Towards Fair Face Recognition: Mitigating Racial Bias via Generative Deep Learning

SEYMA YUCER-TEKTAS

Towards Fair Face Recognition: Mitigating Racial Bias via Generative Deep Learning

Seyma Yucer

A Thesis presented for the degree of
Doctor of Philosophy

Department of Computer Science
Durham University
United Kingdom
October 2023

Abstract

Facial recognition is one of the most academically studied and industrially developed areas within computer vision, where we readily find associated applications deployed globally. This widespread adoption has uncovered significant performance variation across subjects of different racial profiles leading to focused research attention on racial bias within face recognition spanning both current causation and potential future solutions. However, still the use of ill-defined racial categorisations, a lack of both consideration of the broader context of historical and social factors and contemporary evaluation methods hinder collaborative efforts towards mitigation of racial bias within face recognition. In support, this thesis firstly provides an extensive taxonomic review of research on racial bias within face recognition, covering topics from problem definition and racial grouping strategies to every aspect and all stages of the face recognition processing pipeline. Moreover, a comprehensive discussion within the review reveals the potential pitfalls and limitations of contemporary mitigation strategies that need to be considered within future research endeavours or commercial applications alike.

Accordingly, the prior literature has identified a need for alternative evaluation methodologies, particularly in the context of assessing racial bias. In response to this need, a phenotype-based racial bias analysis methodology is introduced via the use of a set of observable characteristics of an individual face where a race-related facial phenotype is hence specific to the human face and correlated to the racial profile of the subject. Subsequently, a commonplace lossy image compression algorithm impact at the initial stage of face recognition processing pipeline, image and dataset acquisition, concerning the racial characteristics of the subject, is investigated by adopting the proposed evaluation methodology. The results reveal the disparate performance decrease on specific racial phenotype categories and show improvement of the use of compressed imagery during training and removing chroma subsampling on the performance of specific racial phenotype categories more affected by lossy compression. Furthermore, a novel adversarial-derived data augmentation methodology is presented by aiming to enable dataset balance at a per-subject level via image-to-image transformation for the transfer of sensitive racial characteristic facial features to improve performance variation among racial and phenotype-based cate-

gories. The proposed approach decreases the performance variations between four racial groups by 15.81%.

Consequently, a novel GAN framework to enable fine-grained control over individual race-related phenotype attributes of the facial images is introduced. The proposed framework achieves both higher image quality and controllability on race-related facial phenotype attributes without requiring any synthetic or 3D data. Within the chapter, we introduce the CelebA-HQ-Augmented-Cleaned dataset, which is the first semi-synthesised, manually-cleaned, high-quality dataset encompassing over 26,500 images with a diverse distribution. Finally, this thesis concludes with an extensive discussion with insights draw from the literature, proposed approaches, and experiments presented throughout the thesis and outlines future directions for addressing racial bias within face recognition.

Declaration

The work in this thesis is based on research carried out at the Department of Computer Science, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

All glory and recognition go to Allah, the Almighty and Sovereign of the universe, for connecting me with exceptional supervisors, friends, family, and many more, enabling me to successfully complete this thesis.

This PhD would have been unimaginable without its supervisors, Toby Breckon and Noura Al Moubayed, who have also been its co-authors. Toby Breckon is not merely for his extensive technical knowledge and guidance but for his patience, calm, and open-minded behaviours towards all of his students from different backgrounds and cultures in the first place. Whilst he was always accessible to support me in many ways, he was also a great example of having a perfect balance between guiding his students and allowing them to take the initiative. His professionalism left a lasting impact on me who had the privilege of working with him. Noura, the reason for this PhD topic, not only did she provide guidance and support on the specific problems and questions that I brought to her, but she also helped me to ease the matters that I did not even mention. As researchers, as professional scholars and a perfect combination of supervisors who are nothing but great encouragers of this PhD, I could not have asked for more.

I am deeply thankful for the unrelenting love, support, and inspiration provided by my mother and father, which has manifested in the form of their prayers, overweight baggage of full of food via Istanbul trips, and frequent motivational conversations from a distance of over 2000 miles. Their influence has contributed significantly to me embrace of my Muslim identity, which values the pursuit of knowledge and personal refinement. I extend my heartfelt appreciation to my sisters, Rumeysa and Edibe, who are also living far away and pursuing similar paths to mine, as well as my best friends and sincere critics, for their support, and lighthearted phone calls.

I am grateful for the camaraderie and humor my office sisters and brothers, Yona Binti-abd-Gaus, Neelanjan Bhowmik, Brian Isaac-Medina, Ghada Alosaimi, Jack Barker, Judge Liu, Tom E, Luis Li, and Yixin Sun, bring to our workplace. Without them, our office would not be the lively and enjoyable place that it is. I would like to extend my sincere gratitude to my colleagues, Samet Akcay, Amir Atapour-Abarghouei, and Matthew Poyser for their invaluable collaboration as co-authors in our published works. Addition-

ally, I would also like to express a special thanks to my dear friend, Muna Almushyti, who has shown me that true friendship transcends language barriers. I am deeply grateful for every moment shared with her.

I am forever grateful to my husband, Furkan Tektas, who has been my lifelong collaborator, greatest gift, my rock, and my biggest supporter throughout my life journey. You have been an incredible partner in every sense of the word -a great listener who understands my thoughts and emotions, a fantastic co-author who helps me to bring my ideas to life, a masterful aesthetic critique who helps me to refine my work, and a helpful hand who supports me in every situation. I am truly blessed to have you by my side and to have your love and support. From the depths of my soul, I thank you for every lunchbox and dinner you have prepared, every illustration crisis you have handled, and every motivation-driven action you take, whether successful or not. Most importantly, thank you for recognising the value in this research work.

Contents

List of Figures

List of Tables

List of Symbols

Discriminator	D
Generator	G
Encoder	E
Image	x
Subject identity label	y
Race or race-related label	s
Image set	$X = \{x_1, \ldots, x_N\}$
Label set	$Y = \{y_1, \ldots, y_N\}$
Race set	$S = \{s_1, \ldots, s_N\}$
Image dataset	$I = \{X, Y, S\}$
Total number of images in the set	N
Mapping function	f
Mapping function space	Ω
Feature embedding	z
Mapped feature embedding	w
Loss function	L
Distance function	d
Set of learnable filters in each layer of DCNN	$W = \{W_1, \ldots, W_k\}$
Set of learnable biases in each layer of DCNN	$B = \{b_1, \ldots, b_k\}$
Number of layers in DCNN	k
Element-wise nonlinear transform	σ
Margin of ArcFace	m
Mutual Information	MI
Expected value over the distribution	$E_{p(z_i, z_j)}$
Joint Probability distribution	$P = p(z_i, z_j)$

Dedication

To the People of Palestine;

CHAPTER 1

Introduction

Numerous machine learning applications utilising facial attributes have proliferated in recent years as autonomous decision-making processes have become widely adopted by companies and governments [16]. A growing number of applications based on face recognition for surveillance [17], recruitment [18], and health-care [19] have increasingly become integrated into our daily lives. However, the generalisation of such research and applications is problematic due to the prevalence of bias within face recognition [20]. The imbalance in specific demographic groups occurring with varying phenotype attributes, including race, age or gender, poses a challenge for current causation and future potential solutions for facial recognition applications.

The most prevalent problem arises from the existence of disparate real-world performance on the race and race-related groupings which is referred to as racial bias within face recognition. This thesis aims to contribute to the racial bias literature by providing the first literature survey of the field providing an information spectrum from grouping definitions to their adoption through to the associated processing of racial groupings used in various literature studies. Accordingly, this thesis proposes a novel racial bias evaluation methodology, and two generative adversarial network-based bias mitigation framework which aim to reduce race-related bias within face recognition. As we show, such bias is often

inherited across consecutive stages of the face recognition pipeline resulting in increasingly bias decision making as an end result. By enhancing the variety of the face samples used in training, and considering different racial and race-related facial phenotype, our methods both improve the overall performance and decrease the performance variation across racial groups.

1.1 Motivation

Over several decades, the objective of developing face recognition systems has gathered significant pace across research, and industry alike [21–23]. Companies, nonprofits, and governments have deployed an increasing number of face recognition systems to make autonomous decisions for millions of users [24]. Such systems have been used across various application areas, such as within employment decisions, public security, criminal justice, law enforcement surveillance, airport passenger screening, and credit reporting [25–27]. However, such wide-scale adoption within real-world scenarios heightens public concern about the potential for abuse and the adverse effect face recognition may have on some individuals due to the presence of bias [28, 29]. The most prevalent problem pertaining to such bias arises within the race and race-related groupings and is referred to as racial bias within face recognition [30].

However, the presence of racial bias within face recognition is not a new thing and is not in itself limited to technological means. *Own-race bias* has been previously established in psychology [31] by showing that humans are less capable of recognising faces from other races than their own. The prolonged societal experience humans generally have with their own-race, especially during their formative years with biological family members, results in biased human perceptual expertise. More specifically, [32] showed how the use of facial feature descriptors varies across participants from different racial groupings. For example, the study shows that darker skin tone participants use face outline, eye size, eyebrows, chin and ears, while lighter skin tone participants use hair colour, texture, and eye colour. Overall, it concludes that lighter-skin-toned participants use less varied descriptors than darker-skin-toned participants [32]. Similar to the *own-race bias*, the conversely named *other-race effect* is also studied by a series of studies in social psy-

chology [33–35] to establish social implications of biased facial processing and feature selection in erroneous jury decisions, eyewitness identification.

Accordingly, the first technological study [36] that explores the other-race effect within the context of face recognition algorithms was developed by East Asian and Western-based research groups that inherently use datasets gathered locally. The study demonstrates that algorithms trained on a locally gathered facial datasets from the Western based group achieve superior performance on Caucasian faces when compared to performance on East Asian faces, and vice versa. Further studies provide extensive evidence about the influence of demographics, including race, gender and age, on both commercial and non-commercial face recognition algorithm performance [37, 38]. Subsequently, the Gender Shades study [39] drew significant attention to gender and skin tone bias within commercial algorithms for gender classification by revealing a 34% performance discrepancy between darker skin tone female and lighter skin tone male subjects. Consequently, growing research on faces has emerged to understand and mitigate racial bias within face recognition [9, 40, 41]. These efforts and associated evidence of bias have forced several commercial and academical research to withdraw products, algorithms, or datasets due to the different forms of disparities, distortions or biases [42–44].

However, face recognition remains a long-standing research topic and a common use case within computer vision that comprises multiple stages of processing, a multitude of downstream tasks and large-scale facial datasets in order to achieve high accuracy. With the availability of such large-scale data resources and the advent of Deep Convolutional Neural Networks (DCNN), the accuracy of face recognition algorithms has now excelled the perceived accuracy requirements for use by the general populous. However, every stage of face recognition, from initial face image acquisition to final performance evaluation, requires attention and investigation to address racial bias, which may otherwise result in disparate outcomes across a diverse user population. Unfortunately, despite the increasing attention to racial bias within face recognition, we are yet to see truly collaborative or tractable solutions emerge from the the global research base [11, 13, 45, 46] that could readily address these issues in real-world system deployments. Moreover, facial data itself is a private biometric capable of identifying a given individual based on their appearance alone, giving rise to obvious operational privacy and ethical concerns in rela-

tion to its processing [47]. Although previous studies on algorithmic bias and fairness in machine learning [48–50] and face recognition in computer vision and biometrics [23,24] exist, many aspects remain under-studied in relation to the specifics of racial bias within face recognition. Face recognition is a fast emerging field of research and applications alike that spans multiple more traditional fields, including machine learning, biometrics, statistics, sociology, psychology and anthropology. Therefore, this thesis aims to address the aspects of the racial bias problem definition, in addition to the race conceptualisation and race-related performance evaluation methodologies, and provides different approaches towards achieving the goal of mitigating racial bias within face recognition using conventional computer vision techniques.

1.2 Thesis Contribution

The work presented in this thesis contributes an advancement in knowledge in the following areas:

- A first comprehensive critical review of prior research on the topic of racial bias within face recognition. It provides a comprehensive coverage of the racial bias problem with respect to each and every stage of the face recognition processing pipeline whilst also highlighting the potential pitfalls and limitations of contemporary mitigation strategies that need to be considered within future research endeavours or commercial applications alike (Chapter 2).

- A novel racial bias analysis methodology via facial phenotype attributes for face recognition without reliance upon any potentially protected attributes or ill-defined grouping strategies (Chapter 3).

- An investigation of the impact of commonplace lossy image compression on face recognition algorithms with regard to the racial characteristics of the subject and the specific impact of chroma-subsampling on bias performance by comparing recognition performance with and without chroma-subsampling within lossy compressed facial imagery (Chapter 4).

- A novel adversarial derived data augmentation methodology that transfers racial attributes of a human face whilst preserving identity features in the face recognition datasets samples making face recognition algorithms more robust and less race-dependant (Chapter 5).

- A novel framework that factorises the latent space and explicitly control facial phenotype on a given face imagery by using only 2D imagery and related 2D metric-based parameters during training and only 2D imagery during inference (Chapter 6).

1.3 Publications

The research undertaken as part of this thesis has been published or is under review in the following peer-reviewed publication venues:

- Seyma Yucer, Amir Atapour, Noura Al Moubayed, and Toby P. Breckon., Disentangling Racial Phenotypes: Fine-Grained Control of Race-related Facial Phenotype Characteristics, The International Joint Conference on Neural Networks, IJCNN, 2024.

- Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby P. Breckon., Racial Bias within Face Recognition, ACM Computing Surveys, ACM CS, 2024, (Under Review).

- Seyma Yucer, Matthew Poyser, Noura Al Moubayed, and Toby P. Breckon., Does lossy image compression affect racial bias within face recognition?, IEEE International Joint Conference on Biometrics, IJCB, pp. 1-10, 2022.

- Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby P. Breckon., Measuring Hidden Bias within Face Recognition via Racial Phenotypes., IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, pp. 995-1004 2022.

- Seyma Yucer, Samet Akcay, Noura Al Moubayed, and Toby P. Breckon., Exploring Racial Bias within Face Recognition via per-subject Adversarially-Enabled Data Augmentation., IEEE/CVF Computer Vision and Pattern Recognition Workshops, CVPRW, pp. 18-19, 2020.

1.4 Thesis Structure and Scope

We present this thesis based on our work in the field and its contribution to the current state-of-the-art. Chapter 2 gives an extensive literature review where we formalise the problem definition with the corresponding evaluation, fairness criteria and discuss standard race and race-related grouping terminology. The discussion provides an information spectrum from grouping definitions to their adoption to the associated processing of grouping labels used in literature studies. Consequently, we provide a general development schema for face recognition systems and summarise the prior work in the field by aligning it to each development stage.

Chapter 3 introduces an alternative racial bias analysis methodology via facial phenotype attributes for face recognition. We define a set of observable characteristics of an individual face where a race-related facial phenotype is hence specific to the human face and correlated to the racial profile of the subject. Subsequently, we propose categorical test cases to investigate the individual influence of those attributes on bias within face recognition tasks. We compare our phenotype-based grouping methodology with previous grouping strategies and show that phenotype-based groupings uncover hidden bias without reliance upon any potentially protected attributes or ill-defined grouping strategies.

Chapter 4 examines the impact of commonplace lossy image compression on face recognition algorithms with regard to the racial characteristics of the subject. With the adoption of our racial phenotype-based bias analysis methodology, we measure the effect of varying levels of lossy compression across racial phenotype categories. Additionally, we determine the relationship between chroma-subsampling and race-related phenotypes for recognition performance.

Chapter 5 proposes a novel adversarial derived data augmentation methodology that aims to enable dataset balance at a per-subject level via the use of image-to-image transformation for the transfer of racial characteristic facial features. The aim is to automatically construct a synthesised dataset by transforming facial images across varying racial domains, while still preserving identity-related features, such that racially dependant features subsequently become irrelevant within the determination of subject identity.

Chapter 6 proposes a novel framework that leverages 2D imagery and related metric-

based parameters to control race-related facial phenotypes, as proposed in Chapter 3, including skin colour, hair colour, nose, eye, and shape. The objective of this framework is to enable explicit control of these race-related facial phenotype parameters using only 2D images. To achieve this, we utilise ConfigNET (Controllable Neural Face Image Generation) [51] and StyleGAN2 [52] and formalise race-related phenotype attributes using metric-based evaluations to map them into latent space. Our results demonstrate the efficacy of this approach in generating synthetic faces that exhibit specific race-related phenotypes with high fidelity.

Finally, Chapter 7 concludes with a summary of the techniques and their contributions and limitations within the field, along with the potential directions for future work.

1.5 Ethical Considerations

Intent: This PhD thesis intends to provide a comprehensive coverage of the topic: racial bias within face recognition. Our proposed novel racial bias analysis methodology, in Chapter 3, via facial phenotype attributes for face recognition avoids the need for researchers to use potentially protected or ill-defined subject attributes and instead introduces racial phenotype attributes to explore racial bias in face recognition.

Denotation of Facial Phenotypes: We denote race- related phenotype attributes according to the studies of [12, 53] to have descriptive naming whilst avoiding causing any unintended offence to individuals.

Use of Face Recognition Datasets: We conduct our experiments on various face datasets including VGGFace2 [3], BUPTBalanced [1], RFW [11], CelebA [54], CelebA-HQ [7], FFHQ [55] which are publicly available for research use only. The reader is directed to the original source publication in corresponding chapters and the associated research organisation for access to these datasets. We make available supplementary labels for VGGFace2 [3] and RFW [11] datasets in order to facilitate the use of our proposed methods and evaluation strategy by other researchers, with the aim of furthering our stated intent above.

Face Editing and Generation: Our main purpose in synthesising face imagery is to reduce the perpetuation of racial bias caused by imbalanced distributions in face recogni-

tion datasets. To avoid the potential misuse of the synthesised images, we have decided not to publicly share the generated data. Instead, we provide pre-trained models that require users to have access to the datasets (access to which is granted by individual dataset owners) in order to use pre-trained models.

CHAPTER 2

Literature Review

This chapter both summarises the current state of the art and gives a comprehensive critical review of prior research on the topic of racial bias within face recognition. In addition, this chapter aims to make the reader pertinently aware as to the subtleties, and potential areas of ambiguity, with regard to how the racial bias problem within face recognition itself is defined.

Furthermore, we identify which parts of the problem have been studied effectively to date and which directions remain open for future contributions to mitigate racial bias within the face recognition domain. In particular, we aim to systematically review each of the stages that are commonplace within contemporary face recognition processing pipelines from a perspective of the potential for racial bias impact: image acquisition (for both dataset collation and deployment), face localisation, face representation, face verification and identification (final decision-making) (see Figure 2.1, right). On this basis, we present this chapter based on our taxonomy of prior work in the field and its contribution to the current state of the art (Figure 2.1). Subsequently, we formalise the problem definition with the corresponding evaluation and fairness criteria (Section 2.1). Next, we discuss standard race and race-related grouping terminology under three categories; race, skin tone and facial phenotypes (Section 2.2). This discussion provides an

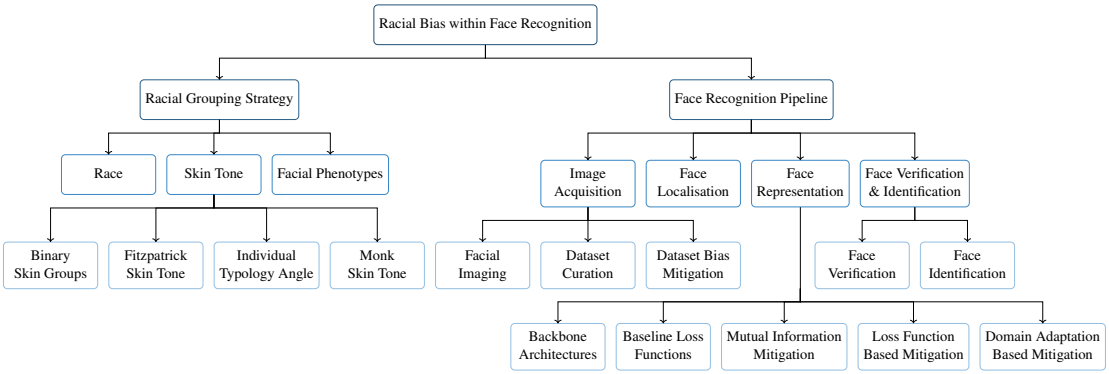


Figure 2.1: Taxonomy of sections of Racial Bias within Face Recognition literature review.

information spectrum from grouping definitions to their adoption to the associated processing of racial groupings used in literature studies. Consequently, we provide a general development schema for face recognition systems and summarise the prior work in the field by aligning it to each development stage (Section 2.3). Within this section (Section 2.3), we firstly give an outline description of the general face recognition processing pipeline using consistent notions and symbols.

Secondly, we cover image and dataset acquisition processes for face recognition showing the risks and investigations within this stage. Thirdly, we extend our analysis to face localisation as it is a mandatory stage where the possible biased localisation results propagate within the following face recognition stages. Penultimately, in the face representation stage, we categorise the proposed racial bias mitigation approaches based on machine learning techniques. Finally, we cover face identification and verification tasks and show the impact of the methodological decisions effects on racial bias. Consequently, we summarise the main critical points of the work and highlight the essential steps that need to be considered within any future research endeavours or commercial applications that aim to mitigate bias or develop fairer face recognition systems (Section 2.4).

The material presented in this chapter of the thesis has been submitted to the following peer-reviewed journal publication:

Seyma Yucer, Furkan Tektas, Noura Al Moubayed, Toby P. Breckon., Racial Bias within Face Recognition: A Survey., ACM Computing Surveys, ACM CS, 2023.

2.1 Preliminaries

Statistical methods are essential for supervised learning problems, including face recognition, which concerns generating a representative and distinctive feature embedding vector z for a subject y given an observed face image x. A mapping function f^* is a particular function among infinite function space Ω ($f^* \in \Omega$) that provides optimal performance over a given training dataset D_{train}. Preferring certain functions over others is denoted as inductive bias in the seminal work by Mitchell [56] and remains a central concept in statistical learning theory. The expression *inductive bias* (also known as learning bias) refers to the optimal selection process of f^*. Due to its importance for generalisation on unseen large-scale datasets, inductive bias is essential for any genre of machine learning approach. On the other hand, the broader societal, historical meaning of the term *bias* instead refers to the unfair treatment of a subset of the populous based on their origins, ethnicity or ideology. While *inductive bias* is necessary for model generalisation, *societal bias* implies negative implications that should ideally be avoided [57]. In order to avoid the obvious potential for confusion, the prior work of [58] prefers to use fairness instead of bias when referring to aspects of demographic criteria in both statistics and machine learning. Subsequently, research on algorithmic fairness and statistical bias has introduced various formal definitions of fairness, and their relationships with each other [58–60]. Before we fully detail these fairness criteria, we first provide a brief explanation of a generic face representation learning and evaluation pipeline to facilitate the introduction of the required notation, which we will subsequently use for the remainder of this review.

