly effective at assigning subjects to the correct age category.

A larger difference is observed in the MAE values, where MiVOLO [61] achieved an error of 6.24 years compared with 5.25 years reported by Bešenić et al. Although both methods produce comparable age-group classifications, the lower MAE reported by Bešenić et al. suggests greater precision when estimating the exact age of individual subjects.

One possible explanation for this difference lies in the age estimation models themselves. While MiVOLO [61] is a general-purpose age estimation model designed to operate across a wide range of datasets and demographic conditions, the approach proposed by Bešenić et al [10]. combines a MobileFaceNet backbone [16] with the DLDLv2 [109] age estimation framework. Differences in architecture, training data, and optimization objectives may therefore contribute to the lower MAE reported by Bešenić et al [10].

Despite the slightly higher error, MiVOLO [61] was selected for this work because it represents a recent state-of-the-art age estimation model that has demonstrated strong generalization performance across diverse datasets. In addition, MiVOLO [61] provides continuous age predictions, supports straightforward integration into video-processing pipelines, and does not require task-specific retraining for the datasets considered in this dissertation. These characteristics make it particularly suitable for evaluating age estimation under less constrained video conditions.

Overall, the performance gap between the two approaches remains relatively small. The obtained results show that MiVOLO [61] remains competitive with a specialised framework developed specifically for video-based age estimation, supporting its use in the subsequent experiments conducted on the Casual Conversations V2 dataset [107].

4.3.3 Discussion of Results

Figure 10 presents the confusion matrix obtained using 5-year age intervals.

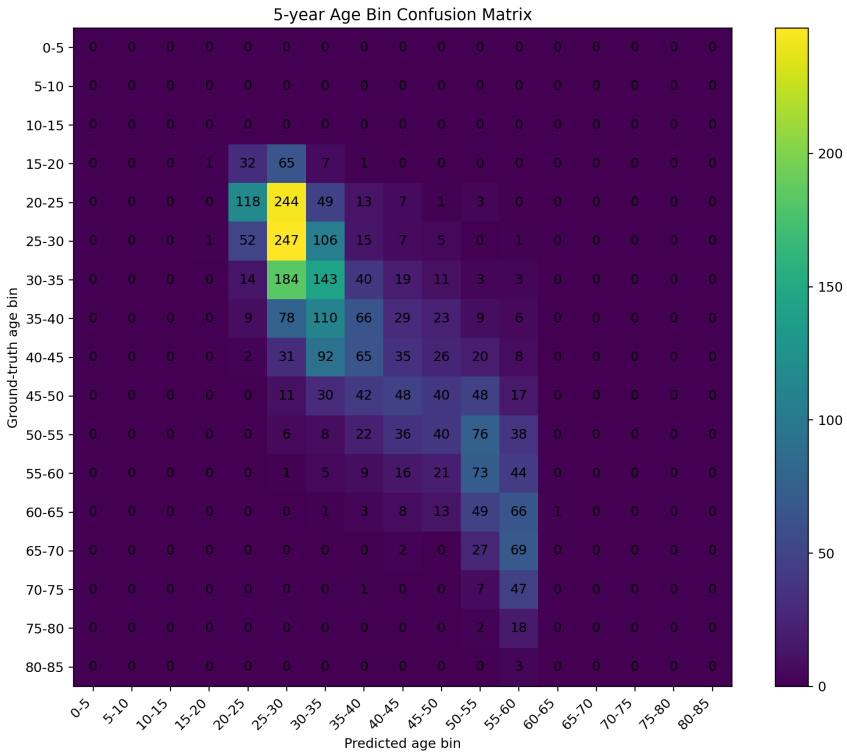


Figure 10: Confusion matrix using 5-year age intervals on Casual Conversations V1.

The confusion matrix reveals that most predictions are concentrated around the main diagonal, indicating that the model generally assigns subjects to the correct age group or to a neighbouring age interval.

A closer inspection of the matrix shows that performance is strongest for young and middle-aged adults, particularly between 25 and 45 years of age, where predictions tend to remain close to the ground-truth age bins. In contrast, the distribution becomes progressively more dispersed for older subjects. Starting around the 55–60 age range, a growing number of individuals are assigned to younger age groups, indicating a tendency to underestimate age at later stages of life.

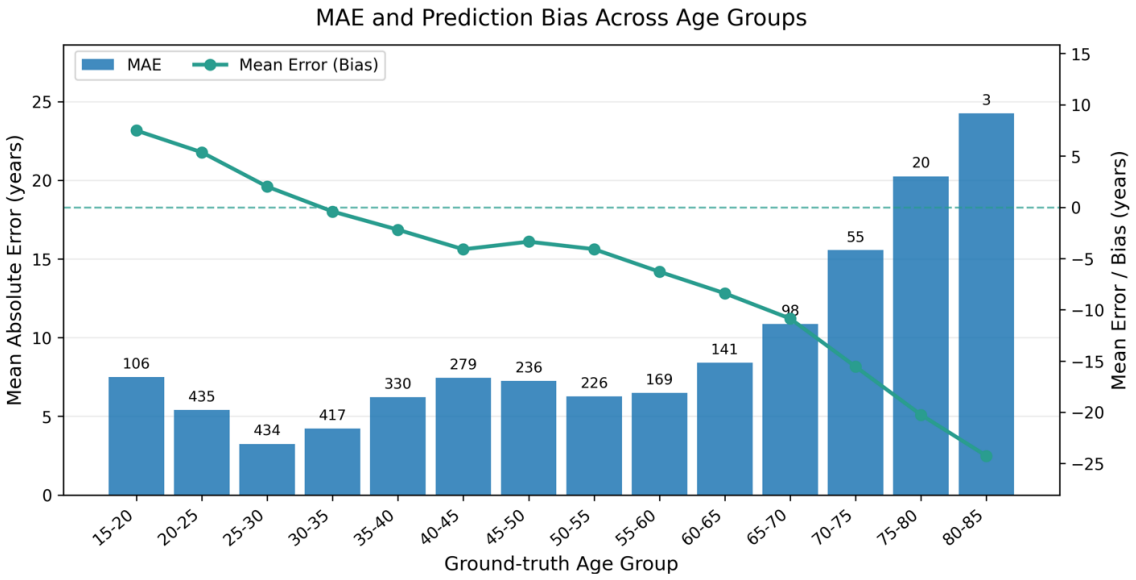


Figure 11: Mean Absolute Error (MAE) and mean prediction error (bias) across 5-year age groups, in Casual Conversation dataset. Numbers above the bars indicate the number of subjects in each age group.

This behaviour is further illustrated in Figure 11, which combines the MAE and the mean prediction error for each age group. The MAE remains relatively low between 20 and 35 years of age, where errors are generally below five years, but increases progressively for older subjects, reaching substantially larger values beyond 65 years of age.

The bias curve provides additional insight into the direction of these errors. For younger age groups, the mean error is positive, indicating a tendency to slightly overestimate age. Around the 30–35 age range, the bias approaches zero, suggesting that predictions are approximately unbiased on average. However, for older subjects the bias becomes increasingly negative, revealing a systematic tendency to underestimate age. This underestimation grows progressively with age and becomes particularly pronounced in the oldest groups.

The combined analysis of MAE and bias helps explain the patterns observed in the confusion matrix. While increasing MAE indicates that predictions become less accurate for older individuals, the negative bias demonstrates that these errors are not random but predominantly occur in the direction of younger age estimates.

This trend is frequently reported in age estimation studies. In many age estimation datasets, younger and middle-aged adults are more heavily represented than elderly individuals, resulting in models that are exposed to fewer examples of advanced ageing during training. In addition, the visual manifestations of ageing become increasingly heterogeneous at later stages of life. Individuals of the same chronological age may exhibit markedly different facial characteristics due to factors such as genetics, lifestyle, health conditions, and environmental exposure. Consequently, distinguishing between older age groups becomes considerably more challenging than distinguishing between younger adults.

Overall, the results indicate that MiVOLO [61] achieves comparable, although slightly lower, performance than the reference approach of Bešenić et al. on Casual Conversations V1, while maintaining similar age-dependent error patterns on the larger V2 evaluation. The observed increase in error and the growing underestimation bias at advanced ages are consistent with findings reported in previous studies, suggesting that this limitation reflects a broader challenge in age estimation rather than a model-specific weakness.

4.4 Age Estimation on Casual Conversations V2

To further evaluate the robustness and generalisation capability of the proposed age estimation framework, an additional assessment was conducted using the Casual Conversations V2 [107] dataset. Unlike the datasets used during model development, Casual Conversations V2 [107] contains videos acquired under a wide range of real-world conditions, including variations in age, ethnicity, recording environments, lighting conditions, and camera quality. Consequently, it provides a challenging benchmark for evaluating age estimation performance in less constrained scenarios. This section first describes the characteristics of the dataset and the evaluation protocol, followed by a detailed analysis of the obtained age estimation results.

4.4.1 Overall Performance

Before selecting the final subject-level aggregation strategy, several approaches were evaluated on the Casual Conversations V2 [107] dataset. Since each subject is represented by multiple sampled frames, frame-level predictions must be combined into a single age estimate before computing subject-level performance. Table 4.6 presents the results obtained for five representative aggregation strategies.

Table 4.6: Comparison of subject-level aggregation strategies on the Casual Conversations V2 dataset.

Aggregation Strategy	**MAE**	**RMSE**	**±5 years**	**±10 years**
Mean Raw	**6.04**	**8.04**	53.4%	**82.6%**
Confidence-weighted Mean	**6.04**	**8.04**	53.3%	**82.6%**
Trimmed Mean Raw	6.05	8.11	**53.8%**	82.5%
Median Raw	6.09	8.22	54.1%	82.4%
Median Top-Confidence	6.18	8.34	53.3%	82.0%

The evaluated aggregation strategies produced broadly similar results, with differences of less than 0.15 years in MAE across all configurations. The best overall performance was achieved by the Mean Raw strategy, which obtained a MAE of 6.04 years and a Root Mean Squared Error (RMSE) of 8.04 years. The Mean Smoothed strategy achieved the same MAE while providing

slightly higher accuracy within both the $\pm$5-year and $\pm$10-year error margins. However, the improvements were marginal and accompanied by a slightly higher RMSE.

