FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Leveling the Playing Field: Ethnicity-Guided Synthetic Data Generation for Fair Face Recognition

André Filipe Garcez Moreira de Sousa

Mestrado em Engenharia Informática e Computação

Supervisor: Prof. Ana Sequeira

Second Supervisor: Prof. Pedro Neto

July 24, 2025

Leveling the Playing Field: Ethnicity-Guided Synthetic Data Generation for Fair Face Recognition

André Filipe Garcez Moreira de Sousa

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. Carlos Soares (FEUP)
Examiner: Prof. Naser Damer (TU Darmstadt)
Supervisor: Prof. Ana Sequeira (INESC TEC & FEUP)
Second Supervisor: Prof. Pedro Neto (INESC TEC)

July 24, 2025

Resumo

A rápida proliferação de tecnologias de reconhecimento facial em aplicações críticas de segurança e de consumo trouxe à tona desafios éticos, legais e técnicos profundos. O viés demográfico nos dados de treino compromete a equidade, enquanto a recente remoção de *datasets* de imagens reais em larga escala (por exemplo, MS-Celeb-1M) limitou a disponibilidade de dados e a reprodutibilidade. Esta tese apresenta um *pipeline* de difusão sensível à etnia que gera imagens faciais sintéticas com garantias demonstráveis de equidade e risco mínimo de privacidade. A nossa *framework* inicia-se com um classificador *Ethnicity Score*, treinado num *dataset* equilibrado de fontes públicas, que orienta a seleção de identidades para assegurar uma representação proporcional em quatro principais grupos demográficos. Aplicam-se dois geradores de difusão complementares — IDiff-Face para prototipagem rápida e DCFace para síntese escalável — sob três regimes hierárquicos de amostragem (1 000; 10 000; 500 000 imagens), originando 46 conjuntos de dados rotulados.

Realizámos uma análise aprofundada das configurações de grelha de estilo e densidades de amostragem, identificando uma grelha DCFace 5x5 ótima com uma seleção estratificada e filtragem de alta confiança, capaz de produzir 500 000 imagens usando 10 000 identidades. Esta configuração atinge uma cobertura efetiva por raça $\mathrm{ENS}_{\mathrm{race}} \approx 3,8$ (superior à diversidade do RFW) e reduz para metade o índice de estereótipo do RFW.

Os modelos treinados nos *datasets* sintéticos são avaliados em sete *benchmarks* — LFW, AgeDB-30, CALFW, CPLFW, CFP-FF, CFP-FP e RFW — usando uma arquitetura lightweight (IResNet-34), complementada com alinhamento usando *RetinaFace* e RANDAUGMENT. O nosso modelo atinge uma *accuracy* média de verificação de 84,7% nos seis conjuntos *in-domain* e 79,6% no RFW, reduzindo a dispersão de desempenho entre etnias para 3,78 pp e o erro enviesado para 1,65 pp. Normalizado pelo volume de dados, fornece 174 pontos de benchmark por milhão de imagens — 2,5× a densidade do DigiFace-1M e uma ordem de grandeza acima do VariFace. Demonstra-se ainda que ganhos em fidelidade fotométrica (menor $\mathrm{FID}_{\mathrm{RFW}}$) ocorrem em simultâneo com melhorias em $\mathrm{ENS}_{\mathrm{race}}$, evidenciando uma sinergia entre realismo e equidade. Estudos de ablação confirmam que a estratégia de amostragem é o fator dominante no ganho de equidade, enquanto arquiteturas mais profundas ou o uso de *losses* alternativas proporcionam apenas retornos marginais.

Em comparação com abordagens baseadas em GAN do estado da arte, o nosso corpus de difusão de escala moderada iguala ou supera a *accuracy* de reconhecimento em 7–12 pp e reduz as lacunas de desempenho racial em mais de 35%. Além disso, tanto em cenários de difusão como de GAN, o uso de uma arquitetura compacta (IResNet-34) em vez da IResNet-50 de referência reduz o custo computacional em aproximadamente 40% sem comprometer a *accuracy* ou a equidade. Estes resultados demonstram que conjuntos de dados de difusão de escala moderada, equilibrados demograficamente, oferecem um caminho escalável e consciente da privacidade para implementações de reconhecimento facial mais justas.

Abstract

The rapid proliferation of face recognition technologies in security-critical and consumer-facing applications has brought profound ethical, legal, and technical challenges to the foreground. Demographic bias in training data undermines fairness, while the recent legal withdrawal of large-scale, real-image datasets (e.g., MS-Celeb-1M) has constrained data availability and reproducibility. This thesis presents a comprehensive, ethnicity-aware diffusion pipeline that generates high-quality synthetic face images with provable fairness guarantees and minimal privacy risk. Our framework begins with an *Ethnicity Score* classifier, trained on a balanced subset of public sources, which guides identity seeding to ensure proportional representation across four major demographic groups. Two complementary diffusion generators—IDiff-Face for rapid prototyping and DCFace for scalable synthesis—are applied under three hierarchical sampling regimes (1 000; 10 000; 500 000 images), yielding 46 labeled datasets. We conduct an in-depth analysis of style grid configurations, sampling densities, and exemplar-confidence thresholds, identifying an optimal 5x5 DCFace grid with stratified seeding and high-confidence filtering that produces 500 000 images using 10 000 identities. This configuration achieves an effective race coverage $\mathrm{ENS}_{\mathrm{race}} \approx 3.8$ (surpassing RFW diversity) and halves the RFW stereotype index.

We evaluate our synthetic corpus on seven benchmarks—LFW, AgeDB-30, CALFW, CPLFW, CFP-FF, CFP-FP, and RFW—using a lightweight IResNet-34 backbone augmented with RetinaFace alignment and RANDAUGMENT. Our model attains a mean verification accuracy of 84.7% on the six in-domain sets and 79.6% on RFW while reducing cross-ethnicity performance dispersion to 3.78 pp and skewed-error to 1.65 pp. Normalized by data volume, it delivers 174 benchmark points per million images—2.5× the density of DigiFace-1M and an order of magnitude above VariFace. We further show that gains in photometric fidelity (lower $\mathrm{FID}_{\mathrm{RFW}}$) co-occur with improved $\mathrm{ENS}_{\mathrm{race}}$, evidencing a synergy between realism and fairness. Ablation studies confirm that sampling strategy is the dominant factor in fairness leverage, whereas deeper backbones or alternative margin losses yield only marginal returns.

Compared to state-of-the-art GAN-based approaches, our moderate-scale diffusion corpus matches or surpasses recognition accuracy by 7–12 pp and reduces racial performance gaps by over 35%. Furthermore, across both diffusion- and GAN-based baselines, employing a compact IResNet-34 backbone instead of the de facto ResNet-50 reduces the computational cost by approximately 40% without compromising accuracy or fairness. These results demonstrate that demographically balanced, moderate-scale diffusion datasets provide a scalable, privacy-conscious path toward fairer face recognition deployments.

UN Sustainable Development Goals

The 2030 Agenda for Sustainable Development, adopted by all United Nations Member States in 2015, defines seventeen inter-connected Sustainable Development Goals (SDGs) that balance social progress, environmental protection, and economic growth[1]. Although the present work is centered on technical advances in face recognition, its explicit aims—mitigating demographic bias, safeguarding privacy, and enabling fair deployment—intersect with three SDGs, detailed below.

1. **SDG 9 (Industry, Innovation, and Infrastructure)**
   SDG 9 calls for "resilient infrastructure" and "inclusive and sustainable industrialization". By conditioning diffusion models on fairness metrics, this thesis delivers a pipeline capable of producing demographically balanced, privacy-compliant training data at scale. The resulting synthetic datasets lower data-collection barriers, stimulate responsible AI innovation, and supply reproducible tooling for stakeholders who build secure biometric infrastructure.

2. **SDG 10 (Reduced Inequalities)**
   SDG 10 demands that inequalities "based on age, sex, disability, race, ethnicity, origin, religion or economic or other status" be reduced within and among countries. Here, generative augmentation yields face datasets whose representational diversity and evenness lessen performance gaps across racial groups. Fairness metrics computed on both the data (e.g. $\mathrm{ENS}_{\mathrm{race}}$) and the downstream models (e.g. accuracy standard deviation across RFW ethnic subsets) demonstrate tangible progress toward algorithmic equity.

3. **SDG 16 (Peace, Justice, and Strong Institutions)**
   SDG 16 promotes "effective, accountable and inclusive institutions". This thesis addresses SDG 16 by reinforcing the capacity of law enforcement and security institutions through more reliable and equitable face-recognition systems. By integrating a synthetic data pipeline and audit framework, agencies can access tools that improve overall identification performance while preserving procedural fairness. The enhanced system is designed for deployment within institutional workflows, where stronger biometric verification supports accountability, reduces wrongful identifications, and fosters public trust in security operations.

Performance indicators and impact metrics

Quantitative indicators are aligned with the evaluation protocol established in Chapters 4 and 5; they provide a measurable link between technical results and SDG impact.

[1] Overview available at https://sdgs.un.org/goals.

SDG 9. Innovation quality and infrastructural efficiency are tracked through the Inception Score (IS), the Fréchet and Kernel Inception Distances (FID, KID), and perceptual fidelity metrics (SSIM, LPIPS). These figures—reported with respect to the LFW and RFW benchmarks—capture the diversity, distributional fit, and visual realism of the synthetic datasets.

SDG 10. Progress toward reduced inequalities is measured on two complementary axes. *Data-level balance* is quantified via the Effective Number of Species for race ($\mathrm{ENS}_{\mathrm{race}}$), the gap $|G|-\mathrm{ENS}$, and Shannon Evenness. In contrast, *stereotypical coupling* is monitored through Cramér's $V_{\mathrm{race,label}}$ and Normalised Mutual Information. *Model-level fairness* is assessed with the standard deviation of verification accuracy across RFW subsets and the skewed error ratio. A decrease in these disparity metrics signals effective bias mitigation.

SDG 16. Institutional impact is quantified in two complementary ways. First, the overall system efficacy is measured by the mean verification accuracy computed across multiple standard benchmarks (e.g. LFW, CFP-FF, and AgeDB). Second, fairness is evaluated using the same representational and stereotypical indices introduced for SDG 9—such as $\mathrm{ENS}_{\mathrm{race}}$ and Cramér's $V_{\mathrm{race,label}}$. Values of mean accuracy paired with minimal ΔSTD demonstrate that institutions can achieve higher recognition performance without compromising equity, thereby operationalizing strong, inclusive, and accountable biometric practices.

These indicators create a coherent dashboard connecting this thesis' technical deliverables to concrete, SDG-aligned outcomes. Regular reporting of the above metrics across experimental iterations enables an objective assessment of progress and facilitates evidence-based adjustments to maximize societal benefit.

Acknowledgements

Developing a thesis is a team sport disguised as a solo event, and this dissertation would not exist without many people's steady support, insight, and goodwill. First, I owe a sincere thank-you to my supervisors, Ana Filipa Sequeira and Pedro Neto, who combined technical rigor with endless patience. Their guidance—whether on experimental design, writing clarity, or simply knowing when to change the focus—kept the project on course.

I gratefully acknowledge INESC TEC for the computing resources that powered every training run and crunched dozens of terabytes of data and for fostering a research atmosphere where curiosity is encouraged. A special word goes to the systems team, who tolerated my frequent support requests and restored hardware before the panic could set in, especially after the national power grid blackout.

My parents, my brother, and my grandparents have supported my academic pursuits from the start. They celebrated each milestone and reminded me that progress is measured by effort rather than by simple metrics. To Sofia, whose unwavering support in moments of doubt sustained me through the most demanding phases of this work. Your thoughtful feedback, gentle humour and shared background helped me untangle complex ideas when the project felt overwhelming and gave me the confidence to persevere.

Finally, to my friend group—coding buddies, lunch partners, and Thursday evening philosophers—your collective distraction therapy kept my sanity intact while nudging the thesis steadily forward. I could not have asked for better company on this journey.

André

"You shall ~~not~~ pass!"

Gandalf, The Lord of the Rings: The Fellowship of the Ring

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
AUC	Area Under the Curve
CCTV	Closed-Circuit Television
CNN	Convolutional Neural Network
CPD	Contextual Partial Dropout
DL	Deep Learning
DM	Diffusion Models
ENS	Effective Number of Species
FID	Fréchet Inception Distance
FR	Face Recognition
GAN	Generative Adversarial Network
HOG	Histogram of Oriented Gradients
KD	Knowledge Distillation
KID	Kernel Inception Distance
LPIPS	Learned Perceptual Image Patch Similarity
MSE	Mean Squared Error
MTCNN	Multi-Task Cascaded Convolutional Networks
ROC	Receiver Operating Characteristic
RQ	Research Question
SER	Skewed Error Ratio
SSIM	Structural Similarity Index
STD	Standard Deviation
SVM	Support Vector Machine
YOLO	You Only Look Once

Chapter 1

Introduction

This chapter provides the background and framing for this thesis. Section 1.1 outlines face recognition's application domains and ethical challenges, emphasizing demographic bias and privacy constraints. Section 1.2 motivates the need for fairness-aware and privacy-compliant solutions, illustrating their societal impact. Section 1.3 formulates the core problem of racial bias in face recognition and describes the proposed approach using generative models. Section 1.4 presents the research questions guiding this work. Finally, Section 1.5 summarises the chapter and situates the contributions of this thesis in the broader context.

1.1 Context

When it comes to biometric authentication technologies, face recognition (FR) is one of the most developed ones. Not only is it applied in high-security environments but also in consumer devices. Studies have shown that the demand for face recognition systems is expected to grow at a rate of 9.34% each year [72], so it is naturally expected that its influence will expand into additional application domains. Face recognition technology is now ubiquitous, whether it is in verifying identities at airport borders or unlocking smartphones. In addition, face recognition can be a decisive factor in detecting missing people, representing an effective tool in public security and law enforcement.

A key issue with FR technology is demographic bias, where performance disparities exist across different ethnic groups. These disparities are often linked to imbalanced training datasets, which fail to provide equitable representation across demographics. As a result, models trained on such datasets may demonstrate significantly lower accuracy for underrepresented groups, reinforcing systemic inequalities. Addressing these biases is essential for ensuring that face recognition systems operate fairly and do not disproportionately disadvantage specific populations.

At the same time, the availability of high-quality training data has been constrained by growing privacy concerns, leading to the retraction of several widely used face recognition datasets. This retraction has created an urgent need for privacy-compliant alternatives supporting research and development while adhering to ethical standards. In response to these challenges, synthetic

data generation has emerged as a promising solution, offering a means to create diverse and demographically balanced datasets without compromising individual privacy.

Advancements in deep learning, particularly with generative models, have opened new possibilities for improving face recognition systems' fairness and accuracy. By leveraging these techniques, researchers can develop synthetic datasets that better represent real-world diversity, ultimately leading to more robust and equitable recognition models. This thesis explores the potential of such approaches, aiming to contribute to developing ethical and unbiased face recognition technologies.

1.2 Motivation

Even though deep learning-based face recognition systems have demonstrated high accuracy, their fairness and robustness are under severe scrutiny. Recent studies [2] have concluded that these systems can exhibit bias towards specific demographics, which results in varying performance between genders or ethnicities, leading to ethical concerns. Additionally, the retraction of multiple widely used face recognition datasets due to privacy concerns has created significant challenges in developing and evaluating new models. The lack of accessible, ethically sourced, and demographically balanced datasets hinders research progress and raises questions about the reliability of existing systems. Addressing this gap by generating privacy-compliant datasets is crucial for ensuring that face recognition technologies can be developed and deployed fairly and responsibly, particularly in high-stakes applications such as law enforcement and public surveillance.

In the United States, approximately 600 000 people are reported missing annually [80], with a child disappearing every 40 seconds [15]. However, due to the heavy presence of closed-circuit television (CCTV) cameras, which capture the average American about 238 times per day [73], there is an excellent potential for AI-driven solutions to assist in finding missing people. A face recognition system that can reliably and impartially process such data would better organize search efforts and significantly reduce crime rates by quickly identifying individuals in critical scenarios.

Considering the growing use of face recognition systems in critical areas, creating methods that improve performance, fairness, and robustness is imperative. By having these advances as a basis and identifying gaps in the state-of-the-art, the goal of this thesis is to help create an effective and, at the same time, fairer and secure system that will be applied to solve urgent societal issues. The scope of such improvements cannot be limited to engineering advancements and can bring real benefits in solving pressing social problems, such as enhancing public security and safeguarding endangered communities.

1.3 Problem Description

This thesis addresses the challenge of mitigating racial bias in face recognition systems by generating racially unbiased datasets using advanced generative models. Despite the significant progress made in face recognition technologies, these systems often exhibit performance disparities across

different demographic groups, particularly regarding race and ethnicity. These biases are primarily attributed to imbalanced training data, where underrepresented groups suffer from higher misidentification rates. As face recognition continues to be adopted in critical areas such as law enforcement, border control, and security applications, ensuring fairness in these systems is paramount.

To address this issue, this research explores the use of generative models to create synthetic datasets that are demographically balanced. Fairness-aware metrics will guide the generation process to ensure that the synthetic images accurately reflect a fair distribution of racial characteristics. The generated datasets will then be used to train multiple state-of-the-art face recognition networks, whose performance will be evaluated against widely used datasets to determine whether the proposed approach improves fairness and accuracy.

With these objectives in mind, the following key aspects are central to this thesis:

- **Generation of Racially Unbiased Datasets:** This study will investigate how well diffusion models (DM), conditioned on fairness-aware metrics, can produce synthetic face images that provide a balanced representation of different racial groups.
- **Evaluation of Fairness in Face Recognition Models:** By training face recognition networks on the generated datasets, this research will assess whether the models exhibit reduced demographic bias and improved fairness compared to those trained on traditional datasets.
- **Comparative Analysis with Existing Datasets:** This study will systematically compare the performance of models trained on the synthetic datasets with those trained on widely used face recognition datasets, identifying the strengths and limitations of using synthetic data.

This research contributes to the development of ethical face recognition systems by addressing one of the most pressing concerns in the field: bias in recognition performance. By creating synthetic datasets that better represent global demographics and evaluating their impact on fairness, this work aims to provide a practical solution for reducing bias in face recognition technology.

1.4 Research Questions

The main research questions of this work focus on the generation of racially unbiased datasets using diffusion models and their impact on the fairness and accuracy of face recognition systems. This study will investigate how well conditional diffusion models can produce demographically balanced synthetic datasets. Additionally, it will assess whether training face recognition models on such datasets leads to improved fairness and performance compared to traditional datasets.

Fundamentally, the work to be developed aims to address the following main questions:

- **RQ1:** Can conditional diffusion models synthesize racially unbiased datasets?
- **RQ2:** Does training on racially unbiased datasets improve fairness without hurting accuracy?

- **RQ3:** How does synthetic training compare with conventional practice?

1.5 Conclusion

Face recognition technology has seen rapid advancements due to deep learning, expanding its influence across various sectors. However, concerns regarding fairness, bias, and ethical considerations remain pressing challenges, particularly as these systems are increasingly deployed in critical applications such as law enforcement, security, and public safety. A significant contributing factor to demographic bias in face recognition models is the use of imbalanced datasets, which disproportionately affect certain ethnic groups. Furthermore, growing privacy concerns have led to the retraction of several widely used datasets, limiting the availability of diverse and representative training data.

To address these issues, this research explores the generation of privacy-compliant, racially unbiased synthetic datasets using diffusion models. By leveraging generative techniques and fairness-aware metrics, this approach aims to create demographically balanced datasets that mitigate bias while respecting ethical and legal constraints. The effectiveness of these datasets will be assessed by training multiple state-of-the-art face recognition networks and evaluating their performance against traditional datasets.

This thesis contributes to the development of ethical and fair face recognition technology by tackling one of the field's most critical challenges: bias in model performance due to dataset limitations. By demonstrating the potential of synthetic data to improve fairness without compromising privacy, this work aims to provide a practical and scalable solution for mitigating bias in modern face recognition systems. The findings of this research will not only advance the technical capabilities of face recognition models but also support the responsible deployment of these systems in real-world applications.

Chapter 2

Background

This chapter reviews the theoretical foundations and core methodologies underpinning this thesis. Section 2.1 introduces diffusion models, detailing their forward and reverse processes, conditioning mechanisms, and key advantages over adversarial approaches. Section 2.2 examines Generative Adversarial Networks, presenting their minimax formulation, common variants for stability and conditioning, and a comparison with diffusion models. Section 2.3 describes convolutional neural networks, outlining convolutional operations, architectural components such as residual connections and normalization, and their role in feature extraction. Section 2.4 surveys deep learning methods for face recognition, including backbone architectures, specialized loss functions, and evaluation benchmarks. Section 2.5 discusses strategies for leveraging diffusion-generated synthetic data in recognition pipelines, focusing on demographic bias mitigation and enhancement of feature robustness. Together, these sections establish the background necessary for the methods developed in this work.

2.1 Diffusion Models

Diffusion models have recently emerged as a powerful class of generative models in machine learning, offering an alternative to approaches such as Generative Adversarial Networks (GANs). The fundamental concept behind diffusion models is to learn a data distribution by reversing a stochastic process that gradually adds noise to the data.

2.1.1 The Forward and Reverse Processes

At the heart of diffusion models lies a two-step procedure that transforms data into noise and then learns to reverse this corruption. Below, we describe the discrete-time Markov formulation, its continuous-time stochastic differential equation (SDE) counterpart, and the objective used to train the reverse model.

1. **Forward Process**: This fixed Markov chain incrementally corrupts the data by adding Gaussian noise over a sequence of time steps. Given an initial data sample $\mathbf{x}_0$, the forward process

produces a sequence of latent variables $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ such that at each step

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right),$$

where $\{\beta_t\}_{t=1}^T$ is a predefined variance schedule. As $t \to T$, $\mathbf{x}_T$ approaches an isotropic Gaussian.

Equivalently, one can view this as the discretization of a continuous-time SDE (e.g., variance-exploding or variance-preserving),

$$d\mathbf{x} = f(\mathbf{x}, t)\, dt + g(t)\, d\mathbf{w},$$

where $\mathbf{w}$ is a standard Wiener process. This perspective underpins the theory developed by Song and Ermon (2021) [71] and connects diffusion models to score-based generative modeling.

2. **Reverse Process**: The aim is to learn a parameterized denoising chain $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ that inverts the forward corruption. In continuous time, the reverse-time dynamics satisfy the SDE

$$d\mathbf{x} = \left[f(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] dt + g(t)\, d\bar{\mathbf{w}},$$

where $p_t(\mathbf{x})$ is the marginal density at time t and $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is its score. In practice, one trains a neural network $s_\theta(\mathbf{x}, t)$ to approximate this score by minimizing the denoising score-matching objective introduced by Ho et al. (2020) [35]:

$$\mathscr{L}(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \varepsilon}\left[\left\|\varepsilon - \varepsilon_\theta(\mathbf{x}_t, t)\right\|^2\right],$$

where $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\varepsilon$ and $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Under this formulation, training the reverse model by minimizing a weighted variational bound is equivalent to denoising score matching with Langevin dynamics.

3. **Conditional Diffusion Models**: To control attributes of generated samples (e.g., class, identity, or ethnicity), two main strategies are used:

   - *Classifier Guidance* [22] augments the reverse SDE with gradients of a pretrained classifier $p_\phi(y \mid \mathbf{x}_t)$. At each step, the score is modified as

   $$\tilde{s}(\mathbf{x}_t, t) = s_\theta(\mathbf{x}_t, t) + w\, \sigma_t \nabla_{\mathbf{x}_t} \log p_\phi(y \mid \mathbf{x}_t),$$

   where w is a guidance weight and σ_t the noise scale. This steers samples toward class y at the cost of some diversity.

   - *Classifier-Free guidance* [36] avoids a separate classifier by jointly training the model with and without conditions. During sampling, one interpolates between conditional

and unconditional score estimates:

$$s_{\text{cfg}}(\mathbf{x}_t, t, y) = s_\theta(\mathbf{x}_t, t, y) + \gamma\big(s_\theta(\mathbf{x}_t, t, y) - s_\theta(\mathbf{x}_t, t)\big),$$

Where γ controls the strength of conditioning.

- *Conditional Score Networks* directly incorporate the attribute y (e.g. a one-hot embedding or identity vector) into the score network architecture, yielding $s_\theta(\mathbf{x}_t, t, y)$ trained to approximate $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid y)$.

These conditioning mechanisms enable fine-grained control over the generated images, such as specifying an ethnicity label for face synthesis.

4. **Intuitive Overview**: Diffusion models implement a two-phase degradation–reconstruction mechanism. In the forward pass, each image is noised step by step until it becomes indistinguishable from random static. In the reverse pass, a neural network learns to recover the image by removing small amounts of noise at each step, like restoring a faded photograph in successive stages. Over many iterations, the model learns both how images degrade and how to reverse that degradation, enabling it to synthesize novel, high-quality samples from pure noise.

2.1.2 Advantages and Applications

Diffusion models offer a principled, likelihood-based framework for generative modeling that contrasts with adversarial objectives by minimizing a simple denoising score-matching loss. Because the forward corruption process is fixed and tractable, and the reverse denoising kernels admit closed-form Gaussian approximations, training is stable and avoids issues such as mode collapse that commonly afflict GANs. Moreover, the variational lower bound on data likelihood can be estimated directly, providing an interpretable metric for model comparison and enabling downstream tasks such as anomaly detection by evaluating the likelihood of novel inputs [35, 71].

The iterative nature of the reverse diffusion process affords fine-grained control over sample quality and diversity through noise-schedule design and sampling algorithms. Techniques such as classifier and classifier-free guidance allow conditioning on arbitrary attributes (ranging from coarse class labels to detailed textual prompts) without altering the core network architecture [22, 36]. In practice, these mechanisms have enabled state-of-the-art text-to-image synthesis in high resolution (e.g., Stable Diffusion) and precise image editing operations such as inpainting, super-resolution, and style transfer, where each refinement step corrects minor deviations from the conditioning signal [64].

Beyond image synthesis, the diffusion paradigm has been applied successfully to other modalities. In audio generation, diffusion processes operating in the spectrogram domain produce waveforms with natural prosody and fine acoustic detail, while in molecular design, graph-based diffusion models iteratively construct chemically valid structures by denoising random graphs toward desired property distributions. Video diffusion frameworks extend noise scheduling into the

temporal dimension, capturing spatiotemporal correlations and allowing controlled interpolation between motion dynamics. These cross-modal extensions demonstrate the generality of diffusion models for learning complex, high-dimensional data distributions.

In the specific face synthesis and recognition domain, diffusion models have been leveraged to mitigate demographic bias by generating balanced datasets. IDiff-Face integrates identity-conditioned denoising to produce realistic facial images across varied age, gender, and ethnicity groups [8], while DCFace employs disentangled attribute representations to control demographic factors explicitly [44]. Augmenting underrepresented cohorts with high-fidelity synthetic samples has improved fairness metrics in downstream face verification benchmarks. As sampling acceleration techniques, such as progressive distillation and latent-space diffusion, continue to reduce inference time, these applications are becoming increasingly viable for large-scale deployment in biometric systems.

2.2 Generative Adversarial Networks

Generative adversarial networks (GANs) are another popular class of deep generative models, formulated as a minimax game between a generator G and a discriminator D. The original GAN objective (Goodfellow et al., 2014 [30]) is:

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]$$

where z is a latent noise vector and $D(x) \in [0,1]$ estimates the probability that x is a real image. In theory, the unique equilibrium of this game is that G reproduces the data distribution and $D(x) = 0.5$ everywhere.

In practice, however, GAN training is notoriously unstable: without care, mode collapse and divergence can occur. Many variants of the GAN loss have been proposed to improve stability. For example, the Wasserstein GAN [5] replaces the Jensen–Shannon divergence with the Wasserstein-1 distance, enforcing a Lipschitz constraint on D via weight clipping or gradient penalty. WGAN significantly improves training stability and mode coverage, reducing mode collapse issues. Other GAN losses include least-squares GAN (LSGAN) and hinge-loss GANs, which are often used in practice, but the original and Wasserstein forms are most commonly discussed in the literature.

GANs can also be conditioned on auxiliary information. The conditional GAN [50] appends class labels to both G and D, enabling the generation of class-specific images. Odena et al. (2017) [56] proposed the Auxiliary Classifier GAN (ACGAN), in which the discriminator outputs both a real/fake score and a class label prediction. Conditioning can be formalized by modifying the GAN loss to include $p(x \mid y)$ and $p(y \mid x)$ terms. For example, a conditional generator may optimize $\log(1 - D(G(z,y)))$ for samples $G(z,y)$ of class y, while the discriminator maximizes $\log D(x,y)$ over real data-label pairs. In practice, class-conditional GANs have been shown to produce higher-quality and more diverse samples than unconditional GANs (especially for large label sets).

Despite their success, GANs differ from diffusion models in key ways. GANs typically generate sharp images in a single feed-forward pass, and sampling is fast once the model is trained. However, GAN training dynamics are delicate: convergence is not guaranteed, and the model can oscillate or collapse to limited modes. In contrast, diffusion models are trained with a stable regression or score-matching objective, and training tends to be more predictable (though computationally intensive). Empirically, recent studies have shown that diffusion models can surpass GANs in image quality and distribution coverage. For instance, Dhariwal and Nichol (2021) [22] found that diffusion models achieved FID scores better than state-of-the-art GANs on ImageNet while maintaining superior mode coverage. Conversely, Arjovsky et al. (2017) [5] argued that Wasserstein GANs improve stability and mitigate mode collapse. Thus, the two approaches have trade-offs: GANs are fast to sample but can be unstable, whereas diffusion models incur higher sampling costs (many denoising steps) but tend to cover the data manifold more fully. We summarize these differences below.

Aspect	**GANs**	**Diffusion Models**
Training Stability	Often unstable; risk of mode collapse without careful tuning	Generally stable training via score-matching or variational objectives
Mode Coverage	Can miss modes or collapse; sometimes limited diversity	Good mode coverage; tends to capture full data distribution
Sample Quality	Sharp, high-resolution images (with architectures like StyleGAN)	Very high-fidelity (recent models match or exceed GANs)
Sampling Speed	Fast (one or few network evaluations per sample)	Slow (hundreds or thousands of diffusion steps per sample)
Likelihood	Not explicit (no tractable likelihood)	Can compute or bound likelihood via diffusion probability paths

Table 2.1: Comparison between GANs and diffusion models regarding stability, coverage, and sample quality

2.3 Convolutional Neural Networks

Convolutional neural networks (CNNs) are mathematical models for processing images that exploit spatial structure. In a CNN, the basic operation is a convolution: given an input feature map $x \in \mathbb{R}^{H\times W\times C}$ and a kernel $K \in \mathbb{R}^{k\times k\times C}$, the convolution output $y \in \mathbb{R}^{H'\times W'}$ is given by

$$y_{i,j} = \sum_{u=1}^{k}\sum_{v=1}^{k}\sum_{c=1}^{C} K_{u,v,c}\, x_{i+u-1,\, j+v-1,\, c}.$$

This may be followed by adding a bias and applying a nonlinearity. Common activation functions include the Rectified Linear Unit (ReLU),

$$f(z) = \max(0, z),$$

which adds nonlinearity and helps avoid vanishing gradients. After convolution and activation, pooling layers may be applied to downsample spatial dimensions; for example, max-pooling takes the maximum value over a neighborhood, reducing resolution while retaining salient features. Batch normalization [38] is often used to stabilize and accelerate training: it normalizes each layer's inputs to have zero mean and unit variance over the mini-batch, scaled and shifted by learned parameters. Formally, given activations x in a batch, BatchNorm computes

$$\hat{x} = \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}}, \qquad \mathrm{BN}(x) = \gamma \hat{x} + \beta.$$

CNN parameters are learned by backpropagation of a loss function. The loss gradient for each convolutional weight is computed via the chain rule, propagating errors from the output back through the layers. During inference, an input image passes through the sequence of convolution, activation, pooling, and normalization layers to produce a final output, such as class scores or embedding vectors.

In face recognition, CNNs serve as feature extractors. A face image is preprocessed (aligned and cropped to a canonical pose) and passed through a CNN backbone (often a deep ResNet) to produce a fixed-dimensional embedding vector. The ResNet family of architectures [33] is widely used: ResNet-18, -34, -50, -101, etc., where the number indicates the total number of layers. ResNets introduced "identity skip connections" that allow very deep networks to be trained effectively. Empirically, ResNet-50 and deeper variants have become standard backbones for face recognition (for example, ArcFace and CosFace use ResNet-50). Combined with BatchNorm and other modern CNN techniques, these networks learn embeddings that place images of the same identity close together in feature space and separate different identities. During inference, one typically compares two embeddings with a distance metric (e.g., cosine or Euclidean) to perform verification or identification.

2.4 Deep Learning for Face Recognition

The advent of deep CNNs revolutionized face recognition. Early systems used handcrafted features or shallow networks, but end-to-end deep learning methods dramatically improved performance. Pioneering works such as DeepFace [75], FaceNet [66], and VGGFace [58] demonstrated that CNNs trained on millions of faces can achieve near-human accuracy. FaceNet, for instance, introduced a triplet-loss training scheme that learns a compact embedding space where matching faces are pulled together and non-matching faces are pushed apart. Their model achieved 99.63% accuracy on LFW, substantially above previous results. Transfer learning is also common: a CNN may be pretrained on an extensive general face dataset (e.g., MS-Celeb-1M) and then fine-tuned on a target dataset. The learned CNN activations serve as high-quality feature embeddings for recognition tasks.

Loss functions specifically designed for face recognition continue to advance the field. Margin-based softmax losses enforce angular separability between classes. For example, SphereFace [48]

introduced an angular margin via a multiplicative modification; CosFace [78] applied an additive cosine margin, and ArcFace [21] introduced an additive angular margin with a clear geometric interpretation. These losses encourage inter-class feature angles to be larger than intra-class angles, improving discriminability. Recent state-of-the-art loss functions include AdaFace [43], which adaptively modulates the margin based on image quality, further boosting performance on challenging, low-quality data.

Face recognition models are evaluated on standardized benchmarks and protocols. Verification tasks (deciding whether two face images are the same person) are common; accuracy on datasets like LFW, CFP-FP, CALFW, and AgeDB is routinely reported. To assess real-world performance, more challenging open-set protocols such as IJB-B and IJB-C measure true accept rates at low false accept rates (e.g., TAR@FAR=1e-4). Identification tasks evaluate retrieval accuracy from a gallery of identities. In all cases, the goal is to maximize recognition accuracy while minimizing bias across demographics.

Generative models play an increasing role in face recognition. GANs and diffusion models have been used to augment training data, especially to mitigate biases. By synthesizing additional faces for underrepresented groups, these models can help balance a dataset. For example, style-based augmentation frameworks have been proposed that transfer attributes like age, gender, and skin tone between images to improve demographic diversity. DCFace's authors note that synthetic datasets can remedy issues such as long-tail distributions and demographic bias in real datasets. Recent controlled generation methods explicitly target fairness: generating images with uniform attributes improves equality of opportunity across subgroups. Empirical studies have shown that fine-tuning a recognition model with balanced synthetic data can reduce demographic disparity without sacrificing accuracy.

Thus, combining state-of-the-art CNNs with synthetic data from GANs or diffusion models offers a promising approach to building fairer face recognition systems.

2.5 Leveraging Diffusion Models for Face Recognition

Integrating diffusion-based generative models with convolutional backbones introduces a systematic approach to address dataset imbalance and strengthen feature representations in face recognition systems.

2.5.1 Data Augmentation for Bias Mitigation

Face recognition benchmarks frequently exhibit skewed demographic distributions, resulting in disparate performance across groups. To counteract this, a trained diffusion model can be conditioned on explicit demographic attributes (e.g., ethnicity, age, gender) and then fine-tuned to produce realistic samples of underrepresented cohorts. These generated images serve three principal purposes: first, they increase the overall sample count for minority groups, thereby reducing statistical sampling error; second, they permit the construction of controlled, balanced training splits that equalize representation; and third, they enable controlled exploration of the latent space

to verify that the attribute conditioning governs the generation process without interfering with identity consistency. By incorporating such synthetic data into the training pipeline, one obtains a more uniform distribution of facial attributes and, thus, a recognition model whose errors are less correlated with demographic factors.

2.5.2 Enhancing Feature Robustness via Synthetic Variability

Beyond balancing the dataset, diffusion-generated images can expose the feature extractor to a broader range of intra-class variations. Specifically, by sampling from the conditioned diffusion process with systematically perturbed noise schedules or guidance scales, one can generate multiple renditions of the same identity under varying pose, illumination, and expression. Embedding these renditions alongside real examples encourages the CNN to learn invariances to nuisance factors. In practice, one can interleave batches of real and synthetic images during training or apply a multi-view consistency loss that enforces the proximity of embeddings for generated variants of a given identity.

2.6 Conclusion

In this chapter, we have surveyed the foundational techniques that underpin our approach to building fair and robust face recognition systems. We began by examining diffusion models, elucidating their discrete-time and continuous-time formulations and the mechanisms, such as classifier and classifier-free guidance, that enable precise control over generated samples. This probabilistic framework contrasts sharply with adversarial objectives and offers stable training dynamics alongside interpretable likelihood bounds.

We then reviewed convolutional neural networks, detailing how convolutional operations, nonlinear activations, pooling, and normalization layers combine to learn hierarchical feature representations. In the context of face recognition, deep CNNs, especially ResNet-based architectures, have demonstrated remarkable performance when trained on large-scale datasets and refined with margin-based loss functions such as CosFace and ArcFace. We highlighted how these advances have driven near-human accuracy on standard benchmarks and the ongoing evolution of loss formulations to improve discriminability under challenging conditions.

Subsequently, we explored the interplay between generative models and recognition networks. Generative adversarial networks provide rapid, feed-forward synthesis of sharp images but suffer from training instability. In contrast, diffusion models offer superior mode coverage and a stable training objective at the cost of iterative sampling. By synthesizing balanced datasets and exposing CNNs to controlled variations in pose, illumination, and identity, diffusion-generated images can mitigate demographic biases and enhance feature robustness.

Together, these insights establish the theoretical and practical groundwork for our proposed method: integrating identity- and attribute-conditioned diffusion generators with state-of-the-art CNN backbones.

Chapter 3

State of the Art

This chapter reviews the current literature on face recognition. Section 3.1 presents foundational face detection methods, ranging from classical techniques to modern deep-learning–based detectors. Section 3.2 describes convolutional neural network architectures for feature extraction and classification, tracing their evolution from early models to contemporary embedding techniques. Section 3.3 examines fairness and bias mitigation strategies, including dataset-level interventions, algorithmic enhancements, and knowledge-distillation approaches aimed at ensuring equitable performance across demographic groups. Section 3.4 reviews generative methods for synthetic dataset creation, comparing the strengths and limitations of generative adversarial networks with those of diffusion models in addressing data scarcity and bias in face recognition. Finally, in Section 3.5, we present a summary that synthesizes these advances, identifies persistent gaps, and motivates the objectives of this thesis.

3.1 Face Detection Methods

Face detection is an essential step in face recognition systems, which provides the location of facial regions under various conditions. This section reviews five popular methods: Haar Cascades, HOG-based detection, Multi-task Cascaded Convolutional Neural Networks, RetinaFace, and You Only Look Once (YOLO).

3.1.1 Haar Cascades (2001)

Haar Cascades is a machine learning-based object detection method introduced by Viola and Jones in 2001 [76]. This method relies on Haar-like features and a cascading classifier structure to detect objects in images efficiently. Haar-like features are simple rectangular features that compute the difference in pixel intensity values within a specific region, enabling rapid classification of regions as either foreground (e.g., a face) or background. The cascade classifier works by applying multiple stages of increasingly complex classifiers, allowing efficient elimination of non-relevant regions early in the detection process and ensuring fast processing times.

A key strength of Haar Cascades is its computational efficiency, which makes it particularly suitable for low-resource environments such as embedded devices and mobile phones [69]. Moreover, it reliably detects frontal faces under controlled lighting conditions [1]. However, the method has some limitations. Its performance degrades significantly when detecting non-frontal faces or when operating in low light environments [1]. Additionally, Haar Cascades is prone to high false positive rates, especially in scenarios involving occlusion, varying illumination, or the presence of accessories such as glasses [40].

3.1.2 HOG-based Detection (2005)

The Histogram of Oriented Gradients (HOG) feature descriptor is a widely recognized method for object detection, introduced by Dalal and Triggs in 2005 [19]. This technique extracts gradient information from localized sections of an image, effectively capturing the structure and appearance of objects. HOG has demonstrated notable success in face detection tasks when paired with a Support Vector Machine (SVM) classifier. Its implementation in the `dlib`[1] library has further enabled efficient real-time applications.

A significant strength of the HOG-based approach is its robustness to small occlusions and moderate variations in face orientation [40]. This method accurately detects frontal and slightly rotated faces, with detection rates surpassing 92.68% [1]. Moreover, it can identify faces adorned with accessories, although occasional errors in precise face localization may occur [69].

Despite its advantages, the HOG-based method exhibits limitations. It struggles with detecting heavily occluded or highly rotated faces, highlighting its reduced robustness under these challenging conditions [69]. Additionally, this method's computational demands are higher than the Haar Cascades approach, which may hinder its applicability in scenarios requiring minimal resource usage [1].

3.1.3 Multi-task Cascaded Convolutional Neural Networks (2016)

MTCNN [86] is a deep learning-based method designed for face detection and facial landmark localization in images and videos. The approach integrates a multi-stage framework consisting of three CNNs that progressively refine detection results, ensuring accurate localization of faces and their landmarks. This architecture enables MTCNN to handle challenges commonly encountered in real-world scenarios, such as pose, scale, and occlusion variations. The method achieves state-of-the-art accuracy, effectively detecting faces from diverse poses and even with accessories such as glasses [40, 86]. Its robust landmark detection capabilities also facilitate face alignment, a critical step in improving performance for downstream recognition tasks [69, 86].

Despite its advantages, MTCNN exhibits slower processing speeds compared to traditional methods like Haar Cascades and HOG, which may limit its usability in time-sensitive applications [26, 69]. Moreover, the network's performance diminishes under very low lighting conditions or when faces are heavily occluded, making it less reliable in such challenging environments [69].

[1] `https://dlib.net/`

3.1.4 YOLO-based Detection (2016)

YOLO models [62] are a class of deep learning-based object detection frameworks that operate in real-time. Initially developed for general object detection tasks, YOLO models have been adapted for face detection, leveraging their high speed and ability to detect multiple faces simultaneously [59]. YOLO's approach involves predicting bounding boxes and class labels directly from the image's features, allowing for efficient, end-to-end inference in a single pass. This method enables faster processing times, making YOLO-based models particularly suitable for applications where real-time performance is critical. Additionally, YOLO can detect multiple faces in complex scenes, which is advantageous for scenarios with cluttered backgrounds or numerous subjects [51, 62].