A face recognition system comprises a training set D_{train} and a test set D_{test} where any of the datasets can be defined as $D = \{X, Y\}$ where $X = \{x_1, x_2, .., x_N\}$ is a set of face images and $Y = \{y_1, y_2, .., y_N\}$ is a set of subject identity labels corresponding the face images where N is the total number of images. The total number of unique subject identity labels is n such that $n <= N$. In addition, in order to measure the fairness of a face recognition system, a set of corresponding race or race-related grouping labels S is also specified, $S = \{s_1, s_2, .., s_N\}$. Therefore any face dataset can be formed as $D = \{X, Y, S\}$ where X denotes the set of images, Y denotes the set of subject labels, and S denotes the set of sensitive race or race-related labels. Furthermore, a mapping

function f plays a significant role in face recognition systems as it maps any given image x into the feature embedding vector z. f is selected from a function space Ω via a loss function $\mathcal{L}$ which measures the performance of a given training set, D_{train}, for any of the aforementioned face recognition tasks. Typically, a softmax loss is adopted by state-of-the-art face recognition methods [4–6, 61] in order to disentangle the feature representation of individual identities within contemporary training datasets. The inductive representation learning is hence a minimisation of the loss function $\mathcal{L}_{softmax}$, which can be formalised as follows:

$$f^* = argmin(\mathcal{L}_{softmax}(f)), \quad f \in \Omega \quad \text{where} \quad \mathcal{L}_{softmax} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T z_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T z_i + b_j}} \tag{2.1}$$

where z_i is the feature representation of the image $x_i \in \mathbb{R}^{u \times v \times 3}$, u is the weight and v is the height of the x_i, within D_{train} belonging to subject class y_i and the number of samples is N labelled with n classes. W_j is the j^{th} column of the weights, b_j is the j^{th} column of the bias term, and d is the number of neurons in the last fully-connected layer which is mostly 512. Weights and bias term dimensions are $W_j \in \mathbb{R}^{dxn}$ and $b_j \in \mathbb{R}^n$, respectively. Moreover, the selected f^* compresses the intra-class distance and expands the inter-class distance between feature embeddings belonging to the same or different subject identity, respectively. Generally, f provides superior approximation over the statistically most predominant population subset within training set, D_{train}, such that $\mathcal{L}_{softmax}$ is minimised.

Additionally, evaluation metrics can quantify how well the selected f^* performs on D_{test}. The most common evaluation metric in face recognition, *accuracy*, relates to the probability of correctly predicting the subject label of a face image as $P(y_\alpha = \hat{y}_\alpha)$. Accuracy can be defined as follows,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.2}$$

where *true positive (TP)* is the number of the f^* correctly predicts the positive subject label and *true negative (TN)* is the number of the f^* correctly predicts the negative subject label. In contrast, *false positive (FP)* is the number of the f^* incorrectly predicts the positive subject label, and *false negative (FN)* is the number of the f^* incorrectly

predicts the negative subject label. Accuracy measures the consistency between predictions and their ground truth values. In a similar vein, the *True Match Rate (TMR)* estimates the number of correct positive predictions made from all possible positive predictions. For instance, a binary face verification task aims to classify whether an image pair $(x_\alpha, x_\beta) \quad where \quad x_\alpha, x_\beta \in D_{test}$ belongs to the same subject label or not. During testing, the selected f^* predicts the feature representation vectors z_α, z_β for the corresponding images x_α, x_β, respectively. Given images are validated as *"match"* if the similarity between two feature vectors (i.e. *cosine similarity*, $cos(z_\alpha, z_\beta) = \frac{z_\alpha \cdot z_\beta}{\|z_\alpha\| \|z_\beta\|}$) is greater than a given threshold parameter $threshold$, otherwise as *"non-match"*. *TMR* is the ratio of correctly verified match pairs (two different images from the same subject) over the total number of match pairs. However, neither *Accuracy* nor *TMR* is indicative of failure samples. To investigate such samples, the False Match Rate (FMR) measures how many incorrect non-match or negative predictions f^* are made via feature representation vectors. Furthermore, the False Non-Match Rate (FNMR) refers to the probability of samples of the same subject identity is incorrectly matched. All terms, TMR, TNMR, FMR, and FNMR, can be formalised as follows:

$$\text{TMR} = \frac{TP}{TP + FN}, \quad \text{TNMR} = \frac{TN}{TP + FN}, \quad \text{FMR} = \frac{FP}{FP + TN}, \quad \text{FNMR} = \frac{FN}{FP + TN} \tag{2.3}$$

Another facial recognition metric, the *ROC curve*, plots TMR against FMR at different thresholds. Lowering the $threshold$ verifies more items as matched, resulting in an increased *FMR* and *TMR*. Furthermore, the racial bias literature commonly measures the variation in performance, indicated by *accuracy* or *FMR*, among racial groups to highlight disparities within each group. However, calculating this deviation varies across studies, as different definitions of standard deviation are used (i.e. sample, population). In this study, we utilise the sample standard deviation for further analysis.

To this extent, we briefly described the selection process of f using the loss function and evaluation metrics of face recognition. Whilst, loss functions help to understand the behaviour of f on D_{train}, evaluation metrics help to measure how well the selected f^* maps D_{test} into feature embedding representation space. Consequently, statistical fairness

criteria can be considered as a formal property of face recognition systems, including mapping function f^*, training D_{train} and test datasets D_{test}. Accordingly, we give the four most commonly used fairness definitions from [60] that are commonplace within racial bias for face recognition.

Definition 1: ***Fairness Through Unawareness*** requires that a machine learning algorithm have an independent conditional probability P of the output given X from S (racial labels). Subsequently, unawareness criteria can be formalised as $P(Y|X) = P(Y|X, S)$. However, removing dependency is impossible for face recognition algorithms due to the high mutual information between facial and racial features. Even though racial labels are not explicitly introduced to the machine learning algorithm, they will implicitly be used in the face representation (algorithm training) via the facial images.

Definition 2: ***Individual Fairness*** refers to treating similar individuals coequally, meaning that an algorithm is fair if it gives similar predictions to similar individuals. In order to estimate such criteria, two distance metrics are defined by Dwork [59]. These are distance metrics that measure the degree of similarity between individual subjects and measure the difference in the associated prediction outcome between those individual subjects. It can be formalised in face recognition context as if image samples x_α and x_β are similar under a given distance metric $d(x_\alpha, x_\beta)$ depending on s_α, s_β then predictions should be similar $\hat{y}_\alpha \approx \hat{y}_\beta$ where $\hat{y}_\alpha$ and $\hat{y}_\beta$ are the predicted labels from corresponding images x_α, x_β and s_α, s_β are the sensitive race labels respectively. However, [62] discusses how individual fairness is inadequate for ensuring fairness on the grounds of four differing arguments, spanning the insufficiency of similar treatment, systematic bias and arbiters, prior moral judgements, and incommensurability (see [62] for a more detailed discussion).

Definition 3: ***Group fairness (or Statistical parity / Demographic parity*)** enforces the predicted subject labels $\hat{Y}$ to be independent of S which can be denoted $P(\hat{Y}|S = 0) = P(\hat{Y}|S = 1)$ where r is the number of different sensitive race labels in the set. Racial bias literature within the face recognition mostly approaches the problem from a supervised machine learning paradigm by considering it as an *group fairness criteria (demographic parity)* [59], which can be satisfied if the race or race-related intersectional groups perform similarly to each other. Unfortunately, such criteria may not ensure fairness as it heavily relies on equalising the acceptance match percentages even though there is little or no

training data available for a given racial grouping category within D_{train} [63].

Definition 4: ***Equal Opportunity, (or Equalised Odds)*** is satisfied if an algorithm predictions $\hat{Y}$ is independent of S conditioned on Y. If the criteria is defined for binary categories [63], it can be denoted $P(\hat{Y} = 1|S = 0, Y = y) = P(\hat{Y} = 1|S = 1, Y = y), y \in \{0, 1\}$. Subsequently, it is adopted by [64] to multiple class labels. More simply, the constraint requires that any sensitive race label has equal true positive rates and false positive rates across the other sensitive race labels. It also enforces that the *accuracy* is equally high in all sensitive labels, penalising algorithms that perform well solely on the statistically most predominant such labels. Furthermore, [63] discusses how demographic parity is crippled in the typical scenario in which the target variable Y is correlated with only S. On the other hand, *equalised odds* aims to achieve accurate prediction while ensuring predictions are fair concerning a specified sensitive labels, S.

As aforementioned, the literature has mainly used *statistical parity or group fairness criteria* to minimise the variation of *accuracy* or *FMR* across sensitive racial groupings labels on datasets. However, such an aim brings a high dependence on sensitive attributes to be used in fairness criteria above, which may actually increase discrimination [60]. Moreover, little attention has been given to how the sensitive attribute labels, S, are assigned, with regard to the potential for bias in the assignment (i.e. labelling) process, and what that potentially means normative *"unbiased"* presumptions for face recognition system design. In the next section, we address these questions by focusing on race and race-related groupings and their conceptualisation.

2.2 Towards Racial Group Fairness

Most studies on racial bias within face recognition, with a few exceptions [65, 66], use the criteria of *group fairness (demographic parity)* to evaluate and mitigate both data and algorithmic bias. However, *group fairness criteria* relies on sensitive attribute labels such as race, ethnicity or skin tone and uses performance evaluation metrics such as accuracy or FMR. Subsequently, stratification of the complex and multi-faceted concept of race into abstract race-related categories becomes necessary in order to address racial bias *group fairness* as the categories allow us to assess whether the final performance of a

given face recognition system is fair and satisfies *group fairness criteria*. Accordingly, the face recognition literature mainly utilises either race (e.g. African, Asian, etc.) or race-related grouping categories (e.g. skin tones, facial phenotypes etc.). However, with regard to racial stratification, this construction of race or race-related groupings also brings with its and additional set of challenges.

For example, early attempts at the conceptualisation of race itself inherited racial bias, as the way race is defined and understood is influenced by preexisting prejudices and discriminatory beliefs [67, 68]. As a result, the way race is conceptualised may perpetuate and reinforce existing forms of racial inequality [68]. Moreover, exposing or using such racial origin identifies the representation of a particular group and may lead to potential racial profiling and associated inequality [69]. Additionally, race or skin tone grouping strategies can limit the scope of any study as they fail to capture the whole aspect of the racial bias problem within face recognition where it needs to consider both multi-racial or less stereotypical members of such groupings [39, 70]. Hanna [71] discussed treating race as an attribute rather than a structural, institutional, and relational phenomenon and ignoring its multidimensional factors can result in missing important aspects of algorithmic fairness. Finally, many researchers do not provide detailed background about their racial categorisation process [72], which makes such race-related groupings even more insurmountable in effectively addressing racial bias. Published datasets and related research work rarely contain details about how racial groups are determined or how racial bias evaluation metrics are designed [72]. In addition to the aforementioned points, many studies [72–74] highlight the potential risks of omitting the details of the racial categorisation strategy along with the appropriate context for use.

In this section, we delve into racial bias within face recognition (i.e. *group fairness criteria*). We examine how race and race-related grouping categories are constructed, the significance of accurately defining these categories and the potential risks and consequences of using and evaluating them in face recognition systems. We classify groupings under the three most predominantly used categories: race, skin tone, and facial phenotype. We discuss the grouping strategies in each category together with their potential positive and negative impact and describe the details of subcategories where they have been used. Furthermore, we cover the literature on annotation processes of grouping cat-

egories and summarise recent literature along with face datasets by organising them under their grouping strategies in Table 2.1.

2.2.1 Race

Race, as a term for human categorisation based on varying factors, is a controversial concept related to sociology, psychology, biology, ethnology, and cultural anthropology, whose definition varies across different fields and throughout history. Within biology, for example, the race concept has been differentiated into three different kinds: genetic, morphological and psychological, which are all widely disputed [75]. Race was first delineated by European naturalists and anthropologists to establish population-based research on human diversity [76]. In the seminal early scientific work of 1758, Systema Naturae [77], Carl Linnaeus categorises humans into four different groups: {*European white, Americanus rubescens (American reddish), Asiaticus fuscus (Asian tawny), Africanus niger (African black)*} using a combination of continental (geographic) and observational (skin tone) terminology. Subsequently, several attempts were made to classify and group humankind in such a manner in order to use it in societal statistics [68,71,78]. Most of the work was problematic (by the standards of today) or error-prone (even by the standards *of the day*) as it reflected the biased ideologies of researchers, politicians and institutions of that time [68]. However, such definitions and classifications were adopted by the national census infrastructure across many jurisdictions [71]. The work of Khalid Muhammad [79] reveals how anecdotal, hereditarian and pseudo-biological race theories transformed into statistics and social surveys. Furthermore, Zuberi [68] addresses the complicated history of racial stratification and its evident impact on social and natural sciences. Consequently, he defines race as a biological notion of physical difference grounded in an ideology [68].

Within face recognition, subject face images form the primary information source that encapsulates these race-related biological and physical differences, which are then combined with additional information, including gender, age, pose, facial expression and contextual aspects such as scene background, illumination, subject clothing and facial accessories such as glasses, facial hair, jewellery and makeup. On this basis, it becomes possible to adopt any such ideology via the use of racial groupings and classifications that are introduced to face recognition with the aim of quantifying racial bias. However,

despite this potential, an increasing number of face recognition studies instead adopt different variations of racial categorisation [30, 80] without any reference to the underlying critical theory of such categorisation and how they are defined [68, 71, 78]. More worryingly, racial annotation of face imagery has now become the initial step in many proposed face recognition approaches aiming to address racial bias, but the crucial decision-making on how and why a given racial categorisation is defined remains subjective, arbitrary and largely undocumented [81].

Dataset Name	Year	Grouping Categories	Images	Source
		Race		
ColorFERET [82]	1993	White, Asian, Black, Others	14K	Participants' photographs
MORPH [83]	2006	Caucasian, Hispanic, Asian, or African American	55K	Public Records
UTK Face [84]	2017	Asian, Black, Indian, White and Others (Hispanic, Latino, Middle Eastern)	20K	MORPH, CACD, online resources
IJB-C [85]	2018	North American, South America, Western Europe, South West Africa, East Europe, East Africa-Middle East, South East Asia, India, China, East Asia	31K	Public, law enforcement databases, social media
RFW [11]	2019	African, Asian, Caucasian, Indian	45K	MS-Celeb [86]
DemogPairs [87]	2019	Asian, Black, White	10.8K	CWF, VGGFace1-2 [3, 88, 89]
BUPT-Balanced [1]	2020	African, Asian, Caucasian, Indian	1.3M	MS-Celeb [86]
VGGFace2 1200 [9]	2020	African, Asian, Caucasian, Indian	1M	VGGFace2 [3]
FairFace [40]	2021	Black, East Asian, Indian, Latino, Middle Eastern, Southeast Asian, and White	108K	Flickr, Twitter, newspapers, online resources
CASIA-Face-Africa [90]	2021	Hause (Sudan, Chad, Binin, Ivory Coast), Non-Hause	38K	Subjects from Nigeria
DiveFace [91]	2021	(Japan, China, Korea), (Europe, North America, and Latin America) (Sub-Saharan Africa, India, Bangladesh, Bhutan)	120K	MegaFace [92]
		Skin Colour		
IJB-B [93]	2017	1-6 skin tones (increasing in darkness)	1K	1M FreeBase Celebrity List
PPB [39]	2018	Light, Dark skin tones (Fitzpatrick I-III,IV-VI)	68K	Gov. Official Profiles
Fair Face Challenge [45]	2020	Light, Dark skin tones (Fitzpatrick I-III,IV-VI)	152K	Flickr, Twitter, newspapers, online resources
Casual Conversations [94]	2021	Fitzpatrick Skin Tones	45K*	Vendor data
Globalface-8 [95]	2021	ITA base 8 skin tones (Tone I-VIII)	2M	1M FreeBase Celebrity List
Balancedface-8 [95]	2021	ITA base 8 skin tones (Tone I- VIII)	1.3M	1M FreeBase Celebrity List
IDS-8 [95]	2021	ITA base 8 skin tones (Tone I-VIII)	10K	1M FreeBase Celebrity List
		Facial Phenotypes		
Diversity in Faces [10]	2019	ITA 6 skin tone, Craniofacial distance, area, ratio, Facial region contrast	0.97M	YFCC-100M
VGGFace2 [3] - [13]	2018	Fitzpatrick Skin Tones, Nose Shape, Eye Shape, Mouth Shape, Hair Type	3.3M	Google Image Search
RFW [11] - [13]	2019	Fitzpatrick Skin Tones, Nose Shape, Eye Shape, Mouth Shape, Hair Type	45K	MS-Celeb [86]

Table 2.1: Overview of most prominent face recognition datasets categorised by racial groupings, including dataset size and image sources.

Previously, racial categories made an initial appearance within automated facial analysis via the task of race classification. For example, [96] propose feature extraction-based techniques for race classification using the MORPH [83], and FERET datasets [82] to predict $\{$*Caucasian, South Asian, East Asian, and African*$\}$ racial classification. Later studies [83] extend the MORPH dataset for face recognition and analysis tasks (identifi-

cation, recognition, and verification) by providing additional ground truth labels spanning age, gender, race, height, weight, and eye position. Subsequently, DCNN-based methods were introduced for race classification [97–99]. The work of [97] proposes the large-scale VGGFace2 Mivia Ethnicity Recognition (VMER) dataset, composed of more than 3 million face images annotated with four ethnicity categories, namely {*African American, East Asian, Caucasian Latin and Asian Indian*}, and provides comprehensive performance analysis for several contemporary deep network architectures, namely VGG-16, VGG-Face, ResNet-50 and MobileNet v2. Although such race classification techniques are not necessarily used as a proxy for facial image annotations with regard to the study of racial bias within face recognition, these public datasets containing race labels and their associated racial groupings are widely adopted *de facto* by the face recognition research community. As we illustrate in Table 2.1, the most commonplace face recognition datasets containing race labels [11, 39] use three grouping strategies, namely race, skin tone and facial phenotypes. Similar to race classification, broader racial groupings such as {*African, Asian, Indian and Caucasian*} or binary racial groupings such as {*Black, White*} are also commonly followed by many datasets creators [11, 39].

Recently, the most commonly used face recognition evaluation dataset, a subset of MS-Celeb-1M [86] released as the RFW dataset [11], was constructed to measure relative face verification performance across four different racial groupings: {*African, Asian, Indian, Caucasian*}. FairFace [40] is another dataset, again drawn as a subset from the larger YFCC-100M Flickr dataset [100], which supplements this earlier set of four labels with two additional racial groupings, {*Middle East, Latino*} to evaluate racial bias more broadly. In addition, UTKFace [84] is a large-scale face dataset with five different racial groupings, namely {*Asian, Black, Indian, White and Others (like Hispanic, Latino, Middle Eastern)*}, for various tasks spanning face detection, age estimation, and age progression/regression. This variation in racial groupings, illustrated more extensively in Table 2.1, highlights the ambiguity and uncertainty behind the race concept upon which the absence or presence of bias is ultimately being evaluated. Consequently, this inconsistency of racial groupings, its historical and geographic instability within the face recognition research literature and the commonplace adoption of ill-defined race concepts that are littered with a problematic history within social statistical science make effective per-

formance evaluation and quantification very challenging within the racial bias problem space.

Similarly, Khan [101] identify four specific problems with the racial categories: (1) the categories are not clearly defined and are often loosely associated with geographic origin; (2) the categories that are extremely broad, with continent-spanning construction that results in individuals with vastly different physical appearance and ethnic backgrounds being grouped incongruously into the same racial category; (3) the categories narrow down the differences between ethnic groups with distinct languages, cultures, separation in space and time, and phenotype into the same racial category; (4) assigning a single racial category to a face example for performance evaluation of any form of automated analysis, including face recognition, is not an ideal solution as it cannot capture a substantial proportion of the distribution of diversity and variation within the human race.

In parallel with Khan, Raji [73] discusses three ethical tensions when auditing commercial facial processing systems, where there exists a requirement to annotate face imagery with race or race-related categories. *Privacy and Representation:* Collecting a diverse and representative dataset for facial recognition can bring privacy risks for individuals included in the dataset. Furthermore, potential consent violations may arise during the data collection process, for example, for the IBM Diversity in Faces dataset [10], which was sourced from images on the public image-sharing platform Flickr that were uploaded under very permissive licensing terms (Creative Commons). However, it later emerged that the individuals within the photos did not necessarily consent to be included within the face recognition dataset [102]. *Intersectionality and Group-Based Fairness:* Intersectionality is based on the idea that the experience of an individual cannot be fully understood by looking at one aspect of their identity. However, when evaluating group fairness in facial recognition systems, assigning individuals to a racial category and performing disaggregated analysis to account for multiple categories is often necessary. This type of analysis can help to identify and address potential biases, but it may not fully capture how varying components of a face recognition processing pipeline interact to recognise individual features across individuals with multiple marginalised identities. *Transparency and Overexposure:* Although sharing details of the dataset development process and publicly disclosing named audit targets can help to clarify the scope of the audit and the

context in which results should be interpreted. This can also result in targeted over-fitting (i.e. *"cheating"*) in order to optimise system performance on the audit. Moreover, this can also lead to pressure to make the audit more operationally relevant to real-world deployment. For example, some institutions have removed or restricted access to their facial recognition benchmark assets following their inclusion in audits, which can compromise the performance validation of future systems and make it more expensive and difficult for other researchers to evaluate relative performance changes in the field [20].

Finally, although many more studies discuss the possible negative consequences of using racial categories in face recognition datasets, Table 2.1 proves that such racial categories have become commonly used and increasingly contributed within the literature. The lack of work on alternative race-related grouping strategies or fairness criteria that do not rely on any racial categories forces racial bias studies to address racial bias using such commonly defined racial categories. Considering the problems that arise with racial categorisation, the current status of research that uses racial categories (*still*) does not paint an optimistic picture of the global face recognition research community collaboratively tackling the issue of racial bias. As information of racial or ethnic origin remains sensitive [103], from these observations across the face recognition field, we agree with the findings of several major studies [68–70, 73, 74, 79, 104, 105] that already highlight the adverse effects of the use of racial categories and their suggestion that researchers should either avoid revealing such sensitive data or provide an appropriate context for use. Furthermore, transparent provision of the ethical considerations together with any details of the racial annotation process in use and the intended possible use cases, limitations, and risks of the designed solution, should be made by the originating researchers in all cases [46].