Given its simplicity and its superior overall error metrics, the Mean Raw strategy was selected for the remaining experiments presented in this chapter.

Using this configuration, the model achieved a mean absolute error (MAE) of 6.04 years and a root mean squared error (RMSE) of 8.04 years. The difference between these two metrics suggests that while most predictions remained relatively close to the ground-truth age, a smaller number of subjects produced larger estimation errors that increased the overall RMSE. This behaviour is expected in less constrained datasets, where variations in appearance, image quality, facial expressions, pose, and age-related characteristics can introduce additional prediction uncertainty.

In terms of practical accuracy, 53.4% of subjects were predicted within $\pm$5 years of their chronological age, while 82.6% were predicted within $\pm$10 years. Considering the diversity of recording conditions and demographic characteristics present in Casual Conversations V2, these results indicate that the model was generally capable of producing age estimates that remained reasonably close to the true age for the majority of subjects.

Additional insight into the error distribution is provided in Figure 12. The histogram reveals a clear concentration of predictions at lower error values, with the frequency of observations decreasing progressively as the error increases. Most subjects exhibit absolute errors below approximately 10 years, while larger errors occur less frequently and form a long right-hand tail extending towards higher error values. This suggests that the overall performance is largely driven by a substantial number of accurate predictions, with a relatively small subset of challenging subjects contributing disproportionately to the highest errors.

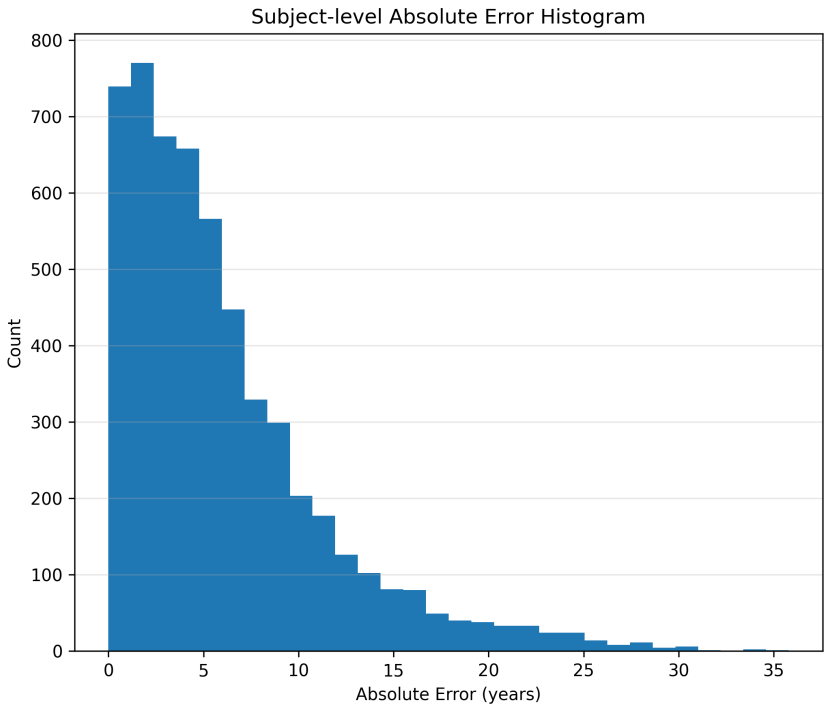


Figure 12: Distribution of subject-level absolute age estimation errors on the Casual Conversations V2 dataset.

Overall, the obtained results demonstrate that the proposed framework achieves consistent age estimation performance on a large-scale dataset collected under real-world conditions. Nevertheless, the presence of a small number of larger errors indicates that performance is not uniform across all subjects, motivating a more detailed investigation of age-specific and demographic effects in the following sections.

4.4.2 Performance Across Age Groups

To investigate the effect of age on prediction accuracy, the mean absolute error (MAE) was computed separately for each 5-year age group. The results are presented in Figure 13.

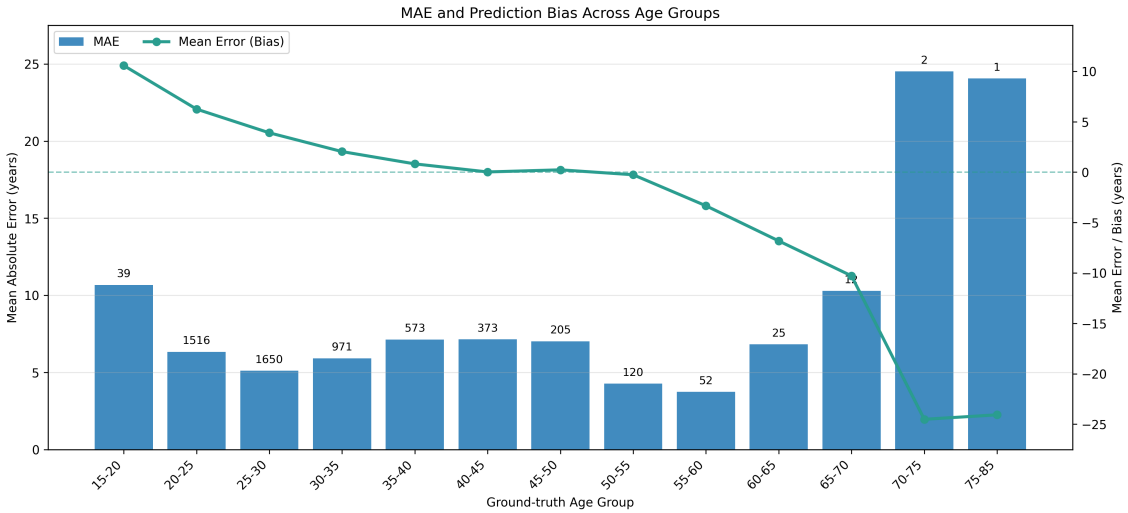


Figure 13: Mean Absolute Error (MAE) and mean prediction error (bias) across 5-year age groups, in Casual Conversation V2 dataset. Numbers above the bars indicate the number of subjects in each age group.

Overall, the model achieved relatively consistent performance across most adult age groups. Subjects between approximately 20 and 65 years of age generally exhibited MAE values ranging from 3.5 to 8 years, indicating stable age estimation performance throughout the majority of the dataset. The lowest errors were observed in the 50–55 and 55–60 age groups, where the average prediction error remained below 5 years.

A different pattern emerged for the oldest subjects. While performance remained relatively stable up to approximately 65 years of age, the MAE increased substantially for the 70–75 and 80–85 age groups, reaching values above 20 years. This behaviour is consistent with observations reported for the Casual Conversations V1 test set and suggests that age estimation becomes increasingly challenging at older ages.

One possible explanation is the limited representation of elderly individuals in many age estimation datasets used during model development. As a result, models often learn more discriminative age-related features for younger and middle-aged adults, while age progression patterns become less distinct at older ages. Furthermore, facial ageing tends to occur at different rates across individuals, increasing the variability of facial appearance within older age groups and making precise age estimation more difficult.

Despite the increase in error among the oldest subjects, the overall trend indicates that the proposed framework maintains relatively consistent performance across most age ranges, with a noticeable degradation occurring primarily in subjects above 70 years of age.

4.4.3 Performance Across Demographic Groups

To investigate the robustness of the proposed framework across different demographic populations, age estimation performance was analysed separately for each country represented in the

Casual Conversations V2 [107] dataset. Only countries with at least five subjects were considered, resulting in seven demographic groups.

Figure 14 presents the mean absolute error obtained for each country. The lowest error was observed for subjects from India (MAE = 4.86 years), followed by the Philippines (5.90 years) and Indonesia (6.13 years). Slightly larger errors were obtained for Brazil, Mexico and the United States, which all produced MAE values close to 7 years. The highest error was observed for Vietnam (12.16 years). However, this result should be interpreted with caution, as only ten subjects were available for this demographic group.

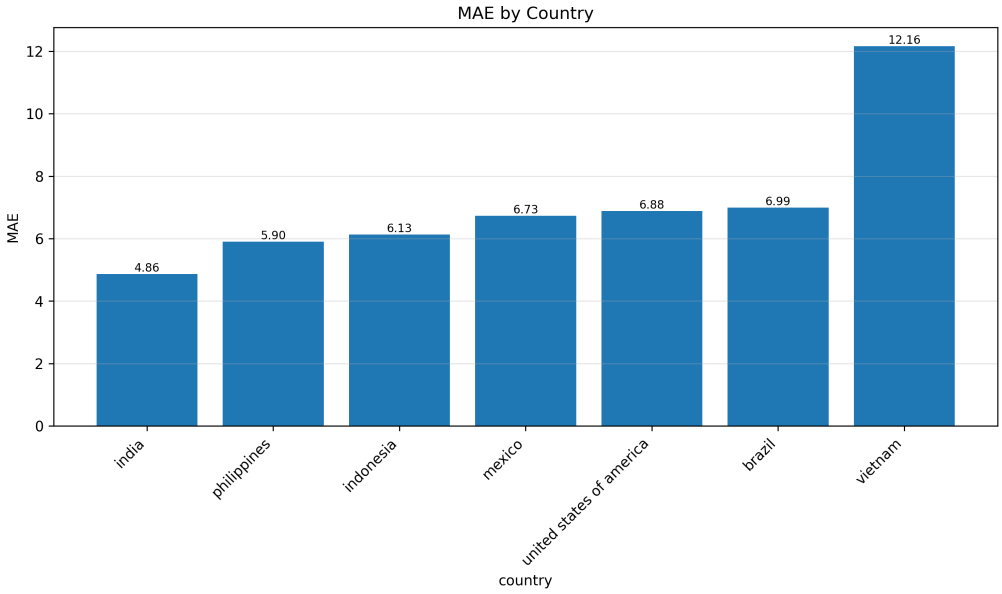


Figure 14: Mean absolute error by country.

The observed differences appear to be influenced, at least in part, by the number of available subjects within each demographic group. Countries with larger populations generally exhibited more consistent performance, whereas groups containing only a few hundred subjects, or fewer, tended to show greater variability in the estimated error metrics. This tendency becomes particularly evident for Vietnam, where the limited sample size coincides with substantially higher error values.