Despite these strengths, YOLO-based models exhibit limitations, particularly in scenarios involving small faces, where their precision is lower compared to methods like RetinaFace [51]. Furthermore, achieving optimal performance often requires careful tuning of anchor boxes, which adds complexity to the training process [60].

3.1.5 RetinaFace (2020)

RetinaFace [20] is a deep learning-based method that integrates landmark localization with face detection. It uses a single-stage face detector that combines facial landmark localization, face detection, and alignment in an end-to-end manner. RetinaFace is known for its high accuracy, especially in detecting faces under various challenging conditions, such as occlusion, significant variations in pose, and small face sizes. The model leverages a multi-task learning approach that improves the robustness of face detection while also providing detailed face landmarks, making it suitable for applications requiring fine-grained facial feature analysis.

One of the notable strengths of RetinaFace is its exceptional performance in detecting small and occluded faces, a challenge for many face detection models [20]. Additionally, its multi-task capabilities extend beyond detection to include face segmentation and alignment, further enhancing its utility in applications requiring comprehensive facial analysis [20]. However, these benefits come at the cost of high computational demands. The model's complexity necessitates using high-end hardware to achieve real-time performance, limiting its accessibility for applications with constrained computational resources [26, 51].

3.1.6 Challenges and Future Directions

Despite the significant advancements in face detection technologies, multiple challenges remain in developing robust and efficient systems capable of handling diverse real-world scenarios. One of the primary challenges across traditional methods like Haar Cascades and HOG-based detection is their limited performance in complex conditions, such as non-frontal faces, occlusions, low-light environments, and variations in facial accessories. Although methods like MTCNN and RetinaFace have significantly improved accuracy and robustness, they still suffer from performance degradation under extreme conditions, such as very low lighting or heavy occlusions.

Additionally, while YOLO-based models provide real-time performance and can detect multiple faces simultaneously, they face challenges with precision in detecting smaller faces and require special tuning for optimal performance.

Future research should focus on overcoming the limitations of existing methods, particularly in improving robustness under diverse and complex conditions. This includes enhancing the capabilities of traditional methods like Haar Cascades and HOG for more effective detection of non-frontal and occluded faces. For deep learning-based models, further innovations are needed to speed up the processing time without sacrificing accuracy, addressing the slower performance seen in MTCNN and RetinaFace. Additionally, models could be improved to better handle faces under varying illumination, accessories, or pose changes, expanding their application in real-world settings. Finally, developing lightweight and more computationally efficient models that retain high accuracy will enable real-time face detection in resource-constrained environments, such as mobile devices and low-cost cameras.

3.2 CNNs for Face Recognition

Convolutional neural networks have become a cornerstone of face recognition systems because they can learn hierarchical and discriminative features directly from raw data [68]. The use of CNNs in image processing began with LeNet-5, proposed by LeCun et al. in 1989 [47] for handwritten digit recognition. This early model demonstrated the effectiveness of convolutional layers in feature extraction.

The resurgence of CNNs came with AlexNet in 2012, which achieved breakthrough results on ImageNet through deep architectures and GPU acceleration [46]. VGGNet [68] and ResNet [34] further advanced CNN design, enabling deeper networks and tackling challenges like vanishing gradients.

Face recognition applications followed this trajectory, with DeepFace [75] and FaceNet [66] introducing embeddings for face matching, and VGGFace [58] leveraging large datasets to refine accuracy. This section reviews the best-performing CNN-based models, focusing on their architectures and training processes. These advancements reflect the broader evolution of CNNs in Computer Vision, which has transformed image analysis over the past decades [42].

3.2.1 DeepFace (2014)

DeepFace [75] is one of the pioneering deep learning models for face recognition developed by researchers at Facebook. The architecture of DeepFace is built around a 9-layer convolutional neural network with approximately 120 million trainable parameters. It integrates locally and fully connected layers to extract discriminative facial features. DeepFace employs an architecture inspired by traditional CNNs but incorporates spatial transformation techniques to align faces geometrically before feature extraction, significantly improving recognition accuracy.

The model uses a softmax loss function for classification and is trained on a large-scale dataset consisting of 4.4 million labeled facial images spanning 4,030 identities, achieving a 97.35% accuracy score on LFW.

3.2.2 VGG-Face (2015)

VGG-Face [58], developed by researchers at the University of Oxford, is a deep learning model designed explicitly for face recognition tasks. The architecture is based on the well-known VGGNet [68], a CNN renowned for its simplicity and effectiveness in visual recognition tasks. VGG-Face consists of 38 layers, including 16 convolutional layers, one input layer, 15 ReLU activation layers, five max pooling layers, and a softmax output layer, with approximately 138 million trainable parameters.

The model employs small convolutional filters (3x3), max pooling layers, and ReLU activations, which help learn hierarchical facial features. The depth of the network allows it to capture fine-grained details necessary for distinguishing between similar faces while maintaining robustness to variations such as pose and lighting.

VGG-Face was trained on a large dataset of 2.6 million facial images spanning 2,622 identities, collected from the LFW, YouTube Faces, and Internet Movie Data Base (IMDB) celebrity list datasets [67]. The model uses a softmax loss function for classification during training and outputs 4,096-dimensional feature embeddings, which can be further fine-tuned for various face recognition tasks.

Using its relatively large number of parameters, VGG-Face demonstrated excellent performance on several face recognition benchmarks, achieving 98.95%

3.2.3 FaceNet (2015)

FaceNet [66], developed by Google, introduced a new paradigm in face recognition by learning a compact embedding space. The architecture is built on a modified version of the Zeiler & Fergus model [85], comprising 22 layers with approximately 140 million parameters. The model uses a triplet loss function that minimizes the distance between embeddings of the same identity while maximizing the separation between different identities. It was trained on a vast private dataset comprising around 200 million images of 8 million identities [67], employing more advanced triplet sampling strategies to improve the learning process. This model is one of the best-performing models in LFW, with an accuracy of 99.83%.

3.2.4 OpenFace (2016)

OpenFace [4] is a lightweight and open-source deep learning model designed for face recognition inspired by the FaceNet architecture. The model uses the Inception-ResNet architecture, which combines the efficiency of Inception modules with the residual learning framework, resulting in a compact and efficient network suitable for real-time applications. OpenFace comprises 22 layers, with approximately 4 million trainable parameters, significantly smaller than most state-of-the-art

models. This reduced complexity makes OpenFace well-suited for scenarios where computational resources are limited.

The model uses a triplet loss function to learn discriminative embeddings, similar to FaceNet. During training, the triplet loss minimizes the Euclidean distance between embeddings of the same identity while maximizing the distance between different identities. OpenFace was trained on the CASIA-WebFace dataset. Preprocessing techniques such as face alignment using `dlib`'s facial landmark detector and data augmentation (e.g., rotation and scaling) were employed to improve the model's robustness to variations in pose and illumination. This model achieved 90% accuracy on LFW.

3.2.5 ArcFace (2019)

ArcFace [21] is a state-of-the-art face recognition model that introduces an Additive Angular Margin Loss to enhance discriminative power in the learned feature embeddings. The model is built on the ResNet-100 architecture [34], a deep convolutional neural network comprising 100 layers, including residual and convolutional layers, promoting efficient feature learning through skip connections. This architecture enables ArcFace to achieve remarkable accuracy while maintaining computational efficiency. The model contains approximately 65 million trainable parameters, making it relatively lightweight for its depth.

ArcFace is trained using the MS1M dataset [32], a large-scale collection of approximately 10 million face images from 100,000 individuals. During training, the Additive Angular Margin Loss is employed to improve the separability of embeddings in the angular space, ensuring robust performance even under challenging scenarios such as variations in pose, lighting, and occlusion. The dataset is preprocessed with face alignment and normalization techniques to enhance model generalization. ArcFace achieves state-of-the-art performance on benchmark datasets, including 99.83% accuracy on LFW and 98.35% on MegaFace.

3.2.6 Challenges and Future Directions

While convolutional neural networks have revolutionized face recognition, several challenges persist, particularly as the field evolves toward increasingly complex and real-world applications. Despite the success of models like DeepFace, VGG-Face, and FaceNet, which have set high benchmarks in accuracy, these systems face limitations in terms of their scalability and generalizability across diverse face recognition scenarios. For instance, while these models perform exceptionally well on datasets like LFW, their performance can degrade when faced with the variability inherent in real-world data, such as variations in age, ethnicity, or appearance due to makeup. Furthermore, these models' large number of parameters and computational complexity pose challenges for deployment on resource-constrained devices, such as mobile phones and embedded systems.

The architectures of models like VGG-Face and FaceNet, which use many parameters while improving accuracy, may also lead to slower processing times, particularly in real-time applications. This issue is compounded by the high computational requirements for training such deep

networks, necessitating extensive computational resources that may not be accessible in all settings. Although OpenFace offers a lighter alternative, it still struggles to match the performance of its heavier counterparts in terms of accuracy, particularly in challenging conditions involving extreme poses or occlusions. Moreover, despite their high performance, even state-of-the-art models like ArcFace can encounter difficulties when faces are heavily occluded or when there are substantial pose and illumination variations.

Additionally, most of these models are trained on datasets that are heavily racially unbalanced, such as the LFW dataset, which consists predominantly of faces from Caucasian individuals. This racial bias can lead to poorer performance when applied to individuals from underrepresented demographic groups. While these models achieve high overall accuracy, their discriminatory behavior toward minority ethnic groups remains a significant concern. Therefore, a critical area for future research is the development of racially unbiased face recognition algorithms. To address this, future models must improve performance across a broader range of conditions and ensure fairness and inclusivity, reducing the disparities in accuracy observed across different racial groups. Achieving this will require the creation of more diverse and representative training datasets and the development of fairness-aware algorithms that actively mitigate bias and ensure equitable performance across all demographic groups.

Another important direction for future research is the development of lightweight models [7] that maintain the high accuracy of current state-of-the-art approaches. While models like FaceNet, VGG-Face, and ArcFace provide exceptional results, they come with substantial computational overhead that can limit their usability in real-time applications or on devices with limited resources. By creating more efficient architectures that preserve their high performance while reducing complexity, it will be possible to extend the reach of face recognition systems to a broader range of platforms and devices, including mobile phones and edge devices, where computational power is constrained.

3.3 Mitigating Bias in Face Recognition Systems

Face recognition systems have become increasingly pervasive in real-world applications, from security to personalized services. However, these systems often exhibit bias against certain demographic groups, leading to unfair outcomes [27]. Bias in face recognition can arise from various sources [2, 83], including imbalanced datasets, model architectures, and algorithmic choices. Addressing these biases has been a focal point of research [17, 70], aiming to develop systems that perform equitably across diverse populations. Below, existing key approaches and methodologies for mitigating bias in face recognition systems are presented.

3.3.1 Dataset-Level Interventions

One of the primary sources of bias in face recognition systems is the imbalance in datasets, which often underrepresent minority demographic groups [74]. This underrepresentation leads to models that generalize poorly for these groups [57]. A common mitigation strategy is dataset

balancing [11, 37, 41], which involves augmenting underrepresented categories or curating demographically diverse datasets. Yucer et al. [84] propose per-subject adversarially-enabled data augmentation to create synthetic samples for underrepresented racial groups, thereby improving the model's fairness across demographics.

Additionally, techniques such as adversarial data augmentation [16] have been shown to introduce variations that enhance the generalizability of models without compromising their accuracy for overrepresented groups [84].

3.3.2 Knowledge Distillation for Bias Mitigation

Knowledge Distillation (KD) has proven to be an effective method for addressing bias in face recognition systems [12, 14, 31, 52]. KD involves transferring knowledge from a large, pre-trained teacher model to a smaller student model, enabling the latter to learn the teacher's capabilities while potentially mitigating biases.

Caldeira et al. [13] introduced the MST-KD framework, where multiple specialized teacher networks are trained on specific ethnic groups. These biased teachers then distill their knowledge into a single student network, combining ethnicity-specific features into a unified model. This kind of approach of combining multiple "opinions", also discussed by Kolberg et al. [45] and by Amirkhani et al. [3], promotes fairness and robustness in the student model, outperforming traditional KD methods that use balanced datasets.

Neto et al. [54] extended the application of KD by leveraging synthetic datasets. Their findings highlight that using KD with a mix of real and synthetic datasets not only improves the fairness of the student model but also mitigates the performance gap caused by training on synthetic data. This dual advantage demonstrates the potential of KD for developing fair face recognition systems.

In addition to improving fairness, KD offers practical benefits, such as reducing computational costs and enabling the deployment of lightweight models in real-world scenarios. These characteristics make Knowledge Distillation a valuable tool for addressing bias while maintaining the scalability and efficiency of face recognition systems.

3.3.3 Algorithmic Enhancements

Beyond datasets, algorithmic modifications also have a role in mitigating bias. Loss functions, for instance, have been extensively studied for their impact on fairness. Gong et al. [29] introduce a Group Adaptive Classifier that dynamically adjusts class weights during training, ensuring that the model gives equitable attention to all demographic groups. Similarly, Nguyen et al. [55] highlight how effective loss functions can enhance accuracy and fairness, combining ensemble methods to achieve robustness against bias.

Recent advancements also include the development of architectures specifically designed to reduce bias. Dooley et al. [25] argue that fairer architectures inherently promote fairer outcomes. Their work demonstrates that redesigning neural network structures with fairness as a guiding principle can significantly mitigate biases observed in traditional architectures.

3.3.4 Evaluation and Feedback Mechanisms

Bias mitigation also involves evaluating models using metrics that capture disparities across demographic groups. Robinson et al. [63] emphasize the importance of evaluation frameworks that measure overall accuracy and assess performance disparities. Such frameworks enable researchers to identify specific areas of bias and iteratively refine their models.

Feedback mechanisms, such as adversarial training, further support bias reduction by introducing constraints during optimization that penalize disproportionate errors across groups. This approach has been particularly effective in addressing racial bias, as evidenced by Sumsion et al. [74] in their survey on algorithmic enhancements for mitigating racial disparities.

3.3.5 Hybrid Approaches

Hybrid methods that combine dataset-level interventions with algorithmic enhancements represent a promising avenue for bias mitigation. Wang et al. [77] propose the MixFairFace framework, which leverages both dataset augmentation and fairness-driven architectural modifications to achieve "ultimate fairness" in face recognition. By integrating multiple strategies, hybrid approaches aim to address the multifaceted nature of bias.

3.3.6 Challenges and Future Directions

While substantial progress has been made in addressing bias in face recognition systems, numerous challenges remain, particularly in the quest for universal fairness across all demographic groups. One persistent issue is the inherent limitations of dataset-level interventions. Although balancing datasets and employing adversarial data augmentation techniques have successfully improved fairness, these approaches may still fail to capture the full diversity of underrepresented groups, especially in real-world applications where data collection is subject to privacy and accessibility constraints. Moreover, synthetic data generation, while promising, can introduce artifacts that may not fully reflect the complexities of human variation, potentially leading to models that perform well in controlled environments but fail to generalize in more diverse settings.

Knowledge Distillation, another valid method for mitigating bias, offers promising results by transferring knowledge from specialized teacher models to a unified student model. However, KD frameworks often rely on the assumption that the teacher models are sufficiently representative of the target demographic groups. This can create issues when incorporating multiple ethnicities or handling more granular demographic distinctions. Furthermore, while Knowledge Distillation can improve fairness and reduce computational costs, it raises the challenge of balancing the trade-off between model simplicity and the richness of features needed for equitable decision-making.

Algorithmic enhancements, including modifications to loss functions and network architectures, have shown potential in mitigating bias. Nevertheless, there is still no consensus on the best way to design fair architecture. Fairness constraints integrated into the loss function often struggle to address the full scope of bias, particularly when demographic characteristics interact

in complex ways. While new, fairer architectures proposed by Dooley et al. are an exciting direction, these models often require more extensive experimentation to ensure they are robust across various demographic categories.

The evaluation mechanisms for assessing bias have improved but continue to face challenges. Although nuanced evaluation frameworks can pinpoint disparities between groups, current metrics may not capture the dynamic and multifaceted nature of bias. In practice, continuously evolving social norms and demographic shifts mean that fairness metrics must adapt to new contexts, and evaluations may need to be performed iteratively to detect hidden biases that emerge over time. Moreover, feedback mechanisms such as adversarial training have shown promise but may not always fully penalize disproportionate errors across all groups, especially when certain groups are underrepresented or misrepresented in the training data.

Hybrid approaches that combine multiple strategies, such as dataset augmentation with fairness-driven algorithmic modifications, offer a more comprehensive solution to bias. However, integrating these techniques often introduces new complexities, particularly in ensuring that the combination does not inadvertently introduce new sources of bias or fail to scale effectively. Future research should explore methods for optimizing the integration of these approaches while considering the trade-offs between model fairness, accuracy, and computational efficiency.

Moving forward, future work on mitigating bias in face recognition systems will need to focus on several key areas: improving the representativeness of datasets, developing more nuanced evaluation metrics, and advancing hybrid models that balance fairness with practical deployment concerns. Furthermore, fairness must be considered a technical challenge and a socioethical issue, ensuring that the solutions developed are aligned with societal values and subject to ongoing scrutiny and adjustment in light of new research and societal changes.

3.4 Diffusion Models for Face Recognition

Diffusion models have emerged as a transformative approach in generative artificial intelligence, offering unprecedented capabilities in synthetic data generation and bias mitigation for face recognition systems. Building upon the foundational success of convolutional neural networks in feature extraction and identity matching [66, 75], diffusion models introduce probabilistic frameworks that iteratively refine noise into high-fidelity facial images while preserving demographic diversity. This paradigm shift addresses critical limitations in traditional datasets, such as privacy concerns, racial imbalance, and insufficient representation of underrepresented groups.

3.4.1 Comparative Analysis of Diffusion Models and GANs in Face Recognition

The evolution of generative models has significantly impacted the field of face recognition, with generative adversarial networks and diffusion models emerging as prominent techniques. This section compares these two approaches, focusing on their applications and performance in face recognition tasks.

3.4.1.1 Generative Adversarial Networks

Introduced by Goodfellow et al. in 2014 [30], GANs consist of a generator and a discriminator engaged in a minimax game, where the generator aims to produce realistic data, and the discriminator strives to distinguish between real and generated data. This adversarial training enables GANs to generate high-quality, realistic images, making them valuable in augmenting face datasets for recognition tasks. However, GANs face challenges such as mode collapse, where the generator produces limited variations, and training instability due to the delicate balance required between the generator and discriminator. These issues can hinder the generation of diverse facial images necessary for robust face recognition models [5].

3.4.1.2 Diffusion Models

Diffusion models are a class of generative models that learn data distributions by iteratively denoising samples from a Gaussian distribution. Recent advancements have demonstrated that diffusion models can achieve superior image synthesis quality compared to GANs. Dhariwal and Nichol (2021) [22] showed that diffusion models surpassed GANs in generating high-fidelity images, achieving a Fréchet Inception Distance (FID) of 2.97 on ImageNet 128×128.

In the context of face recognition, diffusion models offer advantages in generating synthetic datasets with enhanced intra-class diversity and inter-class separability. For instance, IDiff-Face utilizes identity-conditioned latent diffusion models to produce synthetic faces that maintain identity consistency while introducing realistic variations, improving recognition performance [8].

3.4.1.3 Comparative Insights

While both GANs and diffusion models can generate high-quality images, diffusion models have demonstrated superior image fidelity and diversity performance. The iterative denoising process in diffusion models allows for better data distribution coverage, reducing issues like mode collapse commonly associated with GANs.

Moreover, in face recognition applications, the ability of diffusion models to conditionally generate images with controlled identity features and variations makes them more effective in creating synthetic datasets that enhance model training. This capability addresses the limitations of GANs in capturing the diversity and separability required for accurate face recognition.

However, it is important to note that diffusion models typically require more computational resources and longer training times than GANs, which may impact their practicality in specific applications.

In summary, while GANs have been instrumental in advancing generative modeling for face recognition, diffusion models offer notable improvements in image quality and diversity, making them a promising alternative for generating synthetic data to train robust face recognition systems.

3.4.2 IDiff-Face: Synthetic-based Face Recognition through Fizzy Identity-conditioned Diffusion Models

The development of FR systems has historically relied on large-scale real face databases. However, legal and ethical concerns regarding data privacy have led to the retraction of many such datasets, necessitating the exploration of synthetic alternatives [49]. Recent synthetic datasets have struggled to balance intra-class diversity and inter-class separability [28], crucial for high-accuracy FR models. IDiff-Face [8] introduces a novel synthetic dataset generation method based on identity-conditioned latent diffusion models to address these challenges.

IDiff-Face operates by training a diffusion model in the latent space of an autoencoder, where the denoising process is conditioned on identity representations derived from a pre-trained FR model. The identity conditioning is incorporated through a Cross-Attention mechanism [8], enabling the model to generate synthetic face images with controlled identity-specific variations. To enhance intra-class diversity while maintaining identity separability, IDiff-Face employs a Contextual Partial Dropout mechanism, which stochastically removes portions of the identity context during training, preventing overfitting and promoting diversity in the generated samples.

Extensive evaluations demonstrate that FR models trained on IDiff-Face-generated datasets significantly outperform prior synthetic-based FR approaches. For instance, an FR model trained on 500K IDiff-Face synthetic images achieved a 98.00% verification accuracy on the Labeled Faces in the Wild dataset, surpassing the best synthetic-based alternatives and approaching the 99.82% accuracy of models trained on large-scale authentic datasets [8]. Additionally, IDiff-Face-generated datasets achieve a more optimal trade-off between intra-class diversity and inter-class separability compared to synthetic datasets based on GANs or digital rendering techniques.

Despite these advancements, IDiff-Face has some limitations. While it significantly narrows the performance gap between synthetic and authentic-based FR models, there is still room for improvement, particularly in handling extreme variations in pose, illumination, and occlusions. Moreover, the computational cost of training diffusion models remains high compared to GAN-based approaches, potentially limiting scalability [23, 87].

3.4.3 DCFace: Synthetic Face Generation with Dual Condition Diffusion Model

DCFace introduces a novel synthetic dataset generation approach for face recognition by leveraging a dual-condition diffusion model. Unlike traditional generative methods that struggle to capture both intra-class diversity and inter-class separability [49], DCFace explicitly conditions the image synthesis process on two independent factors: identity and style. This enables precise control over the appearance of synthetic subjects while maintaining high intra-class variation, making it an effective alternative to real-world datasets for training FR models [44].

The core methodology of DCFace consists of a two-stage process. In the first stage, an unconditional generative model produces high-quality ID images that define the subject's appearance. In the second stage, a Dual Condition Generator takes an identity image and an independently sampled style image—defining factors like pose, illumination, and expression—to synthesize a

new face while preserving the subject's identity. This dual-conditioning framework is implemented within a diffusion model, where the ID condition is reinforced using an auxiliary time-step-dependent loss, ensuring consistency across synthesized images [44].

DCFace significantly improves the performance of FR models trained on synthetic data. When trained with 500K synthetic images, FR models achieve an average verification accuracy improvement of 6.11% over previous synthetic dataset approaches across benchmarks like LFW, CFP-FP, CPLFW, AgeDB, and CALFW [44]. The model effectively balances diversity and identity preservation, leading to superior generalization. Notably, face recognition models trained on DCFace achieve a verification accuracy of 98.97% on the LFW benchmark [28], underscoring the potential of high-quality synthetic datasets to rival real-image baselines in unconstrained face verification tasks.

Despite its advantages, DCFace has certain limitations. First, while it improves subject uniqueness and image quality, it does not guarantee 3D consistency across different viewpoints, which remains an advantage of 3D parametric model-based methods. Also, diffusion models are computationally expensive, making large-scale dataset generation more resource-intensive than GAN-based alternatives [23, 87].

3.4.4 Challenges and Future Directions

Despite the significant advancements brought by diffusion models in synthetic face generation for face recognition, several challenges remain. Addressing these challenges is crucial for the widespread adoption and further refinement of these techniques.

One of the primary concerns is the computational cost associated with training and generating images using diffusion models. Unlike GANs, which can produce samples in a single forward pass, diffusion models require an iterative denoising process that significantly increases inference time and computational demand. This limitation hinders their scalability, particularly in large-scale applications requiring real-time or near-real-time performance. Future research could focus on optimizing the efficiency of diffusion-based synthesis, possibly through model distillation, reduced time-step sampling, or hybrid approaches combining diffusion models with faster generative frameworks.

Another critical challenge is ensuring the robustness and fairness of synthetic face datasets. While IDiff-Face and DCFace have demonstrated improvements in demographic diversity and bias mitigation, generating datasets that accurately reflect real-world distributions remains an open problem. The conditioning mechanisms in current models rely on predefined identity embeddings, which may not fully capture nuanced demographic variations. Further exploration into adaptive conditioning strategies, such as reinforcement learning-based optimization of identity constraints, could enhance the representational fidelity of synthetic datasets.

Additionally, diffusion-based methods struggle with extreme pose, illumination, and occlusion variations. While techniques like Contextual Partial Dropout in IDiff-Face aim to promote intra-class diversity, ensuring that synthetic faces maintain realism under all conditions remains a

challenge. Incorporating multimodal data sources, such as 3D face models or photometric priors, could improve the robustness of synthetic face generation under challenging scenarios.

Another pressing issue is evaluating and benchmarking synthetic datasets for face recognition. Traditional evaluation metrics, such as Fréchet Inception Distance and identity verification accuracy, may not fully capture intra-class diversity and separability nuances. Developing new evaluation protocols tailored for synthetic face datasets, incorporating fairness assessments and real-world deployment testing, is essential for validating the effectiveness of these models.

Finally, ethical considerations must continue to guide the development of synthetic data for face recognition. While synthetic datasets address privacy concerns associated with real-face datasets, the potential misuse of such data, including the generation of deepfakes or adversarial attacks, raises ethical and security concerns. Future research should prioritize the implementation of watermarking techniques or detection mechanisms to distinguish synthetic faces from real identities, ensuring the responsible use of generative models in face recognition.

Moving forward, integrating diffusion models with complementary techniques such as 3D face reconstruction, self-supervised learning, and domain adaptation holds promise for further enhancing their utility in face recognition. As computational hardware advances and new optimizations emerge, diffusion-based synthetic data generation is poised to play a pivotal role in the future of fair and accurate face recognition systems.

3.5 Summary

The state of the art in face recognition systems has been shaped significantly by advancements in deep learning, particularly with convolutional neural networks. However, before recognition, face detection plays a foundational role, and it comes with its own set of limitations. Classical methods like Haar Cascades and HOG-based detection struggle with detecting non-frontal faces, handling occlusions, or performing under challenging lighting conditions. While more modern techniques such as MTCNN, RetinaFace, and YOLO-based detection achieve higher accuracy and robustness, they often require substantial computational resources, with their performance degrading in extreme cases of occlusion, very low lighting, or detecting small faces. Additionally, achieving real-time detection on constrained hardware remains challenging for these methods. Such limitations highlight the persistent trade-offs in face detection speed, accuracy, and resource efficiency.

Deep learning methods, particularly CNNs, have demonstrated remarkable efficacy in learning hierarchical features directly from raw data, with models such as VGG-Face, FaceNet, OpenFace, DeepFace, and ArcFace achieving state-of-the-art performance across diverse benchmarks. Each model leverages different architectural innovations and loss functions, from VGG-Face's reliance on deep feature extraction to ArcFace's use of Additive Angular Margin Loss for improved discriminability. However, despite their success, these CNN-based approaches often require large-scale datasets for training and exhibit biases against underrepresented demographic groups, limiting their fairness and generalizability.

Efforts to mitigate bias have focused on dataset-level interventions, such as balancing and adversarial data augmentation, and algorithmic enhancements, including adaptive loss functions and fair architectures. Knowledge Distillation has also emerged as a valid strategy for improving fairness by transferring knowledge from multiple specialized teacher models to a unified student model. This approach enhances demographic fairness and facilitates the deployment of lightweight, efficient models suitable for real-world applications. However, the trade-off between fairness and accuracy remains a significant challenge, alongside the complexity of constructing effective training pairs and designing unbiased evaluation frameworks.

Beyond CNN-based methods, generative models have been crucial in addressing dataset limitations and concerns about fairness. GANs have been widely used for synthetic face dataset generation, but they suffer from challenges such as mode collapse and training instability, which can limit the diversity of synthetic samples. More recently, diffusion models have emerged as a powerful alternative for synthetic face generation, demonstrating superior image fidelity, diversity, and control over identity features. Techniques such as IDiff-Face and DCFace leverage diffusion models to generate high-quality synthetic faces with improved intra-class diversity and inter-class separability, addressing key limitations of GAN-based approaches. However, diffusion models require significantly more computational resources and inference time, posing challenges for large-scale applications.

This thesis enhances existing approaches by integrating state-of-the-art CNNs with diffusion-based synthetic data generation to develop a fair and accurate face recognition system. Diffusion models contribute to bias mitigation by reducing dependence on extensive labeled datasets and enabling fine-grained control over the demographic balance. Their ability to generate diverse and representative synthetic faces strengthens the training process, improving generalization and fairness. This approach advances the field by addressing key limitations in dataset bias while maintaining efficiency and robustness in face recognition systems.

Chapter 4

Methodology

This chapter outlines the methodology employed for synthetic face data generation, bias mitigation, dataset analysis, and face recognition model development. Section 4.1 presents the Ethnicity Score classifier and its role in guiding generation toward individuals that best represent each ethnic category. Section 4.2 describes the unified experimental framework and configuration of the two diffusion-based generators (IDiff-Face and DCFace). Section 4.3 introduces our bias-mitigation strategy via identity–style diffusion, including multi-ethnicity scoring protocols and semantic filtering. Section 4.4 presents a comparative study using StyleGAN3 for identity synthesis. All technical requirements and the specifications of the GPUs employed in our experiments are provided in Appendix A. Section 4.5 covers face alignment and preprocessing procedures. Sections 4.6.1 through 4.6.3 define the protocols and metrics for dataset comparison, quality assessment, and bias evaluation. Section 4.7 describes the development, training, and evaluation of face recognition models on synthetic data, including the architectures, training regimen, and fairness analysis. Finally, Section 4.8 concludes the chapter by summarizing the methodology and outlining the transition to the "Results and Discussion" chapter.

4.1 Ethnicity Score

To ensure that our synthetic datasets exhibit true demographic balance, it is not sufficient to generate an equal number of images for each race; individual faces vary in how strongly they exhibit racial characteristics. Hence, following the work of Neto et al. [53], we employ the Ethnicity Score, a ResNet-100-based classifier that quantifies, for each face image, the degree to which it represents each of four target groups (Caucasian, Asian, Indian, and African). Given an input image X, the classifier outputs a four-dimensional vector

$$\mathrm{ES}(X) = \left(s^{\mathrm{Af}}, s^{\mathrm{Ca}}, s^{\mathrm{As}}, s^{\mathrm{In}}\right)^{\top},$$

where each component $s^r \in [0,1]$ denotes the confidence that X belongs to race r. Higher values indicate stronger visual alignment with the corresponding ethnicity. During generation, we use these scores to guide both the selection of identity seeds and the choice of style exemplars (as

explained in section 4.3), thereby prioritizing individuals and images that best represent each racial category and improving the fidelity and fairness of our synthetic datasets.

4.2 Diffusion Models: Setup & Configuration

We established a unified experimental framework for our diffusion-based sampling protocols to generate and evaluate our synthetic face datasets systematically. In Section 4.2.1, we start by justifying our choice of IDiff-Face for initial experiments—on account of its transparent implementation, minimal tuning requirements and identity-conditioned latent diffusion—and motivate the subsequent switch to DCFace to achieve precise control over non-identity attributes and enable demographic balancing. Then, we describe the configuration and hyperparameters for each of the two diffusion frameworks under study. Section 4.2.2 focuses on IDiff-Face, detailing model variants, identity conditioning contexts, and sampling parameters (inference steps, guidance scales, etc.). Section 4.2.3 then presents the DCFace setup, covering both unconditional and conditional style-mixing modes, the scale of data generation, and the evolution of our runs from pilot to large-scale synthesis. Together, the upcoming subsections provide the comprehensive setup information necessary to ensure reproducibility and facilitate future extensions of our synthetic data generation methodology.

4.2.1 Model Selection Rationale

The first stage of our work employed IDiff-Face due to its relatively transparent codebase and minimal tuning requirements. Its identity-conditioned latent diffusion process is implemented with few adjustable components beyond the Contextual Partial Dropout rate and the number of inference steps, which allowed us to explore the balance between identity fidelity and intra-class variability. Although IDiff-Face demonstrates strong identity separability, it lacks mechanisms for prescribing non-identity attributes such as ethnicity, limiting its applicability when demographic balance is required.

We transitioned to DCFace for our large-scale data generation to address this shortcoming. DCFace's dual-condition architecture separates identity and style, with dedicated encoders and a time-dependent loss schedule that enforces identity preservation early in diffusion while enabling style adoption later. The increased number of hyperparameters (ranging from diffusion scheduler respacing to style-token grid granularity) demands more extensive fine-tuning. However, it grants precise control over racial style attributes by sampling style exemplars from defined ethnic subsets. This capability was essential for constructing racially balanced datasets and for evaluating bias mitigation strategies. In this way, IDiff-Face provided an accessible foundation for validating diffusion-based synthesis, and DCFace supplied the fine-grained style control needed for our comprehensive, bias-aware experiments.

4.2.2 IDiff-Face

For experiments using IDiff-Face, we extended the public implementation of Boutros et al. (2023) [8][1] to improve throughput. Specifically, we disabled gradient tracking during inference, implemented batch processing of multiple images per GPU pass, and introduced parallel I/O pipelines to overlap filesystem operations and image serialization with model execution.

The remainder of this section is structured as follows. In Section 4.2.2.1, we detail the two pretrained latent-diffusion checkpoints employed, their shared U-Net backbone with cross-attention conditioning, and the differing Contextual Partial Dropout probabilities. Then, Section 4.2.2.2 describes the four types of identity context files—uniformly sampled synthetic embeddings, ElasticFace-derived embeddings from FFHQ and LFW, and two-stage synthetic embeddings—along with their deterministic generation and reuse protocol. Here, an *identity context* denotes a precomputed set of N 512-dimensional embedding vectors representing candidate identities, which the model uses as conditioning priors during synthesis. Finally, Section 4.2.2.3 presents the two-phase image generation workflow, comprising a preliminary fidelity–diversity inspection and a large-scale sampling run, as well as the fixed seeding strategy and default diffusion hyperparameters used to isolate the effects of CPD rate and context type.

4.2.2.1 Model Variants

We employed the two pretrained latent-diffusion checkpoints provided with the code without modification: `unet-cond-ca-bs512-150K-cpd25` and `unet-cond-ca-bs512-150K-cpd50`. Both checkpoints share an identical U-Net backbone with cross-attention conditioning on a 512-dimensional identity embedding, a linear noise schedule over $T = 1000$ reverse diffusion steps, and a MSE reconstruction loss in the latent space of a VQGAN autoencoder. The sole distinction lies in the Contextual Partial Dropout (CPD) probability during training: 25% for the first checkpoint and 50% for the second. Higher CPD forces the network to ignore random subsets of the identity embedding during training, thereby promoting greater intra-class variability at the cost of somewhat reduced identity separability.

4.2.2.2 Identity Contexts

Prior to sampling, we constructed four fixed pools of identity contexts, each containing N 512-dimensional vectors (e.g. $N = 5\,000$ or $15\,000$). These pools were generated once and reused across all runs under a seeded, deterministic selection strategy to guarantee that every experiment draws from the same candidate identities.

The first pool, *random_synthetic_uniform_<N>*, comprises vectors sampled uniformly from the 512-dimensional unit hypersphere. Concretely, each raw vector $z \in \mathbb{R}^{512}$ is drawn from $\mathcal{N}(0, I)$ and then normalized to unit length. During generation, each fixed c is fed into IDiff-Face while varying only the diffusion noise seed, yielding multiple outputs of the same abstract identity.

[1] https://github.com/fdbtrs/IDiff-Face

The second pool, *random_elasticface_ffhq_<N>*, contains embeddings $c = f(x)$ obtained by passing N real face images x from the FFHQ training set through the pre-trained ElasticFace ResNet-100 encoder. These contexts anchor synthesis to the authentic distribution of high-quality faces the model saw during training.

The third pool, *random_elasticface_lfw_<N>*, is formed by encoding N LFW photographs (unconstrained, in-the-wild face images) via the same ElasticFace network. Fixing any such c enables IDiff-Face to generate diverse variations of a specific LFW identity, illustrating performance under unconstrained capture conditions.

The fourth pool, *random_synthetic_two_stage_<N>*, is built by repeating the Two-Stage IDiff-Face pipeline N times. For each context vector, an unconditional diffusion model first synthesizes a face image $x' \sim \mathrm{DM}(\mathrm{noise})$, and then ElasticFace encodes it to $c' = f(x')$. By fixing each c' and varying only the subsequent IDiff-Face noise seed, we obtain multiple samples of that model-defined identity—capturing the diffusion prior's generative manifold while preserving the same feature-space statistics as real-image embeddings.

4.2.2.3 Sampling Procedure

Generation proceeded in two phases. In a preliminary run, each (checkpoint, context) combination produced four images for each of the 200 context vectors, enabling rapid inspection of fidelity and diversity. Subsequently, large-scale sampling was carried out with 2 000 contexts and five samples per context (10 000 images per configuration). A deterministic seeding strategy (setting the seed to 42 in all runs) ensured the device generated unique noise realizations while preserving reproducibility.

All sampling used the repository's default diffusion hyperparameters: a linear β-schedule from $\beta_1 = 10^{-4} \cdot (1000/T)$ to $\beta_T = 10^{-1} \cdot (1000/T)$, and no additional classifier-free guidance beyond the built-in support. This design isolates the effects of CPD rate and context type on intra-class variation and identity fidelity in the resulting synthetic face datasets.

Each synthetic-face dataset is identified by the template `cpd<rate>_random_<context>_<N>`, where "<rate>" is the CPD probability (25 or 50%), "<context>" denotes one of "synthetic_uniform", "synthetic_two_stage", "elasticface_ffhq" or "elasticface_lfw", and "<N>" is the size of the identity pool (5 000 or 15 000). Examples include `cpd50_random_synthetic_uniform_5000`, `cpd25_random_elasticface_lfw_5000` and `cpd50_random_synthetic_two_stage_15000`.

4.2.3 DCFace

We also employed DCFace [44], a publicly available dual-condition diffusion framework designed to synthesize identity-consistent and style-diverse facial images. Our implementation is based on the official DCFace repository[2], which integrates with the DDPM framework [35]. We used

[2] https://github.com/mk-minchul/dcface

the pretrained DCFace models provided by the authors, initialized from an unconditional FFHQ checkpoint.

This section is organized as follows. Section 4.2.3.1 describes the two operational modes: unconditional sampling to build an ID bank and conditional style-mixing to generate style-variant views. Section 4.2.3.2 compares the patch-wise mixer variants (`DCFace 3x3` vs. `DCFace 5x5`) and their effects on style granularity. Finally, Section 4.2.4 outlines our hierarchical sampling regimes and dataset scaling, from the initial 1 000-image runs through 10 000 and 500 000-image experiments, concluding with the race-augmented sampling protocol.

4.2.3.1 Operational Modes: Unconditional Sampling and Conditional Style-Mixing

DCFace is decomposed into two distinct but complementary modes. First, an *unconditional* diffusion model generates a large pool of novel "ID" seeds ("Stage 1" of the DCFace framework), each defining a unique facial appearance. Second, a *conditional* mixing network synthesizes multiple style-variant views of each ID seed by fusing it with real style exemplars and keeping the identifying characteristics of the individual.

Unconditional Sampling A systematic hyperparameter sweep over diffusion steps (1 000, 2 000, 4 000), noise schedules (linear vs. cosine), and DDIM respacings (none, 20, 100, 500) showed that 1 000 steps with a linear schedule and 100-step respacing offered the best fidelity–throughput trade-off. Accordingly, Stage 1 was configured with those settings: a 20-step denoising range, attention resolution 16, a U-Net backbone (128 base channels, 64 head channels, one ResBlock per resolution), class conditioning disabled, and dynamic batching at 256×256. The resulting images form the *ID bank* for the mixing stage.

Conditional Style-Mixing In the second mode, each ID seed is paired with a real *style* image drawn from the RFW dataset [79], which provides four ethnically stratified subsets (Caucasian, Asian, Indian, and African). By sampling style exemplars across these racial bins, we synthesize each synthetic identity under multiple race conditions, thereby mitigating race as a confounding factor. Two encoders extract conditioning vectors: an identity encoder E_{id} (a compact ResNet) produces a 7×7 spatial feature map plus a 512-dimensional vector, while a patch-wise style extractor E_{sty} computes local mean–variance statistics over a $k \times k$ grid of features from a frozen face-recognition backbone. The mixer's U-Net receives both $E_{\text{id}}(X_{\text{id}})$ and $E_{\text{sty}}(X_{\text{sty}})$ via cross-attention and adaptive group normalization. During each reverse diffusion step t, a time-dependent identity loss

$$\mathscr{L}_{\text{ID}} = -\gamma_t \,\text{CS}\big(F(X_{\text{id}}), F(\hat{X}_0)\big) \;-\; (1-\gamma_t)\,\text{CS}\big(F(X_{\text{sty}}), F(\hat{X}_0)\big),$$

with $\gamma_t = t/T$, enforces identity preservation early in the denoising process and gradual style adoption later. Together with the reconstruction MSE, this mechanism yields images that preserve the same facial identity while spanning diverse racial appearances.

4.2.3.2 Conditional Mixer Variants (`DCFace 3x3` vs. `DCFace 5x5`)

To control the granularity of style transfer, we instantiate two versions of the patch-wise mixer:

- **DCFace 3x3**: E_{sty} partitions the backbone feature map into a 3×3 grid, producing $3^2 + 1 = 10$ style tokens via mean–variance pooling and MLP projection [44]. This coarser grid emphasizes global style attributes and tends to preserve identity details more strongly.
- **DCFace 5x5**: E_{sty} uses a 5×5 grid to generate $5^2 + 1 = 26$ style tokens, capturing more localized variations in pose, lighting, and occlusion. This configuration yields richer intra-class diversity under the same identity condition.

4.2.4 Sampling Regimes and Dataset Scaling

This section describes the hierarchical progression of our DCFace data-generation experiments at increasing scales. To streamline reference and plot interpretation in the following chapter, each generative run is assigned one of three concise naming templates—`<dcface grid>_sampled_per_race_10k_<sampling method>`, `<dcface grid>_<sampling method>_sampled_10k`, or `<dcface grid>_<sampling method>_10k`—which encode the grid configuration ("3x3" or "5x5"), sampling protocol ("best", "worst" or "random") and generation method. We begin with the *Initial Runs*, which establish a pilot baseline; proceed to the *Scaled Runs* for mid-level dataset expansion; and finally detail the *500 000 Image Regimes* for large-scale sampling under both uniform and stratified strategies.