2.2.2 Skin Tone

Human skin tone ranges can vary from saturated black to off-white pale, representing one of the key race-characterising traits. Variations in skin tone among humans have been traditionally used to classify people into race or race-colour identities [106] as skin tone variation caused by genetic differences (also exposure to the sun). Over the past centuries, methods for categorising skin tone have evolved from verbal race-related descrip-

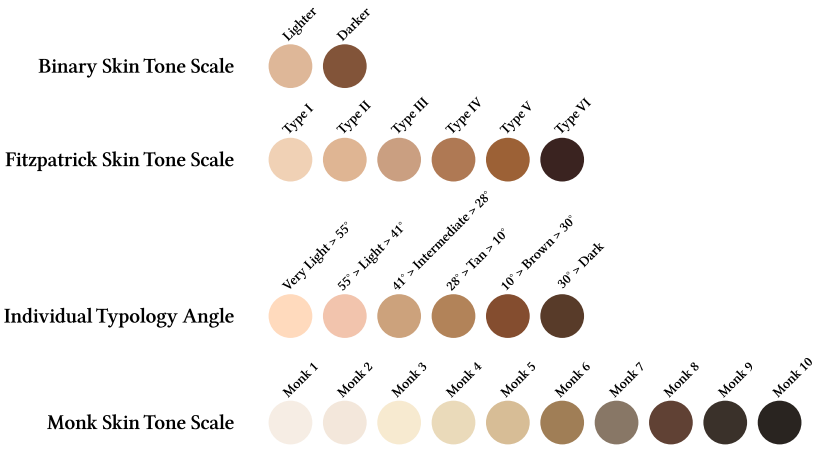


Figure 2.2: Four different skin tone scales used for racial bias analysis within the context of face recognition.

tions (that would potentially be seen as derogatory today) with skin colour categories as *"white", "yellow", "black", "brown"*, and *"red"* [107], to colour-matching-based methods. The colour-matching-based methods compare skin colour based on their similarity to a set of standardised colour samples. The Von Luschan scale, employing 36 coloured glass tiles for skin color comparison, is one of the most common examples of color-matching-based methods, widely utilised for racial categorisation of populations until the mid-20th century [108].

Fitzpatrick Skin Tone Scale

Following the colour-matching methods, the Fitzpatrick Scale, established in 1975, became the most commonly used skin tone scale in dermatology and medicine. The dermatologist Thomas B. Fitzpatrick developed his Fitzpatrick Skin Tone Scale to assess the propensity of the skin to burn during photo-therapy (i.e. the treatment of skin conditions using intense ultra-violet light sources). Initially, four different types ranging from Type I (always burns, does not tan) to Type IV (rarely burns, tans with ease) were released by [109]. Later, he extended his scale to include a broader range of skin types (Type V and VI) [110] in order to offer a more granular representation across darker skin tones. The widespread adoption of this work within medical research studies [111, 112] subsequently influenced early computer vision research studies considering skin tone. Within the racial bias literature, the Gender Shades study [39] was the first to gather attention around the

use of the Fitzpatrick Skin Tone Scale within an automated facial image analysis context. Subsequent studies then released varying datasets, all using the Fitzpatrick scale on this basis [13, 45, 94]. Even recently, the extensive Casual Conversations Dataset [94] containing 45K videos makes use of Fitzpatrick skin tone labels for its racial grouping strategy. However, other researchers have raised concerns about using the Fitzpatrick scale on image-based visual tasks [113]. Primarily, the Fitzpatrick scale was not initially designed for image-based skin tone estimation; hence, its evaluation methodology relies on physical skin measurement. As a result, its use can cause inconsistent skin tone assignment when applied on images [114]. Consequently, [113] observes how challenging it is to robustly assign darker skin tone labels within the Fitzpatrick scale when faced with a significant imaging variance and suggests avoiding the use of such skin tone assignments ascertained from images captured under uncontrolled or unknown conditions.

Individual Typology Angle (*ITA*)

Subsequently, reflectance spectrophotometry and colourimetry methods [115] have become preferential in medical skin tone assessment over earlier methods due to increased accuracy and consistency. Whilst colourimeters quantify the appearance of a tone on the skin, a spectrophotometer measures the spectral characteristics of the skin colour. Such devices convert light reflectance data from the skin into colourimetric values for estimating chromophores in the skin [116]. Subsequently, Individual Typology Angle (ITA) [117] has been proposed by Chardon in 1991 to classify human skin colour using spectrophotometric measurements. This method utilises the reflection of skin light via spectrophotometers that measure LaB colour values of the skin (L: Lightness. a: Red/Green Value. b: Blue/Yellow Value) to represent the intensity of pigments such as carotene, haemoglobins, phaeomelanin, and eumelanin. Accordingly, Chardon proposes six physiologically skin categories: $\{$*very light, light, intermediate, tan, brown, and dark*$\}$ estimated via equation of ITA $ITA = \arctan(\frac{L-50}{b}) \times \frac{180}{\pi}$. *ITA* projects skin colour volume into LaB colour space, and is used to categorise skin angle via the associated *ITA* classification thresholds (see Fig. 2.2) [114]. As the *ITA* solely relies on precise and objective skin tone measurements, it is considered more accurate than traditional visual assessments. Furthermore, it provides a better representation of both the diversity and

contributory factors associated with skin tone [118,119]. On the other hand, the utilisation of *ITA* scores and categories varies in the literature; Wang [95] constructs three large-scale face recognition datasets containing four or eight different skin tone groupings based on *ITA* scores and releases the corresponding skin tone labels for each face image with the datasets. The Diversity in Faces dataset [10] also adapts *ITA* (using six categories) as they find ITA both a more practical and straightforward method for measuring facial skin tone. However, akin to the earlier aforementioned issues with skin tone estimation from digital face images, inconsistent and uncontrolled imaging conditions again impact accurate and reliable *ITA* assessment [114, 118].

Monk Skin Tone (MST) Scale

Most recently, the work of Ellis Monk [120] produced a new extended skin colour scale 10-shade skin tone scale designed to facilitate the construction of more representative datasets for the development of on-line consumer services. Although the associated study discusses the aforementioned limitations of prior work on skin tone groupings such as the Fitzpatrick Skin Tone Scale [110], it does not provide any detail for the practical application of the new 10-shade scale or any additional guidance via the provision of an exemplar dataset [120].

Binary Skin Tone Scale

Lastly, binary skin/racial groupings has been employed in sociological research on race and race relations [121]. Focusing on white-black race relations in the United States brings expensive socio-economic data and analysis around such binary groupings [122]. Accordingly, the adaption of binary skin/racial groupings into computer vision tasks such as skin tone estimation, race classification and racial bias of face analysis systems started from this simple categorisation viewpoint. In order to model skin colour on imagery, several studies [123] proposed quantitative colour-space divisors (i.e. a dark-light pixel colour threshold) and simply grouped skin colours into binary categories. In the racial bias context, many studies adopt such a darker-lighter skin tone grouping by either narrowing the Fitzpatrick scale or dividing subject skin tone variance into binary categories. One of the seminal works in the field, Gender Shades [39], uses darker-lighter skin tone

categories on the Pilot Parliament dataset to demonstrate the algorithmic performance disparities in both gender classification and face recognition tasks. Another example is the Fair Face Challenge study [45], which suggested researchers used a requantised (narrower) set of Fitzpatrick skin tone categories as per Gender Shades [39]. Despite binary skin tone categories are being the most straightforward grouping strategy in terms of automatic image annotation, in practice, it often obscures the complexity of race concept and results in the mis-quantification of the racial bias problem across solutions where the ultimate aim is unbiased performance across any skin tone variant. This is attributable to imaging effects such as skin reflectance, which was shown by Cook [124] to have a very significant net effect on the average biometric performance when considered across three different skin reflectance groupings within face recognition. As such, the use of simple binary groupings is known to result in erroneous or conflicting group interpretations, whilst broader groupings such as Fitzpatrick Skin Types claim to be more robust against this issue [13].

The contrasting examples of these various skin tone scales are illustrated in Fig. 2.2 where we can see a sharp contrast between categorisation in binary, Fitzpatrick, ITA or MST skin tone groupings. However, skin tone scale grouping strategies alone carry various concerns for the mitigation of racial bias within face recognition. We discuss these concerns under three divisions as follows:

Erroneous Skin Tone Annotation: Firstly, most skin tone scales are designed to measure skin tone on physical human subjects in a medical or dermatological context. By contrast, face recognition systems instead used such annotations for digitally captured face images that form part of the training and test data sets. Moreover, such face image samples are commonly yielded from public domain sources (i.e. internet search engine-based image retrieval - *"in-the-wild"*), and as such, this uncontrolled imagery exhibits enormous variation in both environmental and subject conditions at the point of image capture. Similarly, [125] summarises such varying conditions that affect skin-colour detection in the visible spectrum as scene illumination, camera characteristics, demographic characteristics (race, age, gender), and other factors (make-up, wearing glass, hairstyle, head pose). Such varying factors make effective skin tone annotation challenging and result in erroneous skin tone assignment for given subjects/samples. Furthermore, human annotators

often bring subjectivity and inconsistency to the resulting annotation labels far more so than other image labelling tasks (c.f object/scene categorisation), whereas skin tone annotation ideally needs to be objective, consistent, and repeatable [81]. Specifically, [114] highlights the uncertainty within the human-based categorisation of skin tones from digital image and proposes the use of automated skin tone assignment as a means of potentially achieving speed, scalability and consistency. However, the consistent skin tone annotation of a given subject under the aforementioned image variations remains a pertinent issue with such automated solutions - one that in itself presents a circular occurrence of bias within facial processing.

Narrow Representation of Scales: Secondly, the most commonly used skin tone scales used for accessing aspects of racial bias are either too narrow in terms of their discretisation of the skin tone spectrum (e.g. Binary Skins Groups, Fig. 2.2) to facilitate capture of the foundational reasons for bias or alternatively offer the less representative capability for specific groups (e.g. Fitzpatrick Skin Types vs Monk Skin Tone Scale, Fig. 2.2) [113].

Skin Tone as a Single Dimension of Race: Thirdly, race is a multi-faceted concept conflating other phenotypic facial traits such as lips, eyes, hair and face shape. Solely aligning racial grouping with skin tone only transforms the racial bias problem into a single-faceted problem. Moreover, there is no clear evidence that skin tone alone is the primary driver for disparate false match rates within face recognition performance [105]. Accordingly, several studies suggest considering other race-related facial attributes, including lips, eye, and face shape when measuring racial bias in this context [126,127] in order to enable improved interpretation and derivation of bias factors. Accordingly, a consensus is beginning to emerge on skin tone assignment and the appropriate quantification of skin tone within digital facial images as used in face recognition research. Various studies [39,45,94] measure the racial bias in face recognition using either binary skin groupings, the Fitzpatrick Skin Types [110], or ITA [117] as depicted in Figure 2.2.

Overall, this section provides an overview of skin tone characterisation approaches and their associated quantification methodologies spanning both digital imagery and physical dermatological examination. Accordingly, we summarise the most common skin tone scales and discuss the challenges of applying such estimation approaches to the skin tone labelling task within face recognition datasets. Furthermore, we outline all of the face

recognition datasets in the research literature that use varying skin tone scales in Table 2.1. As skin tone-based groupings become widely used for racial bias evaluation studies, many benchmark datasets are unfortunately annotated with varying skin tone scales and with varying levels of labelling robustness. Although utilising skin tone scales as a labelling concept for face recognition datasets avoids otherwise using sensitive or ill-defined racial categories, the subjectivity of human-based skin tone annotation, the inconsistency of facial image capture conditions and most pertinently the fact that the skin tone is only one dimension of race all make it an imperfect mechanism for the quantification of racial bias within face recognition. As a result, we suggest developing a broader strategy based on the use of high-accuracy, consistent and reliable facial phenotypes that can instead analyse the true relationship between facial features and racial bias. Consequently, we believe such approaches enable investigation across every facial trait and hence bring greater granularity to the quantification of racial bias within face recognition whilst avoiding the use of problematic racial categorisation.

2.2.3 Facial Phenotypes

Human phenotypic variation refers to variation over the set of morphological and observable characteristics of an individual, which is the result of both genetic and environmental factors [128]. Such variation is most observable on faces as the face is identified as a *"biological billboard of our identity"* [129]. Subsequently, many studies [130, 131] focus on the impact of human phenotype characteristics (such as morphological attributes) on race. For example, the *Shades of Race* study [12] investigates the marginal effects of phenotypic characteristics, including skin tone, lips, nose, hair and body type on racial categorisation. Moreover, Zhuang [132] considers 21 craniofacial measurements such as face width, length, nose dimensions and eye corner locations in order to show statistically significant differences in facial measurements between four racial grouping, which are $\{$*Caucasian, Hispanic, African, other (mainly Asian)*$\}$. Therefore, a race-related facial phenotypes can be considered to be specific to such facial characteristic attributes, which can then also be correlated to race (*"Phennotopically similar individuals are expected to be genetically more similar as well."*, [133]). On the other hand, facial phenotypes such as skin tone or hair colour do not identify racial categories within themselves, but they

can combine with other attributes to identify a broader racial grouping [134]. Furthermore, this correlation between such facial phenotypes and racial categories may not be readily visible or clearly delineated, which is in fact highly desirable when we aim to curb the continued use of problematic historical racial categorisation approaches and the disclosure of sensitive racial categories [135] (see Section 2.2.1).

Moreover, Maddox [136] explains *racial appearance bias* as a negative disposition toward phenotypic variations in facial appearance. He also discusses how race-conscious social policies may fail to address racial bias with regards to the societal treatment and socioeconomic outcomes of disadvantaged groups [104]. For example, many studies show that individuals with more stereotypical racial appearance suffer from poorer socioeconomic outcomes than those with less stereotypical appearance for their race [104, 137, 138]. Additionally, the sole use of race or skin tone categories to quantify racial bias is limiting as they do not account for multi-racial individuals or those who exhibit less stereotypical racial traits. Within this context, an improved understanding of the role of phenotype variation may complement existing solutions that attempt to address racial bias [136].

Facial Coding	**Description**
Schema 1 [139]	Craniofacial Distances
Schema 2 [140]	Craniofacial Areas
Schema 3 [141]	Craniofacial Ratios
Schema 4 [142]	Facial Symmetry
Schema 5 [143]	Facial Regions Contrast
Schema 6 [117]	ITA-based Skin Tones
Schema 7 [144]	Age Prediction
Schema 8 [144]	Gender Prediction
Schema 9 [145]	Subjective Age & Gender Annotation
Schema 10 [146]	Pose and Resolution

Table 2.2: Summary of facial coding scheme analysis for the DiF dataset [10].

A set of race-related facial phenotype attributes such as skin tone, nose shape, and lip shape are of primary interest for quantifying and addressing racial bias in face recognition. Furthermore, the recent work of [147] show that non-explicit racial attributes (accessories, hairstyles or facial anomalies) conflated with explicit racial attributes (skin

tone, nose shape or eye shape) strongly affect recognition performance. This study discusses the need to investigate each attribute in order to achieve robust, fair and explainable face recognition solutions [147]. Such requirements directly contradict the use of more traditional racial groupings as they remain a high-level, yet impoverished representation to facilitate elaborate performance interpretation [148]. Subsequently, a plethora of work highlighting the shortcomings of race and skin tone-based categorisation push the current direction of research into phenotype-based categories (as discussed in Section 2.2.1 and 2.2.2). One of the example studies, *Diversity in Faces* [10], provides a new large-scale facial data that implements annotations across ten facial coding schemes in order to provide human-interpretable quantitative measures of intrinsic facial features. The study comprises an extensive set of facial annotations spanning intrinsic facial features to include craniofacial distances, areas and ratios, symmetry and contrast, skin tone (*ITA*), age, gender, subjective annotations, head pose and image resolution that are listed in Table 2.2. However, despite its potential to date this *Diversity in Faces* is not publicly available due to increased sensitivity around subject privacy and consent issues (as discussed in Section 2.2.1).

Compared to the prevalence of race or skin tone categories, phenotype-based groupings have received less attention across the racial bias literature to date, as they involve both skilled attribute labelling for dataset construction and a significantly more complex evaluation strategy due to the significant number of phenotype categories, and phenotype combinations present. To these ends, within a phenotype-based grouping strategy the concept of race is not represented by the difference across a single facial phenotype but rather a combination of varying phenotypic differences that differentiate facial characteristics of a given subject from another. As such, subsequently investigating the impact of such differences on face recognition performance becomes both more complex and time-consuming despite the improved comprehensiveness and quantification options that such a phenotype-based approach offers to the evaluation. On the other hand, it is essential to note when used, the correlation of phenotypical categories with more traditional (i.e. historically problematic, see Section 2.2.1) racial categories should be avoided in order to prevent the naturalisation (or popularisation) of such *"headline style"* summation of racial bias evaluation results.

In conclusion, this section presents an alternative methodology for addressing racial bias (group fairness) within face recognition tasks. Whilst the face naturally conveys identity-related biometric information, it also inherently reflects a significant genetic and geographic relationship with race but these secondary relationships with race are not the primary concern for face recognition tasks. Instead, the group fairness objective within face recognition tasks is to ultimately ensure that it equity of performance across all subjects, regardless of subject racial grouping or facial phenotype characteristics. To these ends, it is necessary to avoid the inherited problem of racial and skin tone category usage within face recognition datasets and processing pipelines (Sec 2.2.1 & Sec 2.2.2), and instead adopt a more general option that facilitates quantifiable performance measurement without any explicit reference to such problematic concepts. By contrast, the use of facial phenotypes offers a viable alternative that, whilst not fully independent of earlier racial categorisation, offers significantly more granular insight within the quantification of racial bias spanning both skin tone and numerous other facial characteristics.

2.3 Racial Bias within Face Recognition

Contemporary automated facial recognition encompasses a pipeline of multiple stage processing; image acquisition (for both dataset collation and deployment), face localisation, face representation, face verification and identification (final decision-making) [22, 24].

Image Acquisition covers image capture from a wide range of devices such as smartphone cameras, webcams, high-end DSLR cameras and CCTV-style video surveillance cameras varying imaging conditions that span image resolution and compression, facial occlusion, facial pose, illumination, subject use of make-up/glasses/jewellery and facial expression. Furthermore it includes all stages of initial image pre-processing and formulation such as the demosaicing conversion to per-pixel RGB colour (from the Bayer pattern of the camera CMOS/CCD device), automatic colour and contrast correction (including processes such as automatic exposure control, white balance, automatic focus, brightness correction), pixel quantisation to a given bit-depth (e.g. RGB 8-bit colour) and compression. For data set collation, acquisition is complemented by a data curation such that differing imagery is sampled to select a subset of representative images that are ideally

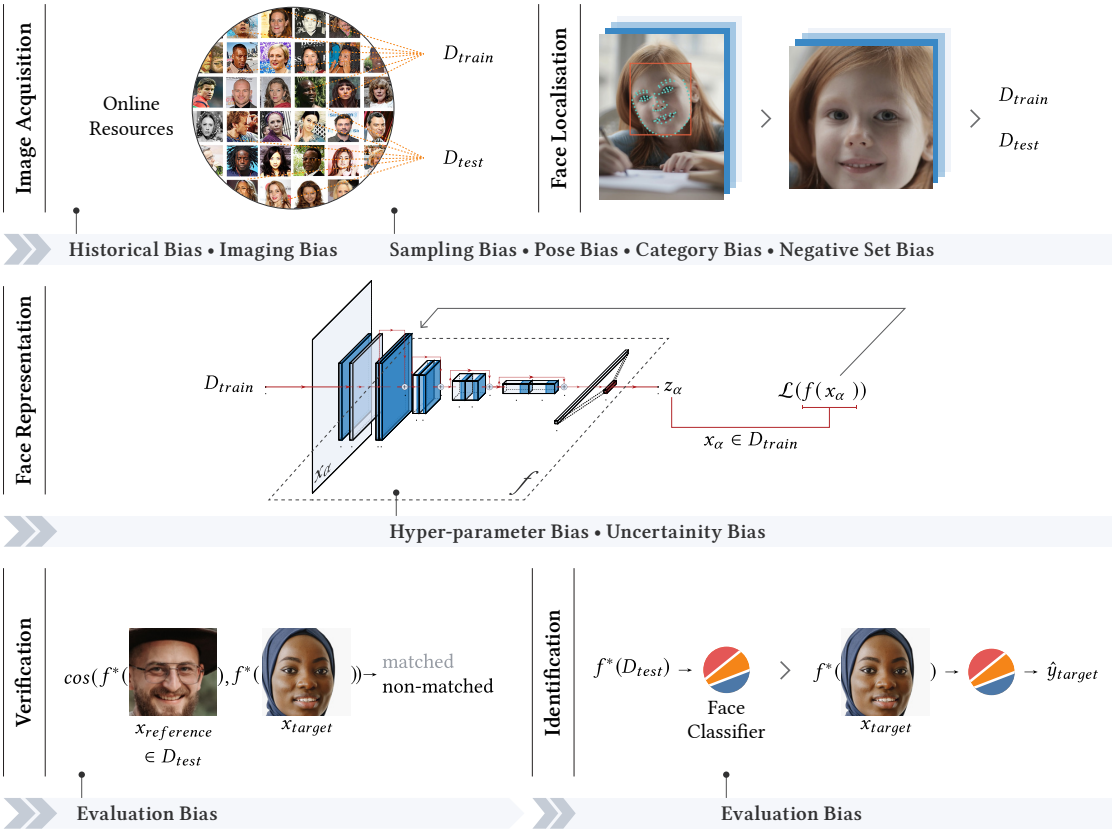


Figure 2.3: Overview of the face recognition processing pipeline and bias attribution.

diverse and challenging enough to capture the full range of faces and imaging conditions that a face recognition system may encounter in real-world (*"in-the-wild"*) deployment. These are then used to form the train I_{train} and test I_{test} datasets for system training and evaluation (as defined in Section 2.1).

Face Localisation consists of two sequential steps to process real-world, in-the-wild images that are captured under uncontrolled conditions and may hence exhibit variation across one or more of the aforementioned imaging conditions (typically: face off centre, rotated and of varying scale relative to the camera). The first step, face detection aims to identify a set of facial landmark locations (e.g. eye, mouth and nose endpoints, face boundaries in width and height) whilst the subsequent step of face alignment aims to correct for positional, rotational and scale variations to obtain a canonical facial image representation. This facial alignment step facilitates the use of the spatial correlation of facial features across both varying subjects and dataset image samples within the subsequent stage of face representation.

Face Representation involves optimisation the mapping function f^* that projects a given face image sample into a feature embedding space, where the feature embedding vectors are both representative and distinctive for each subject. In order to select the optimal mapping function, f^*, a training process is performed via a training dataset, I_{train}, with reference to the minimisation of a loss function that incites the use of a distinctive facial feature mapping (as defined in Section 2.1). Consequently, f^* provides mapping for both the curated training dataset, I_{train}, and unseen images in both test dataset, I_{test}, and any subsequent deployment.

Face Verification and Identification encompass the two most common decision-making (i.e. *"end goal"*) tasks in face recognition. Face verification refers to a one-to-one matching operation to determine whether two facial images belong to the same individual (known subject case), and identification refers to a one-to-many matching operation to conversely identify a given individual against a set of reference images (unknown subject case). The optimal selection of mapping function, f^*, via the training process on training dataset, I_{train}, directly impacts the effectiveness of the feature embedding vectors such that the presence of both improved representational distinctiveness between differing subjects and also the robust representation of identical subjects under varying imaging conditions hence leads to improved face verification and identification performance.

With reference to the formal face recognition problem space definitions of Section 2.1, this four stage conceptual face recognition processing pipeline is illustrated in Figure 2.3 where we additionally highlight the potential sources of bias at each stage. These will be further explored, with reference to related work in the literature on racial bias within face recognition, in the remainder of this section.

2.3.1 Image Acquisition

Image acquisition, spanning the imaging aspects of both initial dataset collation and final real-world deployment. We subdivide this stage into three categories, including *facial imaging*, *dataset curation*, and *dataset bias mitigation*.