The overall pattern nevertheless suggests that the proposed framework maintains relatively consistent performance across the major demographic groups represented in the dataset. Excluding Vietnam, all countries achieved MAE values within a narrow range of approximately 4.8 to 7 years, indicating that the model generalises reasonably well across different populations despite variations in ethnicity, appearance, recording conditions and geographical origin.

4.4.4 Error Analysis

A detailed error analysis was performed to better understand the sources of prediction inaccuracies. Figure 15 presents the confusion matrix obtained using 5-year age bins.

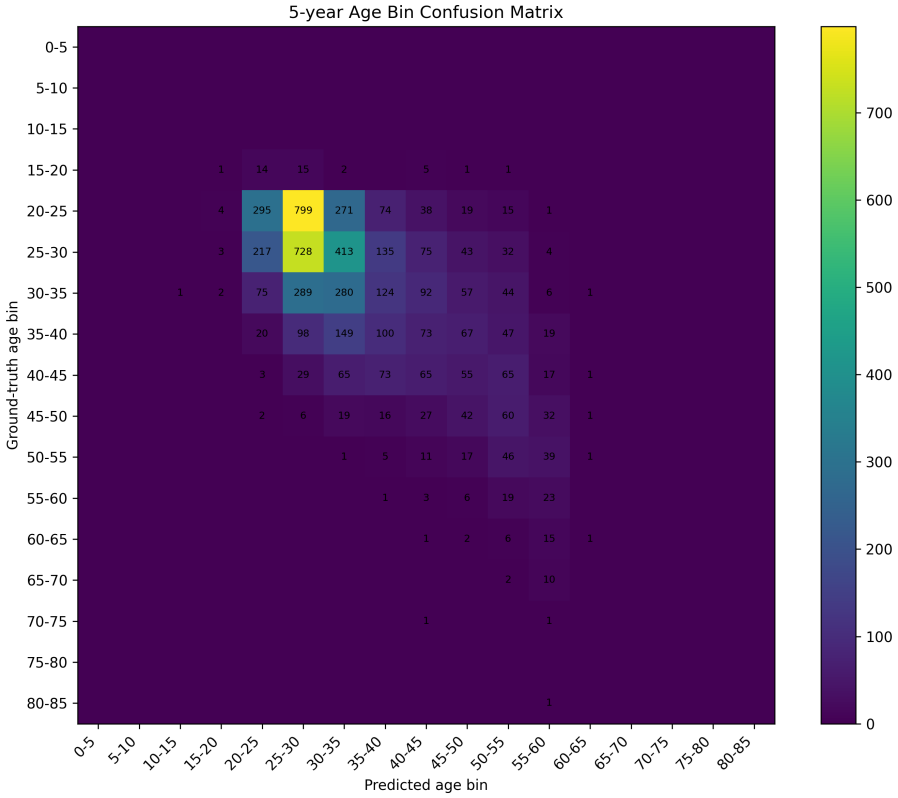


Figure 15: Confusion matrix using 5-year age bins.

Most predictions are concentrated around the main diagonal, indicating that the model generally assigns subjects to the correct age group or to neighbouring age intervals. Large age estimation failures are comparatively uncommon, which is consistent with the previously reported adjacent-bin accuracy exceeding 66%.

A noticeable tendency towards age overestimation can be observed throughout the matrix. Subjects belonging to younger age groups, particularly between 20 and 35 years of age, are frequently assigned to slightly older bins. This behaviour is also reflected by the positive prediction bias observed across all demographic groups, where predicted ages were consistently higher than the corresponding ground-truth ages.

The confusion matrix additionally highlights the difficulty of estimating age in older populations. While most age groups contain predictions distributed around neighbouring bins, subjects above approximately 65 years of age exhibit larger deviations and substantially lower prediction counts. This behaviour is likely influenced by the severe reduction in available training and evaluation samples within the oldest age ranges, as previously shown in the age-distribution analysis.

Further insight into demographic performance differences is provided by Figures 16 and 17, which present the distribution of subject-level absolute errors across age groups for different countries and genders.

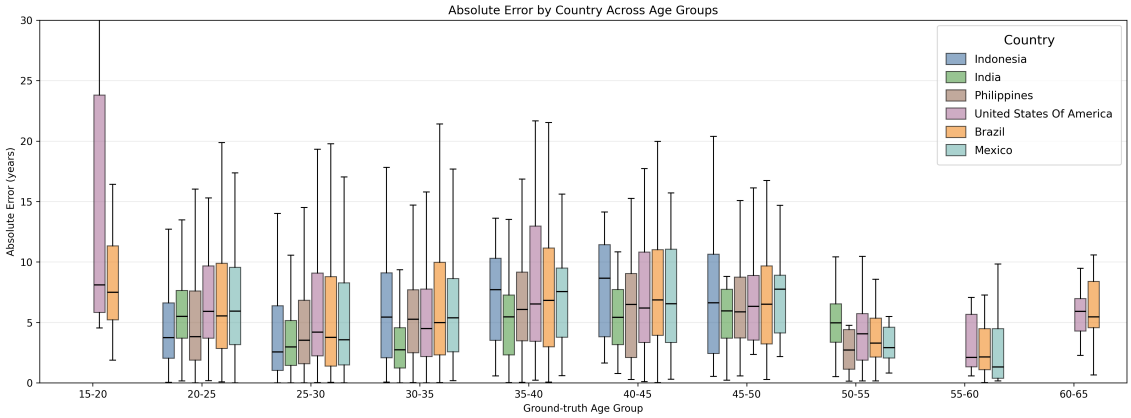


Figure 16: Distribution of absolute errors across countries and age groups.

Figure 16 shows that the relationship between age and estimation error is generally more pronounced than the differences observed between countries. Across most populations, the median absolute error remains relatively low for younger and middle-aged subjects, while larger errors and wider distributions become increasingly common in older age groups. This behaviour is consistent with the age-group analysis presented previously and suggests that age itself has a stronger influence on performance than country of origin.

For the younger age groups, particularly between 20 and 35 years, the distributions are broadly similar across countries, with median errors typically below six years. As age increases, the spread of the distributions also tends to increase, indicating greater variability in prediction quality. Although some differences between countries can be observed, no population consistently exhibits substantially better or worse performance across all age ranges. The larger variations visible in some age-country combinations should be interpreted with caution, as the number of available subjects decreases considerably in the older age groups.

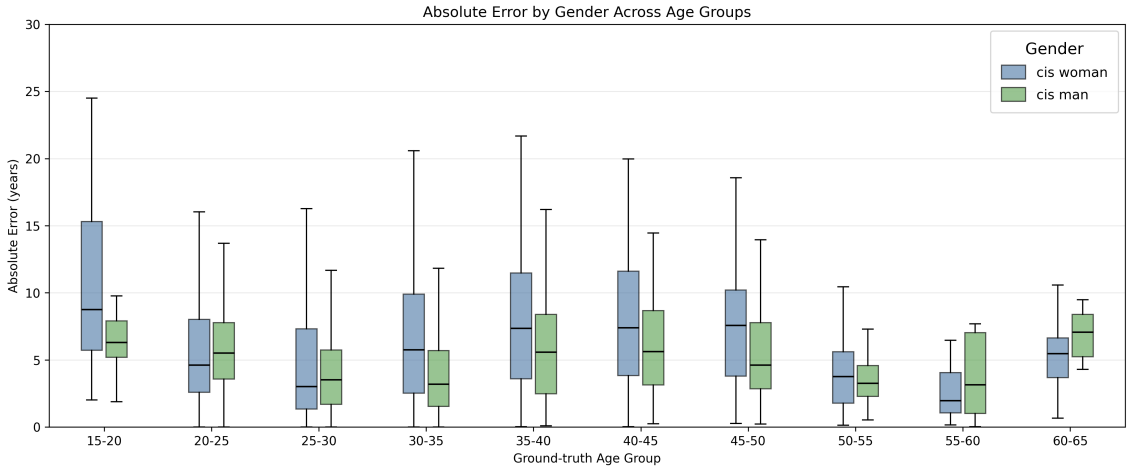


Figure 17: Distribution of absolute errors across age groups for cis women and cis men.

A similar analysis was performed for gender, as shown in Figure 17. To ensure meaningful

statistical comparisons, only the two largest demographic groups, namely cis women and cis men, were considered. The remaining gender categories contained substantially fewer subjects and were therefore excluded from this analysis.

The overall behaviour is broadly similar for both groups, with median absolute errors remaining relatively close across most age ranges. In addition, the interquartile ranges show considerable overlap, indicating that the typical prediction accuracy is comparable for cis women and cis men. As observed previously, the largest errors tend to occur in the older age groups, reflecting the increased difficulty of age estimation at advanced ages.

Despite these similarities, an interesting difference can be observed in the spread of the error distributions. Across most age groups, the upper whiskers associated with cis women extend to higher values than those of cis men, indicating the presence of larger maximum errors and a greater variability in prediction performance. This trend is particularly noticeable between 30 and 50 years of age, where the distribution for cis women exhibits a wider range of errors. While the median performance remains comparable between the two groups, these results suggest that the model encounters a greater number of difficult cases among cis women, leading to occasional larger prediction errors.

Taken together, these results indicate that age remains the dominant factor affecting performance, with error distributions generally increasing in variability for older subjects. Nevertheless, the gender analysis suggests that the model's predictions for cis women are associated with a wider range of errors, even though the typical accuracy observed for both groups remains largely comparable.

4.5 Explainability Analysis

While the quantitative results presented in the previous sections provide insight into the accuracy of the proposed framework, they do not reveal which facial characteristics are being used by the model to estimate age. Understanding the basis of the model's predictions is particularly important in age estimation, where decisions may be influenced by a wide range of visual cues, including facial structure, skin texture, and age-related appearance changes.

To investigate the behaviour of the selected age estimation model, an explainability analysis was conducted using Integrated Gradients [108]. This method enables the identification of image regions that contribute most strongly to the final prediction, providing a visual interpretation of the model's decision-making process. Beyond qualitative inspection, additional analyses were performed to investigate attribution patterns across different age groups and to evaluate the temporal consistency of the generated explanations throughout video sequences.