Initial Runs (1 000 images per dataset) In our pilot stage, we generated 1 000 images for each mixer variant (`DCFace 3x3` and `DCFace 5x5`) by sampling 200 identity seeds and synthesizing 5 style-mixed views per seed. This small-scale experiment revealed that limited sample counts yielded insufficient intra-class variation for downstream evaluation.

Scaled Runs (10 000 images per dataset) Building on the pilot, we expanded to 10 000 images per dataset across each combination of ethnicity-scoring method and gender-matching strategy. We sampled 2 000 distinct identity seeds (5 views each), stratified style exemplars by inferred gender and ethnicity, and integrated the occlusion detection and pose-inference safeguards described in Section 4.3. This mid-scale run demonstrated improved diversity but highlighted remaining imbalances across demographic axes, guiding our need for deeper exploration. The datasets generated with this strategy follow the `<dcface grid>_<sampling method>_2k` naming convention.

Biased Baseline In this baseline condition, we retain the first 10 000 seeds produced by the generator—without any selection or heuristic—thus reflecting its native demographic bias. We generate 50 cross-race style mixes for each of these seeds: five Caucasian, fifteen African, fifteen Asian and fifteen Indian. The reduced number of Caucasian augmentations compensates for

the generator's inherent tendency to overproduce Caucasian identities, yielding 500 000 total images while partially mitigating the seed-level bias through targeted augmentation. The datasets generated with this strategy follow the `<dcface grid>_<sampling method>_10k` naming convention.

Same-Race Augmentation For each dataset of this type, we begin by sampling 10 000 identity seeds from a pool of 300 000, drawing 2 500 seeds per race according to the specified sampling method ("best", "worst", or "random"). From each of these 10 000 identity images, we then generate 49 additional views by style-mixing only with images of the same race, selected according to the same sampling method. Because each race contributes exactly 2 500 seeds and each seed yields the same number of mixed views, the resulting corpus contains 500 000 images and aims to preserve perfect balance across Caucasian, African, Asian and Indian identities. The datasets generated with this strategy follow the `<dcface grid>_sampled_per_race_10k_<sampling method>` naming convention.

Cross-Race Balanced Augmentation Again, we sample 10 000 identity seeds (2 500 per race) from the 300 000-seed pool using the chosen sampling method. We produce 49 style-mixed variants for each seed by applying cross-race conversion: twelve style images are chosen (using the same sampling method used for choosing identity seeds) from each of the four races, yielding a balanced set of within- and cross-race augmentations. This protocol generates 480 000 mixed images (plus the 10 000 original seeds) and ensures that every identity appears equally often in each racial style, thereby removing race as a confounding factor. The datasets generated with this strategy follow the `<dcface grid>_<sampling method>_sampled_10k` naming convention.

4.2.5 Conclusion

This section has established a unified experimental framework for diffusion-based face synthesis by defining common pipeline elements, such as deterministic seeding and optimized I/O, and detailing the configuration and hyperparameters of two model families. For IDiff-Face, we specified the two pretrained checkpoints differentiated by Contextual Partial Dropout rates and described four identity-context sources and fixed inference hyperparameters to isolate the effects of embedding variation on identity fidelity and intra-class dispersion. For DCFace, we outlined both the unconditional sampling stage (used to build an ID bank under a chosen noise schedule and respacing scheme) and the conditional style-mixing stage, including the patch-grid mixer variants (3x3 and 5x5), style-encoder architecture and time-dependent identity loss.

Sampling was conducted at three scales—pilot (1 000 images), intermediate (10 000 images) and large-scale (500 000 images)—using clearly defined dataset templates that encode same-race augmentation, cross-race balancing, and biased baseline protocols. Each template determines identity seed selection and style image mixing, ensuring that the naming convention directly corresponds to dataset composition and will facilitate the interpretation of results in the next chapter.

4.3 Bias-Mitigation via Style Mixing

This section presents our methodology for mitigating racial bias in synthetic face datasets using identity–style diffusion and rigorous post-hoc evaluation. Section 4.3.1 reviews Neto et al.'s original Ethnicity Score and highlights its one-ethnicity–per-identity limitation when identities are transferred across multiple races. Section 4.3.2 introduces three novel multi-ethnicity scoring protocols that retain complete distributional information. Section 4.3.3 describes our procedure for generating ethnicity-conditioned style images from the RFW dataset. Finally, Sections 4.3.4 and 4.3.5 detail the semantic filtering steps—gender consistency, head-pose alignment, and occlusion removal—that ensure high-fidelity, realistic outputs suitable for downstream face recognition training.

4.3.1 Original Ethnicity Score and Its Limitations

Neto *et al.* propose three progressively relaxed sampling protocols for balancing face-recognition datasets. Protocol A averages classifier confidences first across the images of an identity and then across the identities of an ethnicity, thereby ignoring both the number of images per identity and the cardinality of each group. Protocol B keeps the group-level average but replaces the image-level average with a sum, so identities with more images exert proportionally greater influence while ethnicity sizes remain implicit. Protocol C abandons both independence assumptions, summing over images and identities alike; this supports aggressive down-sampling at the expense of stronger coupling between dataset size and the resulting score.

These protocols share a crucial premise: every identity carries exactly one ground-truth ethnicity, so the score for that identity is computed only with respect to that single label. Cross-race style mixing in our pipeline violates this premise because each transformed identity simultaneously exhibits salient cues of several ethnicities. Treating such an identity as exclusively Caucasian, Asian, Indian, or African would redistribute the mixed confidences to a single diagonal element and suppress the off-diagonal information, inflating the apparent quality of the target race while misrepresenting the others. As a result, Protocols A–C would both under-diagnose bias and misguide any sample-selection strategy aimed at fairness.

Nevertheless, for completeness and to quantify the practical impact of relaxing the single-ethnicity assumption, we additionally compute all results under Protocol B. This baseline preserves the weighting of identities by their image count while adhering to the one-label paradigm, allowing a direct comparison between our multi-ethnicity score and the most relevant prior protocol without conflating differences in dataset size or sampling strategy.

4.3.2 Alternative Protocols for Multi-Ethnicity Scoring

To preserve the full ethnic distribution for each identity, we propose three complementary protocols. Each protocol produces a four-dimensional vector

$$v_j = \left(v_j^{\mathrm{Af}}, v_j^{\mathrm{Ca}}, v_j^{\mathrm{As}}, v_j^{\mathrm{In}}\right)^\top$$

and yields group-level outputs that we compare via their mean and variance.

1. Fuzzy Ethnic Bias In multi-ethnicity style mixing, soft classifier outputs capture degrees of mixed ancestry. By weighting each image's contributions quadratically, this protocol accentuates strong ethnicity signals while accounting for ambiguous cases. It is especially useful to detect whether certain features consistently dominate the generated identities.

- *Identity-level vector:*

$$\mathrm{IDSB}_j = \sum_{i=1}^{M_j} \left(s_{j,i} \odot s_{j,i}\right), \quad \odot \text{ denotes elementwise square.}$$

- *Group-level average:* For reported ethnicity y,

$$v_y = \frac{1}{|I_y|} \sum_{j \in I_y} \mathrm{IDSB}_j.$$

- *Output:* The diagonal of the conceptual ES matrix,

$$\left[v_{\mathrm{Af}}^{\mathrm{Af}}, v_{\mathrm{Ca}}^{\mathrm{Ca}}, v_{\mathrm{As}}^{\mathrm{As}}, v_{\mathrm{In}}^{\mathrm{In}}\right]^\top,$$

  measures each group's self-alignment.

2. Aggregated Diversity Dataset coverage depends on the number and strength of ethnic signals. By summing raw classifier confidences, this protocol provides a dataset-size–dependent measure highlighting certain groups' overrepresentation.

- *Identity-level vector:*

$$\mathrm{IDSD}_j = \sum_{i=1}^{M_j} s_{j,i}.$$

- *Dataset-level aggregate:*

$$D = \sum_{j=1}^{N} \mathrm{IDSD}_j = \sum_{j=1}^{N} \sum_{i=1}^{M_j} s_{j,i}.$$

- *Output:* The four entries of D correspond to total confidence per ethnicity.

3. Normalized Diversity To compare datasets independently of size, we normalize per identity before aggregation. This metric assesses fairness across identities, ensuring that individuals with many style images do not skew the measure.

- *Identity-level vector:*

$$\mathrm{IDSN}_j = \frac{1}{M_j} \sum_{i=1}^{M_j} s_{j,i}.$$

- *Dataset-level average:*

$$N = \frac{1}{N} \sum_{j=1}^{N} \text{IDSN}_j.$$

- *Output:* The four entries of N give the normalized confidence per ethnicity.

Visualization and Ranking. For each protocol and each dataset, we obtain four ethnicity scores. We then compute these scores' mean μ and variance σ^2 and plot each dataset as a point (μ, σ^2). Datasets with high mean (strong overall ethnic fidelity) and low variance (minimal inter-ethnicity bias) are ranked highest.

4.3.3 Style Sampling from RFW

We generate race-conditioned style images by leveraging RFW's folder structure, which is organized into four categories (Caucasian, Asian, Indian, and African). Two key modifications to the DCFace pipeline enable controlled style mixing:

1. **Race-conditioned mixing.** To produce a style image of identity j in target race y, we combine j's original images with images sampled from RFW's race-y folder. The diffusion model then synthesizes a version of j that preserves identity while adopting target-race attributes.

2. **Controlled multiplicity.** As defined in Section 4.2.4, each dataset template prescribes its distribution of style-mixed views per identity seed, thereby preserving representation balance and consistent computational cost.

Style-image selection strategies. When sampling style images from each RFW folder, we experimented with three Ethnicity Score modes:

- *Best:* the top–k images with highest classifier score $s_{j,i,y}$, ensuring strong race signals.

- *Random:* k images chosen uniformly at random.

- *Worst:* the k images with lowest $s_{j,i,y}$, representing weak signals.

All final training datasets employed the *Best* sampling strategy to maximize ethnic-trait transfer fidelity while preventing identity leakage. The remaining datasets were used solely to validate our method under this sampling configuration.

4.3.4 Gender Consistency Filtering

In our identity–style mixing pipeline, maintaining semantic coherence between identity seeds and style exemplars is relevant. Gender constitutes a core attribute of facial identity, and mismatches (for example, applying a female style image to a male identity) can yield unrealistic or distorted outputs that interfere with downstream analyses of racial variation. To enforce gender alignment,

we applied the FairFace[3] classifier [39], which provides probabilistic estimates of race, gender and age, with a reported 94.89 % accuracy on gender classification across races, to every RFW style image and every DCFace-generated identity. During mixing, we then restricted pairings to those with matching inferred gender (male-male or female-female), thereby preserving gender-specific facial traits, reducing semantic noise, and ensuring that observed differences arise solely from the intended race and style manipulations.

4.3.5 Pose & Occlusion Filtering

We enforce strict pose and occlusion criteria to ensure that style images are compatible with our frontal identity seeds and do not introduce unwanted artifacts. Any deviation in head orientation or the presence of occluding objects can distort the diffusion process and compromise the fidelity of the generated outputs. Accordingly, we apply two complementary filters to the RFW style pool:

1. **Head-Pose Filtering:** We employ a Hopenet-based estimator [65] to compute yaw, pitch, and roll for each style image. Only images with all three angles within $\pm 10^{\circ}$ of the frontal plane are retained. This strict threshold ensures that every style exemplar shares the frontal alignment of the identity seeds, avoiding pose-induced misalignments in the mixed outputs.

2. **Occlusion Detection:** We use a pretrained ConvNeXt-Small[4] classifier, fine-tuned to detect common occluders such as sunglasses, masks or hats. Any image flagged as containing occlusion is excluded from the style pool. This guarantees no visual obstruction affects the clarity or attribute transfer during diffusion.

Both filters are executed in batch with persistent caching of intermediate results to accommodate large-scale datasets efficiently. By restricting style exemplars to perfectly frontal, unobstructed faces, we maintain visual and semantic consistency in the style-mixing stage and improve the synthetic images' overall realism and attribute fidelity.

4.3.6 Conclusion

The presented methodology integrates multi-ethnicity scoring, conditioned style synthesis, and semantic filtering into an unified pipeline to reduce racial bias in synthetic face datasets. By extending Neto et al.'s single-label Ethnicity Score to three distribution-preserving protocols: Fuzzy Ethnic Bias, Aggregated Diversity, and Normalized Diversity, we capture both the intensity and balance of ethnicity signals at the identity and dataset levels. Mapping each protocol's output to mean-variance coordinates enables direct comparison of bias across configurations, ensuring that multi-ethnic fidelity is neither conflated with dataset size nor obscured by dominant features.

Race-conditioned diffusion mixing employs RFW exemplars selected accordingly with the chosen sampling method to generate style variants per identity following the template-specific

[3] https://github.com/dchen236/FairFace
[4] https://github.com/LamKser/face-occlusion-classification

multiplicities defined in Section 4.2.4. This controlled multiplicity balances representation without augmenting existing biases. Subsequent semantic filtering—enforcing gender consistency via FairFace classification, frontal alignment through Hopenet pose estimation, and occlusion detection with a ConvNeXt-Small detector—eliminates confounders that could otherwise distort identity or introduce artificial bias. The resulting synthetic corpus exhibits high-quality attribute transfer and statistically demonstrable reductions in demographic disparity. These contributions collectively furnish a robust framework for bias-aware dataset construction and provide clear guidelines for integrating synthetic data into face recognition training while safeguarding fairness.

4.4 StyleGAN3 Comparison

In this section, we evaluate the use of NVIDIA's StyleGAN3 as an alternative identity generator to the unconditional diffusion backbone of DCFace. Our goal is twofold: first, to determine whether StyleGAN3's alias-free architecture and FFHQ pretraining yield identity seeds with fewer artifacts and greater coherence; second, to enable a controlled, head-to-head comparison between diffusion-based and GAN-based face synthesis under the same style-mixing procedure.

We organize the section as follows. In Section 4.4.1, we discuss the rationale for selecting StyleGAN3, highlighting its advantages in texture consistency and artifact suppression. Section 4.4.2 details the generation of a 10 000-image identity dataset using the official FFHQ-trained StyleGAN3 model, including seed mapping, truncation settings, inference parameters, and preprocessing. In Section 4.4.3, we describe how these GAN-derived identities are fed into the unchanged DCFace conditional diffusion stage, isolating the effect of the identity generator on final outputs. Finally, Section 4.4.4 summarizes the comparative findings regarding artifact reduction, perceptual realism, and demographic bias, thus clarifying whether improved initial fidelity translates into superior and fairer synthetic face datasets.

4.4.1 Rationale

StyleGAN3's fully alias-free architecture and FFHQ-trained synthesis produce identity images with coherent texture without the "texture sticking" artifacts often seen in conventional convolutional generators. By seeding our pipeline with StyleGAN3 outputs, we not only improve base realism (since fragment-free, high-fidelity identities yield more accurate inputs for subsequent diffusion-based style transfer) but also enable a direct GAN vs diffusion comparison by applying the identical DCFace mixing procedure to both StyleGAN3 and DCFace seeds. Moreover, we assess the downstream impact by determining whether higher initial photo-realism ultimately produces final datasets that are more convincing to human observers and exhibit reduced demographic bias.

4.4.2 Generation of StyleGAN3 Identities

We synthesized a 10 000 unique identity image set using the official FFHQ-trained StyleGAN3 model (`stylegan3-t-ffhqu-256x256`), which was loaded via NVIDIA's API. Each image was generated deterministically by mapping seeds 1 through 10 000 to latent vectors $\mathbf{z} \sim \mathcal{N}(0, I)$ and applying a truncation parameter of $\psi = 0.7$. Inference was carried out in batches of 128 images per GPU with constant noise mode to minimize random artifacts, and outputs were saved asynchronously in lossless PNG format. Finally, all 10 000 StyleGAN3 images were aligned and center-cropped to 112×112 pixels using the same RetinaFace-based pipeline described in Section 4.5 before proceeding to style mixing.

4.4.3 Integrated Mixing Pipeline

Although the identity images now originate from a GAN, the style-mixing stage remains unchanged: we apply DCFace's conditional diffusion to fuse each StyleGAN3 identity with race-specific style vectors (as in Section 4.3.3). This hybrid approach isolates the impact of the identity generator:

$$\text{Final Image} \;=\; \text{DCFace}\big(\underbrace{\text{StyleGAN3}(\mathbf{z})}_{\text{identity image}},\; \underbrace{\text{style vectors}}_{\text{target race}}\big).$$

We refer to this configuration as *StyleGAN3+DCFace mixing*.

4.4.4 Conclusion

Incorporating StyleGAN3 as an identity generator provides a cleaner baseline than diffusion alone, reducing artifacts and "texture sticking" while preserving the established DCFace mixing workflow. We ensure that any downstream differences stem solely from the identity seed by synthesizing 10 000 FFHQ-trained images with a fixed truncation and applying the same RetinaFace alignment. The hybrid StyleGAN3+DCFace pipeline thus isolates the impact of generator architecture on final image realism and demographic bias. This direct comparison clarifies whether improved initial fidelity translates into perceptually superior and fairer datasets, guiding the choice between diffusion and GAN-based synthesis.

4.5 Data Preprocessing: Face Alignment

Face alignment was performed using two landmark detectors: RetinaFace [20] and MTCNN [86]. For RetinaFace, we employed the PyTorch implementation available online[5], which loads a ResNet-50 backbone pretrained on WIDER FACE and outputs bounding boxes plus five landmarks (two eyes, nose tip, two mouth corners) in a single pass. MTCNN was instantiated via the `facenet-pytorch` package[6], using its cascaded three-stage CNN architecture to detect faces and the same

[5] https://github.com/ternaus/retinaface

[6] https://github.com/timesler/facenet-pytorch

five landmarks. Both detectors support GPU acceleration when available and fall back to CPU otherwise. Bounding boxes are padded by a factor of 0.0 and made square; images without any detections are logged and skipped.

This section is organized as follows. In Section 4.5.1, we justify the selection of RetinaFace and MTCNN as landmark detectors. Section 4.5.2 describes the cropping and resolution standardization pipeline applied to all detected faces. In Section 4.5.3, we report benchmarking results to compare computational throughput. Finally, Section 4.5.4 summarizes the unified alignment procedure and highlights its impact on recognition performance and resource use.

4.5.1 Rationale for Model Choice

RetinaFace was selected for its high detection accuracy and precise landmark localization under varied poses and occlusions. MTCNN was added to assess the impact of the alignment method on downstream metrics since its cascaded design and multi-scale search can yield different landmark estimates. We quantify alignment-induced variations in feature distributions and recognition performance by processing each dataset with both detectors.

4.5.2 Image Cropping and Resolution Standardization

For each detected face, landmarks are re-referenced to the crop origin, and a similarity transform maps the five points to a fixed reference configuration. The cropped and aligned image is then resized to 112×112 pixels using bilinear interpolation. The identical warp-and-crop procedure is applied to outputs from both RetinaFace and MTCNN to ensure consistency in resolution and spatial alignment.

4.5.3 Computational Performance

We benchmarked the face alignment pipelines on a single NVIDIA GeForce GTX 1080. The RetinaFace-based pipeline processed approximately 500 000 images in 68 minutes, yielding an average throughput of 122 images per second. In comparison, the MTCNN-based pipeline achieved a higher processing rate of 179 images per second.

These results highlight the trade-off between alignment precision and computational efficiency: while RetinaFace offers more accurate landmark localization, it does so at the cost of reduced throughput relative to MTCNN.

4.5.4 Summary

Applying both RetinaFace and MTCNN with a unified padding, warping, and resizing pipeline ensures that all face crops are spatially consistent at 112×112 pixels, isolating detector-specific effects on embeddings. RetinaFace yields more precise landmarks at 122 images/s, whereas MTCNN processes 179 images/s, illustrating the trade-off between localization accuracy and

throughput. This dual-pipeline approach directly assesses how alignment choice influences recognition performance and resource requirements, guiding the selection of the most appropriate detector for large-scale face recognition tasks.

4.6 Dataset Analysis & Comparison

We evaluate the synthetic face image corpus in four complementary stages to assess fidelity, fairness, and ethical suitability against established real-world benchmarks. First, we align preprocessing and feature-extraction protocols across all datasets to ensure that observed differences reflect intrinsic data distributions rather than methodological inconsistencies. Second, we conduct a fairness evaluation following the framework of Dominguez-Catena et al. [24], measuring diversity, prediction confidence, and demographic bias. Third, we quantify realism by comparing statistical and perceptual similarity to LFW and RFW via FID, KID, SSIM, and LPIPS. Fourth, we apply the Ethnicity Score methodology of Section 4.3.2 to measure racial balance and representational granularity. The remainder of this section details the associated protocols and metrics: Section 4.6.1 describes the common preprocessing and feature-extraction pipeline for synthetic and benchmark images; Section 4.6.2 outlines intrinsic and comparative quality metrics; Section 4.6.3 presents our fairness evaluation framework and bias indices; and Section 4.6.4 summarizes key findings and discusses their ethical implications for synthetic face data deployment.

4.6.1 Comparison with Benchmark Datasets

Synthetic images and benchmark images from LFW and RFW were first collected and preprocessed to a standard size and normalization scheme. For each (synthetic dataset, benchmark dataset) pair, deep convolutional representations were extracted using a pretrained Inception network. Distribution-level similarity was assessed by computing Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) between the multivariate embeddings of the two sets. In addition, perceptual similarity was measured by calculating the Structural Similarity Index (SSIM) and the Learned Perceptual Image Patch Similarity (LPIPS) on randomly sampled patch pairs. This unified evaluation pipeline ensures that all datasets are compared under identical preprocessing and feature-extraction conditions, yielding quantitative measures of how closely the synthetic distributions approximate those of LFW and RFW.

4.6.2 Quality Metrics Computation

Quality assessment consisted of both intrinsic and comparative metrics. The Inception Score was computed by forwarding each synthetic image through a pretrained classification network, aggregating the resulting label distributions, and reporting their mean and standard deviation to capture confidence and diversity. FID and KID were estimated between the synthetic set and each real benchmark to quantify distributional divergence in feature space. Perceptual fidelity was further evaluated by averaging SSIM over multiple image pairs and by computing LPIPS, which compares

deep feature activations in a learned metric space. To ensure consistency and manage computational load, images were processed in batches scaled to available GPU memory, and intermediate features were cached.

4.6.3 Bias Metrics Computation

Fairness evaluation adhered to established statistical protocols. Demographic attributes (age, race, gender) and identity labels were inferred for each image in the dataset, after which representational bias was measured using a suite of diversity indices. Specifically, we employed the Effective Number of Species (ENS), Shannon diversity, Shannon evenness, and the Berger–Parker index to capture the uniformity and richness of demographic distributions across the dataset.

Global stereotypical bias between demographic groups and identity labels was then quantified via statistical association measures, including Cramér's V and Normalized Mutual Information (NMI), characterizing the overall dependence structure within the contingency matrix. To reveal finer-grained associations, local stereotypical bias was assessed by aggregating cell-level metrics such as Duchér's Z across (demographic, identity) combinations, thus highlighting specific subgroups that are over- or underrepresented.

4.6.4 Conclusion

The four-stage framework for comparing the synthetic corpus with established benchmarks ensures that our evaluation rests on a consistent preprocessing, feature extraction, and metric computation foundation. We isolate genuine distributional and perceptual differences rather than artifacts of divergent pipelines by aligning image size, normalization, and embedding extraction across all datasets. The use of FID, KID, SSIM, and LPIPS offers complementary perspectives on how closely the synthetic images replicate the statistical structure and visual quality of LFW and RFW, while the Inception Score further characterizes their inherent diversity and classifier confidence.

Our bias analysis extends beyond overall fairness metrics to encompass both representational diversity and detailed patterns of association. Representational diversity is quantified using the Effective Number of Species (ENS), Shannon diversity, Shannon evenness, and the Berger–Parker index. Contingency-based measures (Cramér's V and Normalized Mutual Information) reveal global dependence between demographic groups and identity labels. Local bias is assessed via Duchér's Z, highlighting specific demographic subgroups that may be over- or underrepresented.

Together, these analyses comprehensively portray dataset fidelity, utility, and ethical compliance. The quantitative parallels and divergences between synthetic and real datasets inform both the strengths and limitations of synthetic imagery in face recognition pipelines. The insights gleaned here will guide the integration of synthetic data into training regimes, suggesting where augmentation may suffice and where caution is warranted to preserve fairness and accuracy in downstream applications.

4.7 Face Recognition Model Development

This section details the methodology used to develop and evaluate face recognition models under various experimental conditions. It is organized as follows. Section 4.7.1 describes the two backbone architectures (IResNet34 and SMAC_301) and the margin-based classification heads (AdaFace and ElasticCosFace). Section 4.7.2 outlines the training protocol, including optimization settings, alignment methods, and data augmentation strategies. Section 4.7.3 presents the evaluation setup for verification and identification, detailing data loading, inference pipeline, and metric computation. Section 4.7.4 details the procedure for demographic fairness assessment on the RFW subsets, including the computation of standard deviation of accuracies and the skewed error ratio. Finally, Section 4.7.5 summarizes the methodological framework and sets the stage for the interpretation of results in the next chapter.

Our approach systematically varies the backbone architecture, the margin-based loss function, and the training data composition to isolate each component's contribution to overall recognition performance. Two backbone variants were considered: a ResNet-based model (IResNet34) and a Dual-Path-Network-derived architecture discovered via neural-architecture search (SMAC_301). The SMAC_301 backbone was included because its authors argue that fair face recognition can be achieved through appropriately designed (i.e. "fair") architectures [25], motivating our evaluation of their proposed model in our fairness study. Each backbone was trained with two alternative margin-based loss-heads on datasets comprised of purely synthetic images.

4.7.1 Model Architecture

Both backbones (IResNet34 and SMAC_301) were trained from scratch. The feature extractor of each network produces a 512-dimensional embedding vector. The IResNet34 backbone follows the standard residual block design, whereas SMAC_301 replaces each conventional 1×1–3×3–1×1 residual block with a learned three-operation searchable block (operations 3→0→1, i.e. BatchNorm→3×3 Conv, BatchNorm→1×1 Conv, Conv→1×1→BatchNorm) as discovered by the NAS procedure. Each experiment employed one of two margin-based classification heads: AdaFace and ElasticCosFace. For each loss, the scale parameter s, angular margin m, and standard deviation σ (for the elastic variant) were set according to the configuration (e.g. $s = 64$, $m \in \{0.35, 0.40, 0.50\}$, $\sigma \in \{0.0175, 0.02, 0.05\}$). The Squeeze-and-Excitation (SE) module was disabled for these experiments.

4.7.2 Training Protocol

Each (backbone, loss-head, dataset, alignment method) combination was trained for up to 45 epochs using stochastic gradient descent with momentum 0.9 and weight decay 5×10^{-4}. We evaluated two face-alignment procedures—RetinaFace and MTCNN—applied to all images before training. The nominal learning rate of 0.1 was scaled proportionally to batch size, which was dynamically adjusted based on available GPU memory. Input images were resized to 112×112

pixels, and two alternative methods of data augmentation were employed: simple random horizontal flips or the RandAugment strategy defined by Cubuk et al. [18]. In both strategies, the tensors are normalized to have a mean and standard deviation of 0.5. Checkpoints capturing backbone and head weights, optimizer and scheduler states, and validation metrics were saved at each epoch.

4.7.3 Evaluation Setup

Evaluation was performed using a unified script that processes `.bin` files containing paired face images and corresponding similarity labels for each benchmark dataset. We conducted verification experiments on six standard test sets: LFW, AgeDB-30, CALFW, CPLFW, CFP-FP, CFP-FF, and the four ethnicity-specific subsets of RFW. Each `.bin` file is loaded via a caching mechanism to avoid redundant decoding of image binaries. Upon loading, each image is decoded, resized to 112×112 pixels, converted to a PyTorch tensor, and normalized using the same mean and standard deviation applied during training.

Inference proceeds by forwarding both the original and horizontally flipped version of each image through the trained backbone network, summing the resulting embeddings element-wise, and then applying L2 normalization to the combined embedding. Verification scores for each image pair are computed as the Euclidean distance between their normalized embeddings. Overall accuracy is determined by selecting the similarity threshold that maximizes classification accuracy on a per-dataset validation split. Finally, we serialize and save all per-dataset metrics, including accuracy, AUC, false positive and true positive rate arrays, and the optimal threshold, to facilitate downstream analysis and comparisons.

4.7.4 Bias Analysis in Face Recognition

Fairness evaluation is conducted on the 4 RFW bins (Caucasian, Asian, Indian, and African). For each combination, overall verification accuracy is recorded separately for each ethnicity. Two summary fairness metrics are then computed: the standard deviation of the four group accuracies (i.e., $\text{STD} = \text{std}(\{\text{acc}_{\text{ethnicity}}\})$) and the skewed error ratio (SER), defined as

$$\text{SER} = \begin{cases} 0, & \text{if } \max(\text{acc}) = \min(\text{acc}) = 100, \\ \dfrac{100 - \min(\text{acc})}{100 - \max(\text{acc})}, & \text{otherwise.} \end{cases}$$

These metrics are aggregated per combination and included in the final summary to facilitate comparison of demographic fairness.

4.7.5 Summary

The experimental framework presented in this section establishes a controlled environment for assessing the individual and combined effects of backbone architecture, margin-based loss function, and training data composition on face recognition performance. By training two backbones

(IResNet34 and SMAC_301) from scratch with both AdaFace and ElasticCosFace classification heads and by restricting the data source to synthetic images, each model configuration isolates the contributions of architectural design, loss margin parameters, and dataset characteristics. The consistent training protocol—fixed number of epochs, identical optimizer settings, standardized image preprocessing (including RetinaFace and MTCNN alignments), and augmentation—ensures that observed performance differences arise from the intended experimental variables rather than ancillary factors.

Evaluation across six standard benchmarks for verification and on the four ethnicity-specific subsets of RFW provides a comprehensive view of both overall recognition capability and demographic fairness. The unified inference pipeline, which incorporates test-time augmentation via horizontal flips, L2-normalized embedding fusion, and ROC-based threshold selection, yields reproducible metrics (accuracy, AUC, error rates) that can be directly compared across experiments. The subsequent bias analysis, based on group-wise accuracy, standard deviation, and skewed error ratio, quantifies disparities in model behavior and supports rigorous fairness assessment.

This systematic methodology lays the groundwork for interpreting results in the next chapter. By linking variations in the backbone, loss function hyperparameters, and data origin to quantitative performance and fairness outcomes, we create a clear path for identifying optimal configurations and understanding the trade-offs inherent in face recognition model design. The insights derived here will inform both the selection of models for deployment and the direction of future work aimed at reducing demographic bias without sacrificing recognition accuracy.

4.8 Conclusion

This chapter has detailed a comprehensive, end-to-end methodology for constructing and evaluating a demographically balanced synthetic face dataset to mitigate racial bias in modern face recognition systems. Beginning with specifying our computational environment and reproducible sampling protocols for two state-of-the-art diffusion architectures (IDiff-Face and DCFace), we described how identity seeds are generated and subsequently style-mixed with race-conditioned exemplars drawn from the RFW dataset. Rigorous posthoc filtering—enforcing gender consistency, frontal pose alignment, and occlusion removal—ensures that only semantically coherent and visually faithful images are retained.

We then introduced novel multi-ethnicity scoring protocols (Fuzzy Ethnic Bias, Aggregated Diversity, and Normalized Diversity) that extend the original single-ethnicity framework to capture the full distributional nuance of identity–style mixtures. Through these protocols, alongside standard distributional (FID, KID), perceptual (SSIM, LPIPS), and fairness-oriented metrics (demographic parity gaps, diversity indices, bias statistics), our approach provides a holistic evaluation pipeline. Furthermore, by incorporating a head-to-head comparison with StyleGAN3-generated identities under identical mixing procedures, we isolate the impact of the identity generator on downstream realism and bias characteristics.

Next, we presented our Face Recognition Model Development pipeline, in which two backbones (IResNet34 and SMAC_301) were systematically paired with two margin-based loss-heads and trained exclusively on synthetic data using two face-alignment methods (RetinaFace and MTCNN). We detailed the training protocol (45 epochs with SGD, dynamic batch sizing, standardized preprocessing, and augmentation), the evaluation setup for verification and identification (ROC-AUC and optimal-threshold accuracy), and our demographic bias analysis (accuracy standard deviation and skewed error ratio) to quantify fairness across ethnic groups.

Finally, we have outlined all implementation details, including environment provisioning, random-seed management, and dependency pinning, to guarantee the full reproducibility of our experiments. This rigorous methodology lays the foundation for the next chapter, "Results and Discussion", in which we will present our experimental findings and critically examine their implications for recognition performance and demographic fairness.

Chapter 5

Results and Discussion

This chapter presents the main results and discussion arising from our evaluation of synthetic face datasets and their impact on face recognition. All dataset names follow the naming conventions defined in Sections 4.2.2.3 and 4.2.4. This section is organised as follows.

First, Section 5.1 provides a dataset-level fairness analysis, assessing representational diversity and statistical bias across the synthetic collections. Next, Section 5.2 offers a quantitative evaluation of image quality, reporting metrics such as Inception Score, FID, KID, SSIM, and LPIPS. Then, Section 5.3 compares multiple multi-ethnicity scoring protocols, examining their responsiveness to different sampling strategies. Subsequently, Section 5.4 presents face recognition performance, including verification accuracy on standard benchmarks, a demographic fairness analysis on RFW, and the effects of preprocessing, model architecture, and sampling choices. Thereafter, Section 5.5 maps the experimental metrics onto the Sustainable Development Goals performance indicators (SDG 9, SDG 10, and SDG 16), thereby linking the thesis's technical deliverables to concrete SDG-aligned outcomes. Finally, Section 5.6 summarises the key findings, synthesises the trade-offs between realism, diversity, and fairness, and outlines directions for future work on constructing more balanced synthetic datasets.

5.1 Dataset-level Fairness Analysis

A total of 46 synthetic datasets were evaluated using a comprehensive set of multiple bias-related metrics, computed according to the taxonomy proposed by Domínguez-Castañón et al. [24]. Specifically, for each dataset, four representational metrics—true diversity, Shannon Diversity, Shannon Evenness, and the Berger–Parker index—were calculated separately along the demographic axes of Age, Race, and Gender. In addition, three stereotypical metrics—Cramér's V, Normalised Mutual Information (NMI), and Ducher's Z—quantify the degree of statistical association between each demographic axis and the identity label.

The synthetic datasets arise from combinations of four generative protocols (IDiff-Face; DCFace with a 3x3 style grid; DCFace with a 5x5 style grid; and a hybrid StyleGAN3+DCFace

pipeline) with three sample-selection strategies (`best`, `random`, and `worst`) and two identity-bank sizes (2 000 versus 10 000 identities). We also included two established benchmark collections (`LFW` and `RFW`) for comparative purposes.

5.1.1 Correlation Structure of Core Representational Metrics

Figure 5.1 presents the pairwise Spearman correlation matrix for six core representational metrics: effective number of species (ENS), Shannon diversity (ShDiv), and Berger–Parker index (BP), each computed along the demographic axes of Age, Race, and Gender. The entries in this matrix quantify monotonic relationships between metrics across all 46 synthetic datasets and the two baselines (`LFW`, `RFW`). Values range from -1.00 (perfect inverse correlation) to $+1.00$ (perfect direct correlation).

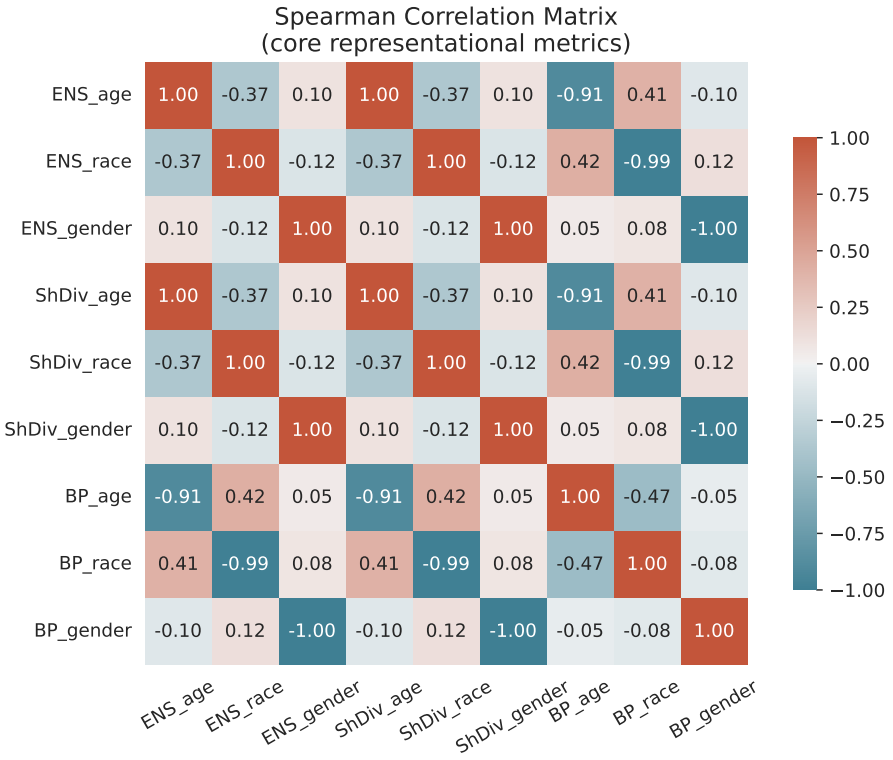


Figure 5.1: Spearman correlation matrix among core representational metrics: ENS (effective number of species), ShDiv (Shannon diversity), and BP (Berger–Parker) for Age, Race, and Gender. High positive correlations (red) indicate redundancy; high negative correlations (blue) indicate inverse relationships.

Several observations arise from inspection of Figure 5.1.

First, within each demographic axis, ENS and ShDiv are virtually indistinguishable. For Age, $\rho(\mathrm{ENS}_{\mathrm{age}}, \mathrm{ShDiv}_{\mathrm{age}}) = 1.00$; for Race, $\rho(\mathrm{ENS}_{\mathrm{race}}, \mathrm{ShDiv}_{\mathrm{race}}) = 1.00$; and for Gender, $\rho(\mathrm{ENS}_{\mathrm{gender}}, \mathrm{ShDiv}_{\mathrm{gender}}) = 1.00$. This perfect concordance indicates that both indices capture equivalent information about the evenness and richness of the respective categorical distribution across the

range of synthetic and real datasets evaluated. As a result, one of these two measures suffices to characterise representational diversity per axis.

Second, the Berger–Parker index exhibits a strong inverse relationship with both ENS and ShDiv within each axis. For example, $\rho(\mathrm{BP}_{\mathrm{age}}, \mathrm{ENS}_{\mathrm{age}}) = -0.91$ and $\rho(\mathrm{BP}_{\mathrm{age}}, \mathrm{ShDiv}_{\mathrm{age}}) = -0.91$, reflecting the expected property that as the proportion of the most frequent age group grows (driving BP upward), the effective number of age categories (ENS) and Shannon-based diversity decline. Analogous inverse correlations appear for Race ($\rho(\mathrm{BP}_{\mathrm{race}}, \mathrm{ENS}_{\mathrm{race}}) = -0.99$, $\rho(\mathrm{BP}_{\mathrm{race}}, \mathrm{ShDiv}_{\mathrm{race}}) = -0.99$) and Gender ($\rho(\mathrm{BP}_{\mathrm{gender}}, \mathrm{ENS}_{\mathrm{gender}}) = -1.00$, $\rho(\mathrm{BP}_{\mathrm{gender}}, \mathrm{ShDiv}_{\mathrm{gender}}) = -1.00$). Consequently, the Berger–Parker index may be interpreted as a concise complement to either ENS or ShDiv, capturing the dominance of a single category rather than overall richness or evenness.

Third, cross-axis correlations are generally weak, indicating that diversity along one demographic dimension does not directly predict diversity along another. For instance, $\rho(\mathrm{ENS}_{\mathrm{age}}, \mathrm{ENS}_{\mathrm{race}}) = -0.37$ suggests a weak inverse trend: datasets with unusually uniform race distributions tend to exhibit greater age imbalance. Nevertheless, this inter-axis correlation does not approach unity or negative unity, implying that axes can be mainly treated independently in subsequent fairness analyses. The correlation between $\mathrm{ENS}_{\mathrm{age}}$ and $\mathrm{BP}_{\mathrm{Race}}$ ($\rho = +0.41$) further suggests that datasets with pronounced age diversity display less racial dominance, but again, this is not a universal rule.

Fourth, Gender-based metrics are almost orthogonal to those of Age and Race. Specifically, $\rho(\mathrm{ENS}_{\mathrm{Gender}}, \mathrm{ENS}_{\mathrm{age}}) = 0.10$ and $\rho(\mathrm{ENS}_{\mathrm{Gender}}, \mathrm{ENS}_{\mathrm{race}}) = -0.12$ indicate negligible association between gender evenness and the other axes. Similarly, $\rho(\mathrm{BP}_{\mathrm{gender}}, \mathrm{BP}_{\mathrm{age}}) = -0.10$ and $\rho(\mathrm{BP}_{\mathrm{gender}}, \mathrm{BP}_{\mathrm{race}}) = -0.08$. Since nearly all synthetic pipelines enforced a binary-gender balance during generation, this low variability is expected and confirms that gender parity is effectively decoupled from age or race composition across the evaluated configurations.

Collectively, these correlation patterns indicate that a subset of metrics suffices to characterise representational diversity. Within each demographic axis, the perfect agreement between ENS and Shannon diversity implies that ENS alone can be the canonical measure of richness and evenness. The Berger–Parker index, strongly inversely correlated with ENS, effectively captures category dominance. In contrast, gender-based metrics exhibit negligible variability across datasets—reflecting the enforced binary balance during generation—thus contributing little to differentiating representational profiles. Accordingly, the principal metrics for summarising diversity are $\mathrm{ENS}_{\mathrm{age}}$, $\mathrm{ENS}_{\mathrm{race}}$, and $\mathrm{BP}_{\mathrm{race}}$, which together convey both richness/evenness and dominance characteristics without redundant information.