Facial Imaging

Biometric data refers to distinctive physical characteristics of the human face, fingerprints, voice, iris, and body. Such biometrics have been used for identification systems for several decades [149] (e.g. fingerprint matching). Commensurately, facial imagery has become a key part of modern biometric tasks due to the proliferation of imaging technologies, which significantly improve facial image quality, accessibility, and quantity. However, the increased prevalence of facial imagery does not necessarily result in improved biometric outcomes across all populations. In addition, collating facial images and annotating them with subject identity or racial category labels at scale have ignited complex discussions around policy and legality due to economic, privacy and ethical implications [150].

We have previously explored the potential risks associated with racial categorisation and the annotation of facial images (Section 2.2). Building upon this, here we focus on the privacy risks and ethical concerns surrounding using facial images as a form of biometric data. Paying attention to such ethical and political considerations on the collation of biometric face imagery becomes particularly important when the presence of racial bias therein directly or indirectly impacts societal fairness. Accordingly, [150] presents a socio-political analysis of face recognition and highlights the distinct challenges and concerns associated with its development and evaluation. The study categorises such concerns into four sections: privacy, **fairness**, freedom and autonomy, and security. Even though the intention of automatic face recognition is not problematic, in practice, it may enable morally unacceptable use cases of such technology. Examining the issue of subject consent, both within dataset collation and in an eventual use-case, is fundamental to that preserving privacy [150]. For example, government use of such technology for racial profiling and racially-targeted restriction in some jurisdictions has been widely reported and investigated [151–154]. In parallel to [150], Prabhu [155] discusses the fundamentals of informed consent, privacy, or agency of the individual in large-scale datasets and shows the fallacy of the commonplace Creative Commons licensing model [156] as a consent-included green flag for large scale dataset curation. They suggest the use of dataset audit cards as an approach to publishing the original research goals, curation procedures, known shortcomings and caveats alongside dataset dissemination [155]. Overall, it must be noted that any erosion of privacy, moral, ethical, or political values will most likely

disproportionately impact minority groups, such as those defined along racial lines.

From a technical standpoint, the ISO/IEC 19794-5 [157] standard and ICAO 9303 guidelines [158] propose both image-based (i.e. illumination, occlusion) and subject-based (i.e. pose, expression, accessories) image quality requirements to ensure facial image quality. Accordingly, facial images should be stored using lossy image compression standards such as JPEG [159] or JPEG2000 [160]; and observable in terms of gender, eye colour, hair colour, expression, properties (i.e. glasses), head pose (yaw, pitch, and roll), and facial landmark positions. However, commonplace *"in-the-wild"* face datasets, that are readily used in face recognition system performance evaluation, do not conform to such requirements. Subsequently, Vangara [161] compares ICAO compliance between African and Caucasian groups in MORPH dataset [83] and found that slightly more than 48% of the African-American images were rated as ICAO compliant, while slightly more than 57% of Caucasian images were rated as ICAO compliant. The most prominent factor contributing to the variation in image quality between the groups is the difference in brightness; the distribution of which differs significantly between the African-American and Caucasian groups. The study argues that the lack of illumination correction with regard to skin tone during image acquisition could be the attributable reason as to why the African-American image group contains a larger number of poorly illuminated images. In parallel, [124] points out the significant impact of skin reflectance across demographic subgroup performance with regard to face recognition and mentions that improved imaging acquisition systems (superior camera specification, lower motion blur, higher image contrast and stricter pose control) may significantly reduce or eliminate performance differences between such subgroups.

Furthermore, prior literature shows that non-ideal imaging conditions, including image blur, noise, distortion, occlusion and lossy compression, all have a considerable impact on the performance of face recognition [162–164]. Recently, [164] examined distorted test imagery impact on gender and skin tone categories (light vs. dark skin tone) using pre-trained DCNN-based face recognition models. As a result, the study [164] finds that the regions of interest used in the models shift towards less distinctive regions in the presence of distortions, resulting in unequal performance degradation among subgroups. Consequently, we refer to these performance disparity effects within face recognition

caused by variable imaging conditions as *imaging bias* as illustrated in Figure 2.3. The limited literature on *imaging bias* within face recognition to date makes it harder to identify the presence of such bias and align it to common underlying factors and conditions. On the other hand, state-of-the-art techniques for robust face recognition such as [165] may help to mitigate such *imaging bias* effects, via the use of a rich set of input variations aligned to phenotypic characteristics, such as skin colour or other common facial phenotype variations [166].

Dataset Curation

The following stage of image acquisition pertains to sampling the captured and processed facial images in order to create representative datasets for face recognition evaluation. Nevertheless, such a sampling process is often affected by sampling bias (also known as selection bias or population bias) [167], which significantly impacts racial bias in face recognition. *Sampling bias*, referring to non-random selection over a population leading to a set of samples that do not fairly represent that population statistically, commonly occurs when facial images are curated from public online image resources, where the available population image distribution may not be representative of the actual societal population that the face recognition system will encounter in deployment. This is attributable to the fact that technology access is not globally or socio-economically homogeneous resulting in a skewed on-line image presence for a subset of the populous. Secondly, the most common approach for face recognition dataset collation is via targeted per-subject search for named individuals (commonly celebrity names from the FreeBase database) using public online image resources [86], which then results in a dataset of millions of subjects who have/had public attention (see Table 2.1).

Even more concerning is that the subsampling decision from the FreeBase celebrity list is most often based on ranking all the subjects by their frequency of occurrence in the media, meaning that celebrities with greater global media coverage are more likely to be included in the dataset. This results in a biased convergence to a specific celebrity group, which is dominated by Western, European and American subjects. Moreover, this impact of sampling bias can be subsequently amplified during the later stage of feature representation learning due to an increased imbalance of phenotypic features which are themselves

aligned to the dominant racial or demographic groupings present from the original dataset curation [57]. For instance, a DCNN-based face recognition model utilising certain features, such as hair, to identify face subjects results in a bias towards a particular hairstyle or hair colour, causing less accurate performance on subjects with different hairstyles, hair colours, or accessories.

Consequently, contemporary face recognition datasets are largely curated to provide large-scale coverage of differing face subjects images under a rich variation of *"in the wild"* imaging conditions, with little consideration of the racially differentiating phenotypes of the underlying subject population. The two most widely used training datasets for face recognition - MS-Celeb-1M [86] and VGGFace2 [3] - contain 10 million and 3.3 million face images respectively, and are curated from the FreeBase celebrity list as shown in Table 2.1. Similarly, the most common benchmark test sets for face recognition - LFW (Labeled Faces in the Wild) [15], CASIA-WebFace [88], and MegaFace [92] - are curated using on-line news (Yahoo), FreeBase celebrity and public on-line photo sharing resources (Flickr), respectively. Despite efforts to overcome sampling bias within face recognition datasets such as the release of new datasets like the CASIA-Face-Africa [90], a large-scale African face image database, or the BUPT-Balanced dataset [1], a large-scale racially balanced training set, the most prominent face recognition datasets used for face recognition evaluation still suffer from sampling bias with regard racial phenotypical population coverage.

Dataset Bias Mitigation

The most common assumption in machine learning is that a training dataset I_{train} and test dataset I_{test} are identical and independently distributed; $P(I_{train}) = P(I_{test})$. However, this assumption is not valid for face recognition systems, and this issue is referred to as "dataset bias" by [168]. Although face recognition datasets should represent the real world to enable face recognition systems to work on real-world applications, they have become closed systems, reflecting the world in a significantly biased way [168]. Accordingly, [168] groups dataset bias under four different types of bias; firstly *selection bias* is similar to *sampling bias* mentioned above. Secondly, *capture bias* occurs because the input imagery has the objects (faces) almost always being in the same direction and po-

sition. Additionally, capture bias can be considered as *pose bias* within face recognition, as there is still a poor pose variance in a specified range (i.e. -30 to 30 left-right, -15 and 15 up-down) generating faces) within face recognition datasets. Thirdly, *category or label bias* poses the ill-definition or mislabelling of subject identities and racial categories. Fourthly, *negative set bias* defines the bias on what the dataset considers to be "the rest of the world". If that set is not representative or imbalanced, that could produce recognition models that are overconfident and misrepresenting. For face recognition, such a set can be considered within the same dataset (rest of the subjects vs a subject), which already inherited different types of bias, leading to poor representation of the whole population.

The latest advancements in Generative Adversarial Networks (GAN) [52, 169–171] have made it possible to generate high-quality face images to mitigate such domain gap between training and test sets. For example, [172] addresses the pose bias by producing synthetic data. Another work transfers the facial images of one race to corresponding images of other races to facilitate data augmentation to balance the ethnic distribution [173]. Moreover, [174] proposes a new data augmentation strategy that imposes the fairness constraint to improve the generalisability of fair classifiers. In particular, they highlight that fairness can be achieved by augmenting interpolated samples between the groups during training. However, generative models produce samples from an underlying training distribution as well, meaning that they can be biased too. Accordingly, [171] conducted an empirical study on the fairness of state-of-the-art pre-trained face synthesis GAN models. They show that a strong correlation between the imbalance degree in the training data and the output of the GAN results in consistently more significant imbalanced GAN outputs meaning that the bias is amplified during GAN training.

This section provides an overview of the issues and bias types that arise in the initial image acquisition stage of the face recognition preprocessing pipeline. The discussion encompasses the impact of the privacy risks and ethical concerns associated with biometric face imagery correlated with racial bias. Additionally, the section addresses various sources of bias that can affect the accuracy and fairness of face recognition systems, such as *imagery bias, sampling bias, pose (capture) bias, category and label bias, and negative set bias*. To illustrate these bias types, Figure 2.3 depicts the corresponding stages of the face recognition pipeline where they occur.

2.3.2 Face Localisation

The face localisation stage of the face recognition pipeline consists of face detection and alignment, thereby enabling the spatially correlated facial features for the subsequent stage of face representation. Prior work has primarily focused on hand-crafted facial feature extraction and classification for face detection. In a notable milestone, Viola and Jones proposed a real-time cascade of simple Haar-like feature classifiers at locally learned image locations [175]. Recently, face detection methods have shifted towards DCNN-based architectures and are categorised into five sub-genres by [176]: Cascade-CNN-based, R-CNN and Faster-RCNN-based, Single Shot Detection, Feature Pyramid Network-based, and other variants. Subsequently, the two most prominent face detectors, Cascade-CNN-based MTCNN [177] and Feature Pyramid Network-based RetinaFace [178], and the face detection benchmark dataset, Wider Face [179], have become widely adopted for face recognition processing pipelines.

The MTCNN face detector is based on a cascading multi-tasking structure [177] with three-stage lightweight DCNN where the Proposal Network (P-Net) generates a set of face regions, or "proposals", at different scales, the Refinement Network (R-Net) subsequently refines such regions to better localise the faces and finally the Output Network (O-Net) performs fine-grained face feature extraction and classification. Subsequently, [178] proposes another multi-level face localisation approach, RetinaFace, encompassing a single-shot detection network, a multi-task branch network that predicts both facial landmarks and attributes, and a bounding box regression network refines the position and size of the detected faces from the facial landmarks and attributes. Both approaches achieve outstanding performance on several benchmarks, including Wider Face [179], which comprises 32,203 images and 393,703 bounding boxes under varying imaging conditions.

Despite the widespread usage of face detectors within the face recognition processing pipeline, only a few studies have investigated racial bias within face detection. Menezes [41], analysis the performances of five state-of-the-art face detectors; DSFD [180], Pyramid Box [181], LFD [182], RetinaFace [178], MTCNN [177] on demographic attributes including age, skin tone, gender. The study randomly samples the Casual Conversation Video Dataset [94] and obtains 550.000 frames for training. The Casual Conversation Video Dataset adapts the Fitzpatrick scale and contains an imbalanced skin tone category

distribution with the percentages of Skin Type 1: 4.0%, Type 2: 28.3%, Type 3: 22.9%, Type 4: 8.4%, Type 5: 15.8%, Type 6: 20.7%. Although Type 1 skin tone has the lowest representation in the training data, LFD, DSFD, and it was found that empirically RetinaFace detectors are more likely to fail to detect faces with skin Type 4. Moreover, the study shows that the highest divergence of FNMR occurs within skin tone (being worse than age and gender grouping) and highlights that three out of five detectors evaluated have a higher likelihood of incorrect detection *(FNMR)* for darker skin tones (Type 5 and 6).

Another study [183] investigates the robustness of three commercial on-line face detection capable systems: Amazon Rekognition, Microsoft Azure, and Google CloudPlatform and evaluates the impact of 15 types of natural noise corruption on the face detection performance of different demographic groups. Similarly to the case of face recognition, they conclude that corrupted data is more likely to cause face detection errors in specific demographic groups. For example, those with darker skin types, older adults, and those with masculine presentation all had higher errors ranging from 20-60%. Subsequently, they compare the performance and robustness of non-commercial approaches (TinaFace [184], YOLO5Face [185], MogFace [186]) with commercial ones [187]. They show that commercial approaches are always as biased or even more biased than non-commercial models, despite relatively larger development investment and supposed commitment to industry-level fairness commitments. More recently, [188] proposes the Fair Face Localisation with Attributes (F2LA) dataset with demographic annotations to detect disparate performance over such demographic groups. The study finds that confounding factors, including facial orientation, illumination, and resolution, can cause such disparate performance among demographic groups. Therefore it is important to analyse the performance of such detection models holistically and not draw conclusions solely based on demographic annotations.

Despite ample evidence indicating the existence of racially disparate performance within face detection, there needs to be further investigation targeting racial bias exploration within face detection. Furthermore, similarly to the image acquisition stage of face recognition (Section 2.3.1), the presence of imaging, sampling and dataset bias within these face detection benchmark datasets again translates through the subsequent stages of

face recognition resulting in skewed overall face recognition pipeline performance.

2.3.3 Face Representation

Facial feature representation has been a prominent area of computer vision research for many decades and several milestones have substantially improved the performance of face recognition today [23]. The first well-known method for estimating the probability of distribution over high-dimensional vector space of face images, Eigenface, was introduced in the early 1990s [189]. Following that, Gabor [190] and LBP [191] provide robust performance by using local filtering to obtain invariant facial features. However, they could not create handcrafted features that were distinctive and compact enough to fully scale to the diversity of large-scale benchmark datasets (and hence the global populous). Although numerous learning-based local descriptors have been developed to tackle various aspects of face recognition [192, 193], higher similarity for intra-class samples and diversity for inter-class samples within face datasets remain challenging. Subsequently, the availability of large-scale dataset resources (2007+) and the proliferation of DCNN (2012+) have now enabled contemporary face recognition architectures to achieve outstanding verification and identification accuracy. Accordingly, this stage involves a mapping operation from face images to face representation vectors which can be performed by a DCNN-based backbone architecture and a loss function, as discussed in Section2.1.

Backbone Architectures

DCNN are multi-layer processing blocks, including convolutional, pooling and fully connected layers. As a central component of DCNN, the convolutional layers extract features from the output of the previous layer, starting from the face image input. Each layer t consists of K kernels with weights $W = W_1, W_2, ..., W_K$ and added bias filters $B = b_1, ..., b_K$. Subsequently, each layer applies an element-wise nonlinear transform (i.e. $\sigma \in \{RELU, tanh, Softmax, \dots\}$ functions) to generate multiple feature map representations and passes the result to the next layer $x^t = \sigma(W_k \cdot x^{t-1} + b_k)$. Moreover, at the end of each layer, a pooling function down-samples the feature maps by taking the maximum or average value of adjacent pixels (patch). Similarly, a fully connected layer applies a linear transformation to the input vector through a weights matrix.

A majority of face recognition methods adopt state-of-the-art DCNN as their backbone architectures, such as the VGG-Net [194], the ResNet [14], and the Inception-ResNet [195]. VGG-Net [194] uses a smaller fixed number of convolutional filters compared to the AlexNet [196] to decrease the total number of trained parameters. On the other hand, ResNet [14] uses skip connections between two consequent layers to avoid the vanishing gradient problem (unstable training of deep networks due to ever decreasing gradients relative to the input). Furthermore, InceptionNet [195] consists of multiple kernels in one layer to grasp salient features at different levels, including global and distributed features.

Baseline Loss Functions

Contemporary, face recognition literature primarily focuses on designing novel DCNN loss functions [3–6] to enhance the distinctiveness and separability of features. Mostly, such loss functions [4–6] operate on the feature embedding vectors of the last fully connected layer of the selected backbone DCNN architecture [14]. Previously, we discussed Softmax loss $\mathcal{L}_{softmax}$ (Eqn. 2.1) which is based on maximising the posterior probability of the ground-truth subject class in order to separate features from different classes. However, a high number of subject identities, n, within training sets increases the size of the linear transformation matrix in the last layer $W \in R^{d \times n}$ leading to high complexity. Moreover, the learned feature embedding vectors of Softmax loss are not distinctive enough to address the open-set face recognition problem [197]. To address these problems, CosFace [5] enforces a larger cosine margin m between the features of different classes and suggests that both norm of the vectors contribute to the posterior probability.

$$\mathcal{L}_{cosface} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\|z\|(\cos(\theta_{y_i,i})-m)}}{e^{\|z\|(\cos(\theta_{y_i,i})-m)}+\sum_{j\neq y_i}^{n}e^{\|z\|\cos(\theta_{j,i}))}} \quad where\cos(\theta_j,i) = {W^T}_j z_i \tag{2.4}$$

where N is the number of training samples, x_i is the ith feature vector corresponding to the ground-truth class of y_i, the W_j is the weight matrix of the jth class, and θ_j is the angle between W_j and z_i. Additionally, the bias term is removed $b = 0$, and the weights W and embeddings z are normalised using L_2 normalisation.

An alternative loss function, ArcFace [6] differs from CosFace [5] based on its distinct margin m. ArcFace has a more accurate geodesic distance because it has a constant linear, angular margin m penalty throughout the interval, while CosFace has a nonlinear angular margin. Similarly, it normalises the weights and embeddings and fixes the bias term to zero. The ArcFace loss function is formalised as follows:

$$\mathcal{L}_{arcface} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\|z\|(\cos(\theta_{y_i,i}+m))}}{e^{\|z\|(\cos(\theta_{y_i,i}+m))}+\sum_{j\neq y_i}^{n}e^{\|z\|(\cos(\theta_{j,i}))}} \tag{2.5}$$

where all definitions are as per Eqn. 2.4. Overall the key Softmax, CosFace [5], and ArcFace [6] differences lie in their use of deep face representation, weight vectors and their margin penalty in the last layer. Consequently, the accuracy of the most popular LFW benchmark has increased from $\sim 60\%$ (Eigenfaces, [189]; 1991) to above $\sim 99\%$ (ArcFace [6]; 2019) further encouraging the broader adaption of face recognition into real-world applications.

The central concept of statistical learning is based on the requirement to choose one generalisation over another in order to be able to classify instances non-arbitrarily beyond those in the training set [56]. Moreover, [56] defines unbiased generalisation as one which makes no prior assumptions about which classes of instances are most likely to occur and bases all its decisions solely on data observation. However, any face recognition system already has dataset bias, meaning that any type of generalisation or observation based on such datasets results in bias. On the other hand, [57] identify two more different type of bias occurs in this face representation stage. The study, first, mentions DCNN *hyper-parameter bias* due to the ubiquitous number of hyper-parameters which are spanning from the choices of number of hidden nodes and layers to type of activation functions made by the user [198]. The strong influence of such chosen parameters on DCNN and their performance makes *hyper-parameter bias* relevant to racial bias as such in the case of *hyper-parameter bias*, certain models may perform better on datasets that are biased towards certain groups leading to potentially perpetuating racial bias. *Hyper-parameter bias* can also be related with aggregation bias (causing selected parameters forming the mapping function is not optimal for specific groups) defined by [199]. Another type of bias, denoted as *uncertainty bias*, is based on the probability values that are often com-

puted together with each produced DCNN architecture. The probability represents uncertainty, and typically has to be above a set threshold for face detection, verification or identification to be performed. For example, a DCNN-based face detection model reports detection predictions via probability values indicating detection confidence. However, this manual selection of the probability threshold can itself create a bias when the threshold is set too conservatively such that faces from underrepresented groups are be more likely to not be detected due to higher uncertainty in the model. Up to this point, we have described the general processes within the face representation stage of a face recognition architecture (Fig. 2.3) and the various forms of bias that may exacerbate racial bias within them. Finally, we complete our discussion of facial representation by exploring current racial bias mitigation strategies and categorise them into three sub-genres:- mutual information mitigation (Section 2.3.3), loss function based mitigation (Section 2.3.3), and domain adaptation based mitigation (Section 2.3.3).

Methods	**Backbone**	**Dataset**	**African**	**Asian**	**Caucasian**	**Indian**	**Avg**	**STD**
Imbalanced Training Sets								
ArcFace [6]	ResNet-34	MegaFace	85.13	86.27	94.78	90.48	89.17	4.39
IMAN-A [11]	ResNet-34	MegaFace	91.42	91.15	94.78	94.15	92.88	1.86
ArcFace [6]	ResNet-34	VGGFace2	87.30	85.47	93.50	87.55	88.46	3.49
ARL+C [200]	ResNet-34	VGGFace2	88.57	87.65	93.48	89.35	89.76	2.57
ArcFace [6]	ResNet-50	BUPT-Global	96.28	96.03	98.22	96.77	96.83	0.98
MV-Softmax [201]	ResNet-50	BUPT-Global	95.83	95.66	99.33	95.83	96.66	1.78
DebFace-ID [202]	ResNet-50	BUPT-Global	93.67	94.33	95.95	94.78	94.68	0.96
CurricularFace [203]	ResNet-50	BUPT-Global	94.93	95.18	97.75	96.07	95.98	1.28
RamFace [204]	ResNet-50	BUPT-Global	96.73	96.17	98.28	96.77	96.99	0.90
ArcFace [6]	ResNet-101	VGGFace2	89.45	87.61	94.71	91.21	90.75	2.91
ArcFace [6]	ResNet-101	BUPT-Global	96.77	96.52	98.55	97.48	97.33	0.91
CurricularFace [203]	ResNet-101	BUPT-Global	96.30	95.98	97.83	96.70	96.70	0.81
RamFace [204]	ResNet-101	BUPT-Global	97.40	96.93	98.65	97.57	97.64	0.73
Balanced Training Sets								
Softmax	ResNet-34	BUPT-Balanced	91.42	91.23	94.18	92.82	92.41	1.19
CosFace [5]	ResNet-34	BUPT-Balanced	92.98	92.98	95.12	93.93	93.75	1.02
ArcFace [6]	ResNet-34	BUPT-Balanced	93.98	93.72	96.18	94.67	94.64	1.10
RL-RBN [1]	ResNet-34	BUPT-Balanced	95.00	94.82	96.27	94.68	95.19	0.73
RamFace [204]	ResNet-34	BUPT-Balanced	95.28	94.83	97.15	96.08	95.84	1.02
GAC-ArcFace [205]	ResNet-34	BUPT-Balanced	94.12	94.10	96.02	94.22	94.62	0.94
Fairness FR [206]	ResNet-34	BUPT-Balanced	95.95	95.17	96.78	96.38	96.07	0.69
ArcFace [6]	ResNet-50	BUPT-Balanced	96.00	95.45	97.57	96.42	96.36	0.90
CurricularFace [203]	ResNet-50	BUPT-Balanced	94.90	94.23	96.38	95.50	95.25	0.91
RamFace [204]	ResNet-50	BUPT-Balanced	96.25	95.50	97.40	96.58	96.43	0.79
GAC [205]	ResNet-50	BUPT-Balanced	94.65	94.93	96.23	95.12	95.23	0.69
Sensitive Loss [207]	ResNet-50	BUPT-Balanced	95.82	96.50	97.23	96.95	96.63	0.62
Fairness FR [206]	ResNet-50	BUPT-Balanced	96.47	95.75	97.08	96.77	96.52	0.57

Table 2.3: Performance of state-of-the-art face verification methods on the RFW dataset [11], with comparison based on sample standard deviation.