4.5.1 Integrated Gradients Methodology

Integrated Gradients [108] was selected as the explainability method due to its strong theoretical foundation and its ability to assign attribution scores directly to the input pixels. Unlike simple gradient-based visualisation methods, which only analyse the local gradient around the original

image, Integrated Gradients estimates feature importance by accumulating gradients along a path between a reference image and the actual input. This makes the resulting attribution maps less dependent on a single local gradient evaluation and more suitable for analysing the contribution of facial regions to the predicted age.

In this work, Integrated Gradients [108] was applied to the MiVOLO [61] age estimation model through a differentiable adapter that exposes the age prediction output as a scalar tensor. Each detected face crop was resized to 224×224 pixels and normalised using ImageNet-style mean and standard deviation values. The reference input, or baseline, was defined as a black image. Attributions were then computed along a straight-line interpolation path between this baseline and the original normalised face crop. A total of 32 integration steps was used, with gradients computed in batches of 8 interpolated samples.

For each interpolated image, the gradient of the predicted age with respect to the input pixels was computed. The final attribution map was obtained by multiplying the average gradient along the interpolation path by the difference between the original input and the baseline. Since the objective was to identify the regions that most influenced the age prediction, independently of whether their contribution increased or decreased the predicted value, the absolute attribution values were used. The RGB-channel attributions were then summed into a single two-dimensional map. To reduce the influence of isolated extreme values, the attribution map was clipped at the 99th percentile and normalised to the range $[0, 1]$ before visualisation.

To facilitate interpretation, the resulting attribution maps were projected onto the original facial images. In addition, the facial region partition shown in Figure 18 was introduced to enable a more structured analysis of attribution patterns. The face was divided into six approximate regions corresponding to the forehead, eyes, nose, cheeks, mouth, and chin. These regions were obtained using MediaPipe Face Mesh [110] landmarks and should therefore be interpreted as practical facial-region masks rather than precise anatomical segmentations. This partitioning allowed attribution scores to be aggregated and compared across facial areas, rather than being analysed only at pixel level.

Figure 18: Facial region partition used for the quantitative analysis of attribution maps.

4.5.2 Qualitative Analysis of Attribution Maps

Figure 19 presents a representative example of an attribution map generated using Integrated Gradients. Warmer colours indicate regions with a stronger contribution to the age prediction, whereas cooler colours correspond to regions with lower influence.

Figure 19: Example of an Integrated Gradients attribution map overlaid on the original facial image.

Visual inspection of multiple attribution maps revealed that the model consistently concentrates attention on facial regions commonly associated with ageing. In particular, the forehead, cheeks, mouth region, and areas surrounding the nose frequently exhibit strong attribution values. These regions contain information related to skin texture, facial contours, wrinkles, and other appearance characteristics that naturally evolve with age.

Importantly, the model does not focus on a single facial feature. Instead, the attribution maps suggest that age estimation is based on information distributed across several facial regions. This behaviour is consistent with the multifactorial nature of facial ageing, where no individual feature alone is sufficient to determine age accurately.

4.5.3 Regional Attribution Patterns and Temporal Stability

To complement the qualitative analysis, attribution values were aggregated within the predefined facial regions shown in Figure 18. This allowed the relative importance of different facial areas to be quantified across the entire dataset.

Figure 20 presents the mean attribution assigned to each facial region. The nose, cheeks, and forehead emerge as the most influential regions, while the chin and eye regions receive lower attribution values on average. This suggests that the model relies primarily on central facial structures and surrounding skin characteristics when estimating age.

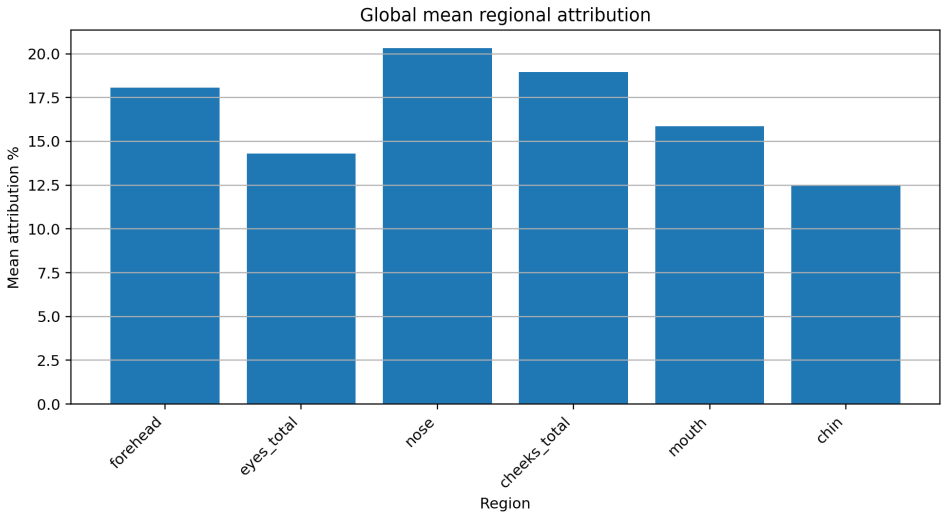


Figure 20: Mean attribution assigned to each facial region across the dataset.

To investigate whether attribution patterns change throughout the ageing process, regional importance was analysed across different age groups. The results are shown in Figure 21.

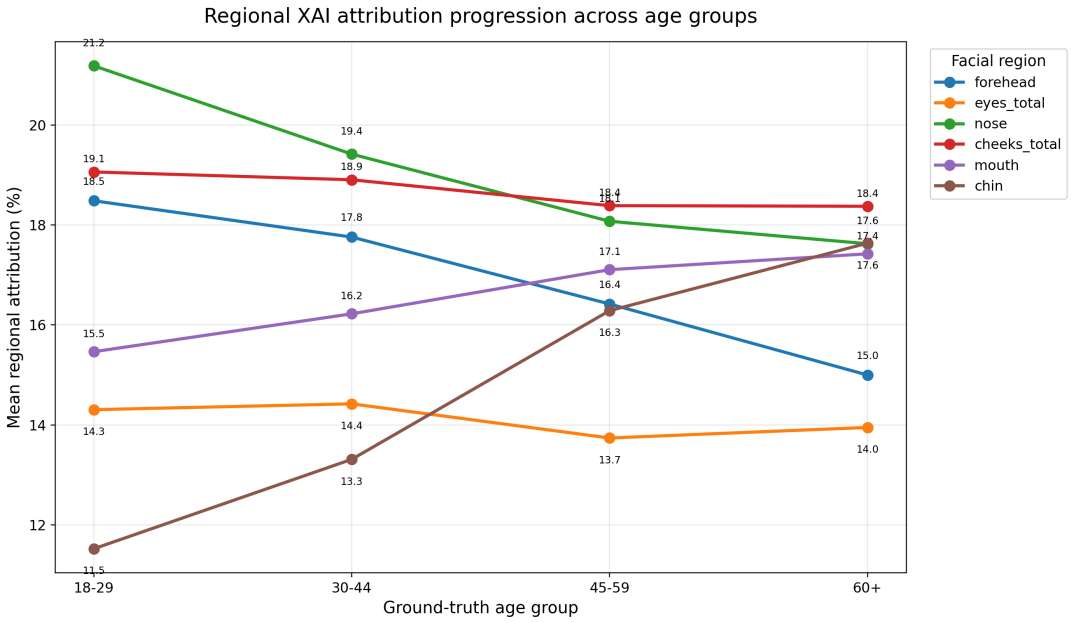


Figure 21: Mean regional attribution as a function of age group.

Several trends can be observed. The importance of the forehead decreases progressively with age, while the contribution of the mouth and chin regions increases. The cheeks maintain relatively stable importance across all age groups. These findings suggest that the model adapts its focus depending on the age of the subject, relying on different facial cues as age-related characteristics become more pronounced in different regions of the face.

Finally, the temporal consistency of the explanations was analysed by computing the cosine similarity between attribution vectors extracted from consecutive frames. Figure 22 shows the resulting distribution.

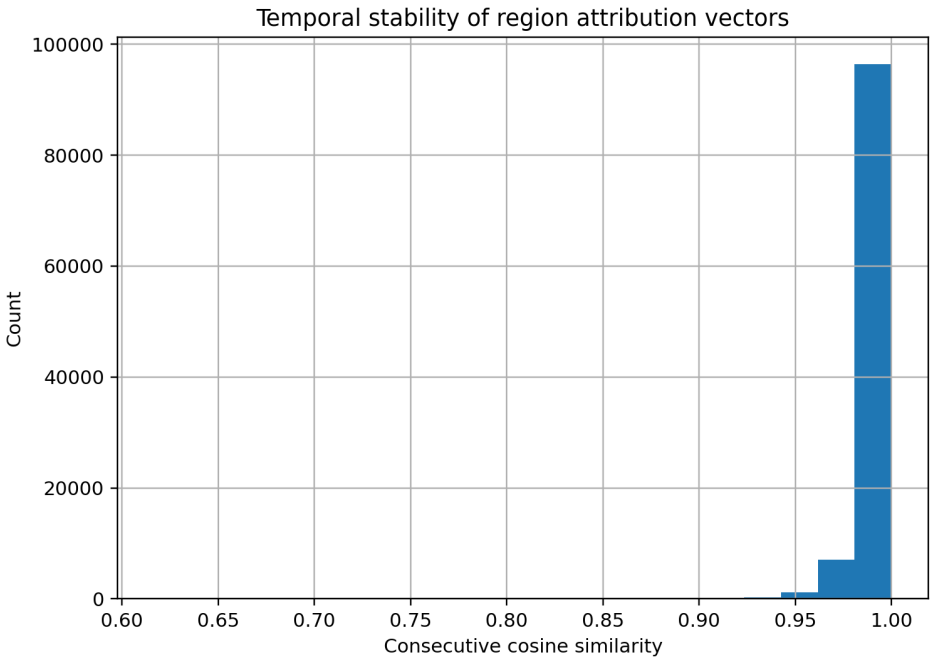


Figure 22: Distribution of cosine similarity between attribution vectors of consecutive frames.