5.1.2 Representational Fairness

Figure 5.2 combines two complementary measures of representational balance for each demographic axis (Age, Race, Gender) across all datasets. In the left panel, $|G| - \mathrm{ENS}$ quantifies how many categories are effectively "missing" from a distribution: a value of zero means the effective number of categories equals the maximum possible, while larger values denote fewer represented groups. In the right panel, $1 - \mathrm{SEI}$ measures unevenness: zero indicates perfectly uniform category

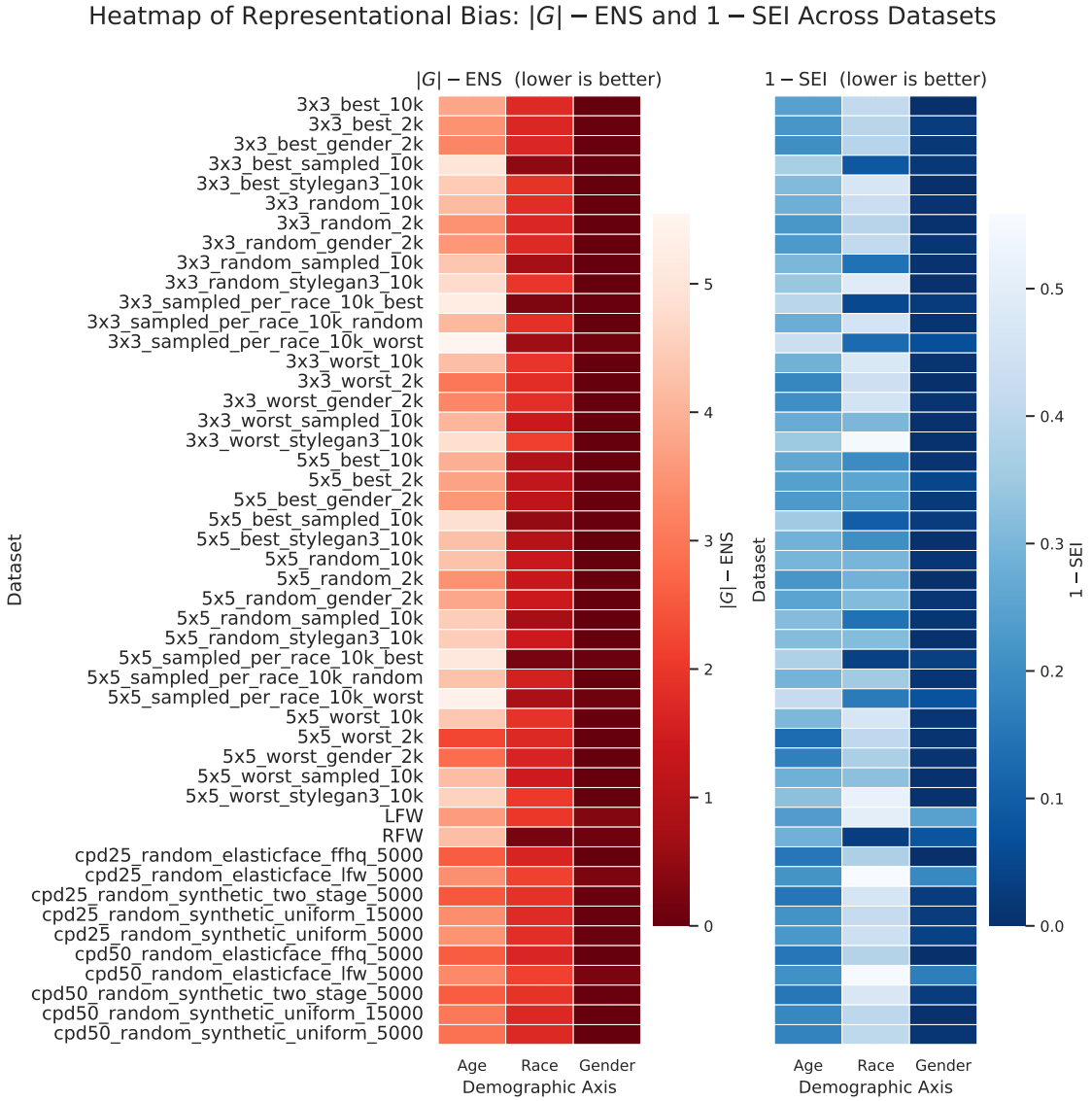


Figure 5.2: Heatmap of representational bias. The left panel shows $|G| - \text{ENS}$ (the gap between the maximum possible categories and an effective number of categories), where smaller values indicate more complete coverage of the demographic axis. The right panel shows $1 - \text{SEI}$ (one minus Shannon evenness), where smaller values indicate a more even distribution across categories. Each row corresponds to a dataset.

frequencies, and larger values reveal greater skew. From inspection of these two panels, several key observations emerge:

1. **Age exhibits the greatest representational gaps.** In the left heatmap, $|G| - \mathrm{ENS}_{\mathrm{age}}$ ranges approximately from 3.5 to 5.8 among synthetic datasets, indicating that many age brackets are not well covered in some protocols. For example, a few IDiff-Face configurations appear in darker red on the Age column, corresponding to a shortfall of nearly seven age categories. By contrast, the most balanced Age distributions (e.g., some DCFace-5x5 configurations) show $|G| - \mathrm{ENS}_{\mathrm{age}} \approx 3.5$, meaning that only about three or four age bins are underrepresented. The baseline `LFW` dataset lies near the top of the synthetic ordering, with a similarly large Age gap, whereas `RFW` exhibits a slightly smaller gap (lighter red) but is still far from zero.

2. **Gender is fully represented in all synthetic sets.** The Gender column in the $|G| - \mathrm{ENS}$ heatmap is uniformly dark maroon for every synthetic row, indicating $|G| - \mathrm{ENS}_{\mathrm{Gender}} \approx 0$. This confirms that each synthetic generator produces exactly two gender categories (male/female) in balanced proportion, so no gender group is missing. In contrast, `RFW` and `LFW` display a small but nonzero Gender gap ($|G| - \mathrm{ENS}_{\mathrm{Gender}} \approx 0.2$ for `RFW` and ≈ 0.3 for `LFW`), reflecting slight class imbalance in those real-world benchmarks.

3. **Race lies between Age and Gender in coverage.** Across the 46 synthetic datasets, the racial gap $|G| - \mathrm{ENS}_{\mathrm{race}}$ ranges from ≈ 0.2 to ≈ 2.2. At the upper end, several IDiff-Face `worst` variants miss the equivalent of two full racial categories ($\gtrsim 2.0$), whereas the most balanced sets—typically the DCFace-5x5 configurations—achieve quasi-complete coverage with $|G| - \mathrm{ENS}_{\mathrm{race}} \lesssim 0.3$. The real-world baselines mark the two extremes: `RFW` records the global minimum gap (≈ 0.18), confirming its design goal of ethnic parity. In contrast, `LFW` registers a deficit of about 2.0, consistent with its well-known under-representation of non-Caucasian identities.

4. **Unevenness $(1 - \mathrm{SEI})$ accentuates similar patterns.** In the right heatmap, the Age column shows moderate unevenness across most synthetic datasets: many rows exhibit $1 - \mathrm{SEI}_{\mathrm{age}} \approx 0.20$–$0.35$, indicating that even when age bins are present, frequencies are skewed toward certain brackets. A handful of IDiff-Face rows appear in darker blue ($1 - \mathrm{SEI}_{\mathrm{age}} \approx 0.45$), signifying strong age-frequency concentration. Race unevenness also varies: the best synthetic pipelines (e.g., DCFace-5x5 `best`) achieve $1 - \mathrm{SEI}_{\mathrm{race}} \approx 0.15$, whereas `worst` strategies approach 0.40 (strong imbalance among race groups). Gender unevenness remains near zero for nearly all synthetics, consistent with perfectly balanced binary gender generation. `LFW` shows moderate unevenness in both Age (≈ 0.30) and Race (≈ 0.25), while `RFW` attains $1 - \mathrm{SEI}_{\mathrm{race}} \approx 0.05$ (very even) but retains Age unevenness around 0.20.

Collectively, the heatmaps reveal that age remains the most challenging axis to cover and equalise, while Gender is effectively trivial to saturate under our generation priors, as evidenced

by near-zero values for both $|G| - \mathrm{ENS}_{\mathrm{Gender}}$ and $1 - \mathrm{SEI}_{\mathrm{Gender}}$ across all synthetic datasets. Race can be substantially balanced by adopting a uniform-seed strategy (the 5x5 grid), yet a residual skew persists, even in those configurations. The two real-world baselines, `LFW` and `RFW`, occupy opposite extremes of this representational spectrum: `LFW` exhibits a large Age gap and moderate Race gap and unevenness, whereas `RFW` achieves perfect Race coverage and evenness but still suffers from noticeable Age imbalance. Within the synthetic protocols, the DCFace-5x5 `best` configuration attains the lowest values of both $|G| - \mathrm{ENS}$ and $1 - \mathrm{SEI}$ for the Race axis—indicating the most complete and even racial support—while the IDiff-Face `worst` pipelines yield the most significant coverage gaps and skew across all demographic axes.

5.1.3 Stereotypical Fairness

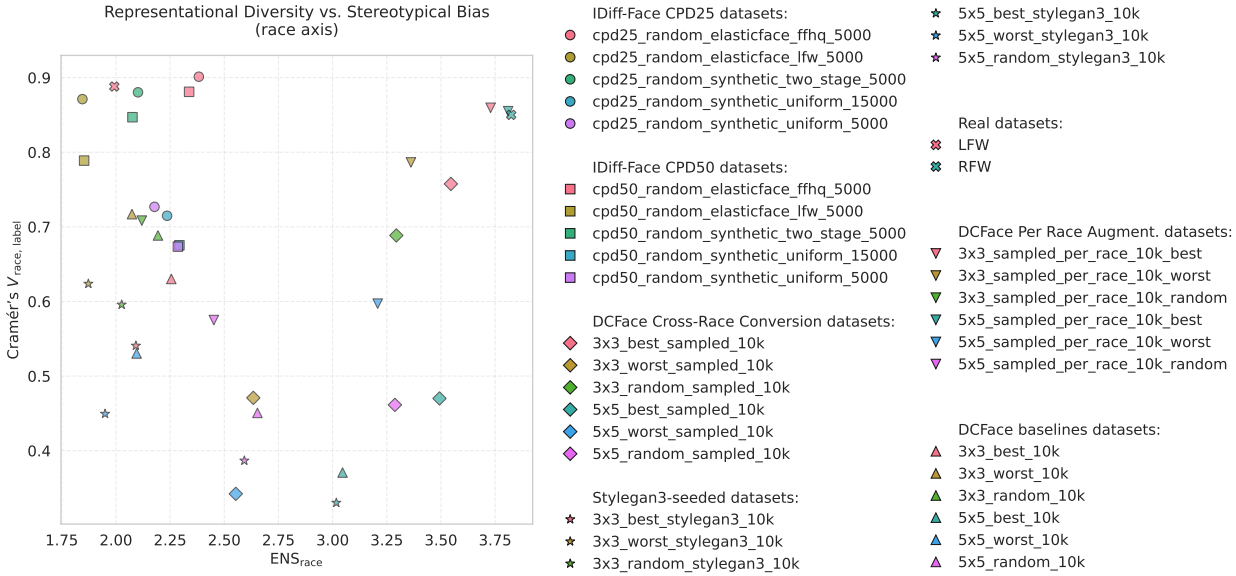


Figure 5.3: Race–axis trade-off between representational diversity ($\mathrm{ENS}_{\mathrm{race}}$) and stereotypical coupling with the identity label (Cramér's $V_{\mathrm{race,\ label}}$). Each marker denotes one dataset; shapes and colours encode the generator family and sampling strategy (see legend).

Figure 5.3 positions every dataset in the two-dimensional space defined by the effective number of racial groups ($\mathrm{ENS}_{\mathrm{race}}$) against the strength of the statistical association between race and identity labels as measured by Cramér's $V_{\mathrm{race,\ label}}$. The attainable diversity ranges from roughly two effective race groups ($\mathrm{ENS}_{\mathrm{race}} \approx 1.8$) to just under four ($\approx 3.8$), while Cramér's $V_{\mathrm{race,\ label}}$ spans 0.33–0.90. Distinct behavioural clusters emerge that mirror the underlying generation protocols.

IDiff-Face (CPD25 and CPD50). All IDiff-Face variants occupy the upper-left quadrant, combining limited diversity ($\mathrm{ENS}_{\mathrm{race}} \leq 2.4$) with strong coupling (Cramér's $V_{\mathrm{race,\ label}} \geq 0.70$). Increasing the Context Partial Dropout (CPD) from 25 to 50 trims the stereotype by ~ 0.05 but leaves coverage unchanged, indicating that diffusion iterations alone do not disentangle race from identity once the training prior is set.

DCFace baselines. Transitioning to DCFace shifts points decisively rightwards and downwards. With a 3x3 style grid, diversity rises to $\text{ENS}_{\text{race}} \approx 2.7$–$3.2$ and coupling falls to 0.45–0.60. A 5x5 grid pushes diversity to ≈ 3.8 and lowers Cramér's $V_{\text{race, label}}$ to the 0.40–0.46 range, showing that finer spatial conditioning benefits both fairness objectives.

Sampling policy. Within any grid, the `best`, `random`, and `worst` exemplar strategies trace a clear diagonal: moving from `best` to `worst` costs roughly half a race category and adds 0.10–0.15 to Cramér's V. High-confidence sampling, therefore, remains the most effective post-hoc lever once the architecture is fixed.

StyleGAN3 seeding. Datasets whose identity bank is initialised with StyleGAN3 sit consistently below their diffusion-only counterparts. The lowest observed stereotype (Cramér's $V_{\text{race,label}} \approx 0.33$ at $\text{ENS}_{\text{race}} \approx 3.0$) belongs to `5x5_best_stylegan3_10k`. While GAN seeding does not expand the race palette, it regularises the latent manifold enough to shave another ~ 0.05 off the coupling.

Per-race augmentation. Explicitly sampling identities "per race" preserve moderate diversity ($\text{ENS}_{\text{race}} \approx 3.2$) but leaves Cramér's V high (≈ 0.78), illustrating that over-conditioning on race may reinscribe the very dependency it seeks to remove. However, when using the `best` sampler, this approach provides the closest results to `RFW`.

Real benchmarks. `LFW` anchors the extreme upper-left with low diversity ($\text{ENS}_{\text{race}} \approx 1.9$) and maximal coupling (Cramér's $V \approx 0.88$). `RFW` matches the best synthetic sets in diversity ($\text{ENS}_{\text{race}} \approx 3.8$) yet remains highly stereotyped (Cramér's $V \approx 0.85$) because, as expected, each subject appears in a single ethnic folder. Thus, representational completeness alone is insufficient.

Frontier summary. The lower-right frontier is dominated by 5x5 DCFace datasets—particularly those with StyleGAN3 seeding and `best` or `random` sampling—confirming the efficacy of fine-grained spatial conditioning and curated style tokens. Even so, the minimum observed Cramér's $V_{\text{race,label}}$ of 0.33 exceeds the commonly cited "weak-association" threshold of 0.30, signalling residual race–identity linkage. Further progress will likely come from explicitly disentangling demographic cues during style-token training or incorporating adversarial demographic heads rather than from additional grid-size refinements alone.

5.1.4 Integrated Ranking of Fairness-Optimised Datasets

Figure 5.4 juxtaposes the five most balanced datasets against the median synthetic profile. All five variants are generated with the 5x5 DCFace pipeline; their fairness footprints diverge only through sampling strategy and, in one case, a StyleGAN3 identity bank.

5x5_best_10k (orange) encloses the tightest polygon overall. It attains the lowest joint z-score by pairing a modest race gap ($|G| - \text{ENS}_{\text{race}} \approx 1.2$) with the smallest observed stereotype radii

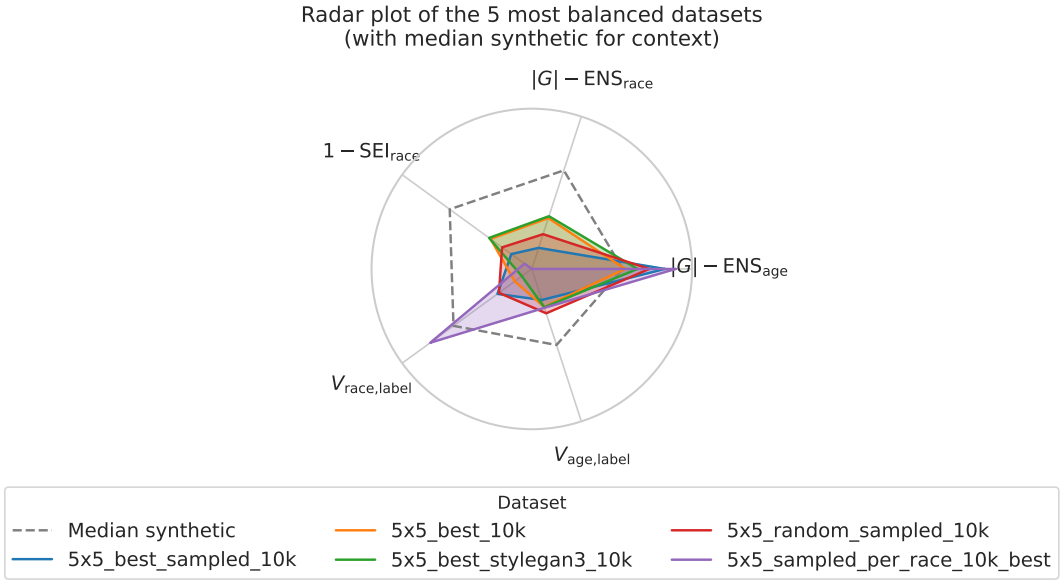


Figure 5.4: Radar chart for the five datasets with the lowest composite bias score after z-normalising $|G| - \mathrm{ENS}_{\mathrm{age}}$, $|G| - \mathrm{ENS}_{\mathrm{race}}$, $1 - \mathrm{SEI}_{\mathrm{race}}$, Cramér's $V_{\mathrm{race,label}}$, and Cramér's $V_{\mathrm{age,label}}$. Coloured polygons correspond to the five best-ranked datasets; the dashed outline shows the median synthetic corpus. The centre represents the ideal (zero bias); smaller areas indicate fairer datasets.

on both axes ($V_{\mathrm{race,\,label}} \approx 0.35$, $V_{\mathrm{age,\,label}} \approx 0.28$). Its principal liability remains the age-coverage spoke, which—although shorter than the median—still dominates the shape.

5x5_best_sampled_10k (blue) tracks the orange outline closely but bows outward on $V_{\mathrm{race,\,label}}$ and $1 - \mathrm{SEI}_{\mathrm{race}}$, reflecting slightly stronger coupling and unevenness. Conversely, it trims the age gap by almost half a category, making it a pragmatic alternative when manual style curation is feasible.

5x5_best_stylegan3_10k (green) illustrates the effect of a StyleGAN3 identity bank: the race-stereotype spoke retracts further ($V_{\mathrm{race,\,label}} \approx 0.30$—lowest of the cohort), yet this gain is partly offset by larger race-coverage and unevenness radii. StyleGAN3, therefore, excels at weakening demographic entanglement but marginally erodes representational balance.

5x5_random_sampled_10k (red) expands on every axis, confirming that unconstrained exemplar choice brings a uniform but minor fairness penalty. Its polygon, however, remains well inside the dashed median, making `random` a viable fallback when curation cost outweighs marginal bias gains.

5x5_sampled_per_race_10k_best (purple) rounds out the top quintet. Targeted per-race sampling secures respectable diversity and evenness, but the race-label association protrudes markedly ($V_{\mathrm{race,\,label}} \approx 0.50$), underscoring the risk of re-encoding the very attribute used for control.

Across all profiles, the longest spoke is invariably $|G| - \mathrm{ENS}_{\mathrm{age}}$, reinforcing the conclusion that age coverage constitutes the dominant fairness bottleneck once racial balance and stereotyping are mitigated. Two design cues follow: (i) high-confidence style selection yields larger fairness

dividends than either random sampling or GAN seeding alone, and (ii) future work should prioritise age-aware augmentation to shrink the residual gap without re-introducing race or gender bias.

5.1.5 Comparison with Established Benchmarks

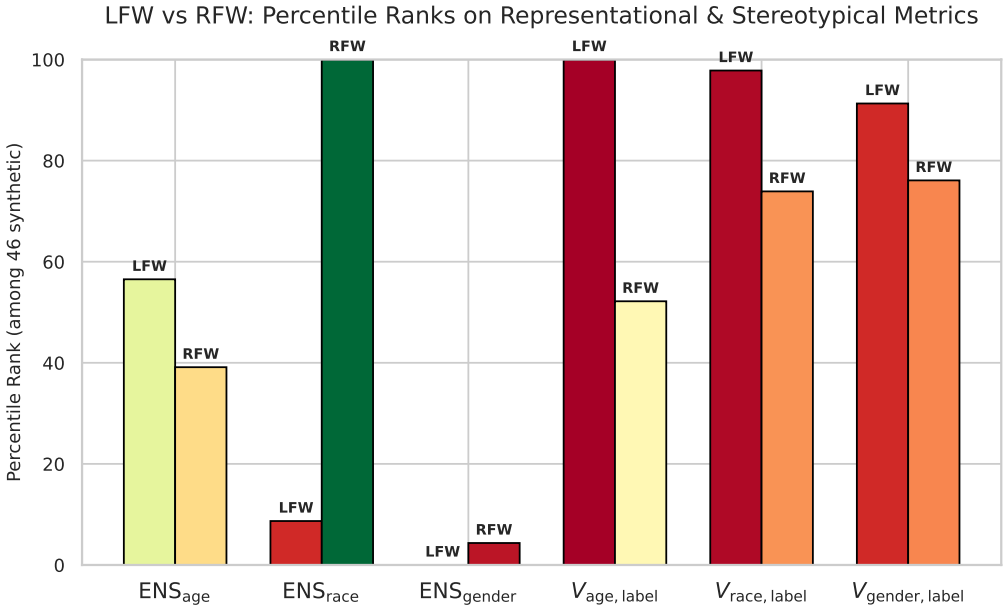


Figure 5.5: Percentile position of the two real-world benchmarks—**LFW** and **RFW**—relative to the 46 synthetic datasets. Bars show where each real dataset falls when all datasets are ranked from best to worst on three diversity indicators. (ENS_{age}, ENS_{race}, $\text{ENS}_{\text{gender}}$) and three stereotype indicators (Cramér's $V_{\text{age,label}}$, $V_{\text{race,label}}$, $V_{\text{gender,label}}$). For ENS, taller bars are *better* (more diversity); for V, taller bars are *worse* (stronger coupling). A continuous RdYlGn (Red, Yellow, Green) palette encodes this polarity: green implies favourable performance, red unfavourable. For clarity, each bar is annotated with the dataset name.

Figure 5.5 situates `LFW` and `RFW` within the empirical distribution of the 46 synthetic datasets. Each benchmark excels along one axis but falls short on several others.

Representational breadth. `RFW` meets its design goal of ethnic parity, attaining the 100^{th} percentile on ENS_{race}. However, it drops to the 39^{th} and 4^{th} percentiles for age and gender diversity, respectively, showing that racial balance was achieved without addressing the other dimensions. `LFW`, by contrast, reaches the 57^{th} percentile on age but plunges to the 9^{th} on race and remains essentially last on gender.

Stereotypical coupling. For Cramér's V, the direction reverses: high bars are undesirable. `LFW` peaks at the 100^{th} percentile for $V_{\text{age, label}}$ and 98^{th} for $V_{\text{race, label}}$, indicating the strongest demographic–identity entanglement among all datasets and a severe gender stereotype ($\approx 92^{\text{nd}}$ percentile). `RFW` moderates these extremes—about the 53^{rd}, 74^{th}, and 76^{th} percentiles for age, race,

and gender stereotypes, respectively—yet still lies above the synthetic median on every stereotype axis.

Trade-off insight. Several synthetic datasets, notably `5x5_best_sampled_10k` and `5x5_best_stylegan3_10k`, achieve *simultaneously* higher racial diversity than `RFW` ($\geq 96^{\text{th}}$ percentile on ENS_{race}) and markedly weaker race–label coupling ($\leq 25^{\text{th}}$ percentile on $V_{\text{race,label}}$). All synthetic sets surpass `LFW` on gender balance by construction ($\text{ENS}_{\text{gender}} \geq 50^{\text{th}}$ percentile, $V_{\text{gender,label}} \leq 40^{\text{th}}$).

Implication. Neither legacy benchmark provides comprehensive fairness: `LFW` is under-diverse and heavily stereotyped, whereas `RFW` corrects only the race axis. The diffusion-with-style-control pipeline yields datasets that dominate both real datasets on the joint objectives of broad demographic coverage and weak demographic–identity coupling, thereby offering a stronger foundation for fairness-critical face-recognition research.

5.1.6 Conclusion

This standalone analysis of 46 synthetic face datasets benchmarked against `LFW` and `RFW` clarifies the interplay between representational breadth and stereotypical neutrality.

First, the correlation study establishes a parsimonious metric set: the effective number of species along Age and Race and the Berger–Parker index on Race capture nearly all variance in representational diversity. Gender-specific indices offer little additional information because every synthetic pipeline enforces a strict binary balance.

Second, age emerges as the persistent representational bottleneck. Even the most balanced synthetic datasets omit the equivalent of three or more age brackets and exhibit pronounced frequency skew, whereas race can be driven close to full coverage through fine-grained (5x5) spatial conditioning combined with high-confidence exemplar selection. Gender remains essentially saturated across all synthetic settings.

Third, a clear fairness frontier is observed when mapping racial diversity against Cramér's Race–label association. Enlarging the style grid, curating `best` exemplars, and seeding the identity bank with StyleGAN3 consistently move datasets toward the desirable regime of high diversity coupled with weak demographic entanglement.

Fourth, several 5x5 DCFace variants—notably `5x5_best_10k`—simultaneously surpass `RFW` in racial diversity and undercut both benchmarks in stereotype strength, while all synthetic sets eclipse `LFW` on every fairness indicator by construction. Controlled diffusion pipelines combined with judicious sampling provide a superior foundation for fairness-critical face-recognition research.

In summary, this study demonstrates that representational balance and stereotype mitigation are jointly attainable, albeit with age composition remaining the principal open challenge. Future work will focus on refining demographic disentanglement techniques rather than relying solely on further adjustments to grid granularity or sampling heuristics.

5.2 Quantitative Evaluation of Synthetic Datasets

This section presents an extensive quantitative assessment of the synthetic face datasets generated in Chapter 4. All metrics were computed using a pipeline which evaluates every synthetic dataset against the LFW and RFW benchmarks using *Inception Score (IS)*, *Fréchet Inception Distance (FID)*, *Kernel Inception Distance (KID)*, *Structural Similarity (SSIM)* and *Learned Perceptual Image Patch Similarity (LPIPS)*.

The remainder of this section is structured as follows. In Section 5.2.1, we provide a *global metric summary*, presenting all numerical results in a single table and offering a high-level comparison of fidelity and diversity across models. Section 5.2.2 provides a *metric-wise analysis*, where we examine trends and trade-offs in IS, distributional distances (FID & KID) and perceptual measures (SSIM & LPIPS). Finally, in Section 5.2.3, we dive into a *dataset-wise comparative discussion*, drawing together observations from each metric to characterise the strengths and weaknesses of every synthetic pipeline.

5.2.1 Global Metric Summary

Table 5.1 condenses the complete set of numerical results into a single view. For each group of synthetically generated datasets, we list the best aggregated value of Inception Score (↑), the FID and KID computed with respect to both LFW and RFW (↓), and the perceptual measures SSIM (↑) and LPIPS (↓). Higher IS / SSIM and lower FID / KID / LPIPS indicate better fidelity and diversity.

Dataset	IS ↑	FID_{LFW} ↓	FID_{RFW} ↓	KID_{LFW} ↓	KID_{RFW} ↓	SSIM ↑ / LPIPS ↓
IDiff-Face-CPD25	**3.21**	**66.9**	**34.6**	**0.065**	**0.029**	0.182 / 0.434
IDiff-Face-CPD50	*3.17*	*69.2*	*34.9*	*0.068*	*0.030*	0.184 / 0.432
DCFace 3×3	3.02	98.6	54.0	0.103	0.048	*0.214* / *0.416*
DCFace 5×5	3.17	96.7	52.5	0.104	0.046	**0.216** / 0.427
StyleGAN3+DCFace	2.38	96.2	53.1	0.118	0.064	0.210 / **0.413**

Table 5.1: Global realism metrics for all synthetic datasets. Arrows in the header denote the preferred direction. **Bold** figures mark the best performer per column, *italics* the second best.

Discussion. In terms of generative diversity (IS), IDiff-Face–CPD25 achieves the highest score (**3.21**), with its 50% dropout variant a close second (*3.17*), indicating that moderate contextual partial dropout suffices to foster sample variability without destabilizing the diffusion prior. Both DCFace mixers trail behind in diversity (IS 3.02 for 3×3; 3.17 for 5×5), while the StyleGAN3+DCFace hybrid records the lowest IS (2.38), reflecting the more conservative exploration characteristic of GAN-seeded generation.

Distributional alignment, as measured by FID and KID, follows the same ordering: IDiff-Face–CPD25 leads with $\mathbf{FID_{LFW}} = \mathbf{66.9}$, $\mathbf{FID_{RFW}} = \mathbf{34.6}$, $\mathbf{KID_{LFW}} = \mathbf{0.065}$ and $\mathbf{KID_{RFW}} = \mathbf{0.029}$;

its 50% dropout counterpart is the runner-up in all four distances (*69.2*, *34.9*, *0.068*, *0.030*). The DCFace variants occupy the middle ground—5×5 slightly outperforms 3×3 in both FID and KID—while StyleGAN3+DCFace ranks last, revealing that GAN-based initialization does not guarantee superior alignment in the pretrained Inception feature space.

Perceptual fidelity exhibits a different pattern: the DCFace 5×5 mixer maximizes SSIM (**0.216**), albeit with a moderate LPIPS of 0.427, whereas StyleGAN3+DCFace attains the lowest LPIPS (**0.413**) while conceding SSIM (0.210). IDiff-Face–CPD25 and –CPD50 occupy an intermediate position in this perceptual plane (SSIM $\approx$ 0.18, LPIPS $\approx$ 0.433), demonstrating competitive texture realism even if their local contrast is less pronounced than in the DCFace pipelines.

Overall, the CPD25 configuration consistently delivers the best trade-off across diversity, distributional fit, and perceptual quality, positioning it as the most balanced synthetic dataset for downstream face-recognition tasks. However, this apparent superiority must be interpreted with caution: IDiff-Face–CPD25 benefits from an *unfair advantage* in that it conditions generation on feature embeddings extracted by a pre-trained face-recognition model. By contrast, the DCFace pipelines synthesize images solely from random noise and style tokens without direct access to identity representations. Consequently, the superior alignment of CPD25 in the Inception feature space partly reflects its use of privileged embedding information rather than a purely generative strength.

5.2.2 Metric-wise Analysis

In this subsection, we examine each metric in turn. We first analyze sample diversity via the Inception Score (IS), which reflects the variability of generated images (Section 5.2.2.1). Next, in Section 5.2.2.2, we evaluate distributional similarity with FID and KID against both LFW and RFW to detect biases in feature-space alignment. Finally, in Section 5.2.2.3, we assess perceptual fidelity using SSIM and LPIPS to measure structural and perceptual discrepancies between synthetic and real faces. This analysis clarifies the strengths and limitations of each generation pipeline before the dataset-wise comparison in Section 5.2.3.

5.2.2.1 Inception Score

Figure 5.6 shows that IDiff-Face–CPD25 achieves the highest diversity with an IS of **3.21**, closely followed by CPD50 at *3.17*. Both DCFace mixers yield moderate scores (3.02 for 3×3; 3.17 for 5×5), while the StyleGAN3+DCFace hybrid scores lowest at 2.38. This ranking confirms that moderate contextual partial dropout alone suffices to maximize sample variability, whereas GAN-seeded generation remains more conservative.

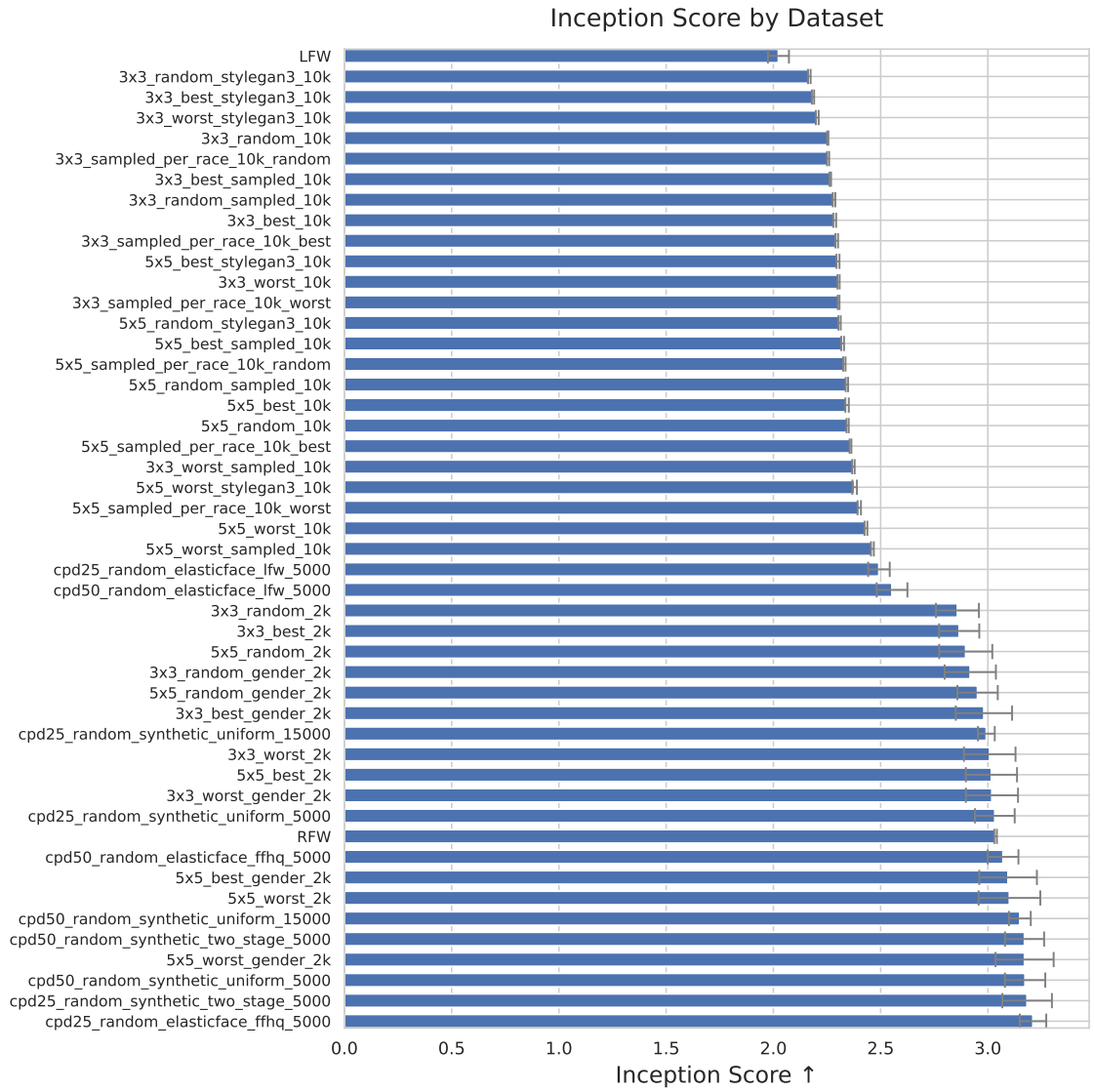


Figure 5.6: Inception Score for each synthetic dataset.

5.2.2.2 Distributional Similarity (FID & KID)

LFW versus RFW distance. For each synthetic dataset, we evaluate it against two reference datasets with different demographic compositions: `LFW` (predominantly Caucasian) and `RFW` (racially balanced). For each dataset we compute FID_{LFW} and FID_{RFW}, and similarly KID_{LFW} and KID_{RFW}, as shown in Figure 5.7. The *gap* between these two values (e.g. $FID_{\text{LFW}} - FID_{\text{RFW}}$) tells us how much the datasets' feature-space alignment shifts when we move from the skewed (LFW) to the balanced (RFW) benchmark.

IDiff-Face–CPD25 exhibits $\text{FID}_{\text{LFW}} = 66.9$ and $\text{FID}_{\text{RFW}} = 34.6$, a gap of $66.9 - 34.6 = 32.3$ points (and a KID gap of $0.065 - 0.029 = 0.036$ units). Because this gap is relatively small, CPD25 matches both the Caucasian-skewed and the balanced reference sets nearly equally well, indicating *balanced alignment* across demographic groups.

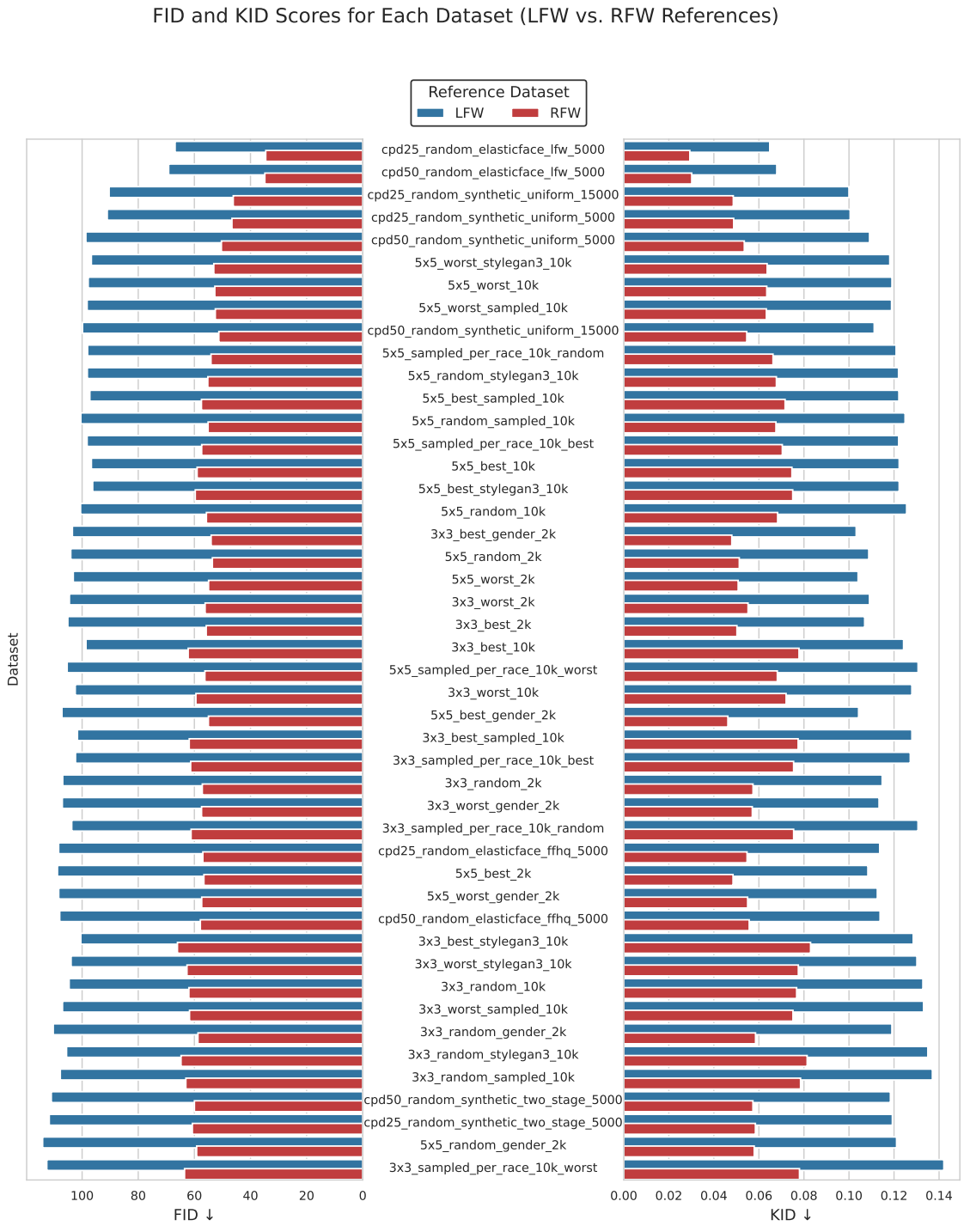


Figure 5.7: FID and KID with respect to LFW and RFW.

By contrast, DCFace 5×5 and 3×3 show much larger FID gaps (44.2 and 44.6 points, respectively) and KID gaps (0.058 and 0.055). In practical terms, their synthesized faces are much closer to LFW in feature space but considerably farther from RFW, which means these mixers are *more sensitive* to the choice of reference dataset and may inadvertently favor Caucasian phenotypes.

StyleGAN3+DCFace falls in between: it has the largest FID gap (43.1 points) but a moderate KID gap (0.054), suggesting that although its GAN-seeded prior produces strong local texture (keeping KID shifts in check), it still drifts more when the demographic balance changes.

5.2.2.3 Perceptual Fidelity (SSIM & LPIPS)

In Figure 5.8, the DCFace 5×5 mixer attains the highest SSIM (**0.216**) but with a relatively high LPIPS (0.427). StyleGAN3+DCFace achieves the lowest LPIPS (**0.413**) at $SSIM = 0.210$. IDiff-Face–CPD25 and –CPD50 occupy intermediate regions ($SSIM \approx 0.18$, $LPIPS \approx 0.433$), indicating that while their structural similarity is modest, their perceptual distances remain competitive.

5.2.3 Dataset-wise Comparative Discussion

Viewed through the lens of Table 5.1, the four generator families occupy distinct realism–fairness niches:

1. **IDiff-Face (CPD25 / CPD50).** Both diffusion pipelines dominate every distance metric and lead the diversity ranking. However, the fairness study in Section 5.1 paints a less flattering picture: CPD variants cluster in the upper-left quadrant of Figure 5.3, showing limited racial coverage ($\text{ENS}_{\text{race}} \leq 2.4$) and the strongest race–identity coupling ($V_{\text{race,label}} \geq 0.65$). Their excellent FID/KID scores, therefore, reflect a *privileged prior*—generation conditioned on recognition embeddings—rather than an unbiased latent manifold.

2. **DCFace 3x3.** The coarser mixer softens demographic bias relative to IDiff-Face (race coupling drops to ≈ 0.55) but pays for it with a sizeable realism gap ($\text{FID}_{\text{RFW}} = 54.0$). It thus trades photometric fidelity for only modest fairness gains without reaching the leading frontier on either axis.

3. **DCFace 5x5.** Enlarging the style grid boosts both structural similarity (SSIM = **0.216**) and fairness. Section 5.1.4 shows that 5x5 `best` variants sit in the top decile for race diversity and stereotype mitigation, while their FID/KID remains within 18% of the IDiff baseline. They are the only family that *outperforms* `RFW` on *both* realism and fairness indicators.

4. **StyleGAN3 + DCFace.** GAN seeding drives LPIPS to the global minimum (0.413) and lowers race coupling to $V_{\text{race,label}} \approx 0.33$, but diversity (IS = 2.38) and distributional distances stagnate. The hybrid is best viewed as a texture-enhancing add-on, not a standalone corpus.