Mutual Information Mitigation

The high mutual information between facial identity and underlying racial features within face images generally transfer into the learned feature embedding of contemporary DCNN based techniques and hence results in an unsatisfied *fairness through unawareness* criteria (i.e. the constraint of not retaining information related to s when estimating y as the the formalised problem statement of Section 2.1). A myriad of studies [91, 202, 208–213] attempt to decrease this mutual information in order to debias the performance of face recognition approaches. For example, [208] provides a general framework with a regularisation strategy such that a model trained on a dataset that is known to be bias *a priori* can be trained in to avoid the selection of biased features therein. The information bottleneck in the model distills the biased features (such as texture, background) and correctly learns to focus on relevant features (such as shape, e.g. within biased MNIST [208]). Moreover, [209] proposes a Flexibly Fair VAE (FFVAE) algorithm concerning demographic parity among multiple sensitive attributes. FFVAE learns the encoder distribution from input and sensitive attributes and disentangles prior structure in latent space by enforcing low mutual information. On the other hand, adversarial-debiasing approaches become applicable in disentangling race-related information on faces within generative generator-discriminator models such as GAN [202, 213]. For example, the Protected Attribute Suppression System (PASS) [213] discourages the generator from encoding information related with sensitive attributes via discriminator. Furthermore, [210] uses a feature mapping network to unlearn biased sensitive attributes in order to disentangle the mutual information between identity and sensitive characteristics. Similarly, [91] suppress the presence of sensitive information to enforce the learning of privacy-preserving embeddings (for any sensitive feature we want to protect) and hence equality across such sensitive attributes in any subsequent decision-making algorithms based on these embeddings. Their results show that it is possible to reduce the performance of gender and ethnicity detection by 60-80% on a given facial image embedding, while face verification performance over the same embedding is only impacted by 5% .

Other recent works on mitigating racial bias introduce a knowledge distillation module for face recognition [214–216]. Accordingly, [215] observes that the face recognition networks attend to different spatial regions in faces according to the category of an at-

tribute label (e.g. light skin vs. dark skin tone). Firstly, in order to eliminate differences in the representations, they propose a teacher-student network that enforces to student network to generate teacher-like representations. Whilst the teacher network is trained on light skin tone images, the student network is trained on dark skin tone images. However, forcing student networks to attend only teacher networks spatial regions does not give fairer results than attending both spatial regions. As a result, they achieve less biased results in face verification and perform better than state-of-the-art adversarial debiasing approaches. Another study, [216] applies knowledge distillation from teacher to student to avoid dataset bias which is identified as an imbalance distribution between either class labels or between easy and hard dataset samples. The imbalance between samples decreases the uniformity of the data, which subsequently makes the data distribution far from uniform. As image datasets are usually collected ad-hoc without any inherent uniformity consideration, they propose two different sampling methods, extrinsic sampling (before training) and intrinsic sampling (during training), to ensure the success of knowledge distillation. On the other hand, some experiments empirically demonstrate that the use of race related facial feature increases overall face classification performance and improves extracted feature discriminability [217].

Loss Function Based Mitigation

Another area of study [95, 200, 204] focuses on setting adaptive margins to tackle racial bias. For previous face recognition baselines [5, 6], the margin between classes was set at a fixed value to maximise accuracy. However, the training distributions of demographic groups and their feature embedding vectors inherently differ from each other meaning that a global margin is essentially a best fit to the largest demographic group in the training dataset. While such a constant global margin may result better performance across one demographic, that same margin may conversely cause inferior performance for another.

Recently, [204] proposes *Race Adaptive Margin (RAM) Loss* using a new compact margin instead of using an ArcFace-style fixed margin, m (Eqn. 2.5), approach. Consequently, they define intra-subject compactness μ^r_{intra} for each racial group, $\{$*African, Asian, Indian, Caucasian*$\}$, in the RFW dataset in order to assign the margin to be an identity-related parameter. As such, the final RAM Loss (denoted ramface loss, [204]) is;

$$\mathcal{L}_{ramface} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\|z\|(\cos(\theta_{y_i,i}+m_r))}}{e^{\|z\|(\cos(\theta_{y_i,i}+m_r))}+\sum_{j\neq y_i}^{n}e^{\|z\|(\cos(\theta_{j,i}))}}$$

$$where \quad m_r = \beta \times Z_r \times \hat{\mu}^r_{intra}, \quad \mu^r_{intra} = \frac{1}{B_r}\sum_{y=1}^{B_r}\frac{1}{M_{y_j}}\sum_{i=1}^{M_{y_j}}\cos\theta^r_{z_i,c_{y_j}}. \tag{2.6}$$

where B_r is the number of subject identities in the race group, M_{y_j} is the number of the samples with subject class y_j, Z_r is the race classification accuracy as the weight indicator in the adaptive margin loss, and β is the scaling parameter to constrain the upper bound of m_r. As per ArcFace loss, Eqn. 2.5, z_i is the feature representation of image x_i. Consequently, they benefit from racially-aware supervision to increase the distinctiveness of the learned feature representations and simultaneously decrease the potential for racial bias within that same representation. RamFace Loss achieves both high accuracy on face verification and appears to successfully mitigate racial bias (see Tab. 2.3).

Another study, [200] proposes an *Asymmetric Rejection Loss*, which aims to reduce the racial bias within trained face recognition models by taking advantage of unlabelled images of under-represented groups. The study utilise unlabelled images collected from online sources where the number of subject identities present is always much greater than the average images per subject. Subsequently, they consider each unlabelled image as a separate class and design an asymmetric learning procedure for those labelled and unlabelled images. Their proposed *Asymmetric Rejection Loss* (denoted *arl*) is defined as:

$$\mathcal{L}_{arl} = \mathcal{L}_L + \lambda_U\mathcal{L}_U + \lambda_C\mathcal{L}_C \quad where \quad \mathcal{L}_C = \frac{\sum_{i,j}cos(z_i,z_j)^2}{N_t}, 0 < cos(z_i,z_j) < t \tag{2.7}$$

where t is the upper bound of the penalty interval, and N_t is the number of feature representation vectors pairs whose cosine similarity lies within the interval $(0, t)$. L_L and L_U are similar to ArcFace loss equation 2.5 operating on labelled and unlabelled images respectively. Simultaneously, λ_U and λ_C are two loss weights. *Asymmetric Rejection Loss* achieves improved performance on under-represented demographic groups whilst perfor-

mance on well-represented groups remains unaffected when compared to other state-of-the-art approaches (Tab. 2.3).

Domain Adaptation Based Mitigation

Following from the discussion of Section 2.3.1 on the out-of-distribution problem, domain adaptation techniques have recently been introduced as a method to address racial bias issues [11, 218–221]. These techniques use multiple labelled source domains with different distributions to improve generalisation to new target datasets. One of the first examples of domain adaptation for racial bias, [11] prove the domain gap between racial groupings and propose a deep information maximisation adaptation network (IMAN) architecture to address this. Subsequently, [219] propose a novel face recognition methodology via the use of meta-learning named Meta Face Recognition (MFR). The meta-optimisation objective of MFR first synthesises the source/target domain. Subsequently, it forces the model to learn effective representations of both synthesised source and target domains. In another example in face recognition, [220] introduces Cross-Domain Triplet (CDT) loss based on the triplet loss [195] and uses similarity metrics from one domain to learn compact feature clusters of identities by incorporating them into another domain. Relative performance for both CDT and MFR on the RFW dataset are shown in Table 2.3.
This section presents a brief overview of face representation learning, including the potential sources of biases and mitigation studies within this stage of the face recognition processing pipeline (Fig. 2.3). In support of this review of prior work on racial bias mitigation a summary table of related work is provided to compare overall relative performance on the RFW dataset [11] (Table 2.3).

2.3.4 Face Verification and Identification

The overarching concept of *face recognition*, whereby an identity confirmation decision is made for a given subject based on facial images, can itself be subdivided into two discrete problems:- Face Verification (i.e. one-to-one facial comparison, Section 2.3.4) and Face Identification (i.e. one-to-many facial comparison, Section 2.3.4).

Face Verification

Face Verification refers to one-to-one facial comparison to verify the identity of a subject by comparing a hitherto unseen facial image against another *a priori* image of the same or different subject. This is commonly used in access control systems for both physical locations (e.g. government sites, border control) and digital assets (e.g. smart phones, digital banking applications) hence representing the most common occurrence of a *face recognition* technology encountered by the general public in contemporary society. Typically, face verification performance is measured in terms of accuracy (see Eqn. 2.2) and matching rates (see Eqn. 2.3) over pairs of identical/non-identical subject images in order to evaluate the number of correct identities matches over all the set of all paired images presented. In order to confirm a match, the feature embedding vector z_{target} from a presented unseen subject image instance x_{target}, and those of a subject image $x_{reference}$ held on record *a priori*, $z_{reference}$, are compared using a distance or similarity score across the learnt feature embedding space (e.g. cosine similarity). Subsequently, an a priori $threshold$ is used to make a decision on the similarity of $z_{target} \approx z_{reference}$ such that a verified identity can be confirmed or not. Several studies demonstrate significant performance on face verification on public benchmark datasets [15,85] where the racial diversity within these datasets is often limited, biased and overlooked [222]. Accordingly, the Labelled Faces in-the-wild Dataset (LFW) [15] contains 13233 images of 1680 subjects, and 6000 specific pairs of images of subjects to measure 1:1 verification performance have become widely adopted. Subsequently, prior work [5, 6] has reached over 99.5 % verification accuracy on LFW.

Face Identification

Face identification refers to a one-to-many facial comparison to identify an unknown facial query image by matching it to against a set of known facial images. Prototypically, law enforcement agencies use it to identify suspects in criminal investigations, track individuals in public spaces and search for missing persons. The process involves comparing an obtained query face image x_{target} with a large database of reference images $X_{enrolment}$. Unlike face verification, which is used to verify the identity of a known individual, face identification is used to identify unknown individuals by matching their facial image to a

reference image within the enrolment set for which the identity is known *a priori*. Face identification tasks can be sub-categorised as either closed-set, when the target is always in the enrolment set ($x_{target} \in X_{enrolment}$), or open-set, when the target may or may not be in the enrolment set ($x_{target} \in X_{enrolment}$ or $x_{target} \notin X_{enrolment}$). Whilst the closed-set face identification task is limited to identifying only the subjects in its enrolment set, the more challenging task of open-set face identification is able to determine unknown faces that are not in the enrolment set. In order to perform a closed-set face identification task, a multi-class classifier is used to identify the target image x_{target} via the use of feature embedding vector z_{target} over $Z_{enrolment}$. Furthermore, for an open-set face identification task an additional threshold becomes necessary in order to ascertain an unknown target that is not present in the enrolment set. As for face identification, [92] provides two large-scale face identification benchmark datasets under various imaging conditions.

Furthermore, [199] defines *evaluation bias* when the benchmark dataset used to post-training performance evaluation is not accurately representative of the target population (in deployment). The most common face recognition benchmark datasets [15, 223] illustrate examples of such evaluation bias, encouraging the development of models that only perform well on the specific racial groupings as the per distribution of the dataset (see Section 2.3.1). Evaluation bias is also related to the decisions made at this stage of the face recognition pipeline, including pairing selection, threshold optimisation, distance and normalisation functions. For example, the selected threshold can vary across datasets, and final model performance is often susceptible to the changes in these thresholds [224]. Studies have found that a single fixed threshold often causes higher variance across demographic groups than an adaptive threshold per-group threshold [224]. Another example, [225], investigates template-based face verification and identification and the effects of template size, negative set construction and classifier fusion on performance. They find that performance is highly dependent on the number of images available in a template. Subsequently, [105] compares the accuracy for African-Americans and Caucasians, in a scenario in which a fixed decision threshold is used for all subjects only to find that African-Americans have a higher FMR and Caucasians have a higher FNMR.

Accordingly, many studies provide verification protocols and a new set of pairings based on racial groupings to address racial bias. For example, the study of [11] released

the RFW dataset with a similar protocol to LFW [15] with the same number (6000) of pairings for each of the four racial groups {*African, Asian, Caucasian, Indian*} with separate thresholds. Moreover, [226] proposes the Adversarial Gender De-biasing algorithm (AGENDA) to train a shallow network that removes the gender information of the embeddings extracted from a pre-trained network. The authors of [213] extend this work with PASS to deal with any sensitive attribute and proposed a novel discriminator training strategy. Subsequently, [227] (2020a) proposed the Fair Template Comparison (FTC) method, which replaces the computation of the cosine similarity score by an additional shallow neural network trained using cross-entropy loss, with a fairness penalisation and L2 penalty term to prevent over-fitting. While this method reduces model bias, it results in an overall decrease in accuracy and requires training and tuning of the shallow neural network. Another work, [80], proposes a group-specific threshold (GST) in which the sensitive attributes themselves define its calibration sets. Another study, [228] proposes the Fair Score Normalisation (FSN) method, which is essentially GST with unsupervised clusters. FSN normalises the scores by requiring the model FMR across unsupervised clusters to be the same predefined global FMR. Salvador, [229] proposes a Fairness Calibration (FairCal) method that applies the K-means algorithm to the image feature representation vectors Z and makes partitions of the embedding space into K clusters. For each set, it calculates separate calibration map scores to cluster-conditional probabilities of the set. If the pair of images belong to the same subject cluster, the algorithm uses the score; if not, it uses the weighted average of the calibrated scores in each cluster of corresponding image features. Consequently, they achieve better overall accuracy, reducing the discrepancy in the FMRs while not requiring the use of the sensitive attribute.

Similar to face verification, open-set face identification requires a threshold to report a match or non-matched decision over test target imagery. Accordingly, [105], highlight the importance of two types of errors in face identification false-non-matched identification and false-matched identification together with their dependency on a threshold that defines the minimum similarity required to report a match. Consequently, there is a need for the design and application of open-set tests for face identification using more diverse benchmark datasets and novel evaluation strategies to measure racial bias robustly under varying conditions.

Designing an ideal evaluation strategy is yet another crucial step in the face recognition processing pipeline. This step becomes particularly important in order to address racial bias within face recognition, as every decision made at this stage can have a significant impact on the overall performance and performance across different groups. In each decision, whether related to verification or identification tasks, there is a risk of misguiding the direction of research, particularly with regards to the development of face representation models, which can result in increased racial bias. Accordingly, we summarise the related literature addressing alternative evaluation methods within this stage and illustrate the corresponding stage and source of bias in Fig 2.3.

2.4 Summary

This chapter provides a comprehensive critical review of research on racial bias within face recognition. Firstly, we discuss the racial bias problem definition formalising the notions of the face recognition evaluation process and elucidate the prominent fairness criteria associated with face recognition. Subsequently, we highlight the racial grouping requirement of current fairness criteria and discuss standard race and race-related grouping terminology under three categories; race, skin tone and facial phenotypes and compare the most prominent grouping strategies across face recognition datasets. The high reliance of prior work on racial categories brings additional challenges as the race concept is defined and understood via the influence of pre-existing prejudices and discriminatory beliefs. Furthermore, skin tone remains only one trait of a comprehensive and multi-faceted race concept. Although a broader facial phenotype approach provides a more objective and granular evaluation strategy, ensuring that racial interpretations are not reduced to only facial phenotypes whilst also considering the broader context of historical and social factors, they remain important and under-explored research topics within the broader goal of achieving more accurate and fairer face recognition performance across increasingly more diverse populations.

Furthermore, we explore the contemporary automated facial recognition multiple-stage processing pipeline providing references to related work in the literature. In each stage, we cover the outline with a related baseline, standard procedures, a potential source

of bias that can exacerbate racial bias and bias mitigation solutions. Firstly, the *Image acquisition* stage consists of sources of bias (*imagery bias, dataset bias*) that can affect the accuracy and fairness of face recognition systems. Such sources of bias within this initial stage will be transferred into the following stages and amplify racial bias in the final performance. Secondly, we consider the *face localisation* stage in terms of racial bias, where there is little attention indicating the existence of racially disparate performance, but further investigation is explicitly needed targeting racial bias within face detection itself. Thirdly, we review the most fundamental works spanning the central stage of the face recognition pipeline, *face representation*, under three sub-genres:- mutual information mitigation, loss function-based mitigation, and domain adaptation-based mitigation, providing an extensive supporting performance comparison across the RFW dataset. Finally, we investigate the final decision-making of the face recognition pipeline, *face verification and identification* and reveal the impact of decision-making within this stage on overall and group-wise face recognition performance.

Overall we observe that racial bias is present at each and every technical stage of the face recognition pipeline such that the cumulative effect remains under-explored mainly in the literature. Furthermore, we observe continued bias within the evaluation strategies employed to measure the presence of this bias themselves that directly contradict the technological needs of a modern, diverse global society.

Building upon these themes, this thesis addresses racial bias and the underlying reasons behind the performance disparities observed among different racial groups within the face recognition processing pipeline. In the next chapter, we introduce a novel methodology for evaluating racial bias using race-related facial phenotypes, eliminating the need for explicit racial grouping labels.

CHAPTER 3

Phenotype-based Racial Bias Analysis Methodology

As discussed in Chapter 2, recent work reports disparate performance for intersectional racial groups across face recognition tasks: face verification and identification. However, the definition of those racial groups has a significant impact on the underlying findings of such racial bias analysis. Previous studies define these groups based on either demographic information (e.g. African, Asian etc.) or skin tone (e.g. lighter or darker skins). The use of such sensitive or broad group definitions has disadvantages for bias investigation and subsequent counter-bias solutions design.

By contrast, this chapter introduces an alternative racial bias analysis methodology via facial phenotype attributes for face recognition. Subsequently, we use the set of observable characteristics of an individual face where a race-related facial phenotype is hence specific to the face and correlated to the racial profile of the subject. Finally, we propose categorical test cases to investigate the individual influence of those attributes on bias within face recognition tasks. We compare our phenotype-based grouping methodology with previous grouping strategies and show that phenotype-based groupings uncover hidden bias without reliance upon any potentially protected attributes or ill-defined grouping strategies. Furthermore, we contribute corresponding phenotype attribute category labels for two dataset: RFW (face verification) and VGGFace2 (test set) (face identification).

The material presented in this chapter of the thesis has been published in the following peer-reviewed publication:
Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby P. Breckon., Measuring Hidden Bias within Face Recognition via Racial Phenotypes., IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, pp. 995-1004 2022.

3.1 Introduction

An increasing number of automated face recognition systems have been deployed by companies, nonprofits and governments to make autonomous decisions for millions of users [24]. Such wide-scale adoption within real-world scenarios brings with it valid concerns on the potential abuse of face recognition due to the presence of data and algorithmic bias [28, 29]. The most common issue pertaining to such bias arises in racial groups [30]. Subsequently, the research community have been focused on methods that rely on demographic or skin tone group annotations drawn from public face recognition benchmark datasets [3, 86]. This provides algorithmic performance on such predefined groupings to measure bias. However, current grouping annotations and related bias evaluation strategies may lead to unintended negative implications (as it is discussed in Chapter 2). In parallel, this chapter proposes a phenotype-based evaluation strategy for racial bias within face recognition. We now briefly illustrate our motivation in four key points.

Ambiguous Definition of Race: The historical and biological definitions of race vary and racial context is not fixed over time [230]. Such ambiguity becomes more problematic for the face recognition literature, as many researchers do not provide any related background about the details of their racial categorisation design process [72]. However, racial groupings are critical to the effective evolution of face recognition methodologies as they often represent the all-important means of quantitative evaluation. As in any recognition task, poorly defined groupings result in skewed mean and standard deviation measures of relative performance due to the ill-posed boundary conditions on membership of each group that can cause a given an example to justifiably transit from one group to another.

Privacy of Protected Attributes: Auditing benchmark datasets can cause potential pri-

vacy and consent violations [73] for dataset subjects. For example, exposing demographic origin may enhance the representations of a group under threat, leading to the potential for racial profiling and associated targeting [69]. As information of racial or ethnic origin is sensitive [103], researchers should either avoid revealing such sensitive data or provide an appropriate context for use [73].

Confined Groupings: Skin or racial grouping strategies such as binary $\{$*light vs. dark; black vs. white*$\}$ for evaluating racial bias limits the scope of any study as they fail to capture the whole aspect of the bias problem where it needs to consider both multi-racial or less stereotypical members of such groups instead [39, 70] use Fitzpatrick skin tone groupings to evaluate racial bias, but one such skin-tone based racial grouping contains multidimensional traits including nose, hair type, eye, and lips [231]. Leveraging all such traits together instead brings improved interpretations and derivations to address racial bias.

Racial Appearance Bias: Maddox [136] explains racial appearance bias as a negative disposition toward phenotypic variations in facial appearance. He also [104] discusses how race-conscious social policies may fail to address racial biases in the treatment and outcomes of disadvantaged groups. Many studies show that individuals with more stereotypical racial appearance suffer poorer outcomes than those with less stereotypical appearance for their race [104, 137, 138]. On the other hand, a better understanding of the role of phenotypic variation complements solutions for both racial bias [136]. By way of phenotype, we mean the set of observable characteristics of an individual face where a race-related facial phenotype is hence specific to the human face and correlated to the racial profile of the subject.

Accordingly, we propose using race-related facial (phenotype) characteristics within face recognition to investigate racial bias. We categorise representative racial characteristics on the face and explore the impact of each characteristic phenotype attribute: skin tones, eyelid type, nose shape, lips shape, hair colour and hair type. We audit these attributes for two different publicly available face datasets: VGGFace2 (test set) and RFW. We assess the impact of both attribute-based and subgroup-based evaluations on racial bias of face recognition tasks. We utilise two different training protocols for face verification to compare performance disparities between imbalance and racially balanced

training datasets. We compare our phenotype-based evaluation strategy with race or skin tone based grouping evaluation. We show that our strategy provides a more elaborate perception of bias without revealing any potentially protected or ill-defined information. This chapter presents a new evaluation strategy using facial phenotype attributes to investigate and measure racial bias with greater granularity within face recognition tasks.
Our key contributions are as follows:

- We propose a new evaluation strategy that uses facial phenotype attributes rather than race labels to measure racial bias within both attribute-based and subgroup-based performance of state-of-the-art face recognition algorithms.
- We contribute additional facial phenotype attribute labelling for the VGGFace2 (face identification) and RFW (face verification) benchmark face datasets.
- We uncover the potentially hidden source of bias within the evaluation of racial groups, which is supported by quantitative evidence.