The distribution is heavily concentrated near one, indicating that attribution patterns remain highly stable throughout video sequences. This suggests that small variations in pose, expression, or illumination do not substantially alter the regions used by the model for age estimation. Such behaviour increases confidence in the reliability of the generated explanations and indicates that the observed attribution patterns reflect consistent model behaviour rather than random fluctuations between frames.

4.5.4 Analysis of Incorrect Predictions

Although the overall performance of MiVOLO [61] was comparable to previously reported results, a small number of subjects produced substantially larger errors than the dataset average. Analysing these cases provides additional insight into the limitations of the model and the challenges associated with age estimation in less constrained video data.

Figure 23 presents the twenty subjects with the largest absolute age estimation errors. While most subjects in the dataset were associated with considerably smaller errors, a limited number of outliers exhibited deviations approaching or exceeding thirty years.

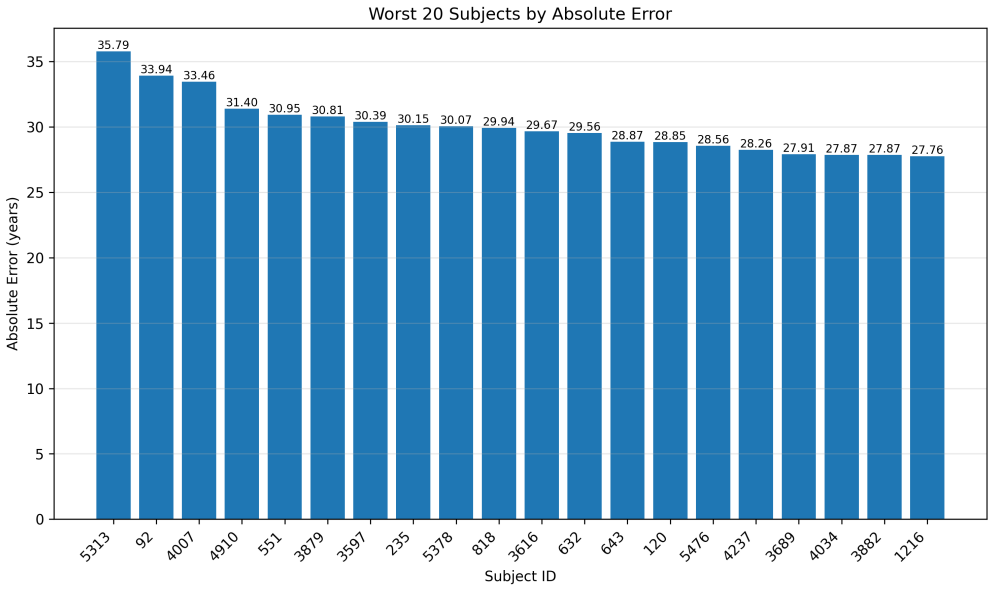


Figure 23: Twenty subjects with the largest absolute age estimation errors.

A closer inspection of these cases revealed several factors that may contribute to prediction failures. Many of the most challenging samples were captured under poor illumination conditions, contained noticeable image noise, or exhibited facial occlusions caused by glasses, shadows, or limited facial visibility. Such conditions reduce the amount of age-related information available to the model and can make reliable estimation considerably more difficult.

An example corresponding to the largest observed error is shown in Figure 24. The subject was recorded in a low-light environment, resulting in reduced facial detail and contrast. Despite the incorrect prediction, the Integrated Gradients [108] attribution map indicates that the model continued to focus on meaningful facial regions, particularly around the nose, cheeks, and central facial area. This suggests that the prediction error was not caused by attention being directed towards irrelevant background content, but rather by the limited visual information available in the image itself.

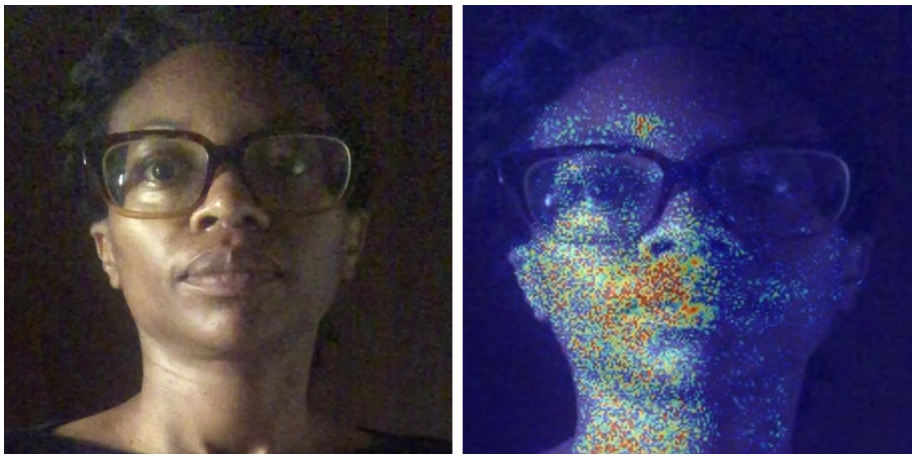


Figure 24: Largest age estimation error in the dataset (ground truth: 18 years; prediction: 53.8 years). Left: original image. Right: Integrated Gradients attribution map.

Another important consideration is the nature of the Casual Conversations [95, 107] dataset. Ground-truth ages are based on information voluntarily provided by the participants during dataset collection. Consequently, the reported age labels cannot be independently verified and may occasionally contain inaccuracies. While there is no indication that such cases are common, the possibility of incorrect self-reported ages should be considered when interpreting the largest prediction errors.

Overall, the analysis suggests that the most severe failures are primarily associated with difficult visual conditions and a small number of extreme outliers. Furthermore, the corresponding attribution maps indicate that the model generally continues to rely on plausible facial cues even when producing incorrect age estimates, supporting the observations presented in the previous explainability analyses.

4.6 Validation of Explanations

While attribution maps provide a visual indication of which image regions influence a model's prediction, visual inspection alone does not guarantee that the highlighted features are genuinely important for the decision-making process. To assess the reliability of the generated explanations, insertion and deletion tests were performed following a perturbation-based evaluation strategy.

The underlying principle is straightforward. If the regions identified by Integrated Gradients [108] contain information that is truly important for age estimation, removing those regions should substantially degrade the prediction, whereas reintroducing them should rapidly recover the model output. Comparisons against random perturbations provide a baseline for determining whether the observed behaviour is meaningful or merely a consequence of modifying the image.

4.6.1 Pixel Deletion Test

The deletion test evaluates whether pixels identified as important by Integrated Gradients [108] are genuinely required for the model's prediction. Starting from the original image, pixels were progressively removed in decreasing order of attribution magnitude, and the resulting prediction confidence was monitored after each perturbation step.

Figure 25 presents the average deletion curve obtained across the analysed samples.

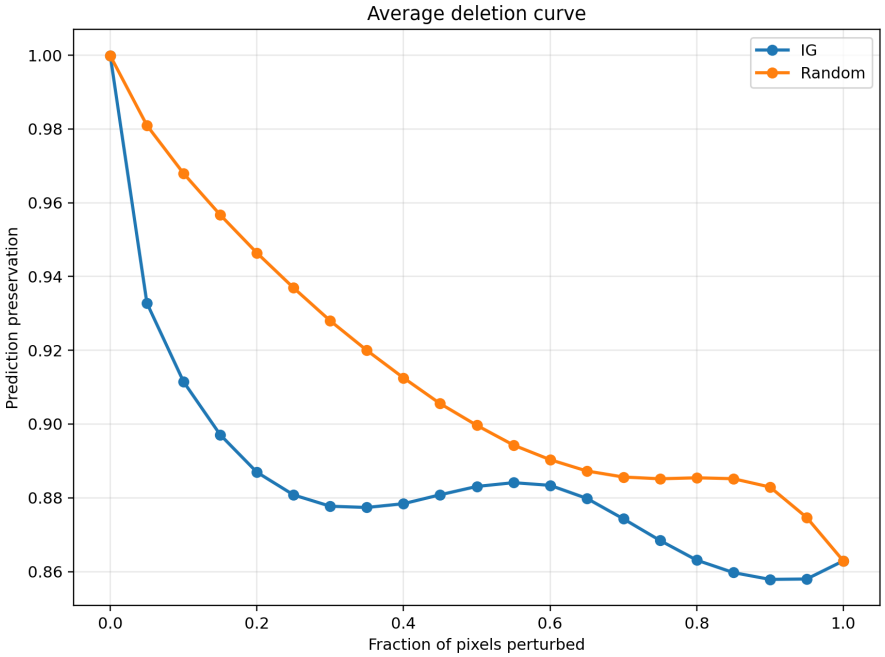


Figure 25: Average deletion curves for Integrated Gradients and random perturbations.

A substantially faster decrease in prediction preservation can be observed when pixels are removed according to the Integrated Gradients [108] ranking. In contrast, random perturbations produce a considerably slower decline. This behaviour indicates that the regions identified by Integrated Gradients contain information that is genuinely important for the age estimation process.

The sharp drop observed during the initial perturbation stages is particularly relevant, suggesting that a relatively small subset of highly attributed pixels contributes disproportionately to the final prediction.

4.6.2 Pixel Insertion Test

The insertion test provides a complementary perspective. Instead of removing pixels from the original image, the process starts from a heavily perturbed image and progressively restores pixels according to their attribution importance.

Figure 26 shows the average insertion curves obtained for Integrated Gradients [108] and random pixel restoration.