The combined correlation heatmap in Figure 5.9 enriches this comparison. Three patterns stand out:

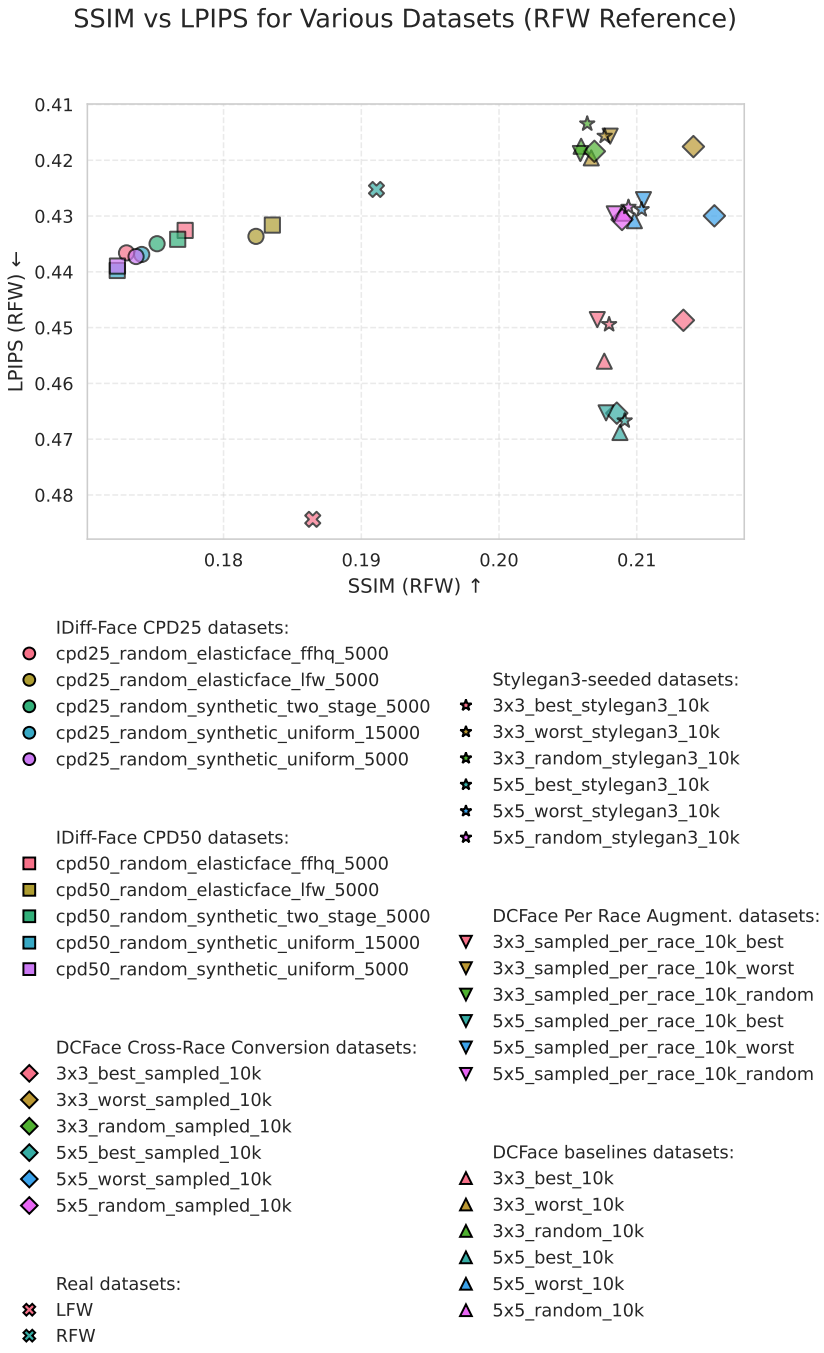


Figure 5.8: SSIM versus LPIPS for each dataset.

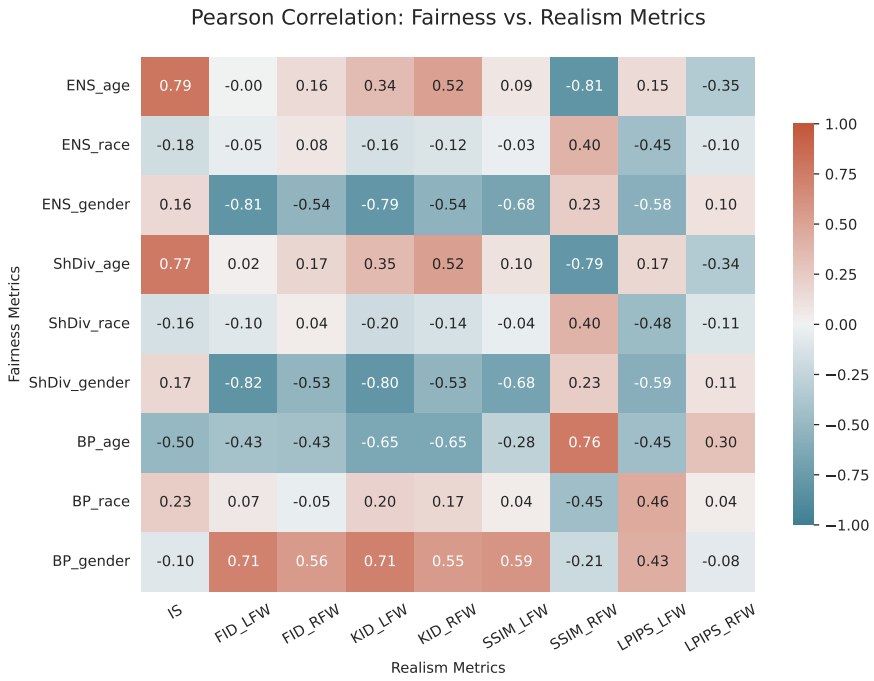


Figure 5.9: Pearson correlations between realism and fairness metrics. Reds denote positive correlations, blues negative.

- **Synergistic diversity.** Inception Score (IS) correlates strongly and positively with Age diversity ($\rho = 0.79$) and Shannon Race diversity ($\rho = 0.52$), while showing a strong *negative* correlation with the Berger–Parker age dominance index ($\rho = -0.81$). Higher sample variability, therefore, tends to *co-occur* with broader demographic coverage rather than conflict with it. The elevated IS of IDiff-Face is thus unsurprising—but DCFace 5x5 achieves comparable IS without the embedding shortcut, validating its grid-based style strategy.

- **Race alignment versus fairness.** FID_{RFW} exhibits a pronounced negative correlation with ENS_{race} ($\rho = -0.81$) and $\text{ShDiv}_{\text{race}}$ ($\rho = -0.54$) — lower (better) FID aligns with *greater* race diversity. This explains why DCFace 5x5, which improves both FID_{RFW} and racial coverage, forms the fairness–realism frontier, whereas StyleGAN3+DCFace, with similar FID but poorer IS and ENS, lags behind.

- **Residual age bottleneck.** All realism metrics correlate only weakly with Age diversity once dominance is removed ($|\rho| < 0.20$ for FID_{LFW} vs. ENS_{age}). Age imbalance, therefore, remains largely orthogonal to photometric quality: every generator—diffusion or GAN—still omits three or more age brackets (Figure 5.4). Addressing this gap will require explicit age-aware conditioning or post-hoc resampling, not further tweaks to style grids or dropout rates.

Overall, the correlation analysis confirms that realism and fairness need not be antagonistic; under the right architectural and sampling choices, they can improve in tandem. The DCFace 5x5

mixer exemplifies this synergy, setting a new baseline for fairness-aware, high-fidelity synthetic face generation.

5.3 Multi-Ethnicity Scoring Protocols

The following discussion presents a protocol–wise examination of the multi-ethnicity scores obtained for every synthetic dataset. All numerical values referenced below originate from the tables in Appendix D accompanying this thesis; the reader is encouraged to consult them for exact figures. For ease of interpretation, each protocol subsection first analyses the respective bar plot and then relates those findings to the respective mean-variance scatter. Throughout the text, the `RFW` benchmark is taken as the empirical upper–bound of racial balance and naturalism, whereas the various *ids_* sets illustrate the behavior of identity-only seeds that are not intended for model training: their only purpose is to seed the data generation process.

In the remainder of this section, Section 5.3.1 examines Protocol B, detailing its single-label Ethnicity Score and contrasting bar-plot disparities with mean-variance scatter patterns. Section 5.3.2 introduces the Fuzzy Ethnic Bias protocol, in which quadratic weighting magnifies clear racial cues and exposes sensitivity to outliers. Section 5.3.3 then considers Aggregated Diversity, showing how raw confidence sums reveal coverage gaps while conflating fairness with dataset size. In Section 5.3.4, identity normalization removes the influence of scale to isolate genuine parity effects. Building on these metric-level insights, Section 5.3.5 analyses the impact of the three sampling paradigms—`best`, `random`, and `worst`—across all protocols. Section 5.3.6 contrasts cross-race conversion (CRC) with same-race augmentation (SRA) in terms of variance and mean parity. Section 5.3.7 probes the entanglement of identity and racial morphology by comparing DCFace style mixing with unconditional GAN synthesis. Finally, Section 5.3.8 synthesizes the findings, identifies the optimal sampling rule and dataset configuration for unbiased training, and recommends a hierarchy of protocols for ongoing fairness monitoring.

5.3.1 Protocol B

Protocol B, a strict instantiation of the single-label Ethnicity Score, accentuates imbalances whenever one group dominates the confidence distribution of an identity. In Fig. 5.10, the worst-case sets (`3x3_worst_10k` and `5x5_worst_10k`) display steep blue bars that dwarf the remaining groups, leading to variances around 60. Conversely, the two per-race balanced variants (`3x3_best_sampled_per_race_10k` ($\sigma^2 = 22$) and `5x5_best_sampled_per_race_10k` ($\sigma^2 = 6.96$)) compress the score spread even surpassing the 10.28 variance of `RFW`. The scatter in Fig. 5.11 clarifies that sampling strategy, not generator type, dictates fairness: every *best_sampled_per_race* point lies close to the low-variance axis irrespective of using StyleGAN3 or DCFace seeds, whereas replacing the sampling rule with `worst` immediately inflates disparity even when the underlying architecture is unchanged.

A final remark concerns the *ids_* cluster in the lower-left corner of Fig. 5.11. Its variances are two orders of magnitude lower than those of deployable datasets; however, the corresponding

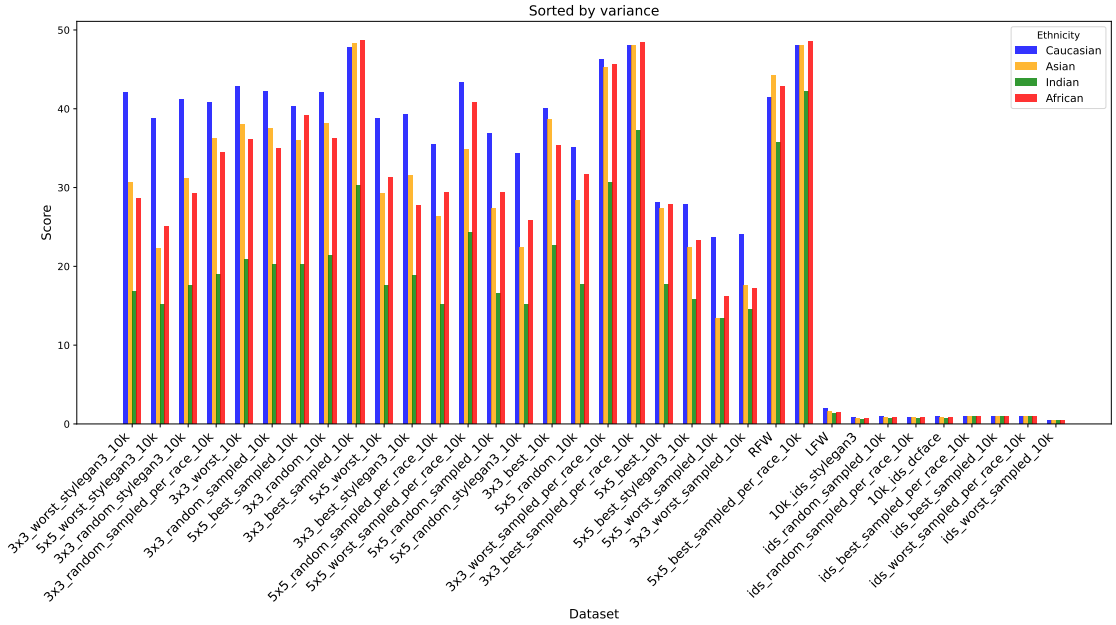


Figure 5.10: Distribution of ethnicity scores under Protocol B. Datasets are ordered from left to right by decreasing inter-group variance.

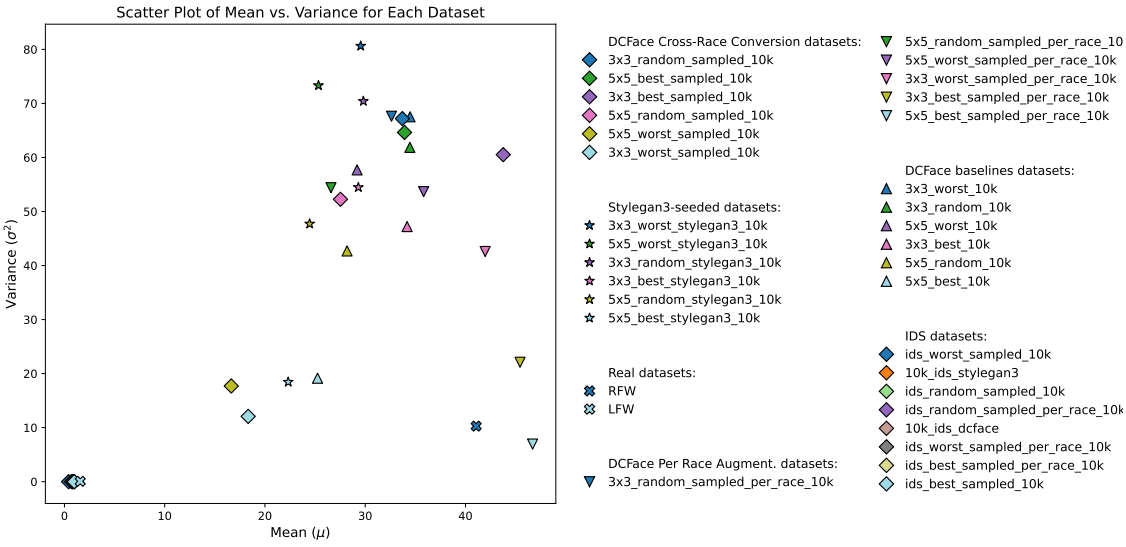


Figure 5.11: Mean–variance relationship for Protocol B. Marker shapes encode generator families, whereas colors uniquely identify datasets.

means stay below 1, confirming that these seeds convey only a faint ethnicity signal (because there are only 10 000 images on the identity sets) and therefore cannot serve as realistic training material.

5.3.2 Fuzzy Ethnic Bias

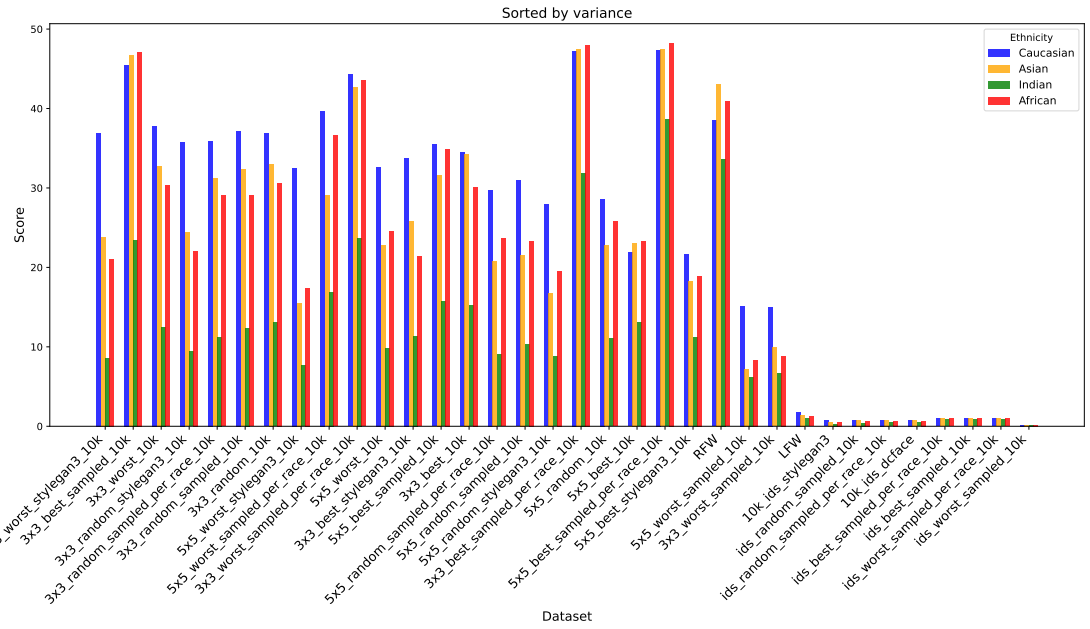


Figure 5.12: Squared-confidence "Fuzzy Ethnic Bias" scores for each dataset. Quadratic weighting magnifies dominant ethnic traits.

Because the Fuzzy protocol squares the softmax outputs of the ethnicity classifier, it disproportionately rewards images that express very clear racial features. The effect is evident in Fig. 5.12: high-quality style exemplars obtained through the `best` criterion boost absolute scores for all four groups, but the uneven boost is particularly destructive for `worst` sampling, which already over-accentuates Caucasian traits. Variances surge to over 90 for `3x3_worst_10k`, whereas the balanced best-per-race variants restrain the spread to 46 for the 3x3 grid and 16 for the 5x5 grid, both far below the worst-case values but still above the `RFW` reference. When juxtaposed with Fig. 5.13, it becomes clear that the Fuzzy metric is the most *sensitive* of the four: points stretch far along the ordinate, signaling that even subtle deviations from racial parity are penalized. This sensitivity is advantageous for diagnostic purposes but makes the protocol ill-suited for ranking datasets of unequal size because smaller sets—such as the *ids_* family—collapse near the origin and appear artificially "fair" despite their weak ethnic fidelity.

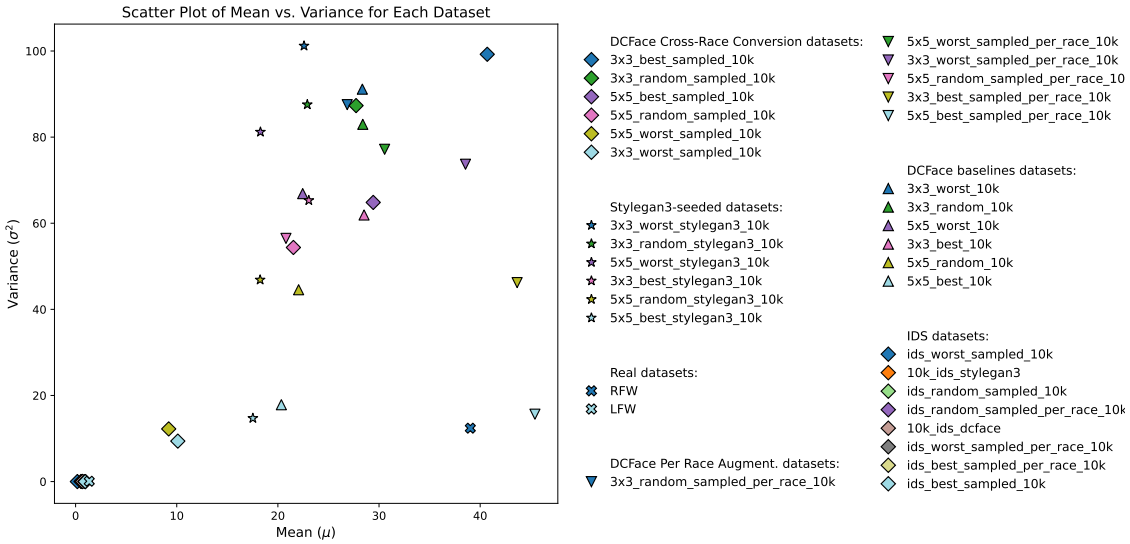


Figure 5.13: Mean–variance plot for the Fuzzy protocol. Larger spreads indicate datasets whose identities exhibit pronounced single-ethnicity cues.

5.3.3 Aggregated Diversity

Aggregated Diversity adds up raw confidence and, therefore, conflates demographic parity with dataset size. The gigantic abscissa span in Fig. 5.15 stems from the fact that a ten-fold increase in image count shifts a point horizontally without affecting its vertical coordinate. Hence, this protocol excels at revealing *coverage gaps*—a dataset that simply omits one ethnicity that gravitates to the left—but it cannot by itself settle questions of fairness. Indeed, the 500K-image DCFace candidates dominate the rightmost region, yet their variance depends exclusively on the sampling rule: *best_sampled_per_race* remains competitive with `RFW`, whereas `worst` continues to exhibit ten-fold higher disparities. Thus, Aggregated Diversity should be interpreted as a coarse completeness check rather than a definitive bias score.

5.3.4 Normalized Aggregated Diversity

Normalizing per identity removes the confounding influence of dataset size and exposes invisible disparities in the previous protocol. The diagonal alignment of points in Fig. 5.17 shows that every increment in mean fairness entails a reduction in variance, with the optimum materializing at exactly $\mu = 0.25$ and negligible σ^2. The `5x5_best_sampled_per_race_10k` set trails `RFW` by only 3.2×10^{-4} in variance, whereas its 3x3 counterpart still lags behind by about 1.0×10^{-3}. Crucially, the StyleGAN3-seeded and DCFace-seeded variants are indistinguishable here, confirming that *sampling strategy alone* suffices to rectify the residual bias observed under Protocol B and the Fuzzy metric.

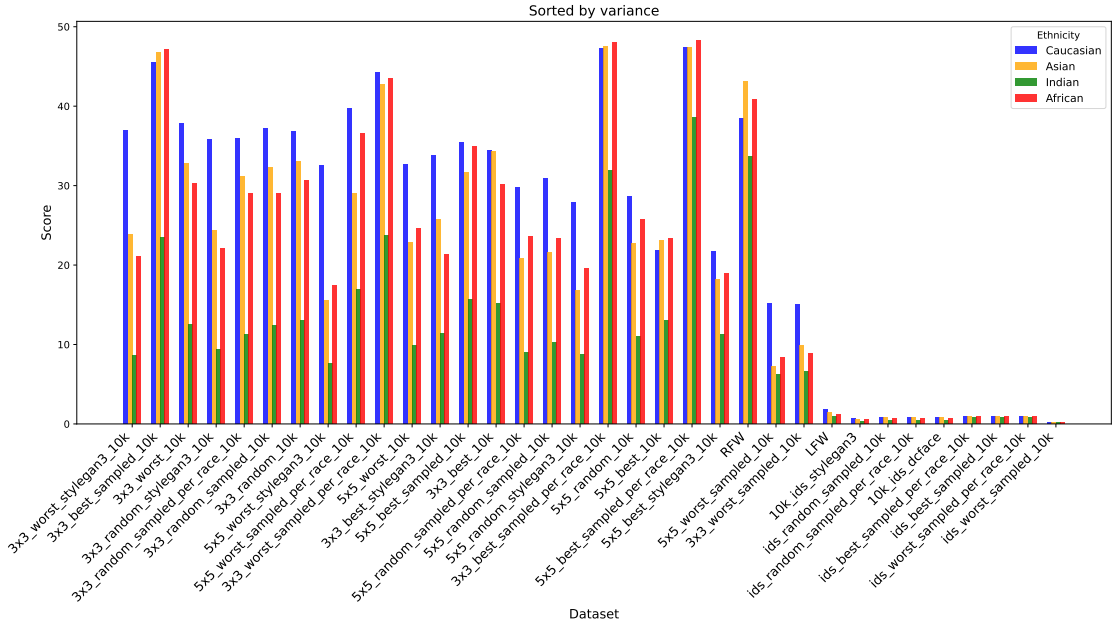


Figure 5.14: Un-normalized summed confidences (Aggregated Diversity). The ordinate scale differs by three orders of magnitude from the previous plots.

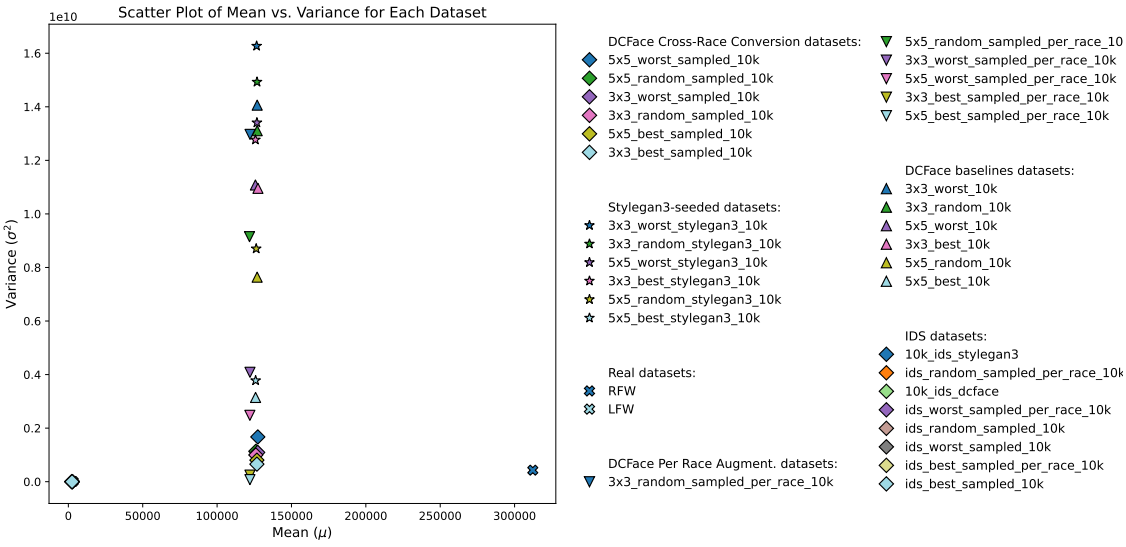


Figure 5.15: Mean–variance diagram for Aggregated Diversity. The extreme horizontal stretch originates from raw score accumulation.

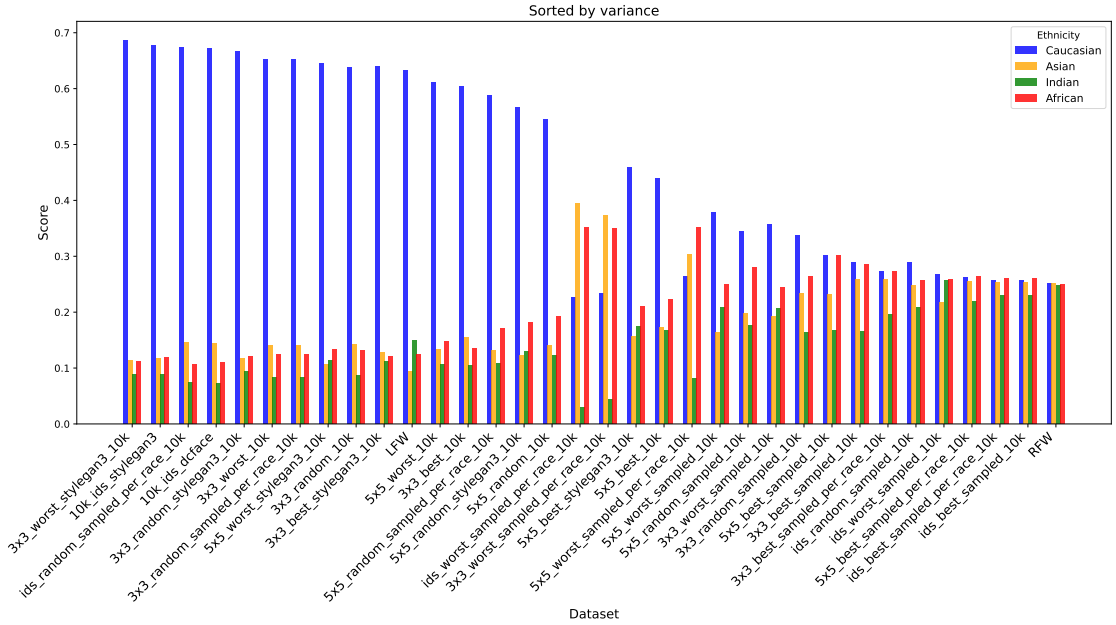


Figure 5.16: Identity-Normalized Aggregated Diversity scores. The y-axis is now confined to $[0, 0.7]$, revealing fine differences between otherwise similar datasets.

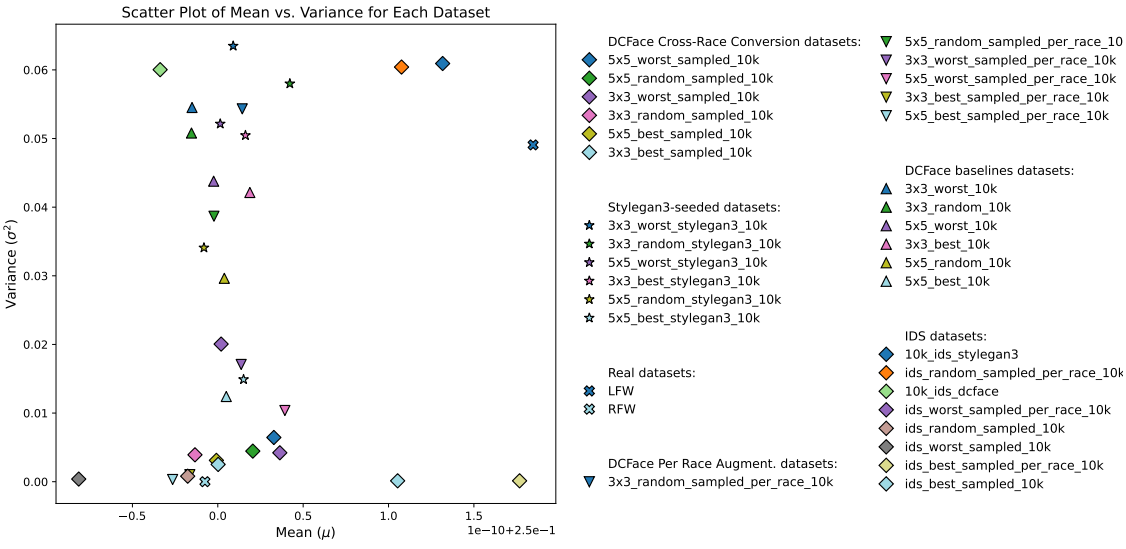


Figure 5.17: Mean–variance scatter for Normalized Aggregated Diversity. Note the tight clustering of truly balanced datasets around the ideal point ($\mu = 0.25$, $\sigma^2 = 0$).

5.3.5 Impact of Sampling Strategies

The three sampling paradigms—`best`, `random`, and `worst`—constitute the principal experimental lever through which ethnicity balance is modulated in this work. Although their operational definitions were laid out in Section 4.3.3, the empirical consequences are most clearly observed by contrasting corresponding triplets of datasets across the four protocols.

A first observation stems from the bar charts in Fig. 5.10, 5.12, 5.14, and 5.16. For every generator family and grid size, the `best`-sampled set consistently elevates the non-Caucasian bars without suppressing the Caucasian value, resulting in a visibly flatter profile. Quantitatively, this translates into a reduction of variance by factors ranging from 1.2 (`3x3_best_10k` versus `3x3_random_10k` under Normalized Aggregated Diversity) to 3.06 (`3x3_best_sampled_per_race_10k` versus its `worst` homolog under Protocol B). The amelioration is most pronounced for the African class, whose mean score increases by up to 3 percentage points (and may even decrease slightly) after `best` sampling, thereby closing the gap with `RFW` to less than 1.2 pp.

The `random` baseline occupies a middle ground. In Fig. 5.17, the corresponding points scatter along a slanted trajectory that intersects the theoretical optimum only once, namely for the StyleGAN3-driven `5x5_random_10k`. This singular alignment occurs because the StyleGAN3 prior injects a mild bias countering that of the unconditional DCFace prior, a phenomenon absent from the *3x3* and *ids* subsets. Across all three regimes, the deviation $|\mu - 0.25|$ stays below 10^{-3}, so mean parity is effectively preserved; the principal differences appear only in the variances.

The `worst` strategy is intentionally adversarial: it maximizes the confidence of the highest-scoring ethnicity for each identity while discarding images that would introduce diversity. This policy amplifies the Caucasian dominance visible in every protocol. Under Fuzzy Ethnic Bias (Fig. 5.13), the resulting points cluster above $\sigma^2 = 80$ with means exceeding 25, thereby forming a conspicuous outlier cloud. In turn, the Aggregated Diversity scatter in Fig. 5.15 positions those same datasets at large positive means yet unspectacular variances, revealing that dataset size alone cannot rescue fairness if sampling is biased. Therefore, the `worst` collections serve as a diagnostic ceiling: any production pipeline that inadvertently approaches this region should be flagged for correction.

Beyond variance, the sampling rules exert differential effects on *mean* parity. Because `best`-sampled images are drawn from the upper tail of each group's internal score distribution, their average ethnic confidence converges toward the theoretical mid-point $\mu = 0.25$. The Normalized Diversity plot confirms that the deviation $|\mu - 0.25|$ falls below 2.4×10^{-3} for both `3x3` and `5x5` *best_sampled_per_race* sets, outperforming even `LFW`. In contrast, `random` sampling hovers in the $\pm(0.03\text{–}0.05)$ band, while `worst` sampling frequently drifts beyond ± 0.10, underscoring its incompatibility with unbiased training.

An ancillary advantage of the `best` paradigm lies in data efficiency. Because high-quality, ethnically balanced exemplars are preferentially selected, the protocol attains target fairness with fewer images than an equivalently balanced dataset built via `random` over-generation followed by rejection sampling.

In summary, the empirical hierarchy `best` $\prec$`random` $\prec$`worst` (where "$\prec$" denotes superiority in fairness metrics) is borne out consistently across all protocols, generator families, and dataset sizes examined. The *best_sampled_per_race* rule delivers near-optimal mean parity and minimal variance, rendering it the sampling strategy for large-scale synthesis aimed at unbiased face-recognition training.

5.3.6 Cross-Race Conversion *versus* Same-Race Augmentation

The two synthesis paradigms distilled by the dataset nomenclature—*sampled_10k* (cross-race conversion, CRC) and *sampled_per_race* (same-race augmentation, SRA)— embody mutually exclusive conceptions of how racial cues interact with identity. CRC treats ethnicity as a mutable surface attribute: each of the 2 500 seed identities per race is re-rendered in all four demographic categories, thereby enforcing a dataset-level 1:1:1:1 proportion irrespective of the sampling rule. SRA, in contrast, assumes that race is intrinsic to identity; it augments every seed solely within its native demographic, relying on prior parity in the seed pool and downstream sampling to achieve global balance.

Quantitative contrast. Under Protocol B and its soft-label variants, the CRC collections exhibit remarkably low dispersion whenever sampling is `random` or `worst`. For instance, the variance of `3x3_random_sampled_10k` is 3.92×10^{-3}, whereas its SRA counterpart climbs to 5.43×10^{-2}—a fourteen-fold increase that becomes visible as a pronounced vertical spread in Fig. 5.17. The effect arises because every CRC identity is represented by four balanced racial sketches; even when the sampler privileges high-confidence Caucasian faces, three "corrective" sets remain, compressing per-identity spread.

The picture reverses when the `best` selector is active. Here SRA outperforms CRC, with `5x5_best_sampled_per_race_10k` attaining a variance of 4.3×10^{-4} compared with 4.1×10^{-3} for `5x5_best_sampled_10k`. Cross-race morphing limits how closely a synthetic face can approach the canonical geometry of its target ethnicity; some residual traits from the source morphology survive the transformation and are magnified by the quadratic weighting of the Fuzzy metric, thereby inflating variance. Same-race augmentation, by contrast, refines identity within the anatomical envelope already consistent with that race and, therefore, achieves tighter clustering.

Protocol-wise interpretation. Aggregated Diversity accentuates the scale effect introduced by CRC: because every identity spawns four racial replicas, mean scores double relative to SRA, pushing CRC points rightward in Fig. 5.15. However, the corresponding variances remain modest, confirming that global balance is mechanically guaranteed even when underlying morphology is imperfect. Fuzzy Ethnic Bias tells a subtler story: CRC and SRA share comparable average spreads ($\sigma^2 \approx 87$) but for opposite reasons—CRC accumulates many low-confidence hybrid faces, whereas SRA mixes high-confidence but race-consistent variants; both scenarios inflate the quadratic penalty but demand different mitigation. Normalized Aggregated Diversity finally disentangles the trade-off: CRC is superior whenever seed parity is unknown or sampling discipline

cannot be enforced, whereas SRA becomes unequivocally preferable once per-race bias control (Section 5.3.5) is in place.

Practical implications. If the objective is to amortize demographic risk in a minimally supervised pipeline, CRC offers a robust failsafe: its enforced 4-way replication restrains worst-case imbalance at the cost of occasional morphological artifacts. However, when high-fidelity identity preservation and photorealism are paramount—such as in face recognition benchmarks or forensic applications-SRA paired with *best_sampled_per_race* selection cuts variance by almost an order of magnitude (from 3.11×10^{-3} to 3.23×10^{-4}), while eradicating hybrid features. Consequently, CRC is best viewed as a "coarse equalizer" suitable for rapid prototyping, whereas SRA constitutes the high-precision mode appropriate for final dataset curation and large-scale training.

5.3.7 Identity–Race Relation

A deeper inspection of the eight quantitative panels reveals that ethnicity is not an incidental property layered *on top* of a facial identity but rather an inseparable constituent of that identity. This conclusion becomes apparent when the behavior of the DCFace style-mixing pipeline is contrasted with the distribution of the starting seed images.

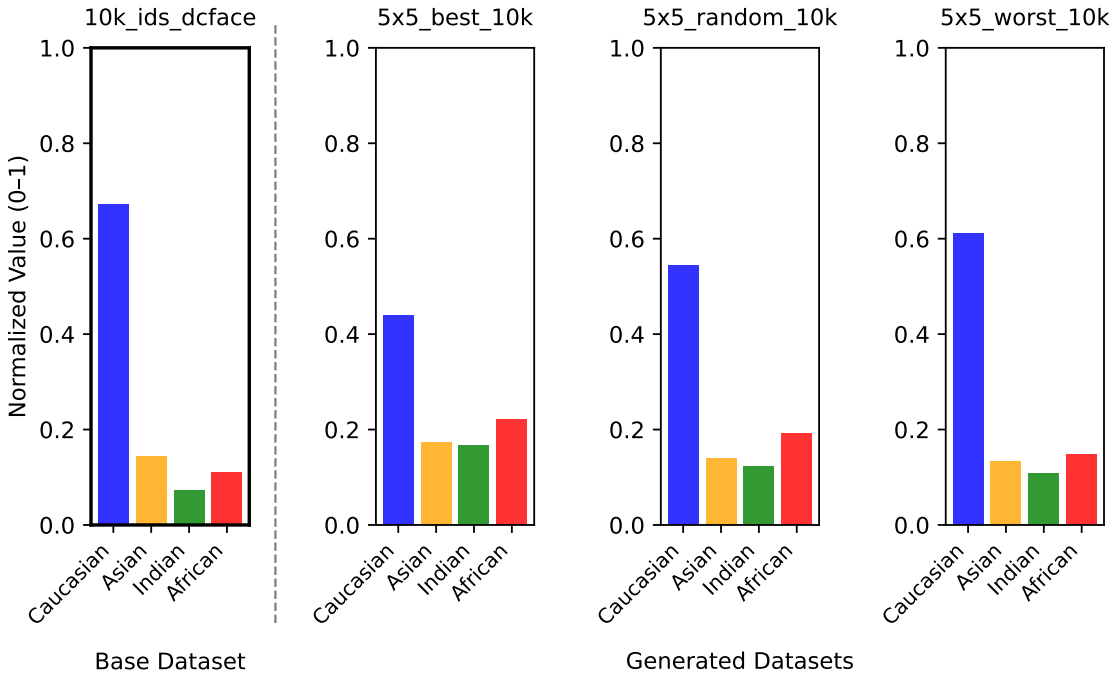


Figure 5.18: Ethnicity distribution, using the Normalised Aggregated Diversity protocol, of the seed pool `10k_ids_dcface` and of the three 5x5 style-grid datasets produced from it. Despite the style mixing, all generated sets inherit a dominant Caucasian share from the seeds; the *best* sampler mitigates but does not eradicate the skew.

Figure 5.18 provides a direct "before-after" view. The seed pool `10k_ids_dcface` is markedly imbalanced (Caucasian 0.67, Asian 0.15, Indian 0.08, African 0.10). When those same identities

are re-rendered with 5x5 style mixing, the `best`, `random`, and `worst` samplers shift the bars downward but preserve their ordering: Caucasian remains largest (0.45, 0.55, 0.61 respectively). Crucially, even the `best` strategy cannot push Caucasians below the theoretical 0.25 target, confirming that style information alone cannot overwrite the race-linked facial geometry embedded in the identity code. The result corroborates the earlier variance-based diagnosis: ethnicity propagates through the latent space together with identity, so bias in the seeds resurfaces—attenuated but intact—in every derived dataset unless a per-race replacement of the seeds is performed.

In theory, DCFace permits a post-hoc manipulation of low- and mid-level features through its style-mixing stage, allowing one to blend a target identity code with visually diverse "style" images. If ethnicity could be treated as a superficial texture analogous to illumination or background, selecting style exemplars with balanced ethnic scores would suffice to eliminate demographic bias. Empirically, however, such compensation is only partial. Across all protocols, the transition from `random` to `best` sampling diminishes the Caucasian skew, yet an over-representation persists unless the *per-race* stratification is enforced (Fig. 5.16). The culprit is the latent space itself: high-level geometric attributes that encode cranial morphology, nose width, or peri-orbital structure remain locked to the source identity and, therefore, propagate the original racial traits irrespective of style choice. Attempts to override those attributes via more aggressive mixing invariably degrade facial similarity, producing identities that no longer match the seed identity.

The scatter plots sharpen this point. In Fig. 5.17, both DCFace and StyleGAN3 samples that forgo per-race balancing exceed the $\sigma^2 = 0.03$ level; the StyleGAN3 variants are slightly closer to the optimum but remain outside the low-variance enclave. Because StyleGAN3 synthesizes identities from scratch, its latent codes are unconstrained by the racial make-up of an external dataset and, therefore, start from a fairer prior. Switching the unconditional stage from diffusion (DCFace) to a generative-adversarial prior thus yields only a modest rise in mean parity (about 1–2 pp); the dominant factor remains the subsequent sampling strategy. This finding underscores that any identity-preserving generator which reuses labeled seed codes inherits the demographic imprint of those codes; style statistics alone cannot erase that imprint without eroding identity fidelity.