3.2 Racial Phenotypes on Face Images

Quine [75] presents three possible definitions of the race concept: a genetic variation between humans, morphological attributes, and genetically determined psychological characteristics. These morphological attributes are the primary interest for resolving racial bias in face recognition. For morphological attributes, studies [130, 131] focus on the impact of human phenotype characteristics over race estimation. They categorise the attributes by considering biological traits. The study of Shades of Race [12] investigates the marginal effects of phenotypic characteristics including skin tone, lips, nose, hair and body type on racial categorisation. Zhuang [132] considers 21 anthropometric measurements such as face width, length, nose breadth and length, eye corner points. He finds statistically significant differences in facial measurements between four racial/ethnic groups, which are $\{$*Caucasian, Hispanic, African, other (mainly Asian)*$\}$ as discussed in Section 2.2.

We adopt such groupings and measurements for face recognition by considering two limitations. Firstly, effectively evaluating face recognition tasks requires tight cropped (e.g. 112×112 px) low-quality images containing occlusion, shadows, and illumination

variations for both the training and test stages. This makes phenotype attribute detection on the specific characteristics of face dataset images more difficult when compared to real-world human faces. Secondly, the broader categorisation increases the number of potential groupings, making bias evaluation inefficient for face recognition systems. Correspondingly, we decide to use 6 primary attributes that define the phenotype groupings for our methodology: skin tone, eyelid type, nose shape, lip shape, hair type and hair colour[1]. Subsequently, we have 21 different attribute categories under the 6 primary attributes as listed in Table 3.1.

Skin Tone: In chapter 2, we highlight that the use of simplistic binary groupings can lead to erroneous or conflicting interpretations for racial bias. To address this issue, we adopt Fitzpatrick Skin Tones for the proposed race-related phenotype grouping strategy as a more robust alternative skin tone scale. Fitzpatrick Skin Tone Scale [110] provides six different skin tone categories including $\{$*Type 1, Type 2, Type 3, Type 4, Type 5, Type 6*$\}$.

Eye Shape: The appearance of the human eye has been grouped by its position, shape and settings in many cosmetic industry guidelines [232]. However, they have either no scientific background or solid relation with race. Instead, we look into epicanthal folds and check eyelid difference as it is a more distinctive attribute for racial bias [233]. We categorise eye shapes into two categories: $\{$*Monolid, Other*$\}$, based on whether or not an individual has a monolid. We acknowledge that a single attribute category can be observed in multiple race groups (i.e. individuals of non-East Asian ancestry can also have monolid eye shapes.). However, our main concern is identifying the most observable and convenient race-related phenotype attributes on images to evaluate the bias (see Table 3.1).

Nose Shape: Nasal breadth refers to the distance between the two nasal bones at the widest point of the nose, usually measured at the base of the nasal bridge. It has been used as an important anthropometric measurement [132]. Although, there is a relationship between nasal breadth and race, nasal breadth can vary significantly among different racial and ethnic groups. Nevertheless, studies have shown that individuals of African

[1]We note that hair information is still present in the tightly cropped images, and it may be helpful for future automated facial analyses tasks.

and some Southeast Asian ancestry tend to have wider nasal breadths on average than individuals of European ancestry [12]. Accordingly, for the appearance of the nose, we use two categories, $\{$*Wide and Narrow*$\}$, by examining the nasal breadth [132].

Lip Shape: Studies provide evidence that individuals of African ancestry and indigenous American ancestry tended to have thicker lips compared to individuals of European and East Asian ancestry [234]. To capture this phenotypic variation in lip shape, we include the lip shape attribute with two distinct categories: $\{$*Full and Small*$\}$.

Hair Type and Colour: Hair texture is labelled into eight categories using the frequency of twists, waves, and curve diameter metric by [235]. However, we need to narrow categorisation to make annotation possible for attribute categorisation on low-quality images. Here we utilise eight categories and group them into three main hair texture types: $\{$*Straight, Wavy, Curly, Bald*$\}$. Despite being the most artificially manipulable attribute, we retain hair colour as it is related to skin tone [236] the categories for hair colour we use: $\{$*Red, Grey, Black, Blonde, Brown*$\}$ (see Table 3.1).

Phenotype Attributes	Categories	RFW	VGGFace2
Skin Tone	Type 1 / 2 / 3 / 4 / 5 / 6	0.71	1.14
Eyelid Type	Monolid / Other	0.80	1.09
Nose Shape	Wide / Narrow	0.24	0.18
Lip Shape	Full / Small	0.28	0.63
Hair Type	Straight / Wavy / Curly / Bald	0.70	1.11
Hair Colour	Red / Blonde / Brown / Black / Grey	1.23	0.67

Table 3.1: Facial phenotype attributes and their categorisation based on [12] along with normalised standard deviations σ/μ.

Finally, we present a collection of example images demonstrating the race-related phenotype attributes and their corresponding categories in Figure 3.1.

3.3 Annotation of Racial Phenotypes

In order to obtain race-related facial phenotype attribute category labels, we require a platform that enables annotating facial images from datasets. Although we have made attempts to use 3rd party annotation platforms and companies, a lack of prior knowledge in the random annotators results in erroneous annotations. Accordingly, we built a platform,

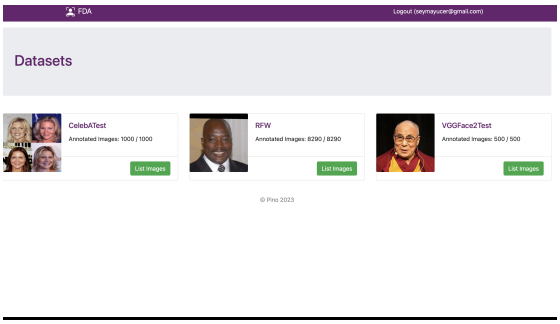
Figure 3.1: A collection of example images illustrating the race-related phenotype attributes and their corresponding categories.

which is explained in detail in this section, to address this issue.

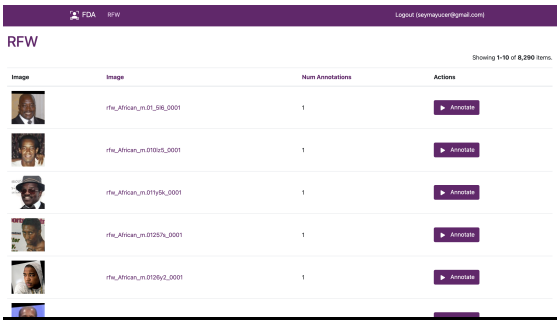
3.3.1 Annotation Platform

We build a web-based annotation tool platform, Face Dataset Annotator (FDA), that enables annotators to audit multiple facial phenotype attributes on face images from multiple datasets. The tool is designed to be easy to use via touch screen devices such as tablets as well as desktop computers. FDA supports simultaneous annotation by multiple experts, with revision support for both dataset images and annotations. Each image can be annotated by multiple experts, and annotations can be exported with their timestamped metadata.

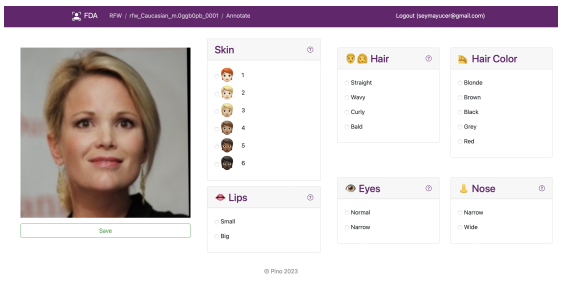
The platform includes a login page for annotators to securely access the platform. Once logged in, annotators can view and annotate images from various datasets. Moreover, within the platform, guidance pages are provided for each phenotype attribute to help annotators identify and classify the attributes accurately. These guidance pages shows examples and explanations for each phenotype attribute, ensuring that annotators have a clear understanding of what they are annotating. This helps to ensure consistency in the annotation process and improves the accuracy of the resulting annotations. We provide an illustrative set of exemplar screenshots from the user interface of the annotation platform showing datasets screen, annotation screen and annotation guidance screen (Figure 3.2).



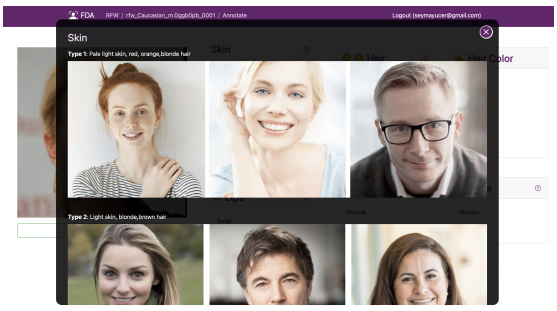
Datasets screen for annotations.



List of samples within the selected dataset.



Annotation screen for a sample.



Annotation guidance screen for skin tone.

Figure 3.2: Exemplary screens from Face Annotation Platform.

3.3.2 Annotation Process

Previously in Section 3.2, we explain how we define racial phenotype attributes and their categories. Before the annotation process, we choose the most established face recognition datasets to validate our proposed methodology. For the face verification task, we choose the RFW dataset [11] as it provides a relatively broader racial distribution of subjects where each subject contains 3-5 images. For face identification, we use the VGGFace2 closed-test set [3], which contains at least 300 images per subject. For both

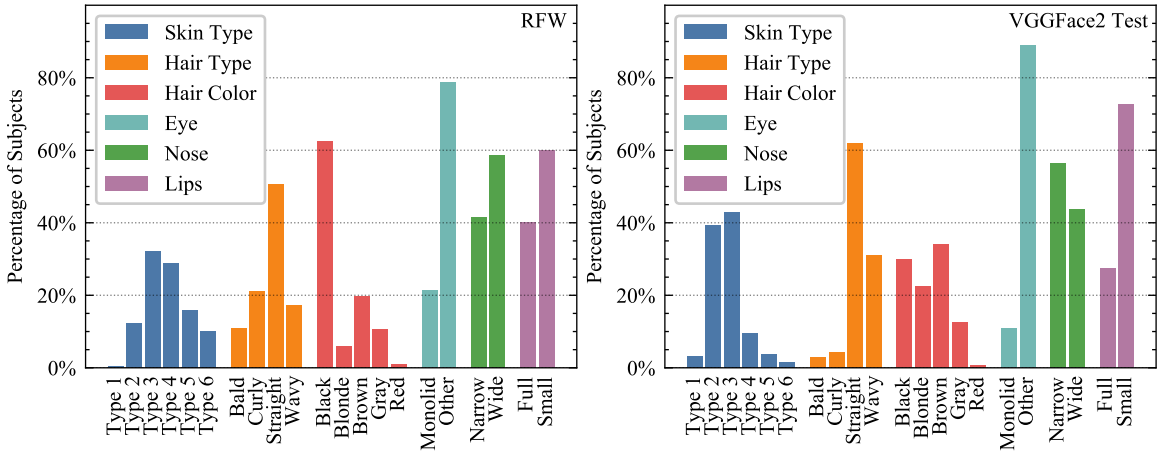


Figure 3.3: The distribution of facial phenotype attributes of RFW (left) and VGGFace2 Test (right) datasets.

datasets, we design an annotation interface to make the annotation process both user-friendly and robust. Each subject is presented with attribute category selectors next to a set of face images within the annotation interface. Subsequently, an experienced annotator who has experience in morphological differences among races annotates each subject using the interface.

We obtain 11654 subjects annotations from the RFW and VGGFace2 benchmark datasets. Each annotation took 10-20 seconds, and overall annotation took 12 days (i.e. annotator working at a maximum of 6 hours per day with regular breaks). The result of this annotation process, the phenotype attributes distributions for the RFW and VGGFace2 benchmark datasets, are shown in Figure 3.3 left/right, respectively. We also present the normalised standard deviations (Coefficient of Variance), σ/μ, among attribute categories of benchmark datasets to show the level of imbalance within these categories in Table 3.1. For both datasets, we can observe that the dominant phenotype attribute categories are skin tone 3, Straight Hair, Narrow Nose, Other (non-monolid) Eyes, Small Lips, which correlates to the dominant presence of Caucasian faces based on the analysis of Figure 3.3.

3.4 Results and Discussion

In this section, we analyse the performance of our phenotype-based grouping methodology for face recognition tasks. We provide a public reference implementation, dataset reference links and pre-trained models [2].

3.4.1 Training Protocols

Protocol 1 (Imbalanced Training Data): We train ArcFace [6] with a ResNet100 [14] on the VGGFace2 benchmark datasets that contains 8631 subjects where subject distribution is racially imbalanced. Here, our specific choice of VGGFace2 is due to investigate the impact of imbalanced training data that includes data bias on our proposed evaluation strategy.

Protocol 2 (Racially Balanced Training Data): We use a ResNet34 [14] backbone architecture with the Softmax loss [4] trained on the BUPT-Balanced benchmark dataset [1] that contains 28000 face subjects. The BUPT-Balanced has racially balanced distributions among four groups $\{$*African, Asian, Indian, Caucasian*$\}$ with 7000 face subjects each. The primary purpose of protocol 2 is to assess the impact of a racially balanced training dataset on results over the bias using our proposed phenotype-based methodology. We compare how much a racially balanced training dataset improved the performance difference compared to protocol 1.

3.4.2 Face Verification

Face verification, also known as one-to-one verification, is the task of comparing two different facial images to estimate whether they belong to the same individual subject. We follow two pairing strategies to explore the impact of single attribute (attribute-based) and appearance-based facial groupings (subgroup-based) on the evaluation performance of face verification.

Attribute Name	Protocol 1 Accuracy %	Protocol 2 Accuracy %
Blonde Hair	97.02	96.63
Red Hair*	96.33	96.83
Type 2	96.22	95.83
Gray Hair	94.85	95.83
Bald	94.75	95.70
Wavy Hair	94.32	95.50
Brown Hair	94.25	94.83
Type 6	93.77	94.77
Narrow Nose	92.92	94.77
Type 5	92.15	94.38
Curly Hair	92.02	93.63
Small Lips	91.92	94.98
Type 3	91.72	93.77
Type 1*	91.31	89.51
Straight Hair	91.25	94.32
Wide Nose	90.68	91.02
Full Lips	89.98	93.23
Type 4	89.90	93.55
Other Eye	89.88	93.75
Black Hair	89.88	91.42
Monolid Eye	88.27	89.73
σ	2.44	2.06
σ^*	2.39	1.77

Table 3.2: Attribute-based face verification performance of RFW. σ represents the standard deviation of all attribute category accuracies, whilst including red hair and type 1, σ^* represents the standard deviation excluding these specific attribute cases.

Attribute-based Face Verification

Firstly, we generate pairs from images containing the same attribute category,for example of facial images from people who all have monolid eyes. Consequently, we compare individual attributes performance using both training protocols for face verification.

For attribute-based face verification, we randomly select 20k positive and 20k negative pairs from all possible pairs of each attribute. We calculate the cosine similarity of feature encoding of all selected negative and positive pairs to obtain the most challenging pairs. Subsequently, we select the most similar 3000 pairs from the negative samples and the

[2] https://github.com/seymayucer/FacialPhenotypes

least similar 3000 pairs from the positive samples for each attribute category in Table 3.2. Due to the low statistical occurrences of the Type 1 skin tone and red hair colour such that we do not have enough samples to generate 6000 pairs, we instead produce 602 pairs (301 positive, 301 negative) for Type 1, 1200 (600 positives, 600 negative) pairs for red hair.

In this way, we measure each face attributes accuracy using on face verification performance. We use both training protocols to show how much standard deviation (σ) changes between balanced and imbalanced training data. We present the performance variation across the attribute-based sample groups in Table 3.2 as a standard deviation of accuracy both excluding the low sample occurrence attributes of red hair and Type 1 attribute accuracy both ($\sigma*$) and including them (σ) . It is clear from Table 3.2 that for both protocol 1 (imbalanced training data) and protocol 2 (racially balanced training data), accuracy is lower for monolid eyes, black hair, full lips, and wide nose than the other eye, blonde hair, and small lips, and narrow nose respectively. We also do find a slight correlation between darker skin tones and higher false matching rates when we pair from the same attribute categories (Table 3.3). Moreover, although the imbalanced training protocol results a bigger performance difference (σ) , the amount of difference between two protocols is small, meaning that a racially balanced dataset distribution is not sufficient to overcome performance bias.

Additionally, NIST [30] suggests providing false matching rates of pairing combinations between each grouping in the dataset as it is necessary for real-world scenarios. Therefore, we pair each attribute category with all other attribute categories to assess cross-attribute pairing performance. Subsequently, we evaluate false matching rates between any attribute category pair combination in Figure 3.4. We randomly generate 10000 pairs for each category pairings; in total, we have 441 (21×21) pairings. For example, each cross-attribute pairings means 10000 pairs between blonde hair - monolid eye, type 3 - wide nose or wavy hair - full lips etc. As a result of this, we clearly show that Type 5, Type 6 and monolid eyes pairings have higher false matching rates among all attribute categories in Figure 3.4 using training protocol 1. Consequently, the impact of the dark skin tones on performance increases for cross-attribute pairings compared to the attribute-based pairings.

	Protocol 1			Protocol 2		
Attribute Name	**F1**	**FNMR**	**FMR**	**F1**	**FNMR**	**FMR**
Blonde Hair	96.85	1.40	3.83	96.04	2.53	3.13
Red Hair	96.60	2.83	4.00	96.48	2.83	2.83
Type 2	95.98	3.10	3.90	95.25	3.77	3.90
Bald	95.32	5.00	5.37	95.44	3.23	4.97
Gray Hair	95.00	3.70	5.87	95.93	2.53	5.13
Brown Hair	94.46	6.40	4.63	94.08	5.43	4.43
Type 6	94.42	4.10	7.30	95.01	5.53	4.67
Wavy Hair	93.96	3.27	7.53	95.42	4.97	3.77
Narrow Nose	93.05	6.77	7.20	94.29	4.40	5.87
Type 5	92.72	4.07	10.33	94.45	5.63	5.47
Curly Hair	92.51	5.47	9.67	93.58	6.87	5.70
Small Lips	92.36	5.80	8.37	94.29	5.03	4.70
Type 1	92.08	5.99	6.45	90.14	6.67	9.41
Type 3	91.80	8.63	7.73	93.59	5.93	6.13
Straight Hair	91.19	9.17	6.87	93.97	4.30	6.53
Other Eye	91.16	7.23	7.27	93.76	7.43	4.47
Wide Nose	90.99	7.23	7.43	89.78	7.10	5.27
Full Lips	90.73	6.60	10.17	93.43	7.13	5.77
Type 4	90.45	8.30	8.53	93.50	5.70	6.93
Black Hair	90.12	7.77	8.73	90.50	6.83	5.83
Monolid Eye	88.84	9.53	13.03	90.62	8.47	6.93

Table 3.3: Attribute-based face verification F1, FNMR, FMR scores of RFW dataset on both training protocols.

Furthermore, we present attribute-based face verification scores including False Non-Match Rate (FNMR), False match rate (FMR) and F1 score in the Table 3.3. We use the same pairings and protocol presented in the Section 3.4.2 for Table 3.2 [15]. Whilst F1 scores are correlated with Table 3.2 accuracies, for the imbalanced training protocol 1, the false matching ratio is higher on attributes like Monolid Eye, Type 6/5/4/3, Wide Nose, Full Lips than the different categories under the same attribute. Moreover, we observe that the balanced training protocol 2 improves the FMR while increasing the FNMR for the attribute categories with higher accuracies and F1 scores.

Subgroup-based Face Verification

Secondly, we create various subgroups with different phenotypic attribute combinations in the dataset. For example, one such subgroup consists of subjects with skin tone 3,

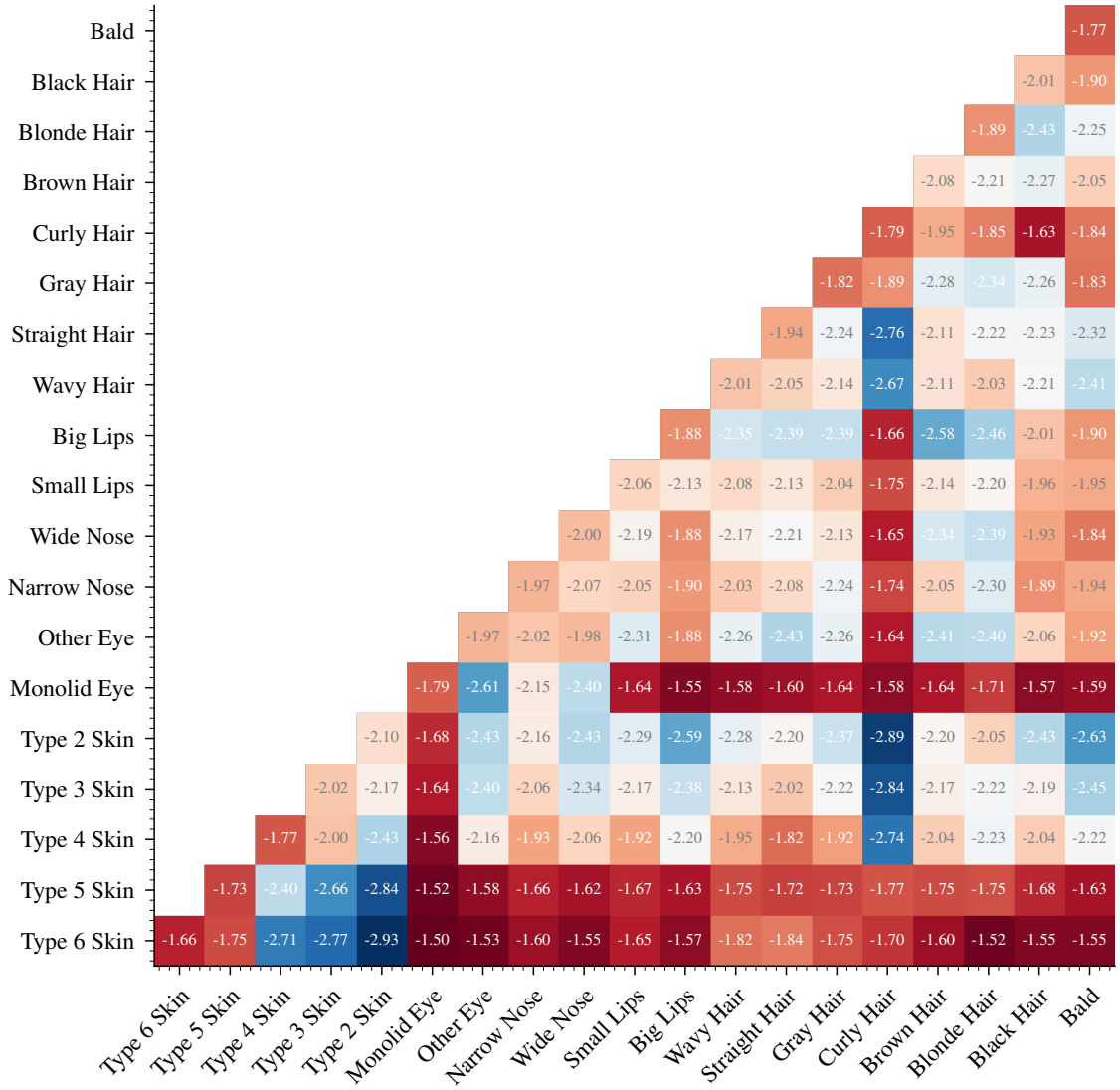


Figure 3.4: False matching rates (FMR) of cross-attribute based pairings for 21 attribute categories using training protocol 1. Each cell depicts FMR on a logarithmic scale which is $log_{10}(FMR)$ with lower negative values (close to zero) encoding superior false match rates.

monolid eyes, straight hair, wide nose, and small lips. Our main purpose of such pairing is to show the effects of single attribute changes over a given grouping. For instance, what would change when only skin gets darker, but other attributes remain the same?