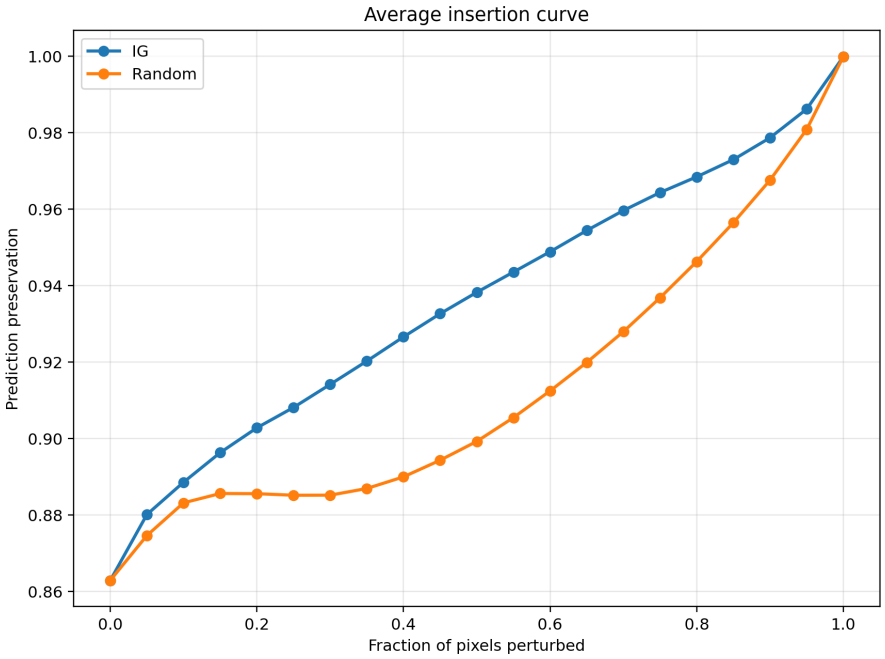


Figure 26: Average insertion curves for Integrated Gradients and random perturbations.

The prediction confidence recovers more rapidly when pixels are reintroduced according to the Integrated Gradients [108] ranking than when pixels are restored randomly. This result indicates that the highlighted regions contain information that is sufficient to reconstruct the model's prediction and further supports the relevance of the generated attribution maps.

Together, the insertion and deletion experiments provide complementary evidence that the explanations capture meaningful features used by the model during age estimation.

4.6.3 Discussion of Explanation Reliability

Figure 27 summarises the average behaviour observed across both perturbation experiments.

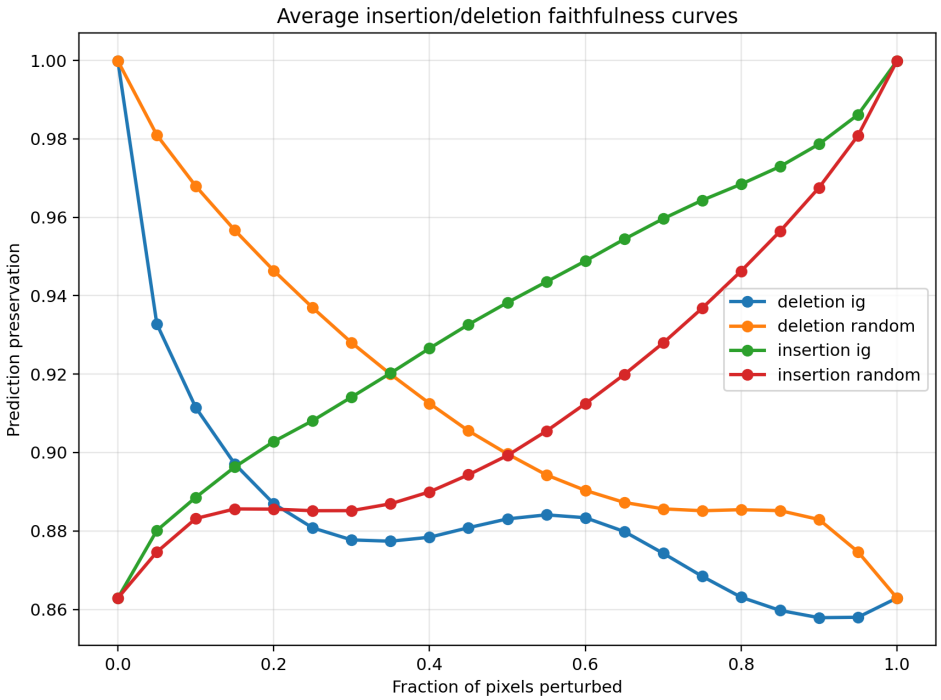


Figure 27: Average insertion and deletion faithfulness curves.

The consistent separation between Integrated Gradients [108] and the random attribution baselines in both validation experiments indicates that the generated explanations capture non-random model behaviour. Regions receiving high attribution scores tend to influence the prediction more strongly than randomly selected image regions, suggesting that Integrated Gradients identifies features that are relevant to the model's decision process.

The insertion and deletion curves are not perfectly monotonic because the effect of perturbing one image region is not independent of the remaining visual information. Age estimation depends on multiple correlated facial cues, including local texture, shape, and appearance, which are processed jointly by the network. Removing a highly attributed region can therefore change the relative importance of the remaining features, while restoring additional pixels can similarly redistribute the evidence supporting the prediction. These interactions can produce small local fluctuations during the perturbation process. Nevertheless, the global trends remain consistent: deleting highly attributed pixels generally degrades the prediction faster than random deletion, while restoring them generally recovers the prediction more effectively than random insertion.

The Area Under the Curve (AUC) analysis provides a quantitative measure of this behaviour. Relative to random perturbations, Integrated Gradients achieved an average deletion AUC gain of 0.0295 and an insertion AUC gain of 0.0231. Positive gains were observed in approximately 72.5% of samples for deletion and 66.6% for insertion. This indicates that, for the majority of samples, the ordering produced by Integrated Gradients prioritizes image regions that have a greater effect on the model's prediction than randomly selected regions.

The validation results are not uniform across all samples. In some cases, the Integrated Gradients ordering does not outperform the random baseline, which may reflect the distributed nature of age-related facial evidence and variations in image quality, pose, occlusion, or prediction confidence. Therefore, the results should not be interpreted as demonstrating perfect faithfulness for every individual explanation. Instead, they provide quantitative evidence that Integrated Gradients identifies prediction-relevant regions more consistently than random attribution across the evaluated dataset.

To illustrate this behaviour qualitatively, Figure 28 presents a representative example together with the corresponding insertion and deletion curves.

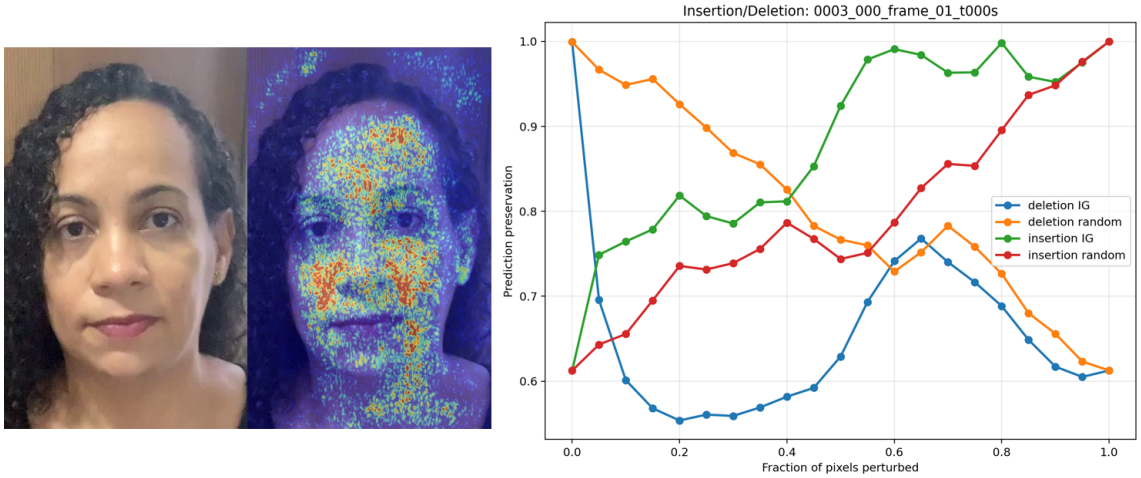


Figure 28: Example of faithfulness validation. The attribution map highlights regions that strongly influence the age prediction, while the corresponding insertion and deletion curves demonstrate that perturbing these regions produces a substantially larger effect than random perturbations.

In this example, removing highly attributed pixels causes a rapid deterioration of the prediction, whereas restoring the same regions quickly recovers the model output. The observed behaviour is consistent with the expected response of a faithful explanation method and provides additional evidence that the generated attribution maps reflect features genuinely used by the model rather than visually appealing but unrelated patterns.

Overall, the insertion and deletion experiments support the faithfulness of the Integrated Gradients [108] explanations generated throughout this work. While no explainability method can provide a complete representation of a deep neural network's internal reasoning, the obtained results indicate that the highlighted regions correspond to information that meaningfully influences the age estimation process.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This dissertation investigated face detection and age estimation from video under less constrained conditions, with particular emphasis on model evaluation, demographic analysis, and explainability. The work combined a review of the relevant literature with a systematic experimental study based on publicly available datasets and state-of-the-art deep learning models.

The first objective was to identify a suitable face detector for integration into a video-based age estimation pipeline. Several recent detectors were evaluated on the WIDER FACE dataset, with results showing that YOLOv12m achieved the highest overall performance across the Easy, Medium, and Hard subsets. In particular, the model obtained the strongest results on the Hard subset, which contains a large number of small and challenging faces. Based on these results, YOLOv12m was selected as the face detector used throughout the remainder of this work.

The second objective focused on age estimation. An initial evaluation was performed on the Casual Conversations V1 dataset to compare the selected approach with previously published work. The results obtained with MiVOLO achieved a level of performance comparable to that reported in the reference study, although some differences remained in terms of prediction error. Nevertheless, the overall results supported the suitability of the model for further evaluation on larger and more diverse datasets.

A more comprehensive assessment was subsequently conducted on the Casual Conversations V2 dataset. The results showed that age estimation performance varies considerably across age groups. Lower errors were generally observed for younger and middle-aged adults, whereas prediction accuracy decreased progressively for older individuals. The error analysis further revealed that prediction bias is not uniform across age ranges, with both overestimation and underestimation effects being observed depending on the age group considered. These findings are consistent with the broader challenges reported in the age estimation literature, particularly for subjects at more advanced ages.

Additional analyses were performed across demographic groups, allowing a broader assessment of model behaviour beyond aggregate performance metrics. While differences between

countries and genders were observed, the results suggested that age had a greater influence on performance than the demographic factors considered. The country-level analysis did not reveal a consistently superior or inferior performance for any specific population, whereas the gender analysis indicated similar median errors for cis women and cis men, although larger error variability was observed for cis women in several age groups.