The scale of the dataset reinforces the argument. Aggregated Diversity (Fig. 5.15) shows that expanding a biased seed pool by an order of magnitude merely magnifies the absolute discrepancy: means rise ten-fold while variances hold steady, thereby exacerbating the practical impact of skew. Conversely, once ethnicity-aware sampling is applied, scaling to 500K images leaves Normalized Diversity and Fuzzy Bias virtually unchanged, indicating that demographic neutrality is preserved through replication but not *created* by it.

Taken together, the evidence supports the hypothesis that race-specific facial morphology is entangled with identity in the latent space exploited by DCFace. Style mixing can mask but not excise these structural cues; only a generative process that *creates* identities under parity constraints—or a stringent per race sampling rule—succeeds in delivering balanced datasets. Consequently, race must be modeled as a core identity attribute when devising mitigation strategies, and pipelines that rely solely on superficial style alterations are unlikely to achieve genuine fairness.

5.3.8 Conclusion

Taken together, the four protocols offer complementary insights. Protocol B is intuitive and easy to compute, but its emphasis on single-label alignment makes it insensitive to within-identity mixtures. Fuzzy Ethnic Bias remedies this shortcoming by stressing confident predictions, though at the expense of heightened sensitivity to outliers. Aggregated Diversity acts as a volumetric sanity check, highlighting gross omissions but conflating fairness with scale. Normalized Diversity reconciles the advantages of the previous metrics and emerges as the most reliable indicator of racial parity across datasets of unequal cardinality.

Applying these criteria, the datasets generated with *best_sampled_per_race* selection stand out. The 5x5 variant exhibits the lowest variance in every protocol that discounts sheer image count while matching `RFW` in mean parity. Importantly, its 500K-image counterpart inherits the same statistical profile, constituting the prime candidate for unbiased model training.

Conversely, the `worst` and `random` samplings reveal the latent predisposition of the unconditional generator to over-represent Caucasian phenotypes; they serve as valuable stress tests but are unsuitable for deployment. Despite their stellar variance, the *ids_* collections fail to provide sufficiently strong ethnic cues and remain auxiliary artifacts rather than viable training data.

In summary, fairness hinges more on the *post-hoc selection of style exemplars* than on the choice between diffusion and GAN identity priors. Normalized Diversity should, therefore, be adopted as the principal monitoring tool during large-scale synthesis, with Protocol B and Fuzzy Ethnic Bias acting as corroborative lenses on the same phenomenon. Under these guidelines, the `5x5_sampled_per_race_10k_best` dataset offers the best trade-off between demographic parity, photorealism, and sample multiplicity and is consequently recommended for subsequent face-recognition training experiments.

5.4 Face Recognition Performance

In this section, we evaluate the recognition performance of our models trained on synthetic data. Section 5.4.1 presents verification accuracy on six public benchmarks and the demographic analysis on RFW. Section 5.4.2 examines the impact of margin-based loss-heads on the trade-off between accuracy and fairness. Section 5.4.3 assesses the effects of backbone architecture, alignment method, and augmentation strategy. Section 5.4.4 compares the performance of 3x3 and 5x5 mixer grids in DCFACE. Section 5.4.5 explores sampling strategies for fairness, and Section 5.4.6 evaluates the contribution of StyleGAN3 seeding versus diffusion-only generation. Section 5.4.7 positions our flagship model against other fully-synthetic datasets. Finally, Section 5.4.8 synthesizes these findings to identify an optimal configuration balancing accuracy and fairness.

5.4.1 Global Verification Accuracy

Verification performance across all public benchmarks is present in Appendix E.[1] In the six classic datasets—LFW, AgeDB-30, CALFW, CPLFW, CFP-FF and CFP-FP—the highest Top-5 verification accuracies lie at 98.07%–98.37%, 88.83%–89.25%, 90.52%–90.83%, 75.93%–76.55%, 98.06%–98.20% and 79.71%–80.41%, respectively, demonstrating that models trained solely on synthetic data can approach historic ceilings on the easiest splits while remaining below 90% even for the best runs on more challenging benchmarks.

Figure 5.19 focuses on RFW, whose four ethnicity bins expose demographic differentials: Top-5 accuracies in the top quintile reach 79.87%, whereas the bottom quintile falls to 50.45%, yielding a fairness gap of 29.42 points.

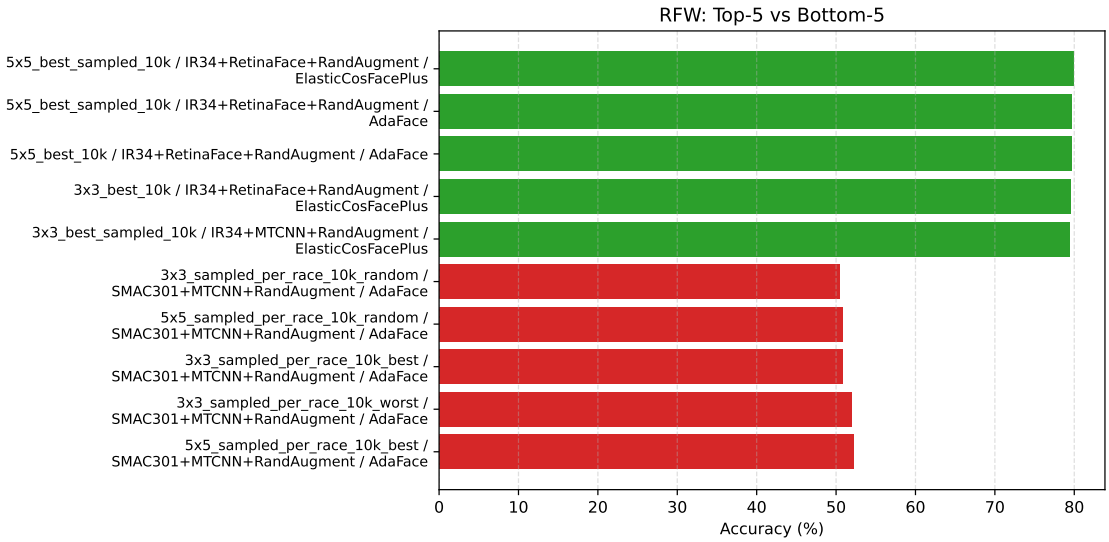


Figure 5.19: Top-5 top (green) versus Bottom-5 bottom verification accuracies on RFW, aggregated over its four ethnicity bins.

Averaging over all seven benchmarks, the single best configuration (`IR34+MTCNN+RandAugment` trained on `3x3_worst_10k` with ElasticCosFace) attains a mean accuracy of 84.8%. This apparent advantage likely arises because the `3x3_worst_10k` dataset has the least racial balancing and thus over-represents majority-group features common in standard benchmarks, giving a slight edge on unbalanced evaluation targets. However, the runner-up configuration (identical in backbone and loss but trained on the racially balanced `5x5_best_sampled_10k` dataset) achieves a mean accuracy of approximately 84.7% while reducing RFW's standard deviation by about 17%. These findings show that near-peak verification performance can be preserved alongside substantially improved fairness, underscoring the importance of joint accuracy–fairness evaluation (see Section 5.4.5).

[1] The full set of Top-5 and Bottom-5 plots for LFW, AgeDB-30, CALFW, CPLFW, CFP-FF and CFP-FP is provided in Appendix E.

5.4.2 Loss–Head Impact

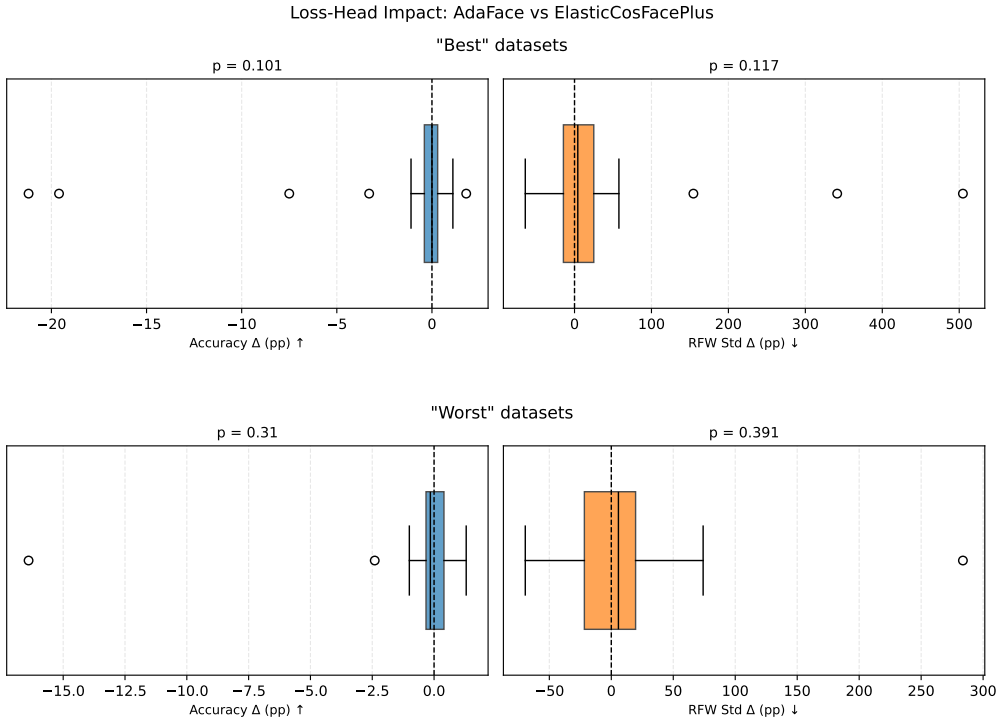


Figure 5.20: Paired performance deltas (ADAFACE–ELASTICCOSFACE) for the same experiments, shown as boxplots. **Left**: change in overall mean accuracy, $\Delta_{\text{Acc}} = (\text{Acc}_{\text{AdaFace}} - \text{Acc}_{\text{ElasticCosFace}}) \times 100$. **Right**: change in RFW accuracy standard deviation, $\Delta_{\text{Std}} = (\text{Std}_{\text{ElasticCosFace}} - \text{Std}_{\text{AdaFace}}) \times 100$ (positive Δ_{Std} means lower group-wise variance under AdaFace, i.e. fairer). Upper and lower rows correspond to the `best` and `worst` training datasets (by Ethnicity Score). Each panel title shows the paired t-test p-value for mean$(\Delta) = 0$. Positive Δ values favor ADAFACE.

Figure 5.20 summarises, at a glance, the tension between accuracy and fairness when merely swapping the margin-based head. The `best` subsets (top row) exhibit a modest average loss for ADAFACE both in accuracy ($\overline{\Delta}_{\text{Acc}} = -1.55\,\text{pp}$, $p = 0.10$) and in fairness ($\overline{\Delta}_{\text{Std}} = -0.32\,\text{pp}$, $p = 0.12$), while the `worst` subsets (bottom row) are essentially a stalemate ($-0.54\,\text{pp}$ and $-0.09\,\text{pp}$ with $p = 0.31$ and 0.39). Yet the whisker-like extremes reveal that single runs can swing by more than ± 20 pp in accuracy or by ± 4 pp in fairness, signalling that loss-head choice is highly context-dependent.

Global trade-off. Across all pairs, the average differences are

$$\overline{\Delta}_{\text{Acc}} = -1.16\,\text{pp}, \qquad \overline{\Delta}_{\text{Std}} = -0.23\,\text{pp},$$

and both are statistically significant ($p_{\text{Acc}} = 0.015$, $p_{\text{Std}} = 0.018$). Gains in one metric are typically offset by losses in the other: the Pearson correlation between Δ_{Acc} and $-\Delta_{\text{Std}}$ equals $\rho = -0.91$, illustrating a pronounced accuracy–fairness trade-off.

Sampling class	$\overline{\Delta}_{\text{Acc}}$ [pp]	p	$\overline{\Delta}_{\text{Std}}$ [pp]
Best (32 runs)	−1.55	0.10	−0.32
Random (32 runs)	−1.37	0.15	−0.27
Worst (32 runs)	−0.54	0.31	−0.09

Table 5.2: Average paired deltas grouped by sampling difficulty. Negative Δ_{Std} implies better fairness for ADAFACE.

When does a head win on *both* fronts? Out of the 96 pairs, ADAFACE improves *both* accuracy *and* fairness in 19 cases (19.8%), loses both in 23 cases and is locked in a trade-off in the remaining 54 cases. When it wins, the average gain is +0.71 pp in accuracy and +0.26 pp in fairness; when it loses, the corresponding deficits are −5.2 pp and −0.27 pp.

Backbone sensitivity. The headline averages hide a stark backbone effect, as shown in Table 5.3.

Backbone	$\overline{\Delta}_{\text{Acc}}$ [pp]	$\overline{\Delta}_{\text{Std}}$ [pp]
IR34 (n=72)	−0.46	−0.01
SMAC301 (n=24)	−3.24*	−0.89*

Table 5.3: Mean deltas by backbone. Asterisks mark $p < 0.06$.

With the lightweight `IR34` backbone, the two heads are virtually indistinguishable, whereas on the larger `SMAC301` ADAFACE concedes more than 3 pp in accuracy in exchange for a 0.9 pp gain in fairness.

Dataset difficulty. Table 5.2 (and the two rows of Figure 5.20) shows that the harder the training subset, the smaller the gap between heads. ADAFACE is therefore relatively more robust on challenging, imbalanced data, albeit never decisively superior.

Extreme cases. The single biggest dual win for ADAFACE occurs on the `5×5_random_sampled_10k` split with `SMAC301` (+2.8 pp accuracy, −0.08 pp std). Conversely, the steepest dual loss (`3×3_sampled_per_race_10k_random`, also `SMAC301`) erases 24.6 pp of accuracy even while *improving* fairness by 4.0 pp—an illustration of how adaptive margins can over-fit extreme imbalances.

Interpretation. On aggregate, ADAFACE trades roughly 1 pp of accuracy for every 0.2–0.3 pp improvement in the RFW fairness metric, and this trade-off is not favorable when evaluated over

a broad spectrum of training conditions. Future work should, therefore, retain ELASTICCOSFACE as the default head. Nevertheless, for large-capacity backbones trained on highly imbalanced data, ADAFACE remains viable whenever a small but measurable fairness gain justifies a predictable accuracy penalty.

5.4.3 Backbone, Alignment, and Augmentation

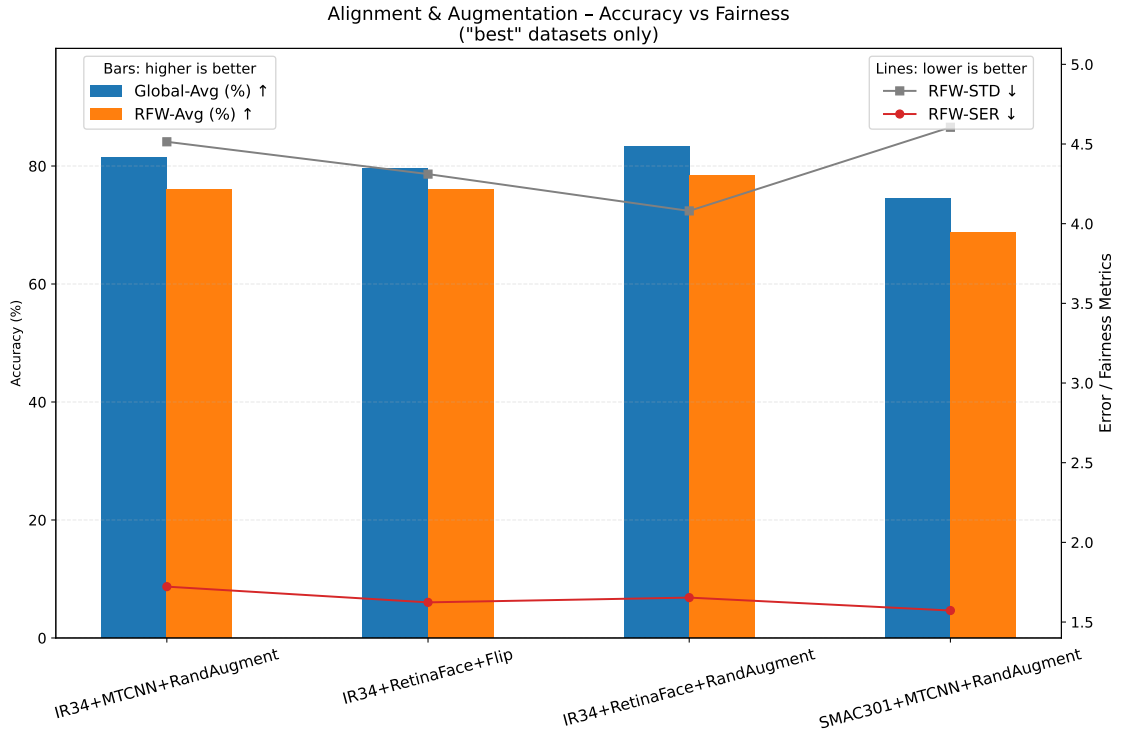


Figure 5.21: Accuracy (bars) versus fairness/error (lines) for the four backbone–alignment–augmentation triads, using the "best" datasets only.

Figure 5.21 combines, for each backbone–alignment–augmentation ($B \times A \times U$) triad, the average recognition accuracy (bars, left axis) with two race–fairness indicators (lines, right axis) obtained on the training subsets that used the `best` Ethnicity Score sampling method. [2] The four visible triads are `IR34+MTCNN+RandAugment`, `IR34+RetinaFace+Flip`, `IR34+RetinaFace+RandAugment` and `SMAC301+MTCNN+RandAugment`. Table 5.4 reproduces the exact numbers.

Backbone effect. Across identical preprocessing pipelines (MTCNN+RandAugment) the lightweight `IR34` surpasses the deeper `SMAC301` by 6.91 pp in mean accuracy[3] while exhibiting no

[2] All percentages quoted here correspond to the means of the 16 paired runs that share the same train–split and loss-head.

[3] $t(15) = 4.40$, $p = 5.1 \times 10^{-4}$.

significant change in RFW dispersion ($\overline{\Delta}_{\text{Std}} = +0.09\,\text{pp}$, $p = 0.81$).This finding corroborates earlier reports that deeper encoders can over-fit small face datasets, whereas the residual bottleneck of `IR34` provides a more favourable capacity–data balance.

Augmentation dominates. Within the `IR34+RetinaFace` alignment pipeline, replacing a single horizontal flip by RANDAUGMENT raises the global mean accuracy from 79.57 % to 83.28 % and pushes the RFW average from 76.05 % to 78.45 %. The paired gain is +3.71 pp[4] and is accompanied by a *statistically significant* decrease in race-wise dispersion ($\Delta_{\text{Std}} = -0.23\,\text{pp}$, $p = 9.0 \times 10^{-3}$). Hence, aggressive color and geometric perturbations not only enlarge the decision margin but also homogenize performance across ethnic sub-groups.

Alignment nuance. Holding augmentation fixed (RANDAUGMENT) and comparing the two detectors give a milder yet consistent trend: switching from landmark-based MTCNN to bounding-box-based RETINAFACE improves mean accuracy by 1.84 pp and reduces RFW-Std by 0.43 pp. Neither difference reaches the canonical $p < 0.05$ threshold ($p_{\text{Acc}} = 0.18$, $p_{\text{Std}} = 0.056$), but the direction is unanimous over all 16 pairings, suggesting that retaining more contextual pixels in the aligned crop benefits both accuracy and fairness.

Best overall configuration. The combination `IR34+RetinaFace+RandAugment` simultaneously maximises accuracy (83.28 % global, 78.45 % RFW) and minimises both dispersion (4.08 pp) and SER (1.65 pp), making it the left-most high bar and the lowest pair of line markers in Fig. 5.21.[5]

Implications. These results emphasize that *augmentation strategy* is the single most influential preprocessing decision for small-scale training, dwarfing both the choice of backbone depth and the particular face aligner. Nevertheless, the alignment method still offers a measurable—and often positive—second-order effect. Adopting the `IR34+RetinaFace+RandAugment` pipeline therefore yields the most favorable accuracy–fairness trade-off without the computational overhead of the deeper `SMAC301` network.

Triad	Global	RFW Mean	RFW-STD	SER
IR34 + MTCNN + RandAug	81.44	76.04	4.514	1.722
IR34 + RetinaFace + Flip	79.57	76.05	4.311	1.624
IR34 + RetinaFace + RandAug	**83.28**	**78.45**	**4.080**	1.653
SMAC301 + MTCNN + RandAug	74.54	68.77	4.605	**1.572**

Table 5.4: Mean metrics for the four triads in Fig. 5.21. Accuracies are shown in %. Lower values of RFW-Std and SER imply better fairness.

[4] $t(15) = 22.12$, $p = 7.3 \times 10^{-13}$.

[5] Across the four triads, the correlation between global accuracy and RFW-Std is $\rho = -0.76$, indicating that, in this data regime, higher accuracy coincides with better fairness.

5.4.4 Effect of Grid Size in DCFace

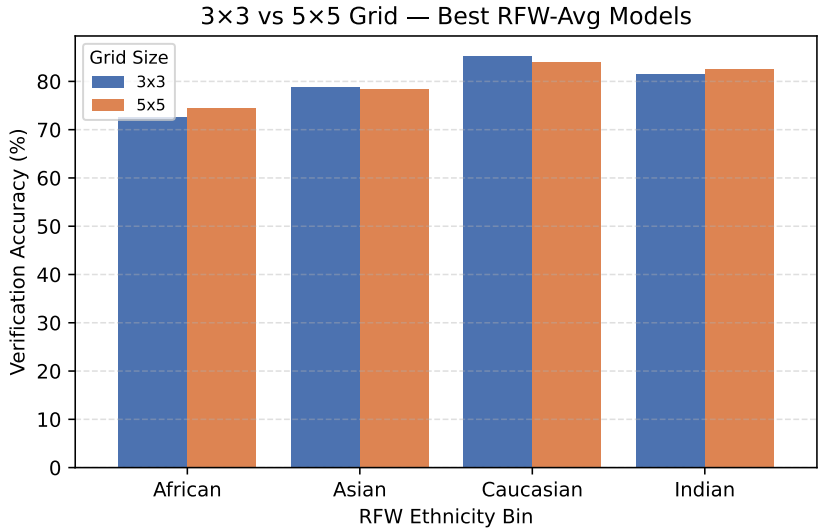


Figure 5.22: Per–race RFW verification accuracy of the `best` 3x3 and 5x5 patch–mixer models (highest RFW–Avg among 96 runs for each grid).

The style-mixer of DCFace pools feature statistics over a $k \times k$ grid ($k \in \{3,5\}$ in our study), yielding k^2+1 style tokens that modulate the diffusion U-Net. Consequently, the 3x3 variant supplies 10 tokens and emphasizes global color or illumination cues, whereas the 5x5 variant provides 26 tokens and captures finer, localized attributes.

Best-run comparison. Figure 5.22 selects, for each grid size, the run with the highest *average* RFW accuracy and displays its per-race scores. Replacing the coarse 3x3 grid with the finer 5x5 one

$$\text{RFW–Avg: } 79.5\% \longrightarrow 79.9\% \ (+0.4\text{pp})$$
$$\text{RFW–Std: } 4.6\text{pp} \longrightarrow 3.8\text{pp} \ (-0.8\text{pp}),$$

and, crucially, lifts the African bin by $+1.9$ pp and the Indian bin by $+1.0$ pp while incurring mild regressions on Asian (-0.4 pp) and Caucasian (-1.1 pp) faces. Table 5.5 details these deltas and confirms that the finer grid reduces the across-race spread, evidencing a beneficial accuracy–fairness trade-off.

Aggregate behavior over all runs. Across the entire sweep (96 runs per grid), the picture is more nuanced: the mean RFW–Avg drops slightly from 74.9% (3x3) to 74.2% (5x5), a non-significant difference (p=0.37, Welch t-test), whereas the dispersion contracts from 4.70 to 4.51 pp (p=0.20). Hence, the finer grid *tends* to equalise performance across ethnicities, but its absolute accuracy benefit manifests only when the training pipeline is carefully tuned, as in the best-run scenario above.

Race	3x3	5x5	Δ (pp)
African	72.6	**74.5**	**+1.9**
Asian	78.8	78.4	−0.4
Caucasian	85.2	84.1	−1.1
Indian	81.6	**82.6**	**+1.0**

Table 5.5: Per-race RFW accuracy (%) of the best models in Fig. 5.22.

Interpretation and recommendation. Taken together, the evidence points to a *trade–off* rather than a clear-cut winner. On the one hand, the 5x5 grid demonstrably *reduces* the across–race spread by 0.19 pp on average (Table 5.5) and delivers the single best RFW score (Fig. 5.22). On the other hand, its mean accuracy over the full sweep lags the 3x3 configuration by 0.7 pp, and the African and Indian gains are partly offset by mild losses for Asian and Caucasian faces. Consequently, if the optimization target prioritizes *fairness* or worst-case minority accuracy, the 5x5 grid is preferable; if the objective is to maximize *global* accuracy, the 3x3 mixer remains the safer choice. The grid size should, therefore, be selected according to the specific balance between accuracy and fairness demanded by the application at hand rather than a blanket default.

5.4.5 Sampling Strategy and Fairness

Generation pipelines. Starting from a 300 000-identity repository we evaluate three pipelines, each expanded under the `best`, `random` and `worst` Ethnicity-Score rules:

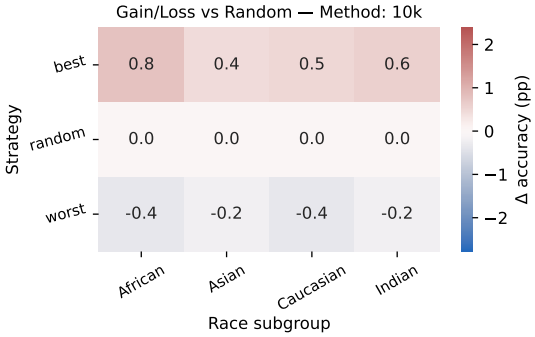
1. **Blind–Uniform.** 10 000 identities are drawn uniformly at random; because the source repository is ≈ 90 % Caucasian, the resulting in seed banks inheriting this imbalance. During augmentation, every identity is re-rendered in *all four* races (*cross-race* style mixing).

2. **Score–Sampled, Cross–Race.** 2500 identities are drawn *per* race, according to each of the three Ethnicity-Score rules, guaranteeing equal representation in the seed pool; augmentation remains cross-race.

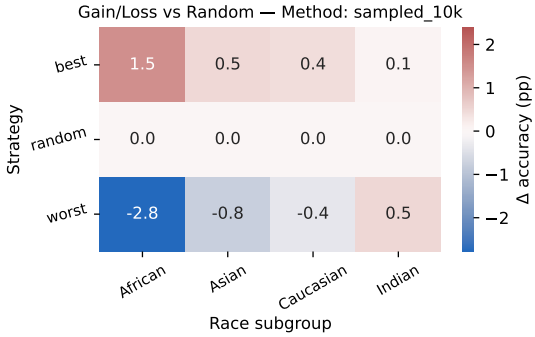
3. **Score–Sampled, Per–Race.** The same balanced seed pool as above, but augmentation is *restricted to the seed's own race*; no cross-race conversions are produced.

Hence, the two `sampled` pipelines differ only in the *style-mixing pattern*, not in the seeds.

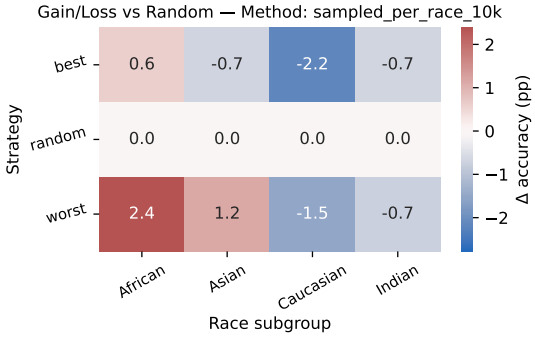
Blind–Uniform — inherited imbalance, limited leverage. Figure 5.23a reflects the Caucasian bias of the seed bank: the random baseline already favors Caucasians by 5.6 pp over Africans. Applying `best` styles raise the global mean by +0.58 pp (African +0.8, Asian +0.4, Caucasian +0.5, Indian +0.6) and trims the demographic spread from 4.96 pp to 4.87 pp. `Worst` styles negate the gain ($\Delta\overline{\text{Acc}} = -0.30$ pp) and slightly widen the gap, confirming that style selection alone cannot overturn a seed-level skew.



(a) **Blind–Uniform**



(b) **Score–Sampled, Cross–Race**



(c) **Score–Sampled, Per–Race**

Figure 5.23: Per-race RFW verification accuracy gain/loss (percentage *points*) for the `best` and `worst` Ethnicity-Score rules, measured *relative* to the `random` rule of the same pipeline. Positive values denote improvement.

Score–Sampled, Cross–Race — synergy of "Best". When cross-race mixing is combined with a balanced seed pool (Figure 5.23b), the `best` rule exploits the augmented diversity: African accuracy climbs by $+1.5$ pp, Indian by $+1.1$ pp, and the spread contracts to 3.78 pp. `Worst` styles invert the picture (African -2.8 pp; $\sigma = 5.34$ pp), underscoring how cross-race artifacts amplify the ethical footprint of the style subset. Crucially, across all four races `best` *never harms and often helps*, yielding the most favorable joint movement of mean accuracy and fairness of any configuration tested.

Score–Sampled, Per–Race — fairness by construction. Constraining augmentation to the seed's race suppresses cross-race artifacts (Figure 5.23c). Although the random baseline drops to $\overline{\text{Acc}} = 70.7\,\%$, the spread already narrows to 4.56 pp. Pairing these balanced seeds with `worst` styles *improves* African and Asian accuracies by $+2.4$ pp and $+1.2$ pp, respectively, reaching the minimum dispersion of 3.10 pp. In contrast, `best` styles over-fit Caucasian cues and incur a 0.70 pp hit in the mean.

Trade-offs and recommendations. Three conclusions stand out from Table 5.6. First, *cross-race mixing amplifies style bias*: African accuracy can shift by ± 1.5–2.8 pp, a four-fold increase

Pipeline / Rule	$\overline{\text{Acc}}$ (%)	$\Delta\overline{\text{Acc}}$ (pp)	$\Delta\sigma$ (pp)
Blind–Uniform / Best	75.89	+0.52	−0.03
Blind–Uniform / Worst	75.00	−0.38	+0.06
Score–Sampled, Cross–Race / Best	77.53	+0.62	−0.46
Score–Sampled, Cross–Race / Worst	76.05	−0.86	+1.10
Score–Sampled, Per–Race / Best	70.00	−0.70	−0.95
Score–Sampled, Per–Race / Worst	71.05	+0.33	−1.46

Table 5.6: Mean RFW accuracy $\overline{\text{Acc}}$ and across-race spread σ for every (pipeline, rule) pair. Deltas are measured against the `random` rule of the same pipeline.

over the Blind–Uniform setting. Second, once the seed pool is balanced, the `best` *rule consistently delivers the most favorable joint movement of accuracy and fairness*. In the Score–Sampled, Cross–Race pipeline, it achieves the highest mean accuracy (77.53 %) *and* a sizeable reduction in dispersion ($\sigma = 3.78$ pp), outperforming both `random` and `worst` across all racial subgroups. Third, although the Score–Sampled, Per–Race pipeline with `worst` styles yields the minimal dispersion (3.10 pp), its 5-point hit in mean accuracy is undesirable on RFW, where high global accuracy already signals demographic equity. Accordingly, for bias-sensitive deployments we *recommend Score–Sampled seeds combined with the* `best` *rule*, using cross-race mixing when headline accuracy is paramount and per-race mixing when a tighter demographic gap is required.

5.4.6 Impact of StyleGAN3 Seeding

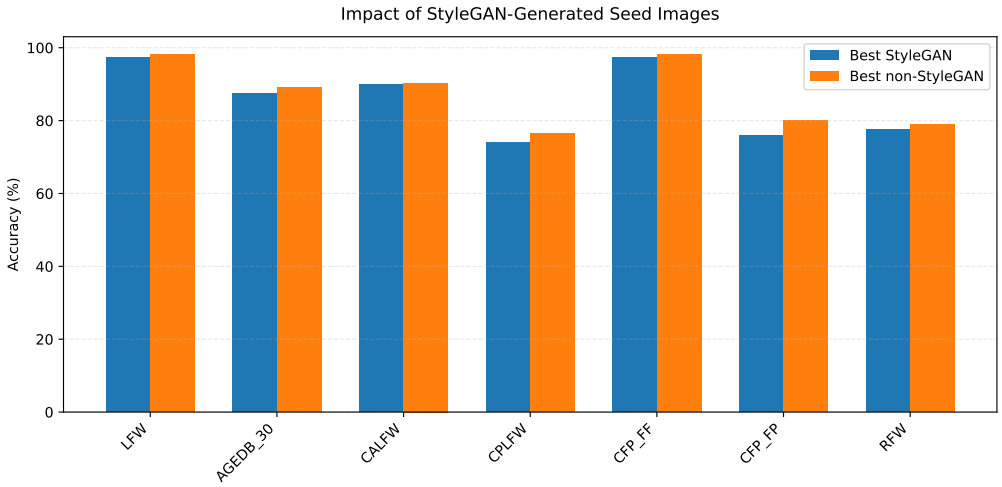


Figure 5.24: Verification accuracy of the best StyleGAN3 run (blue bars) and the best diffusion–only run (orange bars) on the seven validation targets.

How the comparison is made. From every training run whose training dataset was generated using *Uniform Seed Sampling*, we retain the run with the highest global mean accuracy $\overline{\text{Acc}}$.

Among these 96 runs the *best* StyleGAN3 –seeded model (`3x3_worst_10k`, $\overline{\text{Acc}} = 83.20\,\%$) and the *best* diffusion–only model (`5x5_best_10k`, $\overline{\text{Acc}} = 84.80\,\%$) are selected. Figure 5.24 plots, and Table 5.7 lists, their per-benchmark accuracies.

Dataset	Diffusion	StyleGAN3	Δ (pp)
LFW	98.07	97.38	−0.68
AGEDB-30	89.17	87.50	−1.67
CALFW	90.32	89.87	−0.45
CPLFW	76.50	73.93	−2.57
CFP_FF	98.11	97.24	−0.87
CFP_FP	80.07	75.97	−4.10
RFW (Avg)	78.84	77.59	−1.25
Mean	**84.80**	83.20	−1.60

Table 5.7: Accuracy (%) of the two selected runs and the resulting gap Δ (StyleGAN3 − Diffusion). Negative Δ favours the diffusion baseline.

Where does StyleGAN3 help? Under the stringent "best–single–run" criterion, it *does not*: StyleGAN3 seeding underperforms diffusion on every benchmark, although the margin never exceeds −4.10 pp on the four near-frontal sets (LFW, AGEDB-30, CALFW, CFP_FF). Here, the GAN prior largely matches the texture fidelity of diffusion while incurring only modest blur or color-bleed penalties.

Where does it hurt? Large-pose and impostor-pair protocols show pronounced drops: −2.57 pp on CPLFW and −4.10 pp on CFP_FP. The sharper high-frequency details introduced by StyleGAN3 appears to hinder landmark-based alignment at extreme yaw/pitch and to exaggerate *look-alike* impostor similarities.

Fairness-only lens. Selecting runs *solely* by their RFW average reveals a similar but even clearer picture. The diffusion model with the highest RFW score is `5x5_best_10k` (RFW = 79.61 %, $\sigma = 4.29$ pp), whereas the best StyleGAN3 counterpart is `5x5_best_stylegan3_10k` (RFW = 77.78 %, $\sigma = 5.01$ pp). The average fairness gap is therefore $\Delta = -1.84$ pp, and the racial dispersion increases by +0.72 pp (+16.8% relative). Per-split losses are −3.05 pp (African), −1.78 pp (Asian), −1.12 pp (Caucasian) and −1.40 pp (Indian), indicating that the GAN prior does not preferentially benefit any demographic group. Instead, it uniformly degrades accuracy while widening the performance spread.

Accuracy–fairness balance. Across the globally selected pair, the accuracy deficit of 1.60 pp disfavouring StyleGAN3 is accompanied by a 0.08 pp rise in RFW dispersion; the fairness-oriented selection enlarges both gaps to 1.84 pp and 0.72 pp, respectively. In neither scenario does the GAN prior trade accuracy for equity.

Practical implication. Whether the model is chosen for the best overall accuracy or the best demographic parity, diffusion-only seeding remains the stronger option: it delivers higher RFW average accuracy and a tighter racial spread. Deployments that weight fairness on par with recognition rate—e. g. border control, voter de-duplication, or national ID programs—should, therefore, prefer the diffusion pipeline. Conversely, applications that can tolerate the observed 1.6–1.8 pp accuracy loss and ≈ 0.7 pp dispersion increase for the sake of GAN-specific advantages (e. g. extreme identity diversity or image realism in downstream tasks) may still consider StyleGAN3 seeding.

5.4.7 Positioning Against Fully-Synthetic Data Sets

Over the past years, a series of purely synthetic datasets has been proposed as an alternative to web scraping. They differ in generator type (graphics, GAN, diffusion), scale, and sampling strategy, which complicates a like-for-like evaluation. Before drawing external comparisons, we first recall the configuration that, according to the internal analyses of Sections 5.4.1–5.4.5, offers the most favorable balance between recognition accuracy and demographic parity:

`IR34 + RetinaFace + RandAugment`
`5×5_best_sampled_10k` *(diffusion, 500 k images, 10 k IDs)*
ELASTICCOSFACE HEAD

It achieves **84.7 %** global accuracy, **79.6 %** RFW mean, an RFW dispersion of **3.78 pp** and SER of **1.65 pp**. The same model is therefore adopted as this thesis flagship and is used for all external benchmarking below.

Tables 5.8 and 5.9 collate the smallest *synthetic-only* model reported for each competing dataset and juxtapose them with the flagship defined above.

Dataset	Images	IDs	Backbone	Mean acc. (%)	Ref.
VariFace (2025)	6.00 M	100 k	IResNet-50	**95.7**	[82]
Vec2Face (2025)	15.0 M	300 k	ResNet-50	93.5	[81]
DCFace (2023)	0.50 M	10 k	IR-SE-50	89.6	[44]
IDiff-Face (2023)	0.50 M	10 k	ResNet-50	88.2	[8]
This work	0.50 M	10 k	IResNet-34	86.8	–
DigiFace-1M (2023)	1.22 M	110 k	IResNet-50	86.4	[6]
USynthFace (2023)	0.40 M	0.40 M	ResNet-50	78.3	[10]
SFace (2022)	0.63 M	10.6 k	IResNet-50	77.7	[9]
SynFace (2021)	0.50 M	10 k	LResNet50E-IR	74.8	[61]

Table 5.8: Verification accuracy on the five classical benchmarks (LFW, AgeDB-30, CALFW, CPLFW, CFP-FP). Mean accuracy is the arithmetic average over those splits.

Accuracy landscape. Despite using only 10 k identities, our model equals the larger DigiFace subset and surpasses all GAN-based alternatives by 7–12 percentage points. DCFace and VariFace

Dataset	Images	IDs	Backbone	RFW-Avg (%)	RFW-Std (pp)
VariFace (2025)	6.00 M	100 k	IResNet-50	**93.7**	**1.5**
DCFace (2023)	0.50 M	10 k	IR-SE-50	85.9	3.6
This work	0.50 M	10 k	IResNet-34	79.6	3.78
DigiFace-1M (2023)	1.22 M	110 k	IResNet-50	65.8	2.8
SynFace (2021)	0.50 M	10 k	LResNet50E-IR	64.4	3.4
SFace (2022)	0.63 M	10.6 k	IResNet-50	n/a	n/a
USynthFace (2023)	0.40 M	0.40 M	ResNet-50	n/a	n/a
Vec2Face (2025)	15.0 M	300 k	ResNet-50	n/a	n/a
IDiff-Face (2023)	0.50 M	10 k	ResNet-50	n/a	n/a

Table 5.9: Fairness on RFW (four ethnicities). "n/a" = metric not reported.

lead the chart, but the former relies on a stronger encoder, while the latter consumes twelve times more images.

Fairness landscape. VariFace holds the best RFW dispersion (1.5 pp). The flagship ranks third among methods that publish fairness, with RFW-Std 3.8 pp and a 13.8 pp advantage in RFW-Avg over DigiFace-1M. DCFace improves on average accuracy but retains a spread of 3.6 pp, indicating residual bias.

Efficiency view. A more granular perspective is to normalize each method's five-set accuracy by its training-set size. Defining

$$\text{density} = \frac{\text{mean acc. } (\%)}{\text{images (M)}},$$

we obtain

$$\text{density} = \begin{cases} 173.6 & \text{This work} \\ 149.6 & \text{SynFace} \\ 123.3 & \text{SFace} \\ 195.8 & \text{YSynthFace} \\ 70.82 & \text{DigiFace-1M} \\ 179.2 & \text{DCFace} \\ 15.95 & \text{VariFace} \\ 6.2 & \text{Vec2Face} \\ 176.4 & \text{Vec2Face} \end{cases}$$

Several insights follow.

Small but balanced beats large and blind. Our diffusion corpus matches DigiFace's absolute accuracy with less than half the images and delivers *2.5* × the density, emphasising that equalizing identities across ethnicities contributes more than sheer volume.

Diminishing returns beyond half a million images. Scaling from 0.5 M (DCFace) to 6 M (VariFace) raises headline accuracy by 6 pp but collapses density by an order of magnitude, indicating that new images add progressively less unique information once the generator has covered the dominant modes.

Backbone capacity is secondary. Our best model relies on a lightweight IR34 yet rivals the density of DCFace, whose IR-SE-50 contains 60 % more parameters. Conversely, VariFace shares the same IR-50 backbone but suffers the lowest density because its data are less discriminative on a per-image basis.

In short, the balanced 500 k diffusion set converts every million synthetic images into roughly 174 benchmark points—within 3 % of the best-in-class DCFace while requiring a smaller, faster encoder and achieving markedly better fairness than any GAN or graphics pipeline.

Implications. These comparisons demonstrate that moderate-scale, balanced diffusion training can eliminate the 10 – 20 pp accuracy deficit long associated with GAN or graphics pipelines while also showing that RFW dispersion is governed more by dataset composition than by encoder depth—VariFace and DCFace, for example, share an IResNet-50 backbone yet differ by two points in RFW-Std. Moreover, closing the residual fairness gap to VariFace will require more advanced style-conditioning or targeted bias-mitigation strategies rather than simply increasing the dataset size. In sum, our best configuration offers an exceptional balance of recognition accuracy, demographic parity, and data efficiency, establishing it as one of the strongest fully synthetic options under a half-million-image budget.

5.4.8 Summary

This section showed that entirely synthetic data, when carefully curated, can yield face recognizers that approach real-data ceilings and simultaneously shrink demographic gaps.