Furthermore, we generate all possible subgroups with different phenotypic attribute category combinations to investigate subgroup-based performances. However, we need to limit the number of subgroups such that we can present our results efficiently. We first remove the hair colour attribute as it is the easiest race-relevant attribute that individuals can readily modify via styling. Consequently, we merge skin tones into three groups and

Skin	Lips	Eye	Nose	Hair Type	Ratio (%)	Accuracy (%)	Skin	Lips	Eye	Nose	Hair Type	Ratio (%)	Accuracy (%)
{1,2}	Small	Other	Narrow	Straight	3.82	96.53	{3,4}	Full	Monolid	Wide	Straight	1.55	91.63
{3,4}	Small	Other	Narrow	Straight	7.43	96.45	{1,2}	Small	Other	Narrow	Bald	0.28	91.29
{3,4}	Small	Other	Narrow	Wavy	3.67	96.11	{5,6}	Full	Other	Narrow	Curly	1.97	91.23
{1,2}	Small	Other	Wide	Straight	3.03	95.63	{3,4}	Small	Other	Wide	Bald	1.68	91.01
{1,2}	Small	Other	Narrow	Wavy	1.64	95.62	{1,2}	Full	Other	Narrow	Wavy	0.27	90.74
{1,2}	Full	Other	Narrow	Straight	0.70	95.59	{3,4}	Small	Monolid	Wide	Wavy	0.96	90.17
{3,4}	Full	Other	Narrow	Straight	3.59	95.28	{1,2}	Small	Other	Wide	Bald	0.46	89.78
{3,4}	Full	Other	Wide	Straight	4.47	94.98	{5,6}	Small	Other	Narrow	Curly	0.81	89.50
{3,4}	Small	Other	Wide	Wavy	2.95	94.92	{3,4}	Small	Monolid	Narrow	Wavy	1.20	89.35
{3,4}	Small	Other	Wide	Straight	8.83	94.92	{5,6}	Full	Other	Wide	Curly	13.09	89.18
{1,2}	Full	Other	Wide	Straight	0.33	94.87	{3,4}	Full	Other	Wide	Bald	0.80	86.02
{1,2}	Small	Other	Wide	Wavy	0.72	94.56	{5,6}	Small	Other	Wide	Bald	0.99	85.90
{3,4}	Small	Other	Wide	Curly	0.51	93.89	{3,4}	Full	Other	Wide	Curly	0.46	85.38
{3,4}	Full	Other	Wide	Wavy	1.90	93.41	{3,4}	Small	Monolid	Narrow	Bald	0.32	84.10
{3,4}	Full	Other	Narrrow	Wavy	1.94	93.10	{5,6}	Small	Other	Narrow	Bald	0.30	82.81
{3,4}	Small	Other	Narrow	Bald	0.68	92.50	{3,4}	Small	Monolid	Wide	Bald	0.52	82.67
{3,4}	Small	Other	Narrow	Curly	0.31	92.45	{3,4}	Full	Monolid	Narrow	Wavy	0.43	82.04
{5,6}	Small	Other	Wide	Curly	2.81	92.23	{5,6}	Full	Other	Narrow	Bald	0.53	81.24
{3,4}	Small	Monolid	Wide	Straight	6.59	91.93	{1,2}	Small	Monolid	Narrow	Straight	0.47	81.04
{3,4}	Full	Monolid	Narrow	Straight	1.81	91.78	{3,4}	Full	Monolid	Wide	Wavy	0.27	79.47
{5,6}	Full	Other	Wide	Bald	3.62	91.74	{5,6}	Full	Other	Wide	Wavy	0.32	78.94
{3,4}	Small	Monolid	Narrow	Straight	7.95	91.70							
σ													5.07

Table 3.4: Subgroup-based face verification performance of RFW using training protocol 1, sorted by descending order of accuracy.

show them as {1,2} for Type 1 and Type 2, {3,4} for Type 3 and Type 4, and {5,6} for Type 5 and Type 6. Lastly, we remove subgroups with a few or even no samples in the test set, which comprises 3% of all samples. In Table 3.4, we show the performance of each subgroup with its proportion in the original test dataset. To evaluate the performance, we generate 6000 pairs (3k positive and 3k negative) from all possible pairs of subgroups that have enough samples. For the rest, we generate an equal number of negative and positive pairs as much as availability facilitates. From our observation of Table 3.4, we can conclude that groups who have one of the attributes like wide nose, full lips, and monolid eye type always have less accuracy than the other groups with a narrow nose, small lips and other eye (when rest of the attributes are same). Furthermore, whilst the average accuracy of the subgroups with Type {5,6} skin tone is 86.97%, subgroups with Type {1,2} skin tone is 92.56%, but this notably includes other attribute effects.

Moreover, the number of subgroup variations with darker skin tones are much smaller than lighter tones which causes many different evaluation and analysis problems. It lacks sufficient interpretation in the test phase; there are minorities in the global populous with dark skin and monolid eyes or any other less common variations. Benchmark datasets

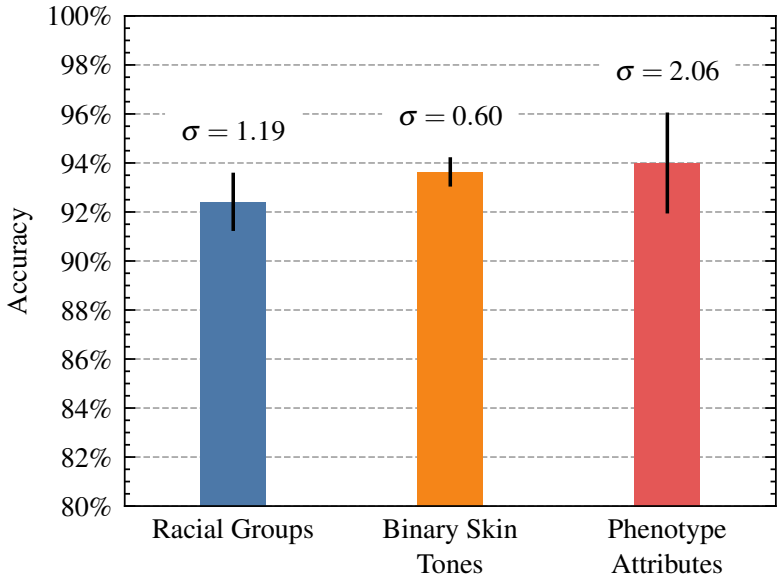


Figure 3.5: Accuracy variations for three grouping strategies. Standard deviation of the groupings reflects the amount of measured bias. Racial groupings {*African, Asian, Caucasian, Indian*} accuracies are obtained from [1]. Binary skin tones {*lighter skin-tone, darker skin-tone*} are the average accuracy of Type 1-3 and Type 4-6 skin tones, respectively.

do not contain enough representation for such minority groups. An improved evaluation dataset would be one that is able to cover more phenotype combinations such that its distribution is an unbiased representation of the global populous.

Lastly, we estimate such disparities among different grouping strategies using training protocol 2. We take racial groupings {*African, Asian, Indian, Caucasian*} and binary skin tone groupings {*lighter skin-tone, darker skin-tone*} as they are very common grouping strategies in the literature. We compare them with our phenotype-based grouping strategy. In Figure 3.5, we show that how accuracy and the standard deviation differs between sub-groups in three different strategies. Higher variation reveals hidden bias, which may be missed in narrow, erroneous racial or binary skin tones grouping strategies. The phenotype-based grouping strategy brings a more granular observation of the variability in performance (i.e. higher standard deviation) and hence a more resolute measure of performance bias.

Attribute	Ratio (%)	Acc (%)	Attribute	Ratio (%)	Acc (%)
Bald	2.80	97.49	Type 6	1.60	96.25
Grey Hair	12.60	97.47	Wide Nose	43.60	96.19
Red Hair	0.80	97.10	Type 3	42.80	96.13
Type 5	3.60	96.87	Brown Hair	34.20	96.05
Type 4	9.60	96.75	Curly Hair	4.40	95.93
Small Lips	72.60	96.56	Wavy Hair	31.00	95.92
Type 2	39.20	96.43	Monolid Eye	11.00	95.73
Black Hair	29.80	96.43	Blonde Hair	22.60	95.52
Straight Hair	61.80	96.35	Full Lips	27.40	95.36
Other Eye	89.00	96.29	Type 1	3.20	92.90
Narrow Nose	56.40	96.26			
σ					0.93

Table 3.5: Face identification performance on VGGFace2 test set using standard linear SVM and features from training protocol 1, sorted by descending order of accuracy.

3.4.3 Face Identification

Face identification as a one-to-many verification is the task of searching for a face across a facial database. There are two scenarios for face identification applications based on whether a queried face is enrolled in a database or not. Open-set identification assumes the database does not necessarily contain the queried face, while closed-set identification always looks for a match in the database. In this chapter, we apply closed-set identification using the test set of the VGGFace2 benchmark dataset on the originally proposed protocol [3] and we extract the image features using training protocol 1 [6]. We apply a 5-fold train-test split where we sample 50 images from each subject as the test set and use the rest as the training set. We train a standard linear SVM on the extracted feature representations and predict the identities for test samples. Our results are shown in Table 3.5 where we can observe that the standard deviation (σ) is much smaller when compared to the earlier attribute-based face verification results of Table 3.2. It shows that the closed-set face identification does not have the same level of bias correlation as we find for face verification. However, in this experiment, we are unable to have the same proportion for each attribute, and we did not measure open-set face identification. As suggested in [30], future work should design and apply open-set tests for face identification on better-distributed benchmark datasets to measure bias extensively.

3.5 Summary

In this chapter, we propose a new evaluation strategy using facial phenotype attributes to assess racial bias in face recognition tasks. Firstly, we annotate facial phenotype attributes on the VGGFace2 and RFW datasets according to proposed attribute categories and provide a public reference implementation, dataset reference links and pre-trained models. Such grouping strategy and attribute category overview are presented in the last rows of Table 2.1 within the Facial Phenotype category for comparison with other grouping strategies.

We elaborate experimental results to show the impact of each phenotype attributes using two different training protocols, including imbalanced and racially balanced training sets. We also provide different pairing strategies for face verification to draw attention to the importance of pairing for comprehensive evaluation.

Furthermore, we reveal apparent performance differences between race-related phenotype attribute categories and subgroups for both training protocols. However, we also uncover more considerable performance disparities among phenotype attributes than racial groups. More specifically, the results reported in Table 2.3 show the standard deviation and average accuracy across racial groups using BUPT-Balanced benchmark dataset [1], ResNet34 architecture [14], and Softmax loss [4] ($std = 1.19, acc = 92.41$) align with Figure 3.5, which visually depicts the accuracy distribution across four racial groups using the racial grouping strategy. Subsequently, Figure 3.5 shows higher variation reveals hidden bias, which may be missed in narrow, erroneous racial grouping strategy. Crucially, our phenotype-based evaluation strategy reveals racial bias in facial analysis models more comprehensively while avoiding exposing potentially protected or ill-defined racial attributes.

CHAPTER 4

On the Impact of Lossy Image Compression on Racial Bias within Face Recognition

This chapter investigates the impact of commonplace lossy image compression on face recognition algorithms with regard to the racial characteristics of the subject. We adopt our previously proposed racial phenotype-based bias analysis methodology, from Chapter 3, to measure the effect of varying levels of lossy compression across racial phenotype categories. Additionally, we determine the relationship between chroma-subsampling and race-related phenotypes for recognition performance. Prior work investigates the impact of lossy JPEG compression algorithm on contemporary face recognition performance [164, 237]. However, there is a gap in how this impact varies with different race-related inter-sectional groups and the cause of this impact. Via an extensive experimental setup, we demonstrate that common lossy image compression approaches have a more pronounced negative impact on facial recognition performance for specific racial phenotype categories such as darker skin tones (up to 34.55%). Furthermore, removing chroma-subsampling during compression improves the false matching rate (up to 15.95%) across all phenotype categories affected by the compression, including darker skin tones, wide noses, big lips, and monolid eye categories. In addition, we outline the characteristics that may be attributable as the underlying cause of such phenomenon for lossy compression

algorithms such as JPEG.

The material presented in this chapter of the thesis has been published in the following peer-reviewed publication:
Seyma Yucer, Matthew Poyser, Noura Al Moubayed, and Toby P. Breckon., Does lossy image compression affect racial bias within face recognition?, IEEE International Joint Conference on Biometrics, IJCB, pp. 1-10, 2022.

4.1 Introduction

Previously, we have discussed that each and every technical stage of the face recognition pipeline is prone to bias (Chapter 2). However, most research focuses on the latter aspects of dataset collation and face representation stages to explore and mitigate such bias [95, 205, 207]. As such, many datasets and annotations have been released [1, 45], generative adversarial networks have been explored to enrich under-represented groups during training [9, 238] and regularisation methods have been proposed to minimise performance differences between subgroups [208]. Furthermore specific evaluation methodologies have been devised to tackle bias collaboratively [13, 239, 240]. Despite this plethora of research, no studies examine the potential impact of image acquisition decisions (imaging bias) when addressing racial bias within face recognition. Any source of bias at this early stage is just propagated and exacerbated within contemporary face recognition processing stages [49].

On the other hand, existing image acquisition standards for face recognition systems such as ISO/IEC 19794-5 [157] and ICAO 9303 [158] propose both image-based (i.e. illumination, occlusion) and subject-based (i.e. pose, expression, accessories) quality standards to ensure facial image quality. Accordingly, facial images should also be stored using lossy image compression standards such as JPEG [159] or JPEG2000 [160]; and identifiable for gender, eye colour, hair colour, expression, facial properties (e.g. wearing glasses), pose angles (yaw, pitch, and roll), and landmark positions. However, common face recognition benchmarks do not conform to the ISO/IEC 19794-5 and ICAO 9303 standards. Moreover, in-the-wild samples are often obtained under the varying camera

and environmental conditions to challenge the proposed solutions. Nevertheless, most facial image samples within such datasets are compressed via lossy JPEG compression [159].

Accordingly, some limited previous work [241–243] focuses on the impact of low-quality, blurred, noisy or distorted imagery on Convolutional Neural Network (CNN) based image recognition or classification. Dodge and Karam [244] highlight a significant decrease in contemporary neural network performance, whilst human examiners remain resilient to such factors. Particularly, Torfason [245] focuses on compression methods and bypasses the decoding phase of image compression. They point out that encoded representations are more advantageous than compressed/decoded images for classification and semantic segmentation. Poyser [163] evaluates the impact of lossy compression algorithms on various CNN architectures, in which they measure the robustness and performance impact of compression for various computer vision tasks. They determine that, in general, CNN architectures can be resilient to the introduction of lossy JPEG compression artefacts if the initial training regime includes the use of compressed images [163]. These results align with the findings of Zanjani [246], who considers the impact of JPEG 2000 compression [160] on CNN for cancer diagnosis systems. Indeed, retraining the CNN architecture on lossily compressed images affords a 59% performance increase for tumour detection within compressed test imagery [246].

Prior literature on image acquisition operations (compression, quality assessment) for face recognition [237] are limited with regard to racial bias and its race-based phenotypic influence, which is where this chapter is focused. The most related work to ours, [164] explores the test image distortion impact on pre-trained face recognition models using binary gender G1 (Male) and G2 (Female), and race R1 (light skin colour) and R2 (dark skin colour) subgroups. As a result, they find that the regions of interest used in the models shift towards less discriminatory regions in the presence of distortions, resulting in unequal performance degradation among subgroups.

In this chapter, we examine whether lossy image compression adversely impacts phenotype-based racial performance bias within face recognition during training and testing. We estimate such impact on phenotype attribute categories individually. Furthermore, we also investigate differing chroma-subsampling rates to assess how this common

lossy compression colour-related trait directly impacts recognition performance across varying phenotype-based categories. More precisely, however, we determine the relationship between the level of compression and chroma-subsampling applied and recognition performance in order to allow us to build a better understanding.

To these ends, we adapt our proposed evaluation methodology [13] that introduces phenotype-based racial bias measurement for face recognition. Furthermore, we determine the effect of varying factors, including the compression levels of lossy JPEG [159] image encoding, chroma-subsampling, and compressed versus non-compressed training on different race-based phenotype categories in order to evaluate the racial bias across multiple face recognition datasets.

In this chapter, our key contributions are as follows:

- We evaluate the impact of lossy image compression on CNN-based facial recognition approaches across different racial characteristics using the phenotype-based methodology.
- We compare several variants of training strategies, including lossy compression, within the balanced/imbalanced training datasets and race-related facial phenotypes.
- We experimentally demonstrate that the use of lossy image compression during inference adversely affects the performance of contemporary face recognition approaches [6] on a subset of race-related facial phenotype grouping (i.e. darker skin tones, monolid eye shape) and that its effect is present regardless of whether compressed imagery is used for model training.
- we investigate the specific impact of chroma-subsampling on bias performance by comparing recognition performance with and without chroma-subsampling within lossy compressed facial imagery.

4.2 Experimental Methodology

In this section, we explain the most widespread lossy image compression process (JPEG, Section 4.2.1), how we evaluate the influence of chroma subsampling on performance

(Section 4.2.2), our compression level selection methodology (Section 4.2.3), and the training strategies used (Section 4.2.4) for the generation of our results (Section 4.3).

4.2.1 Lossy Image Compression

The Joint Photographic Experts Group (JPEG), an international image compression standard [159] for still images, operates within manageable algorithmic space and time complexity whilst offering good reconstruction image quality. The JPEG standard defines four operating modes *(1: Sequential Lossless Mode, 2: Sequential DCT-based Mode, 3: Progressive DCT-based Mode, 4: Hierarchical Mode)*, formed by an encoder and decoder which follow block-based transform coding. The image encoding strategy includes colour space transformation (from RGB to YCrCb), chroma channel subsampling, Discrete Cosine Transform (DCT), quantisation and entropy coding to compress the image [159].

In this chapter, we use ImageMagick Library (version 7.0.11.13) to perform JPEG compression (via libjpeg 8). The implementation switches the JPEG operational modes according to the compression level specified (i.e. quality level q, range: 0 - 100 for JPEG, higher = better image quality, less information loss + larger file sizes). Similar to the mode one operation, it does not down-sample the chroma channels if the compression level is higher than 90 (i.e. there is no colour-based information loss for compression, $q = 90$). It applies the baseline JPEG algorithm between compression levels 90 and 10, which is sequential DCT-based Mode (2). For compression levels, $(q = 90)$, lossy compression is applied to both the luminance channel, Y, and the colour containing chroma channels, Cr, Cb.

4.2.2 Chroma Subsampling

Standard lossy compression algorithms such as JPEG contain a colour space reduction step, as the human eye is less sensitive to chromatic (i.e. colour) changes than changes in illumination (i.e. brightness). In this step, the luminance channel (Y) remains unchanged, but the image colour space (Cr and Cb) is reduced. Subsequently, by default JPEG algorithm employs 4:2:0 chroma subsampling to reduce the colour information of the original image. It takes a 2-by-2-pixel block within each block and assigns the same colour (the

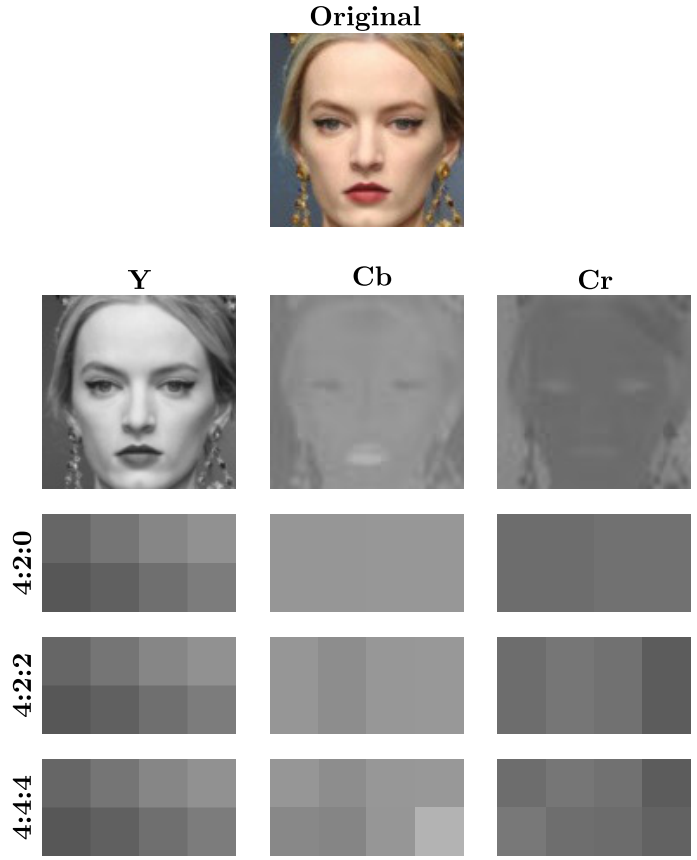


Figure 4.1: Chroma subsampling operation on different rates (4:2:0, 4:2:2, 4:4:4). Each rate differs according to how many pixels will be the same in the block.

colour of the top-left pixel) while the luminance component varies. Alternatively, for less colour information reduction, 4:2:2 with half sampling rate horizontally takes 2 pixels in each row and assigns the same colour. In Figure 4.1, we illustrate the three different sampling ratios (4:2:0, 4:2:2 and 4:4:4 no subsampling) on image pixels. In this first step of compression, chroma subsampling converts the image to YCbCr colour space and then reduces the chroma channels Cb, Cr information by assigning the top-left block pixel value to other pixels in the block. Block size and how many pixel values remain vary according to the sampling ratio.

This evaluation investigates the effect of sampling ratio on phenotype-based face recognition performance. We compare the default 4:2:0 subsampling with the 4:4:4 no chroma-subsampling factor, which keeps luminance and colour information in its entirety (i.e. unchanged). The rationale behind this evaluation is that if chroma subsampling has a profound impact on recognition performance, we can avoid this issue by recommending the use of 4:4:4 (no chroma-subsampling) with only a small impact on compression

performance.