The final part of the dissertation focused on explainability. Integrated Gradients was employed to investigate the facial regions that contributed most strongly to age predictions. The resulting attribution maps indicated that the model generally relied on semantically meaningful facial features associated with ageing. Furthermore, insertion and deletion experiments provided quantitative evidence supporting the faithfulness of the generated explanations, with Integrated Gradients outperforming random perturbation baselines in the majority of evaluated samples.

Overall, the results demonstrate that the proposed framework is capable of producing competitive age estimation performance on less constrained video data while providing interpretable explanations of its predictions. At the same time, the analyses highlight important limitations related to age estimation in older subjects and illustrate the value of explainability methods for improving the understanding of model behaviour in facial analysis systems.

5.2 Limitations

Despite the positive results obtained throughout this work, several limitations should be acknowledged.

First, the age estimation model was evaluated without additional fine-tuning on the target datasets. Although this decision enabled a fair assessment of the model's generalisation capability, dataset-specific adaptation could potentially improve performance, particularly for underrepresented demographic groups.

Second, the adopted framework relies primarily on frame-level age estimation followed by temporal aggregation. While this approach improves prediction stability and remains computationally efficient, it does not explicitly model temporal dependencies between consecutive frames. Consequently, potentially useful temporal information present in video sequences may not be fully exploited.

Another limitation relates to the distribution of the available data. As observed in the experimental results, older age groups are less represented than younger and middle-aged adults. This imbalance contributes to increased prediction errors for elderly subjects and remains a common challenge in age estimation research.

Finally, the explainability analysis was conducted using a single attribution method. Although Integrated Gradients provided meaningful and consistent explanations, different explainability techniques may highlight complementary aspects of model behaviour and could lead to additional insights.

5.3 Future Work

Several directions for future research emerge from the findings of this dissertation.

A natural extension would be the fine-tuning of age estimation models on large-scale video datasets. Such an approach could improve robustness to variations in pose, illumination, motion blur, and other factors commonly encountered in less constrained video environments.

Future work could also explore architectures specifically designed for video analysis. Instead of aggregating independent frame-level predictions, temporal models capable of directly learning relationships between consecutive frames may further improve prediction stability and accuracy.

Another promising direction involves expanding the explainability component of the framework. Comparing Integrated Gradients with alternative attribution methods could provide a broader understanding of model decision-making processes and improve confidence in the generated explanations.

Further investigation of demographic fairness and bias would also be valuable. Evaluating additional demographic attributes and analysing performance across more diverse populations could contribute to the development of more equitable age estimation systems.

Finally, the methodology presented in this dissertation could be extended beyond age estimation to other video-based face analytics tasks, including emotion recognition, facial attribute analysis, and behavioural assessment. Such extensions would help determine whether the conclusions drawn in this work generalise to a broader range of facial analysis applications.

In summary, this dissertation provides a systematic evaluation of face detection, age estimation, and explainability techniques for less constrained video analysis. The obtained results establish a solid foundation for future research while highlighting several opportunities for further methodological and experimental improvements.

References

[1] Paul Viola and Michael Jones. “Robust Real-Time Face Detection”. In: *International Journal of Computer Vision*. Vol. 57. 2. 2004, pp. 137–154.

[2] Kaipeng Zhang et al. “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks”. In: *IEEE Signal Processing Letters* 23.10 (2016), pp. 1499–1503. DOI: 10.1109/LSP.2016.2603342.

[3] Jiankang Deng et al. “RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 5202–5211. DOI: 10.1109/CVPR42600.2020.00525.

[4] Stefanos Zafeiriou, Chen Zhang, and Zheng Zhang. “A Survey on Face Detection in the Wild: Past, Present and Future”. In: *Computer Vision and Image Understanding* 138 (2015), pp. 1–24. URL: https://www.sciencedirect.com/science/article/pii/S1077314215000727.

[5] Mahmoud Hassaballah and Sherif Aly. “Face Recognition: Challenges, Achievements and Future Directions”. In: *IET Computer Vision* 9.4 (2015), pp. 614–626. DOI: 10.1049/iet-cvi.2014.0084.

[6] Zhifei Zhang, Yang Song, and Hairong Qi. *UTKFace: Large-Scale Face Dataset with Age, Gender and Ethnicity Labels*. https://susanqq.github.io/UTKFace/. Accessed: 2025. 2017.

[7] Stylianos Moschoglou et al. “AgeDB: The First Manually Collected, In-the-Wild Age Database”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2017.

[8] Ze Cheng, Yitong Li, and Danni Yang. “Survey on Video-Based Face Recognition: Progress and Challenges”. In: *Pattern Recognition* (2021).

[9] Ankit Rajpal et al. “Xai-fr: explainable ai-based face recognition using deep neural networks”. In: *Wireless Personal Communications* 129.1 (2023), pp. 663–680.

[10] K. Bešenić, I. Pandžić, and J. Ahlberg. “Let Me Take a Better Look: Towards Video-Based Age Estimation”. In: *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*. SciTePress, 2024, pp. 57–69. ISBN: 978-989-758-684-2. DOI: 10.5220/0012376800003654.

[11] Ming-Hsuan Yang, David J. Kriegman, and Narendra Ahuja. “Detecting Faces in Images: A Survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.1 (2002), pp. 34–58. DOI: 10.1109/34.982883.

[12] Gil Levi and Tal Hassner. “Age and Gender Classification Using Convolutional Neural Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2015, pp. 34–42. DOI: 10.1109/CVPRW.2015.7301352.

[13] Shifeng Zhang et al. “RefineFace: Refinement Neural Network for High Performance Face Detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11 (2021), pp. 4008–4020. DOI: 10.1109/TPAMI.2020.2997456.

[14] Gary Huang et al. “Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments”. In: *Tech. rep.* (Oct. 2008).

[15] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. “Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset”. In: *IEEE Transactions on Multimedia* 17.6 (2015), pp. 804–815.

[16] Sheng Chen et al. “MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices”. In: *Chinese Conference on Biometric Recognition*. Springer, 2018, pp. 428–438.

[17] Hu Han, Charles Otto, and Anil K Jain. “Pose Invariant Age Estimation of Face Images in the Wild”. In: *Computer Vision and Image Understanding* 202 (2021), p. 103123.

[18] Hamdi Dibeklioglu et al. “Combining facial dynamics with appearance for age estimation”. In: *IEEE Transactions on Image Processing* 24.6 (2015), pp. 1928–1943.

[19] Abdenour Hadid. “Analyzing facial behavioral features from videos”. In: *Human Behavior Understanding*. Springer. 2011, pp. 52–61.

[20] Wenjie Pei et al. “Attended End-to-End Architecture for Age Estimation From Facial Expression Videos”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 1972–1984. DOI: 10.1109/TIP.2019.2948288.

[21] Beichen Zhang and Yue Bao. “Age Estimation of Faces in Videos Using Head Pose Estimation and Convolutional Neural Networks”. In: *Sensors* 22.11 (2022), p. 4171. DOI: 10.3390/s22114171.

[22] Xiangxin Zhu and Deva Ramanan. “Face Detection, Pose Estimation, and Landmark Localization in the Wild”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012, pp. 2879–2886. DOI: 10.1109/CVPR.2012.6248014.

[23] Guodong Guo and Guowang Mu. “Simultaneous Dimensionality Reduction and Human Age Estimation via Kernel Partial Least Squares Regression”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011, pp. 657–664. DOI: 10.1109/CVPR.2011.5995404.

[24] Hu Han and Anil K. Jain. “Age, Gender and Race Estimation from Unconstrained Face Images”. In: MSU-CSE-14-5. 2014. URL: https://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf.

[25] Matthew Turk and Alex Pentland. “Eigenfaces for Recognition”. In: *Journal of Cognitive Neuroscience* 3.1 (1991), pp. 71–86. DOI: 10.1162/jocn.1991.3.1.71.

[26] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. “Face Description with Local Binary Patterns: Application to Face Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.12 (2006), pp. 2037–2041. DOI: 10.1109/TPAMI.2006.244.

[27] Navneet Dalal and Bill Triggs. “Histograms of Oriented Gradients for Human Detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2005, pp. 886–893. DOI: 10.1109/CVPR.2005.177.

[28] Young H. Kwon and Niels da Vitoria Lobo. "Age Classification from Facial Images". In: *Computer Vision and Image Understanding* 74.1 (1999), pp. 1–21. DOI: 10.1006/cviu.1997.0549.

[29] Yun Fu, Guodong Guo, and Thomas S. Huang. "Age Synthesis and Estimation via Faces: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.11 (2010), pp. 1955–1976. DOI: 10.1109/TPAMI.2010.36.

[30] 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 (CVPR)*. 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682.

[31] Jiankang Deng et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4690–4699. DOI: 10.1109/CVPR.2019.00482.

[32] Rasmus Rothe, Radu Timofte, and Luc Van Gool. "Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks". In: *International Journal of Computer Vision* 126.2 (2018), pp. 144–157. DOI: 10.1007/s11263-016-0940-3.

[33] Khaled ELKarazle, Valliappan Raman, and Patrick Then. "Facial age estimation using machine learning techniques: An overview". In: *Big Data and Cognitive Computing* 6.4 (2022), p. 128.

[34] Rasmus Rothe, Radu Timofte, and Luc Van Gool. "DEX: Deep EXpectation of Apparent Age from a Single Image". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2015, pp. 252–257. DOI: 10.1109/ICCVW.2015.41.

[35] Ahmed Othmani and Abdelmalik Taleb-Ahmed. "Age Estimation from Faces Using Deep Learning". In: *Computer Vision and Image Understanding* 196 (2020), p. 102961. DOI: 10.1016/j.cviu.2020.102961.

[36] Yichun Chen et al. "DarkFace: Face Detection in Low Light Condition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[37] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*)*. 2018, pp. 77–91. URL: http://proceedings.mlr.press/v81/buolamwini18a.html.

[38] Yaron Gurovich et al. "Identifying Facial Phenotypes of Genetic Disorders Using Deep Learning". In: *Nature Medicine* 25.1 (2019), pp. 60–64. DOI: 10.1038/s41591-018-0279-0.