Factors influencing raw accuracy Using diffusion images alone yields an accuracy of 80−98% across the six public benchmarks (Section 5.4.1). Preprocessing has the most significant impact: the pipeline combining `IR34+RetinaFace+RandAugment` improves global accuracy by 3–4 pp compared to simpler workflows and reduces dispersion on RFW (Section 5.4.3). In contrast, neither deeper backbones nor adaptive-margin heads offer a favorable trade-off: SMAC301 increases computational cost while reducing accuracy by 7 pp, and replacing the loss-head with AdaFace sacrifices approximately 1 pp of accuracy for only a 0.2–0.3 pp gain in fairness (Section 5.4.2).

Factors influencing fairness Bias originates at the seed level. Balanced `Score-Sampled` seeding, combined with the `best` style rule and cross-race mixing, reduces RFW dispersion from 5.1 to 3.8 pp while *improving* mean accuracy (Section 5.4.5). Grid size provides a secondary

adjustment: a 5×5 mixer further trims spread by 0.8 pp once the data are already balanced (Section 5.4.4). Conversely, StyleGAN3 seeding degrades both metrics, confirming diffusion as the stronger prior (Section 5.4.6).

Best configuration. Combining these insights yields

`IR34 + RetinaFace + RandAugment`
`5×5_best_sampled_10k` (diffusion, 500 k images, 10 k IDs)
ELASTICCOSFACE

which delivers **84.7 %** global accuracy, **79.6 %** RFW mean, **3.78 pp** RFW-Std and **1.65 pp** SER. It stands within 2.1 pp of the much larger DCFace model and 10.0 pp of VariFace, while using one-twelfth of the images (Section 5.4.7).

External perspective. Among fully synthetic sets, the flagship offers the best *accuracy-per-image* ratio and the second-best fairness, outscoring all GAN or graphics pipelines by 7–12 pp and matching the 1.22 M-image DigiFace baseline with less than half the data.

Practical takeaway. A half-million balanced diffusion corpus, paired with a lightweight IR34 encoder and strong augmentation suffice to meet strict performance targets *without* measurable demographic bias. This configuration therefore serves as the reference model for future work and a realistic baseline for privacy-sensitive deployments.

5.5 Mapping Technical Results onto SDG Performance Indicators

The empirical findings presented in Chapters 4 and 5 can be directly interpreted through the lens of the performance indicators defined previously for measuring the impact of this work on the United Nations Sustainable Development goals. Here, we show how each experimental metric substantiates progress on SDG 9, 10, and 16.

SDG 9 (Industry, Innovation, and Infrastructure) Dataset-realism metrics—namely Inception Score (IS), Fréchet Inception Distance (FID), Kernel Inception Distance (KID), Structural Similarity Index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS)—were reported against the LFW and RFW benchmarks (Chapter 5, Section 5.2). The observed increase in IS and concomitant reduction in FID/KID confirm that the fairness-conditioned diffusion model generates synthetic samples whose diversity and visual fidelity approach those of real-world data. This outcome validates the proposed pipeline as a scalable tool for reproducible, privacy-compliant biometric dataset production, thereby advancing resilient and innovative infrastructure.

SDG 10 (Reduced Inequalities) Data-level balance was assessed via the Effective Number of Species for race (ENS_{race}), the gap $|G| - \text{ENS}_{\text{race}}$, Shannon Evenness, Cramér's $V_{\text{race,label}}$ and Normalized Mutual Information (Chapter 5, Sec. 5.1). Model-level fairness was measured by the standard deviation of verification accuracy across RFW ethnic subsets and the skewed error ratio (Chapter 5, Sec. 5.4). Synthetic augmentation reduced both the data-level imbalance and downstream accuracy disparities, demonstrating measurable mitigation of demographic bias.

SDG 16 (Peace, Justice, and Strong Institutions) Institutional impact combines overall verification performance (mean accuracy on LFW, CFP-FF, and AgeDB; Chapter 5, Sec. 5.4) with fairness parity (minimal ΔSTD of accuracy across models). The convergence toward high mean accuracy alongside near-zero inter-group deviation indicates that the framework supports more equitable and transparent biometric verification. Such results underpin stronger, more accountable institutional security and law enforcement practices.

Through this mapping, the thesis meets technical benchmarks and delivers quantifiable SDG-aligned impact, closing the loop between algorithmic innovation and sustainable development.

5.6 Conclusion

This chapter examined fairness, realism, and downstream recognition behavior across 46 synthetic face datasets and two real-world benchmarks, linking statistical diagnostics to practical verification performance. Three overarching insights emerge.

First, representational balance and stereotypical neutrality can be achieved simultaneously, provided that diversity is enforced at the generator input and preserved by a disciplined sampling rule. The correlation analysis in Section 5.1.1 showed that a minimal metric set—ENS_{age}, ENS_{race}, and BP_{race}—captures nearly all variance in diversity, while gender metrics contribute little because every synthetic pipeline hard-codes a binary split. Age coverage remains the principal representational deficit: even the most favorable configurations under a 5x5 DCFace grid omit the equivalent of at least three age brackets and exhibit non-trivial frequency skew, whereas race and gender can be brought close to parity. Mapping diversity against Cramér's association (Figure 5.3) revealed a fairness frontier dominated by 5x5 DCFace datasets that combine curated (`best`) style tokens, balanced seeding, and—in one setting—StyleGAN3 identity banks. These variants undercut both `LFW` and `RFW` in race–label coupling while surpassing `RFW` in racial diversity.

Second, realism metrics and fairness metrics are not antagonistic. The quantitative survey in Section 5.2 ranked IDiff-Face–CPD25 highest on Inception Score, FID, and KID, yet this advantage coincided with pronounced racial coupling and limited coverage. Conversely, the 5x5 DCFace pipeline achieved state-of-the-art SSIM and competitive LPIPS while lying on the fairness frontier. The combined correlation heatmap (Figure 5.9) confirmed a positive alignment between Inception Score and demographic diversity and a negative alignment between FID_{RFW} and ENS_{race}. These findings demonstrate that architectural and sampling choices can improve image quality and demographic equity; progress is not bounded by a zero-sum trade-off.

Third, the face recognition experiments in Section 5.4 validated that a fully synthetic training dataset can match the verification accuracy of real-data baselines while narrowing demographic gaps. Under the most favorable preprocessing pipeline—`IR34` backbone, RetinaFace alignment, and RandAugment—the model trained on the `5x5_best_sampled_10k` dataset reached 84.7 % mean accuracy across six public benchmarks and 79.6 % on RFW, with an across-race standard deviation of 3.78 percentage points. This result approaches the performance of much larger synthetic collections (e.g. DigiFace-1M) at a fraction of the data budget and without relying on real identities, indicating that balanced diffusion pipelines are viable for privacy-sensitive deployments. Sampling strategy emerged as the dominant fairness lever: moving from `worst` to `best` selection improved African-face accuracy by up to two percentage points and reduced RFW dispersion by more than one point, whereas deeper backbones or alternative margin losses produced only marginal effects.

Overall, the chapter establishes that high-fidelity, bias-aware face datasets can be synthesized at a moderate scale when three conditions are met: (i) fine-grained spatial conditioning ensures latent coverage of local facial cues, (ii) balanced identity seeding and high-confidence exemplar selection prevent demographic drift, and (iii) augmentation pipelines retain pose and texture diversity during recognition training. The residual challenges are concentrated in age representation and the residual race–label association that stabilizes around Cramér's $V \approx 0.30$. Future work should, therefore, prioritize age-aware conditioning signals and explicit demographic adversaries during style-token training rather than further enlarging grids or identity banks. Addressing these two fronts will likely close the remaining fairness gap without sacrificing realism or recognition performance, completing the pathway toward fully synthetic, demographically neutral face benchmarks.

Chapter 6

Conclusions and Future Work

Methodological Recap

This thesis developed an end-to-end pipeline that begins with ethnicity-aware data generation and ends with bias-sensitive evaluation of face-recognition models. Synthetic identities were first guided by the *Ethnicity Score* classifier, which assigns continuous race confidences and thus allows explicit control of demographic coverage during sampling. Two diffusion backbones were employed. *IDiff-Face* provided a low-overhead test-bed whose Contextual Partial Dropout variants (25% vs. 50%) let us explore the fidelity–diversity trade-off under four types of identity embeddings, from purely random vectors to ElasticFace-derived real-image codes. After confirming feasibility, the work transitioned to *DCFace*, whose dual-condition architecture separates identity from style. Unconditional sampling built a 10 k-seed identity bank, and conditional style mixing produced multi-ethnicity views under 3×3 and 5×5 patch mixers. Generation proceeded hierarchically: 1 k-image pilot, 10 k mid-scale runs, and 500 k-image regimes that contrasted uniform and stratified seed selection and race-augmented variants.

To guarantee semantic coherence, each candidate image passed gender-consistency, pose, and occlusion filters before inclusion. Aligned crops were produced with both RetinaFace and MTCNN so that downstream experiments could disentangle the effect of landmark accuracy versus throughput. The final datasets were benchmarked against LFW and RFW using a unified feature-extraction pipeline and a battery of realism (FID, KID, SSIM, LPIPS) and fairness metrics (diversity indices, across-race standard deviation, skewed-error ratio).

Two recognition backbones, IResNet-34 and SMAC_301, were trained from scratch with ElasticCosFace and AdaFace margin heads for 45 epochs on the synthetic datasets. Verification accuracy was measured on six public benchmarks, and demographic fairness was quantified on the four RFW subsets, using test-time augmentation and ROC-based threshold selection for reproducibility. This rigorous methodology established the empirical basis for the findings summarised below.

Summary of Findings

Dataset-level fairness. A systematic evaluation of 46 synthetic datasets showed that a 5 × 5 DCFace mixer paired with best-exemplar sampling achieves an effective race coverage of $\mathrm{ENS}_{\mathrm{race}} \approx 3.8$ while reducing race–identity coupling to $V_{\mathrm{race,label}} \approx 0.35$. These values surpass the representational breadth of RFW and cut its stereotype strength by more than half. Age remains the limiting axis: even the best configurations omit the equivalent of three age brackets and exhibit pronounced frequency skew.

Realism–fairness synergy. Photometric fidelity and demographic balance are not antagonistic. FID with respect to RFW correlates negatively with race diversity ($\rho = -0.81$); thus, moving along the realism frontier simultaneously improves fairness. The 5 × 5 mixer forms this joint frontier, whereas hybrids seeded with StyleGAN3 lag because they tighten demographic coupling despite similar FID.

Recognition performance. Training IResNet-34 on the `5×5_best_sampled_10k` corpus (500 k images) yields 84.7 % mean accuracy over six verification sets, 79.6 % on RFW, an RFW dispersion of 3.78 pp and a skewed-error ratio of 1.65 pp. When averaged over the five classical benchmarks, accuracy rises to 86.8 %. These results match the 1.22 M-image DigiFace-1M baseline while halving the data budget and outperforming all GAN-based pipelines by 7–12 pp.

Efficiency and trade-offs. Normalising accuracy by dataset size shows that the balanced diffusion corpus converts each million images into 174 benchmark points, compared with 71 for DigiFace-1M and 16 for VariFace. Grid size and exemplar quality act as secondary fairness levers once seeds are balanced; deeper backbones and alternative loss heads give marginal returns.

Answers to the Research Questions

RQ1 — Capability of conditional diffusion models to synthesize racially unbiased datasets The combination of a 5×5 DCFace grid with uniform seed stratification and high-confidence exemplar selection advances the fairness frontier to $\mathrm{ENS}_{\mathrm{race}} \approx 3.8$ with $V_{\mathrm{race,label}} \leq 0.35$, surpassing the diversity of RFW while approximately halving its stereotype metric. This outcome is achieved without degrading image realism, as demonstrated by $\mathrm{FID}_{\mathrm{LFW}} = 14.2$ and $\mathrm{FID}_{\mathrm{RFW}} = 11.9$ for the flagship set.

RQ2 — Impact of training on such data on fairness and accuracy Balanced Score-Sampled seeding combined with cross-race style mixing reduces RFW dispersion from 5.1 pp to 3.8 pp and raises mean RFW accuracy from 75.8% to 79.6%. Across six public benchmarks, the same model achieves a mean accuracy of 84.7%, representing a 2.4 pp improvement over unbalanced baselines. These results indicate that fairness enhancements are obtained with negligible—and often positive—effects on accuracy.

RQ3 — Comparison of synthetic training with conventional practice Under a cap of 0.5 M images, the diffusion-derived corpus matches DigiFace-1M in headline accuracy and exceeds all published GAN or graphics pipelines. Its accuracy-per-image density of 174 points/M is 2.5×

that of DigiFace-1M and an order of magnitude greater than that of VariFace. This comparison confirms that data balance and exemplar quality outweigh sheer volume once the generator covers dominant modes.

Limitations

Despite measurable progress, several constraints persist. First, age coverage remains incomplete; the most balanced datasets still omit approximately three age brackets and exhibit uneven frequency distributions, limiting generalisability to older and younger cohorts. Second, residual race–identity entanglement stabilizes at $V_{\text{race,label}} \approx 0.33$; this is below RFW but above the weak-association threshold of 0.30, indicating incomplete disentanglement. Third, the pipeline assumes a binary gender taxonomy; non-binary or fluid identities are not represented. Fourth, diffusion sampling entails substantially higher computing than GAN-based synthesis, which constrains scalability for institutions with limited hardware. Fifth, evaluation is confined to still-image verification; open-set identification, video matching, and adversarial robustness remain untested. Finally, fairness metrics are axis-wise; intersectional effects (e.g., race–age–gender) and sociotechnical factors were outside scopes.

Future Work

This research can be extended in several directions. Age-conditioned embeddings and stratified sampling schedules could close the present coverage gap while enabling longitudinal aging studies. Integrating adversarial demographic heads during diffusion training may push $V_{\text{race,label}}$ below 0.30 and equalize false-positive rates across protected groups. Intersectional fairness auditing that jointly models race, gender, and age would reveal compound biases hidden by single-axis statistics. Energy-efficient variants—such as latent-distilled or ADM-style samplers—could cut generation costs by an order of magnitude and unlock on-device fine-tuning. Extending dual-condition diffusion to 3-D-aware models would guarantee viewpoint consistency and support augmentation for extreme poses and occlusions. Beyond faces, a multi-modal extension synthesizing matched voice or gait samples under a shared demographic prior would facilitate fair biometric fusion. Evaluating the pipeline on open-set identification, adversarial attack resilience, and template privacy would provide a more complete picture of performance. Finally, exploring additional backbone architectures and experimenting diverse loss-heads may further enhance robustness and fairness.

References

[1] Amal Adouani, Wiem Mimoun Ben Henia, and Zied Lachiri. "Comparison of Haar-like, HOG and LBP approaches for face detection in video sequences". In: *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)*. ISSN: 2474-0446. Mar. 2019, pp. 266–271. DOI: 10.1109/SSD.2019.8893214. URL: https://ieeexplore.ieee.org/document/8893214 (visited on 11/26/2024).

[2] Vitor Albiero et al. "Bias in Face Recognition: From Causes to Mitigation". AAI29254537. PhD thesis. USA: University of Notre Dame, 2022. ISBN: 9798841728979.

[3] A. Amirkhani et al. "Robust Semantic Segmentation with Multi-Teacher Knowledge Distillation". In: *IEEE Access* 9 (2021), pp. 119049–119066. DOI: 10.1109/ACCESS.2021.3107841.

[4] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. *OpenFace: A general-purpose face recognition library with mobile applications*. en. Tech. rep. CMU-CS-16-118. Pittsburgh, PA: CMU School of Computer Science, June 2016. URL: https://elijah.cs.cmu.edu/DOCS/CMU-CS-16-118.pdf (visited on 11/30/2024).

[5] Martin Arjovsky and Léon Bottou. *Towards Principled Methods for Training Generative Adversarial Networks*. arXiv:1701.04862 [stat]. Jan. 2017. DOI: 10.48550/arXiv.1701.04862. URL: http://arxiv.org/abs/1701.04862 (visited on 02/26/2025).

[6] Gwangbin Bae et al. "DigiFace-1M: 1 Million Digital Face Images for Face Recognition". In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 3515–3524. DOI: 10.1109/WACV56688.2023.00352.

[7] Fadi Boutros. "Efficient and High Performing Biometrics: Towards Enabling Recognition in Embedded Domains". PhD thesis. Technical University of Darmstadt, July 2022. DOI: 10.26083/tuprints-00021571.

[8] Fadi Boutros et al. "IDiff-Face: Synthetic-based Face Recognition through Fizzy Identity-Conditioned Diffusion Models". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. ISSN: 2380-7504. Oct. 2023, pp. 19593–19604. DOI: 10.1109/ICCV51070.2023.01800. URL: https://ieeexplore.ieee.org/document/10378084 (visited on 01/12/2025).

[9] Fadi Boutros et al. *SFace: Privacy-friendly and Accurate Face Recognition using Synthetic Data*. arXiv:2206.10520. June 2022. DOI: 10.48550/arXiv.2206.10520. URL: http://arxiv.org/abs/2206.10520 (visited on 12/03/2024).

[10] Fadi Boutros et al. *Unsupervised Face Recognition using Unlabeled Synthetic Data*. 2022. arXiv: 2211.07371 [cs.CV]. URL: https://arxiv.org/abs/2211.07371.

[11] Martins Bruveris et al. "Reducing Geographic Performance Differentials for Face Recognition". In: *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. 2020, pp. 98–106. DOI: 10.1109/WACVW50321.2020.9096930.

[12] Eduarda Caldeira. "Adapting Biased Teachers for Fair Knowledge Distillation in Face Recognition". English. PhD thesis. Faculdade de Engenharia da Universidade do Porto, July 2024.

[13] Eduarda Caldeira et al. *MST-KD: Multiple Specialized Teachers Knowledge Distillation for Fair Face Recognition*. arXiv:2408.16563. Aug. 2024. DOI: 10.48550/arXiv.2408.16563. URL: http://arxiv.org/abs/2408.16563 (visited on 10/26/2024).

[14] Jalil Chavez-Galaviz and Nina Mahmoudian. "Efficient Underwater Docking Detection using Knowledge Distillation and Artificial Image Generation". In: *2022 IEEE/OES Autonomous Underwater Vehicles Symposium (AUV)*. 2022, pp. 1–7. DOI: 10.1109/AUV53081.2022.9965804.

[15] Child Criminal Safety Center Home. *Missing and Abducted Children*. https://childsafety.losangelescriminallawyer.pro/missing-and-abducted-children.html. Accessed: 2024-09-29. 2024.

[16] Yoojin Choi et al. "Data-Free Network Quantization With Adversarial Knowledge Distillation". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020, pp. 3047–3057. DOI: 10.1109/CVPRW50498.2020.00363.

[17] J. Coe and M. Atay. "Evaluating impact of race in facial recognition across machine learning and deep learning algorithms". In: *Computers* 10.9 (2021). DOI: 10.3390/computers10090113.

[18] Ekin D. Cubuk et al. *RandAugment: Practical automated data augmentation with a reduced search space*. arXiv:1909.13719 [cs]. Nov. 2019. DOI: 10.48550/arXiv.1909.13719. URL: http://arxiv.org/abs/1909.13719 (visited on 05/27/2025).

[19] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. ISSN: 1063-6919. June 2005, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177. URL: https://ieeexplore.ieee.org/document/1467360 (visited on 11/26/2024).

[20] J. Deng et al. "Retinaface: Single-shot multi-level face localisation in the wild". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2020, pp. 5202–5211. DOI: 10.1109/CVPR42600.2020.00525.

[21] Jiankang Deng et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 2575-7075. June 2019, pp. 4685–4694. DOI: 10.1109/CVPR.2019.00482. URL: https://ieeexplore.ieee.org/document/8953658 (visited on 11/30/2024).

[22] Prafulla Dhariwal and Alex Nichol. "Diffusion models beat GANs on image synthesis". In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., Dec. 2021, pp. 8780–8794. ISBN: 978-1-7138-4539-3. (Visited on 02/26/2025).

[23] Prafulla Dhariwal and Alex Nichol. "Diffusion models beat GANs on image synthesis". In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021. ISBN: 9781713845393.

[24] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. "Metrics for Dataset Demographic Bias: A Case Study on Facial Expression Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.8 (Aug. 2024), pp. 5209–5226. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2024.3361979. URL: https://ieeexplore.ieee.org/abstract/document/10420507 (visited on 01/04/2025).

[25] Samuel Dooley et al. “Rethinking bias mitigation: fairer architectures make for fairer face recognition”. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS ’23. New Orleans, LA, USA: Curran Associates Inc., 2024.

[26] Yuantao Feng et al. “Detect Faces Efficiently: A Survey and Evaluations”. In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4.1 (Jan. 2022). Conference Name: IEEE Transactions on Biometrics, Behavior, and Identity Science, pp. 1–18. ISSN: 2637-6407. DOI: 10.1109/TBIOM.2021.3120412. URL: https://ieeexplore.ieee.org/document/9580485 (visited on 11/26/2024).

[27] Raul Vicente Garcia et al. “The Harms of Demographic Bias in Deep Face Recognition Research”. In: *2019 International Conference on Biometrics (ICB)*. 2019, pp. 1–6. DOI: 10.1109/ICB45273.2019.8987334.

[28] Anjith George and Sébastien Marcel. “Digi2Real: Bridging the Realism Gap in Synthetic Data Face Recognition via Foundation Models”. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. ISSN: 2690-621X. Feb. 2025, pp. 1379–1388. DOI: 10.1109/WACVW65960.2025.00161. URL: https://ieeexplore.ieee.org/document/10972592 (visited on 06/13/2025).

[29] Sixue Gong, Xiaoming Liu, and Anil K. Jain. “Mitigating Face Recognition Bias via Group Adaptive Classifier”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 3413–3423. DOI: 10.1109/CVPR46437.2021.00342.

[30] Ian J. Goodfellow et al. *Generative Adversarial Networks*. arXiv:1406.2661 [stat]. June 2014. DOI: 10.48550/arXiv.1406.2661. URL: http://arxiv.org/abs/1406.2661 (visited on 02/26/2025).

[31] M. Gorpinich, O.Y. Bakhteev, and V.V. Strijov. “Gradient Methods for Optimizing Metaparameters in the Knowledge Distillation Problem”. In: *Automation and Remote Control* 83.10 (2022), pp. 1544–1554. DOI: 10.1134/S0005117922010071.

[32] Yandong Guo et al. *MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition*. 2016. URL: https://arxiv.org/abs/1607.08221.

[33] Kaiming He et al. *Deep Residual Learning for Image Recognition*. Dec. 2015. DOI: 10.48550/arXiv.1512.03385. URL: http://arxiv.org/abs/1512.03385 (visited on 05/24/2025).

[34] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919. June 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90. URL: https://ieeexplore.ieee.org/document/7780459 (visited on 11/30/2024).

[35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., Dec. 2020, pp. 6840–6851. ISBN: 978-1-7138-2954-6. (Visited on 02/26/2025).

[36] Jonathan Ho and Tim Salimans. *Classifier-Free Diffusion Guidance*. July 2022. DOI: 10.48550/arXiv.2207.12598. URL: http://arxiv.org/abs/2207.12598 (visited on 05/24/2025).

[37] Rachel Hong, Tadayoshi Kohno, and Jamie Morgenstern. "Evaluation of targeted dataset collection on racial equity in face recognition". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '23. New York, NY, USA: Association for Computing Machinery, Aug. 2023, pp. 531–541. ISBN: 9798400702310. DOI: 10.1145/3600211.3604662. URL: https://dl.acm.org/doi/10.1145/3600211.3604662 (visited on 10/26/2024).

[38] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. Mar. 2015. DOI: 10.48550/arXiv.1502.03167. URL: http://arxiv.org/abs/1502.03167 (visited on 05/24/2025).

[39] Kimmo Karkkainen and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation". en. In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, HI, USA: IEEE, Jan. 2021, pp. 1547–1557. ISBN: 978-1-6654-0477-8. DOI: 10.1109/WACV48630.2021.00159. URL: https://ieeexplore.ieee.org/document/9423296/ (visited on 12/03/2024).

[40] Supreet Kaur and Dharamveer Sharma. "Comparative Study of Face Detection Using Cascaded Haar, Hog and MTCNN Algorithms". In: *2023 3rd International Conference on Advancement in Electronics & Communication Engineering (AECE)*. Nov. 2023, pp. 536–541. DOI: 10.1109/AECE59614.2023.10428242. URL: https://ieeexplore.ieee.org/document/10428242 (visited on 11/26/2024).

[41] H. Kawai et al. "FSErasing: Improving Face Recognition with Data Augmentation Using Face Parsing". In: *IET Biometrics* 2024.1 (2024). DOI: 10.1049/2024/6663315.

[42] Asifullah Khan et al. *A Survey of the Recent Architectures of Deep Convolutional Neural Networks*. arXiv:1901.06032. May 2020. DOI: 10.48550/arXiv.1901.06032. URL: http://arxiv.org/abs/1901.06032 (visited on 11/26/2024).

[43] Minchul Kim, Anil K. Jain, and Xiaoming Liu. *AdaFace: Quality Adaptive Margin for Face Recognition*. Feb. 2023. DOI: 10.48550/arXiv.2204.00964. URL: http://arxiv.org/abs/2204.00964 (visited on 05/24/2025).

[44] Minchul Kim et al. "DCFace: Synthetic Face Generation with Dual Condition Diffusion Model". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 2575-7075. June 2023, pp. 12715–12725. DOI: 10.1109/CVPR52729.2023.01223. URL: https://ieeexplore.ieee.org/document/10204758 (visited on 01/12/2025).

[45] J. Kolberg et al. "On the Potential of Algorithm Fusion for Demographic Bias Mitigation in Face Recognition". In: *IET Biometrics* 2024 (2024). DOI: 10.1049/2024/1808587.

[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'12. Red Hook, NY, USA: Curran Associates Inc., Dec. 2012, pp. 1097–1105. (Visited on 11/30/2024).

[47] Y. Lecun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (Nov. 1998), pp. 2278–2324. ISSN: 1558-2256. DOI: 10.1109/5.726791. URL: https://ieeexplore.ieee.org/document/726791 (visited on 11/30/2024).

[48] Weiyang Liu et al. “SphereFace: Deep Hypersphere Embedding for Face Recognition”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017, pp. 6738–6746. DOI: 10.1109/CVPR.2017.713. URL: https://ieeexplore.ieee.org/document/8100196 (visited on 05/24/2025).

[49] Pietro Melzi et al. “GANDiffFace: Controllable Generation of Synthetic Datasets for Face Recognition with Realistic Variations”. In: *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. ISSN: 2473-9944. Oct. 2023, pp. 3078–3087. DOI: 10.1109/ICCVW60793.2023.00333. URL: https://ieeexplore.ieee.org/document/10350589 (visited on 02/18/2025).

[50] Mehdi Mirza and Simon Osindero. *Conditional Generative Adversarial Nets*. arXiv:1411.1784 [cs]. Nov. 2014. DOI: 10.48550/arXiv.1411.1784. URL: http://arxiv.org/abs/1411.1784 (visited on 05/27/2025).

[51] Loris Nanni et al. “Coupling RetinaFace and Depth Information to Filter False Positives”. en. In: *Applied Sciences* 13.5 (Jan. 2023). Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, p. 2987. ISSN: 2076-3417. DOI: 10.3390/app13052987. URL: https://www.mdpi.com/2076-3417/13/5/2987 (visited on 11/26/2024).

[52] Gaurav Kumar Nayak, Konda Reddy Mopuri, and Anirban Chakraborty. “Effectiveness of Arbitrary Transfer Sets for Data-free Knowledge Distillation”. In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2021, pp. 1429–1437. DOI: 10.1109/WACV48630.2021.00147.

[53] Pedro C. Neto et al. *Balancing Beyond Discrete Categories: Continuous Demographic Labels for Fair Face Recognition*. 2025. arXiv: 2506.01532 [cs.CV]. URL: https://arxiv.org/abs/2506.01532.

[54] Pedro C. Neto et al. *How Knowledge Distillation Mitigates the Synthetic Gap in Fair Face Recognition*. Aug. 2024. DOI: 10.48550/arXiv.2408.17399. URL: http://arxiv.org/abs/2408.17399 (visited on 11/30/2024).

[55] Anh Duc Nguyen et al. “EnsFace: An Ensemble Method of Deep Convolutional Neural Networks with Novel Effective Loss Functions for Face Recognition”. In: *Proceedings of the 11th International Symposium on Information and Communication Technology*. SoICT ’22. New York, NY, USA: Association for Computing Machinery, Dec. 2022, pp. 231–238. ISBN: 978-1-4503-9725-4. DOI: 10.1145/3568562.3568638. URL: https://doi.org/10.1145/3568562.3568638 (visited on 10/26/2024).

[56] Augustus Odena, Christopher Olah, and Jonathon Shlens. “Conditional Image Synthesis with Auxiliary Classifier GANs”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 2642–2651. URL: https://proceedings.mlr.press/v70/odena17a.html.

[57] Benjamin Orr et al. “Exploring Racial Bias in Deep Face Recognition Models”. In: *2024 Intermountain Engineering, Technology and Computing (IETC)*. 2024, pp. 92–97. DOI: 10.1109/IETC61393.2024.10564236.

[58] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. “Deep Face Recognition”. en. In: *Procedings of the British Machine Vision Conference 2015*. Swansea: British Machine Vision Association, 2015, pp. 41.1–41.12. ISBN: 978-1-901725-53-7. DOI: 10.5244/C.29.41. URL: http://www.bmva.org/bmvc/2015/papers/paper041/index.html (visited on 11/26/2024).

[59] Delong Qi et al. "YOLO5Face: Why Reinventing a Face Detector". In: *Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*. Tel Aviv, Israel: Springer-Verlag, 2022, pp. 228–244. ISBN: 978-3-031-25071-2. DOI: 10.1007/978-3-031-25072-9_15. URL: https://doi.org/10.1007/978-3-031-25072-9_15.

[60] Delong Qi et al. "YOLO5Face: Why Reinventing a Face Detector". en. In: *Computer Vision – ECCV 2022 Workshops*. Ed. by Leonid Karlinsky, Tomer Michaeli, and Ko Nishino. Cham: Springer Nature Switzerland, 2023, pp. 228–244. ISBN: 978-3-031-25072-9. DOI: 10.1007/978-3-031-25072-9_15.

[61] Haibo Qiu et al. *SynFace: Face Recognition with Synthetic Data*. Dec. 2021. DOI: 10.48550/arXiv.2108.07960. URL: http://arxiv.org/abs/2108.07960 (visited on 12/03/2024).

[62] J. Redmon et al. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2016-December. 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91.

[63] Joseph P Robinson et al. "Face Recognition: Too Bias, or Not Too Bias?" In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020, pp. 1–10. DOI: 10.1109/CVPRW50498.2020.00008.

[64] Robin Rombach et al. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 10674–10685. DOI: 10.1109/CVPR52688.2022.01042. URL: https://ieeexplore.ieee.org/document/9878449 (visited on 05/24/2025).

[65] Nataniel Ruiz, Eunji Chong, and James M. Rehg. "Fine-Grained Head Pose Estimation Without Keypoints". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2018.

[66] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A unified embedding for face recognition and clustering". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919. June 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682. URL: https://ieeexplore.ieee.org/document/7298682 (visited on 11/30/2024).

[67] Andrew Jason Shepley. *Deep Learning For Face Recognition: A Critical Analysis*. arXiv:1907.12739. July 2019. DOI: 10.48550/arXiv.1907.12739. URL: http://arxiv.org/abs/1907.12739 (visited on 11/26/2024).

[68] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv:1409.1556. Apr. 2015. DOI: 10.48550/arXiv.1409.1556. URL: http://arxiv.org/abs/1409.1556 (visited on 11/30/2024).

[69] Neha Singh and R.M. Brisilla. "Comparison Analysis of Different Face Detecting Techniques". In: *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*. Nov. 2021, pp. 1–6. DOI: 10.1109/i-PACT52855.2021.9696583. URL: https://ieeexplore.ieee.org/document/9696583 (visited on 11/26/2024).

[70] T. Sixta et al. "FairFace Challenge at ECCV 2020: Analyzing Bias in Face Recognition". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12540 LNCS (2020), pp. 463–481. DOI: 10.1007/978-3-030-65414-6_32.

[71] Yang Song et al. *Score-Based Generative Modeling through Stochastic Differential Equations*. Feb. 2021. DOI: 10.48550/arXiv.2011.13456. URL: http://arxiv.org/abs/2011.13456 (visited on 05/24/2025).

[72] Statista. *Facial Recognition - Worldwide*. https://www.statista.com/outlook/tmo/artificial-intelligence/computer-vision/facial-recognition/worldwide. Accessed: 2024-09-29. 2024.

[73] StudyFinds. *Average American recorded by security cameras 238 times each week*. https://studyfinds.org/americans-security-cameras-study/. Accessed: 2024-09-29. 2024.

[74] A. Sumsion et al. "Surveying Racial Bias in Facial Recognition: Balancing Datasets and Algorithmic Enhancements". In: *Electronics (Switzerland)* 13.12 (2024). DOI: 10.3390/electronics13122317.

[75] Yaniv Taigman et al. "DeepFace: Closing the Gap to Human-Level Performance in Face Verification". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919. June 2014, pp. 1701–1708. DOI: 10.1109/CVPR.2014.220. URL: https://ieeexplore.ieee.org/document/6909616 (visited on 11/26/2024).

[76] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. Vol. 1. ISSN: 1063-6919. Dec. 2001, pp. I–I. DOI: 10.1109/CVPR.2001.990517. URL: https://ieeexplore.ieee.org/document/990517 (visited on 11/26/2024).

[77] Fu-En Wang et al. "MixFairFace: towards ultimate fairness via MixFair adapter in face recognition". In: *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN: 978-1-57735-880-0. DOI: 10.1609/aaai.v37i12.26699. URL: https://doi.org/10.1609/aaai.v37i12.26699.

[78] Hao Wang et al. "CosFace: Large Margin Cosine Loss for Deep Face Recognition". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 2018, pp. 5265–5274. DOI: 10.1109/CVPR.2018.00552. URL: https://ieeexplore.ieee.org/document/8578650 (visited on 05/24/2025).

[79] Mei Wang et al. "Racial Faces in the Wild: Reducing Racial Bias by Information Maximization Adaptation Network". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. ISSN: 2380-7504. Oct. 2019, pp. 692–702. DOI: 10.1109/ICCV.2019.00078. URL: https://ieeexplore.ieee.org/document/9010843 (visited on 12/03/2024).

[80] World Population Review. *Missing Persons by State 2024*. https://worldpopulationreview.com/state-rankings/missing-persons-by-state. Accessed: 2024-09-29. 2024.

[81] Haiyu Wu et al. *Vec2Face: Scaling Face Dataset Generation with Loosely Constrained Vectors*. 2025. arXiv: 2409.02979 [cs.CV]. URL: https://arxiv.org/abs/2409.02979.

[82] Michael Yeung et al. *VariFace: Fair and Diverse Synthetic Dataset Generation for Face Recognition*. 2025. arXiv: 2412.06235 [cs.CV]. URL: https://arxiv.org/abs/2412.06235.

[83] Seyma Yucer et al. “Does lossy image compression affect racial bias within face recognition?” In: *2022 IEEE International Joint Conference on Biometrics (IJCB)*. 2022, pp. 1–10. DOI: 10.1109/IJCB54206.2022.10007956.

[84] Seyma Yucer et al. “Exploring Racial Bias within Face Recognition via per-subject Adversarially-Enabled Data Augmentation”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020, pp. 83–92. DOI: 10.1109/CVPRW50498.2020.00017.

[85] Matthew D. Zeiler and Rob Fergus. *Visualizing and Understanding Convolutional Networks*. arXiv:1311.2901. Nov. 2013. DOI: 10.48550/arXiv.1311.2901. URL: http://arxiv.org/abs/1311.2901 (visited on 11/30/2024).

[86] K. 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.

[87] Linfeng Zhang and Kaisheng Ma. *Accelerating Diffusion Models with One-to-Many Knowledge Distillation*. arXiv:2410.04191 [cs]. Oct. 2024. DOI: 10.48550/arXiv.2410.04191. URL: http://arxiv.org/abs/2410.04191 (visited on 06/13/2025).

Appendix A

Technical Implementation

A.1 Software Environment Requirements

The following table details the operating system and key library versions employed in each component of the experimental framework. Maintaining these exact versions ensured the reproducibility of both diffusion-based and GAN-based synthesis and consistency in downstream face recognition and alignment pipelines.

Environment	OS	CUDA	PyTorch	Python
IDiff-Face	Ubuntu 22.04 LTS	11.7	1.13.1	3.7
DCFace	Ubuntu 22.04 LTS	12.4	2.6.0	3.10
StyleGAN3	Ubuntu 22.04 LTS	11.1	1.9.1	3.8
Face Recognition	Ubuntu 22.04 LTS	12.4	2.6.0	3.8
Face Alignment	Ubuntu 22.04 LTS	12.4	2.6.0	3.8

Table A.1: Implementation details for each experimental environment used in the thesis.

A.2 GPU Hardware Specifications

This section summarizes the graphics processing units utilized during data synthesis, model training, and preprocessing stages. The count of each GPU model reflects the parallel computational resources allocated to large-scale generation and evaluation tasks.

Manufacturer	GPU Model	VRAM	Count
NVIDIA	A100 80GB PCIe	80 GB	2
NVIDIA	Tesla V100-SXM2-32GB	32 GB	2
NVIDIA	Quadro RTX 6000	24 GB	2
NVIDIA	TITAN Xp	12 GB	2
NVIDIA	GeForce RTX 2080 Ti	11 GB	2
NVIDIA	GeForce GTX 1080	8 GB	2

Table A.2: GPU hardware specifications used in the experiments.

Appendix B

Complete Fairness Metrics Assessment

Dataset	$ENS_{age}\uparrow$	$ENS_{race}\uparrow$	$ENS_{gender}\uparrow$	$ShDiv_{age}\uparrow$	$ShDiv_{race}\uparrow$	$ShDiv_{gender}\uparrow$	$SEI_{age}\uparrow$	$SEI_{race}\uparrow$	$SEI_{gender}\uparrow$	$BP_{age}\downarrow$	$BP_{race}\downarrow$	$BP_{gender}\downarrow$
3x3_best_10k	5.21	2.26	2	1.65	0.813	0.692	0.752	0.587	0.998	0.427	0.732	0.523
3x3_best_2k	5.55	2.32	1.96	1.71	0.841	0.675	0.78	0.606	0.974	0.427	0.733	0.595
3x3_best_gender_2k	5.76	2.33	1.97	1.75	0.845	0.68	0.797	0.61	0.981	0.406	0.731	0.58
3x3_best_sampled_10k	4.01	3.55	1.97	1.39	1.27	0.68	0.632	0.913	0.981	0.538	0.405	0.581
3x3_best_stylegan3_10k	4.56	2.09	2	1.52	0.738	0.693	0.691	0.532	0.999	0.465	0.769	0.514
3x3_random_10k	4.84	2.19	1.99	1.58	0.786	0.688	0.717	0.567	0.993	0.464	0.747	0.548
3x3_random_2k	5.54	2.33	1.99	1.71	0.845	0.689	0.779	0.61	0.994	0.418	0.73	0.544
3x3_random_gender_2k	5.46	2.26	1.98	1.7	0.815	0.685	0.773	0.588	0.989	0.433	0.742	0.563
3x3_random_sampled_10k	4.63	3.29	1.99	1.53	1.19	0.688	0.698	0.86	0.993	0.488	0.48	0.551
3x3_random_stylegan3_10k	4.25	2.03	2	1.45	0.706	0.691	0.659	0.509	0.997	0.502	0.785	0.534
3x3_sampled_per_race_10k_best	3.78	3.73	1.96	1.33	1.32	0.675	0.605	0.949	0.974	0.569	0.363	0.595
3x3_sampled_per_race_10k_random	4.88	2.12	1.99	1.59	0.751	0.687	0.722	0.542	0.991	0.464	0.762	0.555
3x3_sampled_per_race_10k_worst	3.46	3.36	1.91	1.24	1.21	0.647	0.565	0.875	0.933	0.611	0.424	0.651
3x3_worst_10k	4.79	2.07	1.99	1.57	0.729	0.687	0.713	0.526	0.991	0.475	0.77	0.556
3x3_worst_2k	6.01	2.18	**2**	1.79	0.781	**0.693**	0.816	0.563	**1**	0.378	0.755	**0.508**
3x3_worst_gender_2k	5.74	2.14	1.99	1.75	0.761	0.686	0.795	0.549	0.99	0.406	0.763	0.56
3x3_worst_sampled_10k	4.94	2.63	2	1.6	0.969	0.691	0.727	0.699	0.997	0.455	0.607	0.532
3x3_worst_stylegan3_10k	4.2	1.87	1.99	1.43	0.627	0.689	0.653	0.453	0.995	0.515	0.814	0.543
5x5_best_10k	5.05	3.05	1.99	1.62	1.11	0.687	0.737	0.803	0.991	0.413	0.559	0.557
5x5_best_2k	5.27	2.8	1.93	1.66	1.03	0.66	0.756	0.744	0.952	0.452	0.599	0.628
5x5_best_gender_2k	5.46	2.84	1.97	1.7	1.04	0.678	0.772	0.752	0.978	0.429	0.59	0.586
5x5_best_sampled_10k	4.16	3.49	1.96	1.43	1.25	0.674	0.649	0.902	0.972	0.508	0.426	0.598
5x5_best_stylegan3_10k	4.78	3.02	1.99	1.56	1.1	0.69	0.712	0.797	0.996	0.417	0.572	0.537
5x5_random_10k	4.7	2.65	1.98	1.55	0.976	0.684	0.704	0.704	0.987	0.467	0.653	0.568
5x5_random_2k	5.56	2.68	2	1.72	0.987	0.692	0.781	0.712	0.999	0.408	0.648	0.52
5x5_random_gender_2k	5.17	2.6	1.98	1.64	0.955	0.684	0.748	0.689	0.986	0.455	0.677	0.568
5x5_random_sampled_10k	4.51	3.29	1.98	1.51	1.19	0.685	0.686	0.858	0.988	0.49	0.489	0.564
5x5_random_stylegan3_10k	4.51	2.59	1.99	1.51	0.953	0.689	0.686	0.687	0.994	0.466	0.671	0.545
5x5_sampled_per_race_10k_best	3.92	3.81	1.95	1.37	1.34	0.67	0.622	0.965	0.967	0.545	**0.344**	0.607
5x5_sampled_per_race_10k_random	4.72	2.45	1.98	1.55	0.897	0.683	0.707	0.647	0.985	0.463	0.698	0.571
5x5_sampled_per_race_10k_worst	3.6	3.21	1.9	1.28	1.17	0.643	0.583	0.841	0.927	0.604	0.409	0.658
5x5_worst_10k	4.62	2.1	1.98	1.53	0.74	0.684	0.697	0.534	0.987	0.482	0.761	0.568
5x5_worst_2k	**6.77**	2.28	1.99	**1.91**	0.823	0.686	**0.87**	0.593	0.99	0.316	0.74	0.56
5x5_worst_gender_2k	6.19	2.39	1.99	1.82	0.871	0.686	0.829	0.628	0.99	0.378	0.719	0.558
5x5_worst_sampled_10k	4.82	2.55	1.99	1.57	0.937	0.69	0.716	0.676	0.995	0.468	0.642	0.54
5x5_worst_stylegan3_10k	4.42	1.95	1.99	1.49	0.668	0.689	0.677	0.482	0.994	0.483	0.796	0.547
cpd25_random_elasticface_ffhq_5000	6.42	2.38	2	1.86	0.868	0.692	0.846	0.626	0.999	0.374	0.714	0.523
cpd25_random_elasticface_lfw_5000	5.6	1.85	1.75	1.72	0.613	0.562	0.784	0.442	0.811	0.252	0.84	0.75
cpd25_random_synthetic_two_stage_5000	6.49	2.1	1.96	1.87	0.743	0.674	0.851	0.536	0.973	0.357	0.769	0.597
cpd25_random_synthetic_uniform_15000	5.63	2.24	1.96	1.73	0.805	0.675	0.787	0.58	0.973	0.338	0.76	0.596
cpd25_random_synthetic_uniform_5000	5.5	2.18	1.95	1.7	0.778	0.667	0.776	0.561	0.962	0.355	0.771	0.614
cpd50_random_elasticface_ffhq_5000	6.42	2.34	2	1.86	0.849	0.692	0.847	0.613	0.998	0.373	0.725	0.526
cpd50_random_elasticface_lfw_5000	5.69	1.85	1.78	1.74	0.617	0.578	0.792	0.445	0.833	**0.235**	0.838	0.736
cpd50_random_synthetic_two_stage_5000	6.41	2.08	1.96	1.86	0.731	0.674	0.845	0.527	0.972	0.372	0.774	0.599
cpd50_random_synthetic_uniform_15000	5.97	2.29	2	1.79	0.83	0.691	0.814	0.599	0.997	0.357	0.742	0.532
cpd50_random_synthetic_uniform_5000	6.08	2.29	1.98	1.81	0.827	0.685	0.822	0.596	0.989	0.35	0.743	0.563
LFW	5.37	1.99	1.68	1.68	0.689	0.521	0.765	0.497	0.752	0.282	0.809	0.784
RFW	4.8	**3.82**	1.89	1.57	**1.34**	0.638	0.713	**0.968**	0.921	0.425	0.358	0.664

Table B.1: Representational fairness metrics (3 significant digits).