[39] Kin Choong Yow and Roberto Cipolla. "Feature-based human face detection". In: *Image and Vision Computing* 15.9 (1997), pp. 713–735.

[40] Roberto Brunelli and Tomaso Poggio. "Face recognition: Features versus templates". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15.10 (1993), pp. 1042–1052.

[41] Kah-Kay Sung and Tomaso Poggio. "Example-based learning for view-based human face detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.1 (1998), pp. 39–51.

[42] Paralect Blog. *Face Detection in Image and Video*. https://blog.paralect.com/post/face-detection-image-and-video. Accessed: 12 November 2025. 2020.

[43] Rainer Lienhart and Jochen Maydt. "An extended set of Haar-like features for rapid object detection". In: *Proceedings of the IEEE International Conference on Image Processing*. IEEE, 2002, pp. I–900.

[44] Davis E King. "Dlib-ml: A machine learning toolkit". In: *Journal of Machine Learning Research*. Vol. 10. 2009, pp. 1755–1758.

[45] Pedro F Felzenszwalb et al. "Object detection with discriminatively trained part-based models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.9 (2010), pp. 1627–1645.

[46] Vidit Jain and Erik Learned-Miller. "FDDB: A benchmark for face detection in unconstrained settings". In: *UMass Technical Report 2010-009*. 2010.

[47] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. "Neural network-based face detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.1 (1998), pp. 23–38.

[48] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. 2012.

[49] Shuo Yang et al. "WIDER FACE: A Face Detection Benchmark". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 5525–5533.

[50] Shaoqing Ren et al. "Faster R-CNN: Towards real-time object detection with region proposal networks". In: *Advances in Neural Information Processing Systems*. 2015, pp. 91–99.

[51] Wei Liu et al. "SSD: Single shot multibox detector". In: *European Conference on Computer Vision*. Springer, 2016, pp. 21–37.

[52] Joseph Redmon et al. "You Only Look Once: Unified, real-time object detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 779–788.

[53] Shifeng Zhang et al. "Recent Advances in Face Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.11 (2020), pp. 2794–2813.

[54] Mahyar Najibi et al. "SSH: Single Stage Headless Face Detector". In: *IEEE International Conference on Computer Vision*. 2017, pp. 4875–4884.

[55] Jian Li et al. "DSFD: Dual Shot Face Detector". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5060–5069.

[56] Yanjia Zhu et al. "Tinaface: Strong but simple baseline for face detection". In: *arXiv preprint arXiv:2011.13183* (2020).

[57] Jian Li et al. "ASFD: Automatic and scalable face detector". In: *Proceedings of the 29th ACM international conference on multimedia*. 2021, pp. 2139–2147.

[58] Ming Yang, Xiang Li, Hong Zhang, et al. "HP-FLDet: A High-Performance Joint Face and Landmark Detector". In: *IEEE Access* 12 (2024), pp. 26483–26496. DOI: 10.1109/ACCESS.2024.3369134.

[59] Jia Guo et al. "Sample and Computation Redistribution for Efficient Face Detection". In: *International Conference on Computer Vision Workshops*. 2021.

[60] Yunjie Tian, Qixiang Ye, and David Doermann. "Yolov12: Attention-centric real-time object detectors". In: *Advances in neural information processing systems* 38 (2026), pp. 78433–78457.

[61] Maksim Kuprashevich and Irina Tolstykh. "MIVOLO: Multi-input Transformer". In: *Analysis of Images, Social Networks and Texts: 11th International Conference, AIST 2023, Yerevan, Armenia, September 28–30, 2023, Revised Selected Papers*. Vol. 14486. Springer Nature. 2024, p. 212.

[62] Shifeng Zhang et al. "Fast Video Face Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 8126–8135.

[63] Alex Bewley et al. "Simple online and real-time tracking". In: *2016 IEEE International Conference on Image Processing*. 2016, pp. 3464–3468.

[64] Shifeng Zhang et al. "A Dataset and Benchmark for Face Detection in Videos". In: *ICCV Workshops*. 2019.

[65] T. Wu et al. "A Survey of Video-Based Face Detection". In: *ECCV Workshops*. 2016.

[66] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. "Simple Online and Realtime Tracking with a Deep Association Metric". In: *IEEE International Conference on Image Processing*. 2017, pp. 3645–3649.

[67] Gunnar Farnebäck. "Two-frame motion estimation based on polynomial expansion". In: *Scandinavian Conference on Image Analysis*. Springer, 2003, pp. 363–370.

[68] Yang Song et al. "A Temporal CNN for Video-Based Face Detection". In: *IEEE Transactions on Image Processing* (2019).

[69] Shuiwang Ji et al. "3D Convolutional Neural Networks for Human Action Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 221–231.

[70] Joon Son Chung and Andrew Zisserman. "Out of Time: Automated Lip Sync in the Wild". In: *Workshop on Multiview Lipreading, ACCV*. 2017.

[71] Andrew Howard et al. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications". In: *arXiv preprint arXiv:1704.04861*. 2017.

[72] Shifeng Zhang et al. "FaceBoxes: A CPU real-time face detector with high accuracy". In: *Proceedings of the IEEE International Joint Conference on Biometrics*. 2017, pp. 1–11.

[73] K. Chen et al. "Super-Resolution Face Detection in Cloud and Edge Environments". In: *IEEE ICIP*. 2018.

[74] Mark Everingham et al. "The pascal visual object classes (voc) challenge". In: *International journal of computer vision* 88.2 (2010), pp. 303–338.

[75] Bin Yang et al. "Fine-grained evaluation on face detection in the wild". In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 1. IEEE. 2015, pp. 1–7.

[76] Brendan F Klare et al. "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1931–1939.

[77] Hajime Nada et al. "Pushing the limits of unconstrained face detection: a challenge dataset and baseline results". In: *2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS)*. IEEE. 2018, pp. 1–10.

[78] H. Han et al. “Demographic Estimation from Face Images: Human vs. Machine Performance”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.6 (2013), pp. 1148–1161. DOI: 10.1109/TPAMI.2014.2362759.

[79] Wei Shen et al. “Deep regression forests for age estimation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 2304–2313.

[80] Rasha Ragheb Atallah et al. “Face Recognition and Age Estimation Implications of Changes in Facial Features: A Critical Review Study”. In: *IEEE Access* 6 (2018), pp. 28290–28304. DOI: 10.1109/ACCESS.2018.2836924.

[81] K. Liu, C. Yan, and X. Tan. “Age Estimation via Deep Learning: A Comparative Analysis”. In: *IEEE Transactions on Cybernetics* 45.10 (2015), pp. 2401–2412. DOI: 10.1109/TCYB.2014.2364555.

[82] Zhifei Zhang, Yang Song, and Hairong Qi. “Age progression/regression by conditional adversarial autoencoder”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5810–5818.

[83] A. Lanitis, C. J. Taylor, and T. F. Cootes. “Toward Automatic Simulation of Aging Effects on Face Images”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.4 (2002), pp. 442–455. DOI: 10.1109/34.993553.

[84] X. Geng, Z. Zhou, and Y. Zhang. “Facial Age Estimation by Learning from Label Distributions”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.10 (2007), pp. 1763–1776. DOI: 10.1109/TPAMI.2007.1110.

[85] G. Guo et al. “Human Age Estimation Using Bio-inspired Features”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2009), pp. 112–119. DOI: 10.1109/CVPR.2009.5206681.

[86] Guodong Guo et al. “A study on automatic age estimation using a large database”. In: *2009 IEEE 12th International Conference on Computer Vision*. 2009, pp. 1986–1991. DOI: 10.1109/ICCV.2009.5459438.

[87] Guodong Guo and Guowang Mu. “Human age estimation: What is the influence across race and gender?” In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE. 2010, pp. 71–78.

[88] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. “Ordinal hyperplanes ranker with cost sensitivities for age estimation”. In: *CVPR 2011*. IEEE. 2011, pp. 585–592.

[89] Z. Niu et al. “Ordinal Regression with Multiple Output CNN for Age Estimation”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4920–4928. DOI: 10.1109/CVPR.2016.532.

[90] Charlene Chu et al. “Strategies to mitigate age-related bias in machine learning: Scoping review”. In: *JMIR aging* 7 (2024), e53564.

[91] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. “Are you really smiling at me? spontaneous versus posed enjoyment smiles”. In: *European conference on computer vision*. Springer. 2012, pp. 525–538.

[92] Zhenyu Ji et al. “Deep age estimation model stabilization from images to videos”. In: *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE. 2018, pp. 1420–1425.

[93] Andreas Lanitis and Tim Cootes. “Fg-net aging data base”. In: *Cyprus College* 2.3 (2002), p. 5.

[94] Karl Ricanek and Tamirat Tesafaye. "Morph: A longitudinal image database of normal adult age-progression". In: *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE. 2006, pp. 341–345.

[95] Caner Hazirbas et al. "Towards measuring fairness in ai: the casual conversations dataset". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4.3 (2021), pp. 324–332.

[96] Ali et al. "P-Age: Pexels Dataset for Robust Spatio-Temporal Apparent Age Classification". In: *arXiv preprint arXiv:2311.02432*. 2023.

[97] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

[98] Riccardo Guidotti et al. "A survey of methods for explaining black box models". In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42.

[99] R. R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

[100] Aditya Chattopadhay et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks". In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2018, pp. 839–847.

[101] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[102] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[103] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[104] Scott Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *NeurIPS*. 2017.

[105] Ann-Kathrin Dombrowski et al. "Explanations can be manipulated and geometry is to blame". In: *NeurIPS* (2019).

[106] Pavel Korshunov and Sébastien Marcel. "Deep Learning Framework for Video-based Age Estimation and Verification". In: *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*. 2022.

[107] Bilal Porgali et al. "The casual conversations v2 dataset". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 10–17.

[108] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[109] Bin-Bin Gao et al. "Age estimation using expectation of label distribution learning." In: *IJCAI*. Vol. 1. 2018, p. 3.

[110] Ivan Grishchenko et al. "Attention mesh: High-fidelity face mesh prediction in real-time". In: *arXiv preprint arXiv:2006.10962* (2020).