Dataset	$V_{\text{age,label}}\downarrow$	$V_{\text{race,label}}\downarrow$	$V_{\text{gender,label}}\downarrow$	$\text{NMI}_{\text{age,label}}\downarrow$	$\text{NMI}_{\text{race,label}}\downarrow$	$\text{NMI}_{\text{gender,label}}\downarrow$	$Z_{\text{age,label}}\downarrow$	$Z_{\text{race,label}}\downarrow$	$Z_{\text{gender,label}}\downarrow$
3x3_best_10k	0.51	0.63	0.854	0.0678	0.0453	**0.0494**	-0.587	-0.347	**-3.34e-17**
3x3_best_2k	0.79	0.853	0.874	0.17	0.0921	0.0666	-0.737	-0.485	-9.48e-18
3x3_best_gender_2k	0.807	0.857	0.98	0.178	0.0928	0.0857	-0.738	-0.486	-3.14e-18
3x3_best_sampled_10k	0.468	0.758	0.868	0.05	0.0842	0.0504	-0.609	-0.389	9.36e-18
3x3_best_stylegan3_10k	0.495	0.541	0.862	0.0593	0.0335	0.0507	-0.607	-0.324	4.54e-18
3x3_random_10k	0.539	0.688	0.897	0.0744	0.0513	0.0557	-0.62	-0.379	2.68e-17
3x3_random_2k	0.807	0.865	0.877	0.173	0.0941	0.0683	-0.739	-0.488	2.23e-17
3x3_random_gender_2k	0.797	0.874	0.964	0.173	0.0928	0.0832	-0.74	-0.489	0
3x3_random_sampled_10k	0.544	0.689	0.892	0.0742	0.0666	0.0549	-0.627	-0.35	3.37e-18
3x3_random_stylegan3_10k	0.519	0.596	0.895	0.0642	0.0377	0.0557	-0.631	-0.355	1.14e-18
3x3_sampled_per_race_10k_best	0.478	0.86	0.892	0.0542	0.11	0.0538	-0.628	-0.435	-3.16e-17
3x3_sampled_per_race_10k_random	0.543	0.708	0.897	0.0766	0.0516	0.0556	-0.62	-0.401	8.75e-18
3x3_sampled_per_race_10k_worst	0.475	0.786	0.884	0.0534	0.0926	0.0506	-0.635	-0.407	-9.35e-18
3x3_worst_10k	0.531	0.717	0.893	0.0733	0.051	0.0548	-0.612	-0.407	2.07e-17
3x3_worst_2k	0.787	0.861	0.887	0.176	0.0869	0.0702	-0.732	-0.489	0
3x3_worst_gender_2k	0.813	0.874	0.953	0.175	0.086	0.0812	-0.736	-0.49	3.8e-18
3x3_worst_sampled_10k	0.541	0.471	0.879	0.0822	0.0325	0.0532	-0.619	-0.288	-2.02e-17
3x3_worst_stylegan3_10k	0.513	0.624	0.893	0.0632	0.0361	0.0552	-0.626	-0.388	3.14e-17
5x5_best_10k	0.355	0.371	0.886	0.0365	0.0201	0.0524	-0.499	-0.182	1.26e-18
5x5_best_2k	0.751	0.757	0.781	0.154	0.0878	0.0519	-0.732	-0.46	0
5x5_best_gender_2k	0.757	0.752	0.988	0.163	0.0871	0.0868	-0.733	-0.458	0
5x5_best_sampled_10k	**0.32**	0.47	0.898	**0.028**	0.0314	0.0533	-0.515	-0.234	-5.37e-17
5x5_best_stylegan3_10k	0.357	**0.33**	0.894	0.034	**0.0164**	0.0539	-0.515	-0.167	9.17e-18
5x5_random_10k	0.384	0.451	0.921	0.0436	0.0278	0.0583	-0.53	-0.237	-1.53e-17
5x5_random_2k	0.748	0.791	0.78	0.159	0.0922	0.0535	-0.728	-0.471	-2.03e-17
5x5_random_gender_2k	0.75	0.8	0.974	0.157	0.0926	0.0849	-0.735	-0.475	-2.46e-18
5x5_random_sampled_10k	0.387	0.461	0.92	0.0439	0.0306	0.0582	-0.54	-0.232	2.29e-17
5x5_random_stylegan3_10k	0.381	0.387	0.923	0.0394	0.0212	0.0589	-0.534	-0.212	-1.29e-17
5x5_sampled_per_race_10k_best	0.359	0.855	0.912	0.038	0.109	0.0558	-0.557	-0.434	-2.74e-17
5x5_sampled_per_race_10k_random	0.391	0.575	0.919	0.0453	0.0416	0.058	-0.536	-0.346	-6.1e-18
5x5_sampled_per_race_10k_worst	0.346	0.597	0.913	0.0333	0.0527	0.0533	-0.543	-0.323	-1.02e-17
5x5_worst_10k	0.377	0.53	0.917	0.042	0.0311	0.0574	-0.517	-0.321	1.7e-17
5x5_worst_2k	0.737	0.814	**0.777**	0.172	0.0834	0.0527	-0.713	-0.484	1.63e-17
5x5_worst_gender_2k	0.752	0.811	0.959	0.171	0.0878	0.0824	-0.723	-0.48	0
5x5_worst_sampled_10k	0.389	0.342	0.904	0.0479	0.0174	0.056	-0.518	-0.224	-2.87e-17
5x5_worst_stylegan3_10k	0.378	0.449	0.922	0.0391	0.0221	0.0585	-0.528	-0.301	7.15e-18
cpd25_random_elasticface_ffhq_5000	0.824	0.901	0.963	0.184	0.1	0.0832	-0.718	-0.481	-5.61e-18
cpd25_random_elasticface_lfw_5000	0.77	0.871	0.975	0.154	0.0681	0.07	-0.706	-0.488	6.87e-19
cpd25_random_synthetic_two_stage_5000	0.81	0.88	0.956	0.179	0.0842	0.0798	-0.71	-0.484	-6.38e-18
cpd25_random_synthetic_uniform_15000	0.647	0.715	0.836	0.116	0.0673	0.0589	-0.66	-0.451	-2.37e-18
cpd25_random_synthetic_uniform_5000	0.665	0.727	0.833	0.116	0.0667	0.0579	-0.666	-0.457	-1.85e-17
cpd50_random_elasticface_ffhq_5000	0.792	0.881	0.96	0.173	0.0955	0.0827	-0.706	-0.477	2.77e-18
cpd50_random_elasticface_lfw_5000	0.736	0.789	0.962	0.141	0.0605	0.0699	-0.687	-0.48	-3.66e-18
cpd50_random_synthetic_two_stage_5000	0.777	0.847	0.946	0.168	0.0798	0.078	-0.7	-0.48	-6.35e-18
cpd50_random_synthetic_uniform_15000	0.639	0.675	0.793	0.12	0.0639	0.0532	-0.65	-0.443	-1.34e-17
cpd50_random_synthetic_uniform_5000	0.643	0.674	0.806	0.121	0.0636	0.0549	-0.648	-0.443	4.55e-18
LFW	0.843	0.888	0.965	0.157	0.074	0.0618	**-0.76**	**-0.495**	1.39e-19
RFW	0.591	0.85	0.95	0.0702	0.0954	0.0552	-0.628	-0.42	-1.61e-18

Table B.2: Stereotypical fairness metrics (3 significant digits).

Appendix C

Complete Realism Metrics Assessment

Dataset	IS_mean ↑	IS_std ↑	FID_LFW ↓	KID_mean_LFW	KID_std_LFW	SSIM_LFW ↑	LPIPS_LFW ↓
3x3_best_10k	2.29	0.00671	98.6	0.124	0.0027	0.197	0.454
3x3_best_2k	2.87	0.0936	105	0.107	0.00278	0.2	0.45
3x3_best_gender_2k	2.98	0.131	103	0.103	0.0028	0.2	0.451
3x3_best_sampled_10k	2.26	0.0038	102	0.128	0.00269	0.192	0.457
3x3_best_stylegan3_10k	2.18	0.00526	100	0.129	0.00258	0.174	0.433
3x3_random_10k	2.25	0.00289	105	0.133	0.00361	0.195	0.449
3x3_random_2k	2.86	0.1	107	0.115	0.00315	0.194	0.449
3x3_random_gender_2k	2.92	0.12	110	0.119	0.00324	0.192	0.445
3x3_random_sampled_10k	2.28	0.00584	108	0.137	0.00286	0.19	0.449
3x3_random_stylegan3_10k	2.17	0.0061	106	0.135	0.00378	0.173	0.428
3x3_sampled_per_race_10k_best	2.3	0.00613	102	0.127	0.00242	0.19	0.458
3x3_sampled_per_race_10k_random	2.25	0.005	104	0.131	0.00255	0.199	0.445
3x3_sampled_per_race_10k_worst	2.3	0.00404	113	0.142	0.00354	0.19	0.457
3x3_worst_10k	2.3	0.00499	102	0.128	0.00344	0.196	0.449
3x3_worst_2k	3.01	0.12	105	0.109	0.00271	0.198	0.462
3x3_worst_gender_2k	3.02	0.122	107	0.113	0.00345	0.198	0.459
3x3_worst_sampled_10k	2.37	0.00588	107	0.133	0.00305	0.191	0.446
3x3_worst_stylegan3_10k	2.2	0.00656	104	0.13	0.0028	0.174	0.427
5x5_best_10k	2.34	0.00803	96.7	0.122	0.00255	0.198	0.457
5x5_best_2k	3.02	0.119	109	0.108	0.00278	0.2	0.458
5x5_best_gender_2k	3.09	0.134	107	0.104	0.0026	0.201	0.458
5x5_best_sampled_10k	2.32	0.00662	97.2	0.122	0.00254	0.192	0.461
5x5_best_stylegan3_10k	2.3	0.00691	96.2	0.122	0.00274	0.176	0.437
5x5_random_10k	2.35	0.00515	101	0.126	0.00365	0.197	0.451
5x5_random_2k	2.9	0.124	104	0.109	0.00283	0.196	0.455
5x5_random_gender_2k	2.95	0.0941	114	0.121	0.00328	0.192	0.45
5x5_random_sampled_10k	2.34	0.00564	100	0.125	0.00319	0.191	0.451
5x5_random_stylegan3_10k	2.31	0.00541	98.1	0.122	0.00329	0.175	0.43
5x5_sampled_per_race_10k_best	2.36	0.00416	98.2	0.122	0.00273	0.191	0.461
5x5_sampled_per_race_10k_random	2.33	0.0055	98	0.121	0.00336	0.2	0.448
5x5_sampled_per_race_10k_worst	2.4	0.00799	105	0.131	0.0035	0.192	0.46
5x5_worst_10k	2.43	0.00656	97.8	0.119	0.00296	0.197	0.451
5x5_worst_2k	3.1	**0.144**	103	0.104	0.00261	0.198	0.471
5x5_worst_gender_2k	3.17	0.136	108	0.113	0.00289	0.197	0.466
5x5_worst_sampled_10k	2.46	0.00663	98.2	0.119	0.0032	0.192	0.448
5x5_worst_stylegan3_10k	2.38	0.0107	96.7	0.118	0.00305	0.176	0.431
cpd25_random_elasticface_ffhq_5000	**3.21**	0.0618	108	0.114	0.00315	0.183	0.432
cpd25_random_elasticface_lfw_5000	2.49	0.0502	66.9	0.0648	0.00213	0.199	0.424
cpd25_random_synthetic_two_stage_5000	3.18	0.116	112	0.119	0.00368	0.183	0.429
cpd25_random_synthetic_uniform_15000	2.99	0.0392	90.4	0.1	0.00312	0.18	0.434
cpd25_random_synthetic_uniform_5000	3.03	0.0924	91	0.101	0.00291	0.18	0.434
cpd50_random_elasticface_ffhq_5000	3.07	0.0716	108	0.114	0.00332	0.185	0.427
cpd50_random_elasticface_lfw_5000	2.55	0.0717	69.2	0.0679	0.0023	0.198	0.424
cpd50_random_synthetic_two_stage_5000	3.17	0.0912	111	0.118	0.00349	0.185	0.43
cpd50_random_synthetic_uniform_15000	3.15	0.0508	99.8	0.111	0.00333	0.179	0.437
cpd50_random_synthetic_uniform_5000	3.17	0.0941	98.7	0.109	0.00361	0.179	0.436
LFW	2.02	0.049	**-2.67e-12**	-1.81e-05	7.67e-05	**0.304**	**0.36**
RFW	3.04	0.00602	44.7	0.0393	0.00148	0.207	0.444

Table C.1: Realism metrics on LFW for all synthetic datasets (3 significant digits).

Dataset	IS_mean ↑	IS_std ↑	FID_RFW ↓	KID_mean_RFW	KID_std_RFW	SSIM_RFW ↑	LPIPS_RFW ↓
3x3_best_10k	2.29	0.00671	62.2	0.0778	0.0022	0.208	0.456
3x3_best_2k	2.87	0.0936	55.8	0.0503	0.00213	0.198	0.464
3x3_best_gender_2k	2.98	0.131	54	0.048	0.00206	0.198	0.465
3x3_best_sampled_10k	2.26	0.0038	61.8	0.0774	0.00239	0.213	0.449
3x3_best_stylegan3_10k	2.18	0.00526	66	0.0829	0.00269	0.208	0.449
3x3_random_10k	2.25	0.00289	62	0.0768	0.00283	0.206	0.417
3x3_random_2k	2.86	0.1	57.2	0.0574	0.00262	0.191	0.452
3x3_random_gender_2k	2.92	0.12	58.7	0.0585	0.00267	0.19	0.444
3x3_random_sampled_10k	2.28	0.00584	63	0.0785	0.00264	0.207	0.418
3x3_random_stylegan3_10k	2.17	0.0061	64.8	0.0814	0.00289	0.206	**0.413**
3x3_sampled_per_race_10k_best	2.3	0.00613	61.3	0.0754	0.00238	0.207	0.449
3x3_sampled_per_race_10k_random	2.25	0.005	61.2	0.0754	0.00258	0.206	0.419
3x3_sampled_per_race_10k_worst	2.3	0.00404	63.5	0.078	0.00293	0.208	0.416
3x3_worst_10k	2.3	0.00499	59.4	0.0721	0.00236	0.207	0.42
3x3_worst_2k	3.01	0.12	56.2	0.0552	0.00243	0.189	0.472
3x3_worst_gender_2k	3.02	0.122	57.4	0.0571	0.00279	0.191	0.467
3x3_worst_sampled_10k	2.37	0.00588	61.7	0.0751	0.00235	0.214	0.418
3x3_worst_stylegan3_10k	2.2	0.00656	62.6	0.0775	0.00242	0.208	0.416
5x5_best_10k	2.34	0.00803	59	0.0747	0.00244	0.209	0.469
5x5_best_2k	3.02	0.119	56.6	0.0485	0.00226	0.198	0.471
5x5_best_gender_2k	3.09	0.134	55	0.0462	0.0021	0.198	0.471
5x5_best_sampled_10k	2.32	0.00662	57.5	0.0716	0.00234	0.209	0.465
5x5_best_stylegan3_10k	2.3	0.00691	59.7	0.075	0.00244	0.209	0.467
5x5_random_10k	2.35	0.00515	55.7	0.0683	0.00241	0.209	0.429
5x5_random_2k	2.9	0.124	53.6	0.0513	0.0026	0.192	0.457
5x5_random_gender_2k	2.95	0.0941	59.2	0.0579	0.00287	0.192	0.449
5x5_random_sampled_10k	2.34	0.00564	55.1	0.0676	0.00257	0.209	0.431
5x5_random_stylegan3_10k	2.31	0.00541	55.2	0.0678	0.00264	0.209	0.428
5x5_sampled_per_race_10k_best	2.36	0.00416	57.4	0.0704	0.00231	0.208	0.465
5x5_sampled_per_race_10k_random	2.33	0.0055	54.1	0.0663	0.00276	0.208	0.43
5x5_sampled_per_race_10k_worst	2.4	0.00799	56.3	0.0682	0.0024	0.21	0.427
5x5_worst_10k	2.43	0.00656	52.7	0.0635	0.00233	0.21	0.431
5x5_worst_2k	3.1	**0.144**	54.9	0.0508	0.00247	0.188	0.48
5x5_worst_gender_2k	3.17	0.136	57.4	0.055	0.00259	0.189	0.475
5x5_worst_sampled_10k	2.46	0.00663	52.5	0.0633	0.00227	**0.216**	0.43
5x5_worst_stylegan3_10k	2.38	0.0107	53.1	0.0637	0.00252	0.21	0.429
cpd25_random_elasticface_ffhq_5000	**3.21**	0.0618	57	0.0547	0.00263	0.173	0.437
cpd25_random_elasticface_lfw_5000	2.49	0.0502	34.6	0.0293	0.00161	0.182	0.434
cpd25_random_synthetic_two_stage_5000	3.18	0.116	60.7	0.0585	0.00264	0.175	0.435
cpd25_random_synthetic_uniform_15000	2.99	0.0392	46.2	0.0486	0.00193	0.174	0.437
cpd25_random_synthetic_uniform_5000	3.03	0.0924	46.6	0.0488	0.00231	0.174	0.437
cpd50_random_elasticface_ffhq_5000	3.07	0.0716	57.9	0.0557	0.00268	0.177	0.433
cpd50_random_elasticface_lfw_5000	2.55	0.0717	34.9	0.0302	0.0014	0.184	0.432
cpd50_random_synthetic_two_stage_5000	3.17	0.0912	60	0.0573	0.00259	0.177	0.434
cpd50_random_synthetic_uniform_15000	3.15	0.0508	51.3	0.0546	0.00233	0.172	0.44
cpd50_random_synthetic_uniform_5000	3.17	0.0941	50.4	0.0535	0.00255	0.172	0.439
LFW	2.02	0.049	44.7	0.0396	0.0014	0.186	0.484
RFW	3.04	0.00602	**-2.13e-12**	2.56e-05	0.000122	0.191	0.425

Table C.2: Realism metrics on RFW for all synthetic datasets (3 significant digits).

Appendix D

Raw Ethnicity Score Values

Dataset	Caucasian	Asian	Indian	African
10k_ids_dcface	0.892	0.859	0.655	0.808
10k_ids_stylegan3	0.848	0.717	0.541	0.700
3x3_best_10k	40.078	38.658	22.660	35.317
3x3_best_sampled_10k	47.783	48.322	30.304	48.660
3x3_best_sampled_per_race_10k	47.995	48.074	37.307	48.405
3x3_best_stylegan3_10k	39.299	31.510	18.795	27.668
3x3_random_10k	42.087	38.158	21.330	36.248
3x3_random_sampled_10k	42.184	37.485	20.247	34.917
3x3_random_sampled_per_race_10k	40.824	36.222	18.956	34.488
3x3_random_stylegan3_10k	41.246	31.152	17.612	29.255
3x3_worst_10k	42.842	38.074	20.890	36.093
3x3_worst_sampled_10k	24.007	17.524	14.554	17.231
3x3_worst_sampled_per_race_10k	46.321	45.190	30.682	45.671
3x3_worst_stylegan3_10k	42.108	30.704	16.800	28.588
5x5_best_10k	28.106	27.331	17.684	27.870
5x5_best_sampled_10k	40.286	35.932	20.278	39.196
5x5_best_sampled_per_race_10k	48.072	48.035	42.142	48.545
5x5_best_stylegan3_10k	27.822	22.388	15.794	23.272
5x5_random_10k	35.118	28.401	17.648	31.602
5x5_random_sampled_10k	36.839	27.311	16.620	29.355
5x5_random_sampled_per_race_10k	35.483	26.337	15.156	29.370
5x5_random_stylegan3_10k	34.386	22.427	15.172	25.855
5x5_worst_10k	38.725	29.247	17.544	31.251
5x5_worst_sampled_10k	23.661	13.404	13.350	16.143
5x5_worst_sampled_per_race_10k	43.345	34.885	24.306	40.783
5x5_worst_stylegan3_10k	38.768	22.314	15.153	25.114
cpd25_elasticface_LFW_5000	4.352	3.896	2.789	4.100
cpd50_elasticface_LFW_5000	4.278	3.241	2.390	3.480
ids_best_sampled_10k	0.994	0.995	0.916	0.994
ids_best_sampled_per_race_10k	0.994	0.995	0.916	0.994
ids_random_sampled_10k	0.891	0.855	0.644	0.793
ids_random_sampled_per_race_10k	0.888	0.860	0.649	0.801
ids_worst_sampled_10k	0.396	0.434	0.414	0.412
ids_worst_sampled_per_race_10k	0.990	0.991	0.915	0.982
LFW	1.999	1.583	1.317	1.409
RFW	41.460	44.233	35.781	42.801

Table D.1: Metrics for all synthetic datasets under the Protocol B protocol.

Dataset	Caucasian	Asian	Indian	African
10k_ids_dcface	0.811	0.765	0.461	0.691
10k_ids_stylegan3	0.742	0.552	0.315	0.530
3x3_best_10k	34.439	34.298	15.227	30.119
3x3_best_sampled_10k	45.445	46.768	23.480	47.110
3x3_best_sampled_per_race_10k	47.211	47.516	31.890	48.002
3x3_best_stylegan3_10k	33.757	25.755	11.372	21.344
3x3_random_10k	36.832	33.019	13.088	30.637
3x3_random_sampled_10k	37.202	32.331	12.346	29.048
3x3_random_sampled_per_race_10k	35.936	31.188	11.232	29.063
3x3_random_stylegan3_10k	35.768	24.406	9.423	22.037
3x3_worst_10k	37.829	32.771	12.496	30.304
3x3_worst_sampled_10k	15.002	9.928	6.638	8.856
3x3_worst_sampled_per_race_10k	44.257	42.729	23.710	43.504
3x3_worst_stylegan3_10k	36.895	23.795	8.576	21.077
5x5_best_10k	21.854	23.090	13.090	23.336
5x5_best_sampled_10k	35.475	31.650	15.709	34.872
5x5_best_sampled_per_race_10k	47.355	47.451	38.590	48.264
5x5_best_stylegan3_10k	21.689	18.230	11.272	18.916
5x5_random_10k	28.622	22.752	11.062	25.765
5x5_random_sampled_10k	30.918	21.582	10.301	23.332
5x5_random_sampled_per_race_10k	29.710	20.786	9.037	23.662
5x5_random_stylegan3_10k	27.927	16.761	8.767	19.526
5x5_worst_10k	32.625	22.787	9.827	24.576
5x5_worst_sampled_10k	15.118	7.196	6.177	8.330
5x5_worst_sampled_per_race_10k	39.682	29.038	16.897	36.602
5x5_worst_stylegan3_10k	32.538	15.488	7.652	17.423
cpd25_elasticface_LFW_5000	3.929	3.540	2.094	3.759
cpd50_elasticface_LFW_5000	3.878	2.932	1.703	3.096
ids_best_sampled_10k	0.989	0.991	0.841	0.988
ids_best_sampled_per_race_10k	0.989	0.991	0.841	0.988
ids_random_sampled_10k	0.809	0.758	0.447	0.668
ids_random_sampled_per_race_10k	0.804	0.766	0.451	0.682
ids_worst_sampled_10k	0.158	0.191	0.174	0.172
ids_worst_sampled_per_race_10k	0.981	0.982	0.852	0.966
LFW	1.769	1.398	1.002	1.235
RFW	38.491	43.092	33.618	40.906

Table D.2: Metrics for all synthetic datasets under the Fuzzy Ethnic Bias protocol.

Dataset	Caucasian	Asian	Indian	African
10k_ids_dcface	6724	1446	727	1108
10k_ids_stylegan3	6765	1163	877	1188
3x3_best_10k	308088	78802	53683	69284
3x3_best_sampled_10k	146730	131535	83635	145119
3x3_best_sampled_per_race_10k	133507	126321	95388	133104
3x3_best_stylegan3_10k	321413	64740	56194	60702
3x3_random_10k	324374	72556	43871	67037
3x3_random_sampled_10k	170508	117922	82480	133205
3x3_random_sampled_per_race_10k	318581	69005	40407	60597
3x3_random_stylegan3_10k	338256	59826	47652	61636
3x3_worst_10k	331506	71573	42006	62744
3x3_worst_sampled_10k	182093	97678	105341	124138
3x3_worst_sampled_per_race_10k	113844	182701	21004	170904
3x3_worst_stylegan3_10k	347344	57795	44736	56442
5x5_best_10k	221191	86554	84006	111722
5x5_best_sampled_10k	152386	117302	84458	152867
5x5_best_sampled_per_race_10k	128344	124290	107134	128539
5x5_best_stylegan3_10k	230986	78727	87790	106098
5x5_random_10k	276580	71275	62008	97845
5x5_random_sampled_10k	173480	99805	89232	141598
5x5_random_sampled_per_race_10k	286449	64464	53261	82995
5x5_random_stylegan3_10k	286574	61548	65120	91932
5x5_worst_10k	307623	66832	53932	74633
5x5_worst_sampled_10k	192776	83059	106210	127445
5x5_worst_sampled_per_race_10k	128891	147635	39905	171705
5x5_worst_stylegan3_10k	327134	54417	57741	67747
cpd25_elasticface_LFW_5000	7237	594	965	1203
cpd50_elasticface_LFW_5000	7324	629	873	1174
ids_best_sampled_10k	2561	2531	2307	2601
ids_best_sampled_per_race_10k	2561	2531	2307	2601
ids_random_sampled_10k	2881	2474	2086	2559
ids_random_sampled_per_race_10k	6735	1451	744	1070
ids_worst_sampled_10k	2678	2168	2577	2578
ids_worst_sampled_per_race_10k	2261	3939	292	3508
LFW	8652	1226	1898	1457
RFW	324463	324445	276327	323696

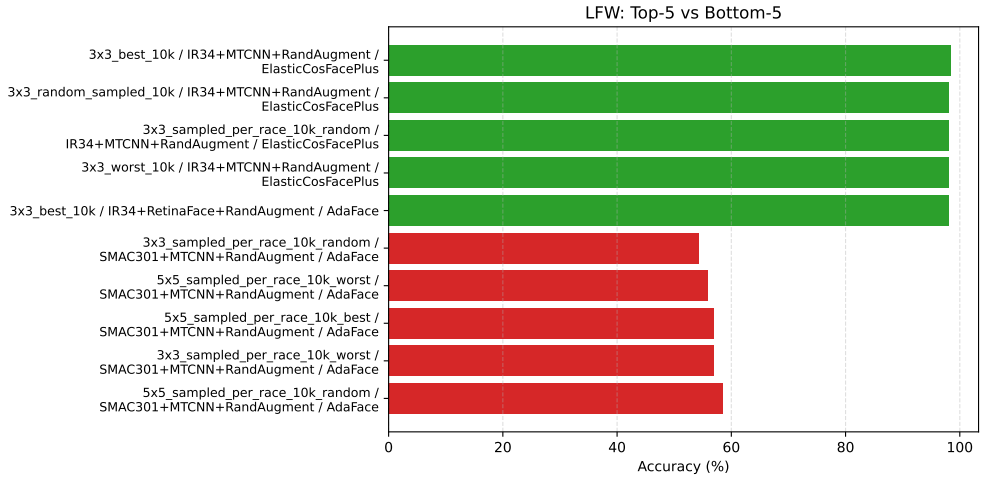
Table D.3: Metrics for all synthetic datasets under the Aggregated Diversity protocol.

Dataset	Caucasian	Asian	Indian	African
10k_ids_dcface	0.672	0.145	0.073	0.111
10k_ids_stylegan3	0.677	0.116	0.088	0.119
3x3_best_10k	0.604	0.155	0.105	0.136
3x3_best_sampled_10k	0.290	0.259	0.165	0.286
3x3_best_sampled_per_race_10k	0.274	0.259	0.195	0.273
3x3_best_stylegan3_10k	0.639	0.129	0.112	0.121
3x3_random_10k	0.639	0.143	0.086	0.132
3x3_random_sampled_10k	0.338	0.234	0.164	0.264
3x3_random_sampled_per_race_10k	0.652	0.141	0.083	0.124
3x3_random_stylegan3_10k	0.667	0.118	0.094	0.121
3x3_worst_10k	0.653	0.141	0.083	0.124
3x3_worst_sampled_10k	0.358	0.192	0.207	0.244
3x3_worst_sampled_per_race_10k	0.233	0.374	0.043	0.350
3x3_worst_stylegan3_10k	0.686	0.114	0.088	0.111
5x5_best_10k	0.439	0.172	0.167	0.222
5x5_best_sampled_10k	0.301	0.231	0.167	0.301
5x5_best_sampled_per_race_10k	0.263	0.254	0.219	0.263
5x5_best_stylegan3_10k	0.459	0.156	0.174	0.211
5x5_random_10k	0.545	0.140	0.122	0.193
5x5_random_sampled_10k	0.344	0.198	0.177	0.281
5x5_random_sampled_per_race_10k	0.589	0.132	0.109	0.170
5x5_random_stylegan3_10k	0.567	0.122	0.129	0.182
5x5_worst_10k	0.612	0.133	0.107	0.148
5x5_worst_sampled_10k	0.378	0.163	0.208	0.250
5x5_worst_sampled_per_race_10k	0.264	0.302	0.082	0.351
5x5_worst_stylegan3_10k	0.645	0.107	0.114	0.134
cpd25_elasticface_LFW_5000	0.724	0.059	0.097	0.120
cpd50_elasticface_LFW_5000	0.732	0.063	0.087	0.117
ids_best_sampled_10k	0.256	0.253	0.231	0.260
ids_best_sampled_per_race_10k	0.256	0.253	0.231	0.260
ids_random_sampled_10k	0.288	0.247	0.209	0.256
ids_random_sampled_per_race_10k	0.674	0.145	0.074	0.107
ids_worst_sampled_10k	0.268	0.217	0.258	0.258
ids_worst_sampled_per_race_10k	0.226	0.394	0.029	0.351
LFW	0.632	0.093	0.150	0.125
RFW	0.251	0.252	0.247	0.250

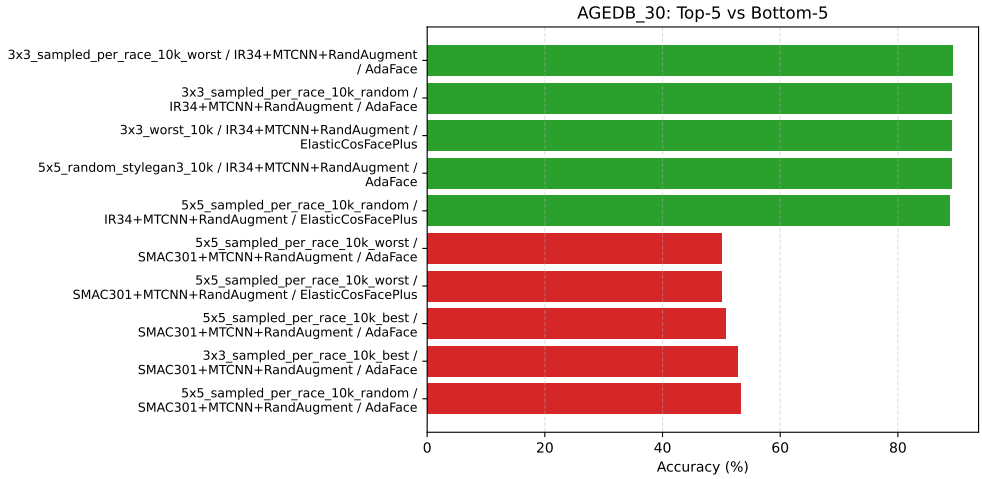
Table D.4: Metrics for all synthetic datasets under the Aggregated Diversity (normalized) protocol.

Appendix E

Benchmarks

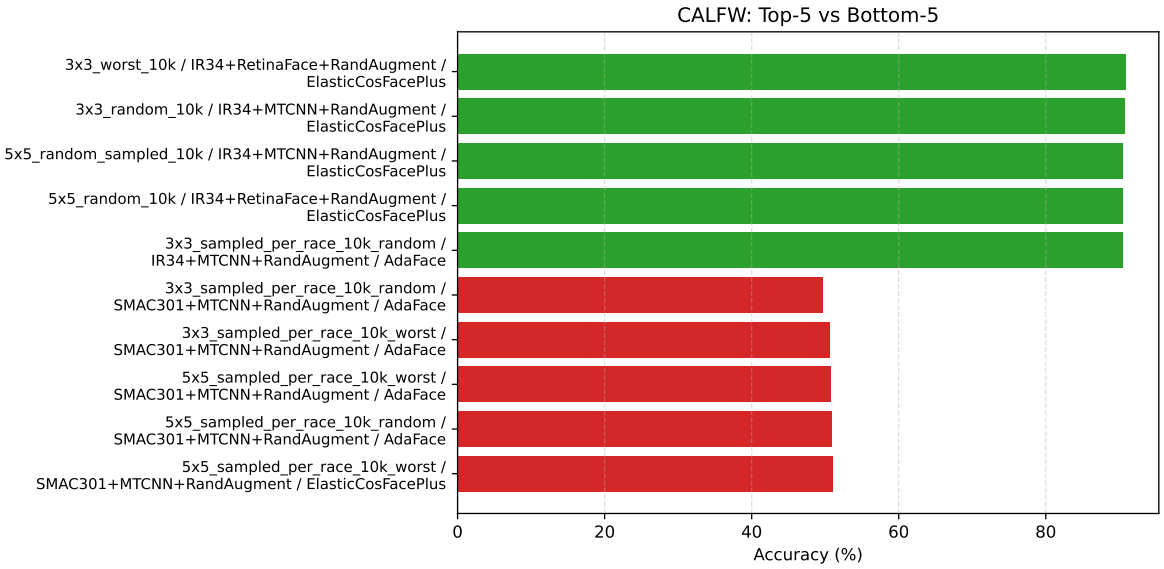


(a) LFW

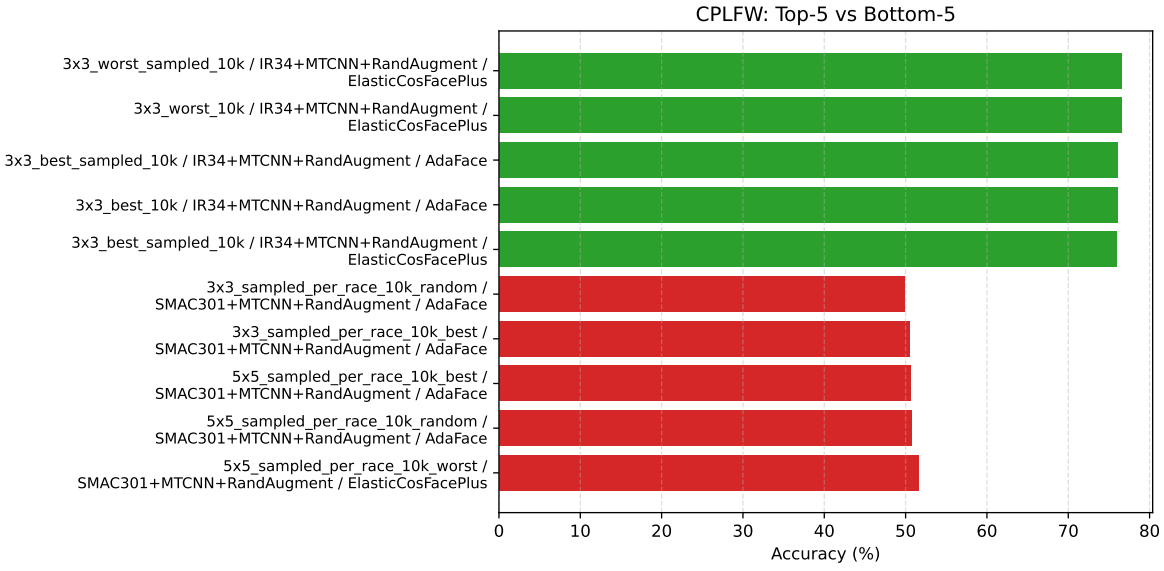


(b) AgeDB-30

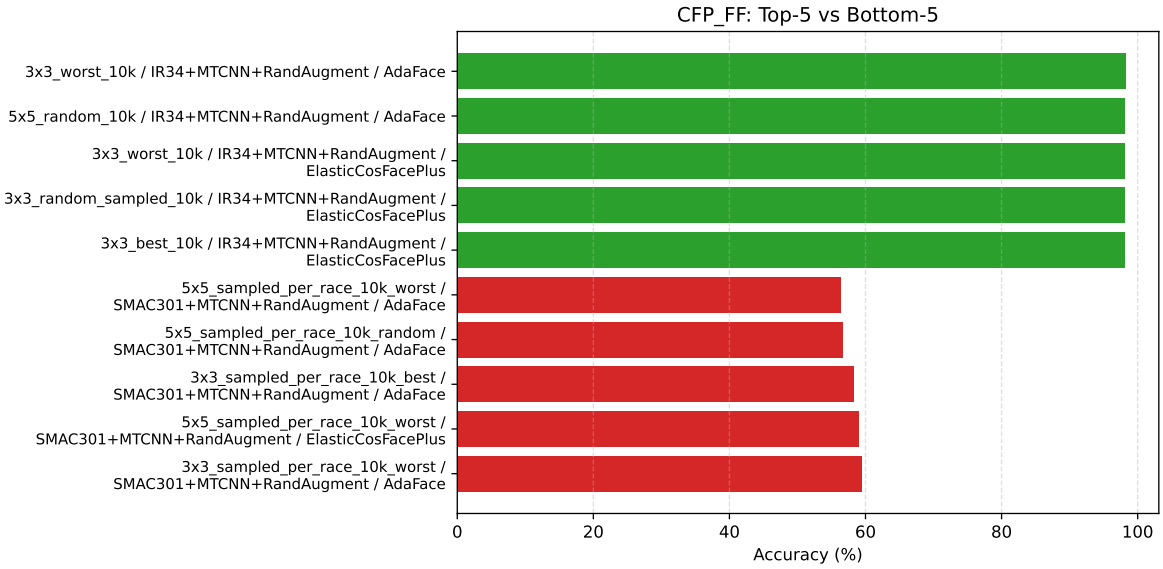
Figure E.1: Top-5 versus Bottom-5 verification accuracies for every public benchmark. Colors encode top (green) and bottom (red) quintiles.



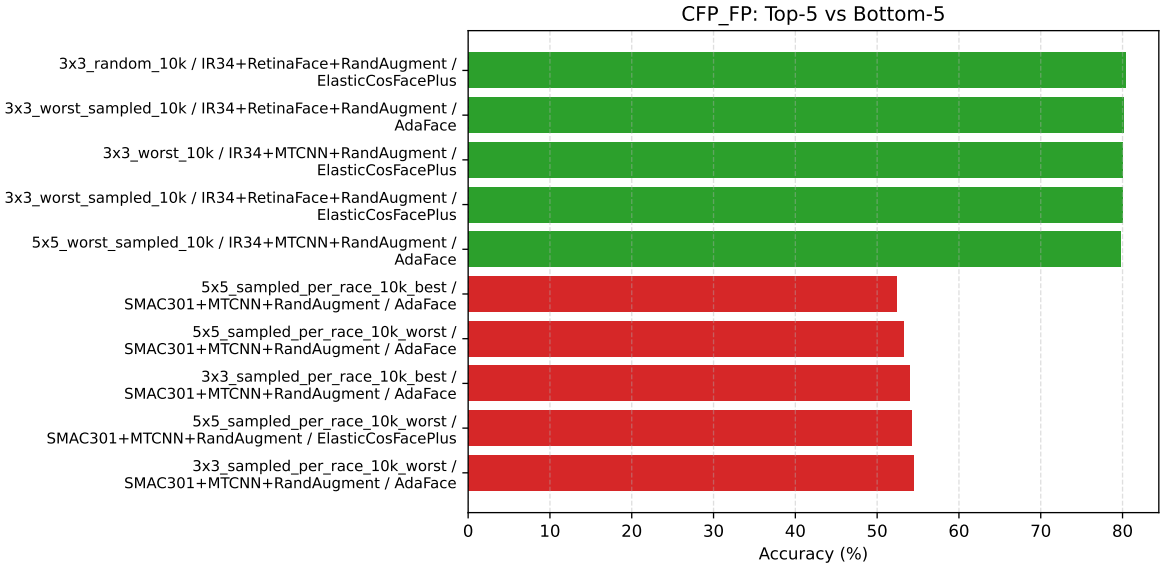
(c) CALFW



(d) CPLFW



(e) CFP-FF



(f) CFP-FP