4.2.3 Compression Level Selection

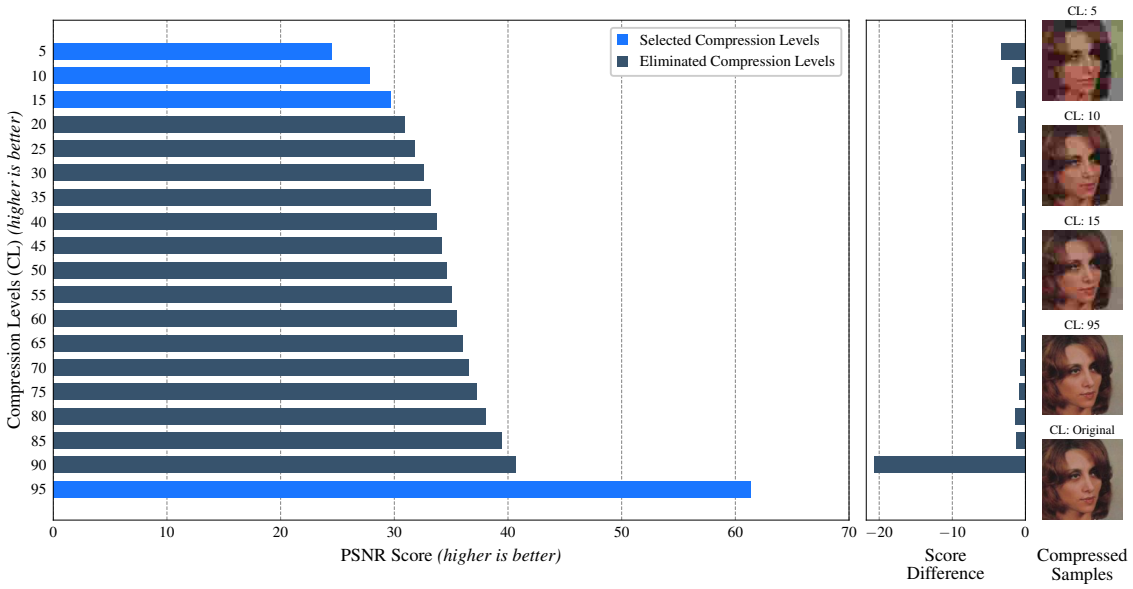


Figure 4.2: PSNR scores of RFW dataset at different compression levels (CL). Relative score difference shows how much the image quality changes at each level due to lossy compression.

In order to ascertain the impact of lossy compression on face recognition performance, we are interested in the resulting reduction in image quality at varying levels of JPEG compression. Consequently, we analyse uniformly distributed compression levels on the RFW benchmark face recognition dataset [11] using PSNR; Peak signal-to-noise ratio [247]. PSNR score is correlated with the quality of reconstruction of lossy JPEG compression. In Figure 4.2, we show the relation between the PSNR score versus the JPEG compression level, q. Firstly, we uniformly select levels $q = \{5...95\}$ in intervals of 5 and compress the whole dataset to each of these JPEG compression levels. Secondly, we measure the PSNR score on all levels and highlight the relative score difference. Based upon this analysis, we downselect the set of JPEG compression levels ($q = \{5, 10, 15\}$), in which quality decrease is most apparent (PSNR score decreases harshly). In addition, we select $q = 95$ as it represents the case where there is no chroma down-sampling used within the lossy compression scheme.

4.2.4 Training Strategies

We design different test scenarios to measure the impact of image compression on face verification performance.

Racially Imbalanced Dataset: Firstly, we train ArcFace [6] with ResNet101v2 [14] on the original aligned VGGFace2 benchmark dataset [3], containing 3.3 million images with 8631 subjects where subject distribution is racially imbalanced. Subsequently, we test using the RFW benchmark dataset [11] with the original (aligned) images and compressed images to each of the previously down-selected JPEG compression levels. We then repeat the training on the VGGFace2 benchmark dataset [3] four times, having first compressed the entire dataset to each of the down-selected JPEG compression levels. This results in four ArcFace models, each trained on image samples at a different JPEG compression level. Subsequently, we measure the performance of each of these four trained ArcFace models using the RFW benchmark dataset [11] that has been compressed to the corresponding JPEG compression level upon which each of the models was trained.

Racially Balanced Dataset: Similar to the imbalanced train set strategy, we train ArcFace [6] with ResNet50 on the original aligned BUPT-Balanced benchmark dataset [1] that contains 28000 face subjects containing balanced racial distributions among four groups $\{$*African, Asian, Indian, Caucasian*$\}$ with 7000 subjects each. Subsequently, we repeat the training on the BUPT-Balanced benchmark dataset [1] four times, having first compressed the entire dataset to each of the same down-selected JPEG compression levels. This way, another four ArcFace models are trained on image samples at a different JPEG compression level. Additionally, we replicate non-compressed and compressed training at level 5 ($q = 5$) by removing chroma subsampling (4:4:4) to measure the impact of the colour reduction step in lossy compression on face verification performance.

4.3 Results and Discussion

This section provides extensive experimental results to understand the impact of chroma subsampling and compressed training imagery using two different dataset training datasets and different compression levels.

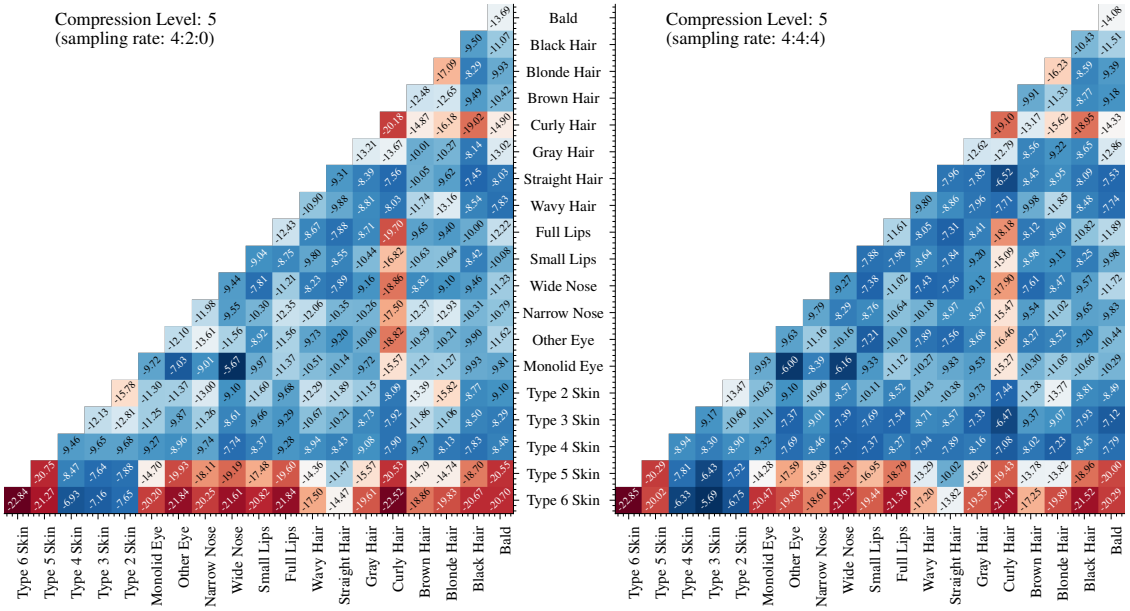


Figure 4.3: BUPT-Balanced non-compressed training set, compressed RFW test set at level 5 (q=5); FMR performance differences of cross-attribute based pairings. Each cell depicts $FMR_{original} - FMR_q$.

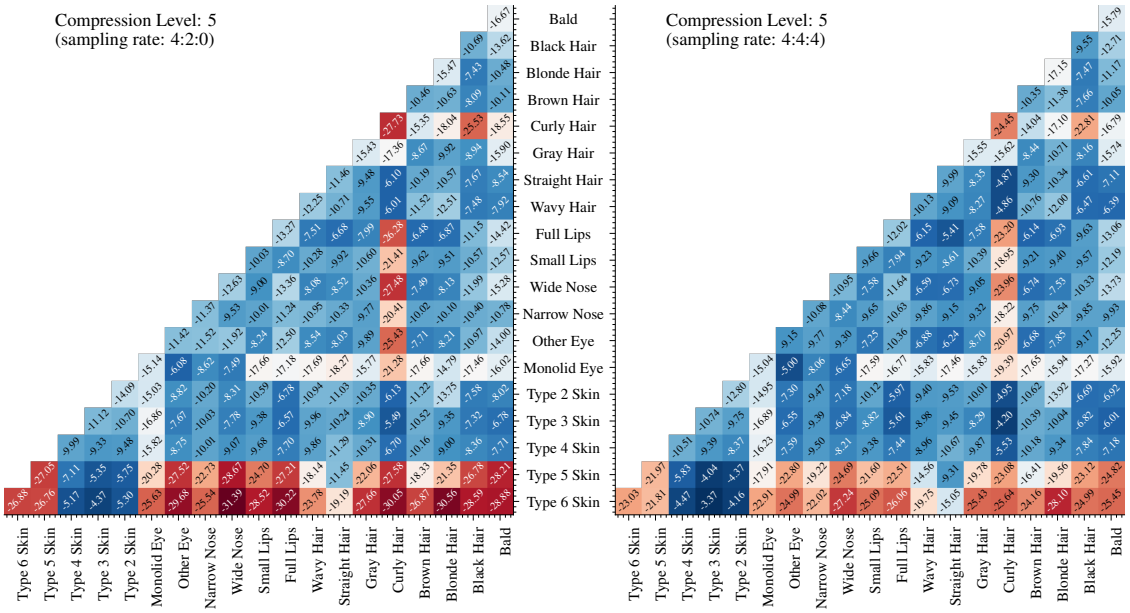


Figure 4.4: VGGFace2 non-compressed training set, compressed RFW test set at level 5 ($q = 5$); FMR performance differences of cross-attribute based pairings. Each cell depicts $FMR_{original} - FMR_q$.

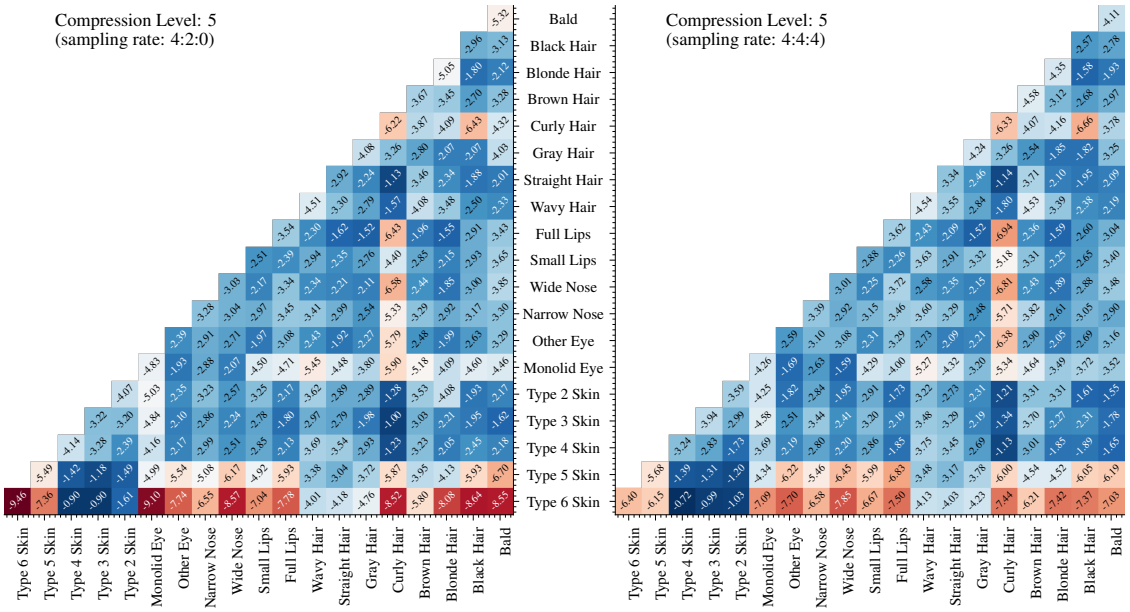


Figure 4.5: BUPT-Balanced compressed training set $(q = 5)$, compressed RFW test set at level 5 $(q = 5)$; FMR performance differences of cross-attribute based pairings. Each cell depicts $FMR_{original} - FMR_q$.

4.3.1 False Verification Matching Rates

In this section, we present False Matching Rate (FMR) differences for each of the proposed training strategies in Section 4.2.4 and the down-selected compression levels (Figure 4.2). FMR is a critical metric, such that any change in performance may result in false facial verification and the associated consequences [248].

Figures 4.3, 4.4, 4.5 show the FMR changes under the varying sampling rates of lossy image compression and how this varies across the racial phenotype labels associated with the dataset. Using the cross attribute pairings provided by [13], we evaluate $FMR_{original} - FMR_q$ where $FMR_{original}$ is FMR of non-compressed training and test imagery. FMR_q is the FMR of compressed or non-compressed training but compressed test imagery at down-selected level q. Smaller (and negative) values indicate a more considerable decline from the original level of performance.

Compression Levels: We observe that for all down-selected compression levels $q = \{5, 10, 15, 95\}$, the FMR increases when additional lossy compression is applied, demonstrating that compression level 5 (the highest compression rate) results in the most significant decrease in FMR performance, whilst compression level 95 (the lowest compression rate) does not result in any noticeable FMR performance differences. We compare com-

pression levels 95, 15, 10 and 5 with baseline results to show how FMR rise at higher compression levels. For additional performance results on different levels, see Supplementary Materials.

Chroma subsampling vs. No-chroma subsampling: We compress all the imagery in the BUPT-Balanced training dataset under two different sampling rates, 4:2:0 (JPEG default) and 4:4:4 on compression level 5 ($q = 5$). The FMR cross-attribute category results are compared in Figures 4.3, 4.4, 4.5. For non-compressed and compressed training, the 4:4:4 sampling rate decreases the FMR for all phenotype categories meaning that removing chroma sampling within the image encoding strategy of the lossy compression technique improves the performance difference and reduces the prevalence of the bias. Accordingly, we evaluate the average FMR for each phenotype category and calculate the standard deviation across all categories. Indeed, for both training strategies in Figure 4.4 and 4.5, using no chroma-sampling improves FMR variation across all categories. For VGGFace2 non-compressed training (Figure 4.4), standard deviation drops from 3.91 to 3.28 ($15.95\% \downarrow$), whilst BUPT compressed training (Figure 4.5) standard deviation drops from from 0.91 to 0.81 ($10.88\% \downarrow$).

Non-compressed vs. compressed training sets: When the model is trained on original/ non-compressed training imagery (Figures 4.3 and 4.4), FMR on darker skin tone (Type 5-6) increases considerably compared to other phenotypes such as lighter skin tones (Types 2-4) with the introduction of lossy compression at test time. At the highest level of compression ($q = 5$), the increase in FMR is greater when both phenotype categories in the pair are correlated with the stereotypically African/Afro-Caribbean racial features [132]. For instance, the Full Lips $\leftrightarrow$ Type 6 pair has the highest FMR among all other pairs higher than Type 2 $\leftrightarrow$ Type 6 skin tone pairings. For compressed training imagery (Figures 4.5 and Supplementary 4.9), we observe improved results for both imbalanced and balanced dataset training. However, darker skin tone and related categories still maintain FMR higher than the other phenotype categories.

Racially balanced vs. imbalanced training sets: Using the racially balanced dataset for training does not ameliorate FMR differences among such pairings. For example, at the highest level of compression ($q = 5$), the average performance decrease of all skin tone Type 5 pairings (Type 5-Bald, Type 5-Black Hair etc.) is 16.06% for imbalanced

	Uncompressed Training Set				Compressed Training Set				
Attribute Name	**95**	**15**	**10**	**5**	**95**	**15**	**10**	**5**	**Original**
Curly Hair	93.10	82.37	75.80	59.53	92.77	87.20	82.90	73.27	93.15
Full Lips	93.37	83.55	77.03	61.37	92.80	87.97	83.62	75.30	93.38
Monolid Eye	93.25	83.43	77.28	63.18	93.48	87.62	85.10	76.95	93.30
Type 5	94.87	85.98	80.17	60.32	94.53	90.22	87.03	76.97	94.85
Type 6	94.85	86.55	79.35	61.75	94.43	90.02	86.20	77.72	94.82
Black Hair	93.70	85.13	79.97	65.83	93.50	89.55	86.87	77.92	93.73
Wide Nose	93.95	85.53	79.97	63.15	93.42	89.57	86.78	78.33	93.98
Other Eye	94.32	86.65	81.10	65.28	93.70	89.57	87.43	78.55	94.38
Type 4	94.05	87.72	83.47	67.28	93.72	89.67	87.45	79.23	94.07
Type 1	92.86	86.88	84.72	72.43	94.19	89.87	88.21	79.57	92.86
Straight Hair	94.18	86.70	81.98	66.15	93.92	89.43	86.28	79.65	94.12
Narrow Nose	94.35	86.30	80.07	66.73	94.60	89.63	87.20	79.77	94.43
Type 3	94.05	86.07	81.03	67.05	94.32	89.48	86.80	79.93	93.98
Small Lips	94.35	87.28	82.03	67.53	95.00	90.63	87.97	81.22	94.37
Wavy Hair	95.87	89.05	84.63	69.53	95.52	92.17	89.33	82.73	95.83
Brown Hair	95.12	88.40	83.33	67.32	95.23	91.85	89.03	82.80	95.15
Bald Hair	96.55	90.43	85.93	67.62	95.88	93.07	90.37	83.13	96.55
Red Hair	96.91	90.57	84.97	71.20	96.33	92.49	89.98	84.89	96.91
Type 2	96.27	89.98	85.98	68.45	96.57	94.27	91.58	85.93	96.33
Gray Hair	96.53	92.47	88.83	72.60	96.42	94.35	91.93	86.75	96.55
Blonde Hair	97.15	92.50	88.52	71.55	97.15	94.83	93.40	87.85	97.15
Acc	94.74	87.31	82.20	66.47	94.64	90.64	87.88	80.40	94.76
STD	1.31	2.76	3.58	3.85	1.27	2.18	2.61	3.81	1.31

Table 4.1: Verification performance on RFW test set using uncompressed (left) and compressed (right) training imagery. Attribute-based pairings are those from the study of [13].

dataset training (Figure 4.3). At the same time, it is decreases by 17.69% (Figure 4.4) from balanced dataset training. However, in racially imbalanced training, the FMR results for pairings with monolid eyes degrade more compared to racially balanced training. As there are significantly fewer monolid eye face samples than other phenotypes in the imbalanced VGGFace2 dataset, we assume that their representation degrades more than other phenotypes as the lossy compression level increases.

4.3.2 Attribute-based Verification vs. Compression Levels

We additionally present attribute-based verification accuracy for the down-selected compression levels applied at training and test time for the BUPT-Balanced benchmark dataset

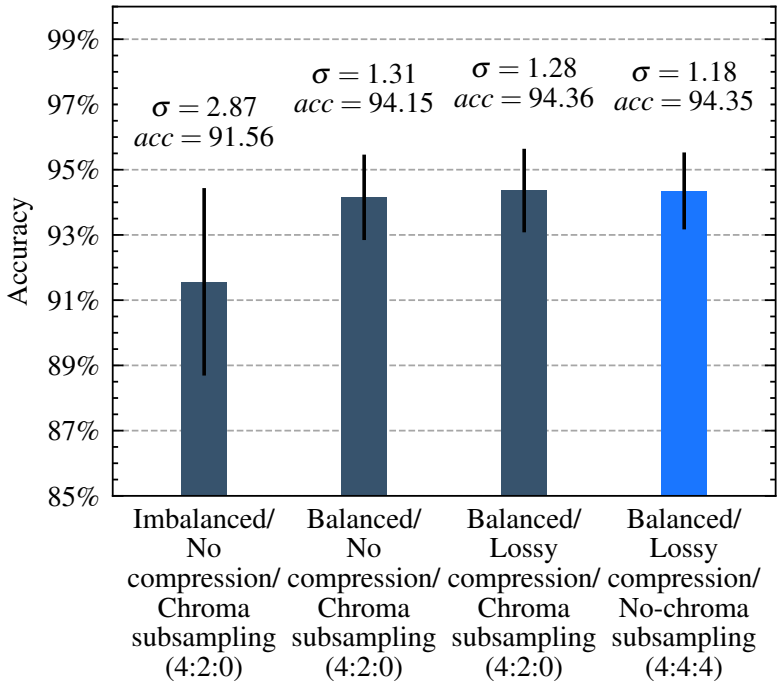


Figure 4.6: Mean Accuracy and standard deviation of all attribute categories and their comparison on different training strategies using compressed $(q = 75)$ RFW test set.

[1]. Moreover, we provide supporting evidence of compressed *vs.* uncompressed training set face verification performance in Table 4.1. We use the same 6000 (3000 positive 3000 negatives) attribute-based image pairings provided by [13]. For both non-compressed and compressed training setups, we show that as the compression increases, the standard deviation across all phenotype categories increases (as a measure of non-uniform performance and bias). Similarly, accuracy decreases for all phenotype categories. However, using uncompressed training imagery (Table 2, left) results in a further decline in performance for darker skin tones Type 5-6, curly hair, full lips and monolid eye, when compared to other facial phenotypes, as the level of lossy compression within the test set is increased. Skin Type 5 attribute pairings accuracy drops from 94.87% to 60.32% $(34.55\% \downarrow)$, while Skin Type 2 attribute accuracy drops from 96.33% to 68.45% $(27.88\% \downarrow)$. Similar to the non-compressed training set, we do observe non-uniform disparate changes in accuracy when the model is trained on compressed imagery (Table 4.1, right). Furthermore, the compressed training set produces a smaller standard deviation in accuracy between phenotype categories.

Lastly, we summarise the relationship between all factors (dataset distribution, com-

pression, chroma subsampling) in Figure 4.6. We evaluate attribute-based pairings accuracy for all phenotype categories and compare different training strategies mean accuracy and standard deviations. We change one factor during training in each strategy and provide corresponding performance results. We use a compressed RFW test set in level 75 ($q = 75$) for all training strategies. Firstly, we show racially imbalanced VGGFace2 datasets training performance, which is lowest in accuracy and highest in standard deviation. A balanced BUPT-Balance dataset provides the most significant improvement in accuracy and standard deviation. Furthermore, while compressed training imagery causes a minor decrease in standard deviation, no-chroma subsampling improves bias performance more significantly. Therefore, removing chroma sampling during compression becomes viable for reducing racial performance bias. We conclude from the aforementioned results that while compressed imagery or racially balanced training data during training improves the overall performance for all race-related categories, disparate results remain for specific phenotype characteristics. Furthermore, we highlight that the reduced retention of the chroma (colour) information affects, due to the use of chroma subsampling in lossy JPEG compression, on darker skin tones to a greater degree than on lighter skin tones. Furthermore, it is likely that the lossy image quantisation disproportionately affects finer image details on the facial region, such as those associated with monolid eye characteristics.

4.3.3 FMRs on Selected Compression Levels

We provide down-selected compression levels differences (additional compression levels ($q = 10, 15, 95$)) for each of the proposed training strategies using cross attribute pairings provided by [13]. As described in the paper, smaller (and negative) values indicate a larger decline from the original level of performance. The FMR increases when the lossy compression increases. In Figure 4.7, 4.8, 4.9 and 4.10, we demonstrate that compression level 5 (the highest compression rate) results in the most significant decrease in FMR performance for all different training strategies. In contrast, compression level 95 (the lowest compression rate) does not result in any noticeable FMR performance differences.

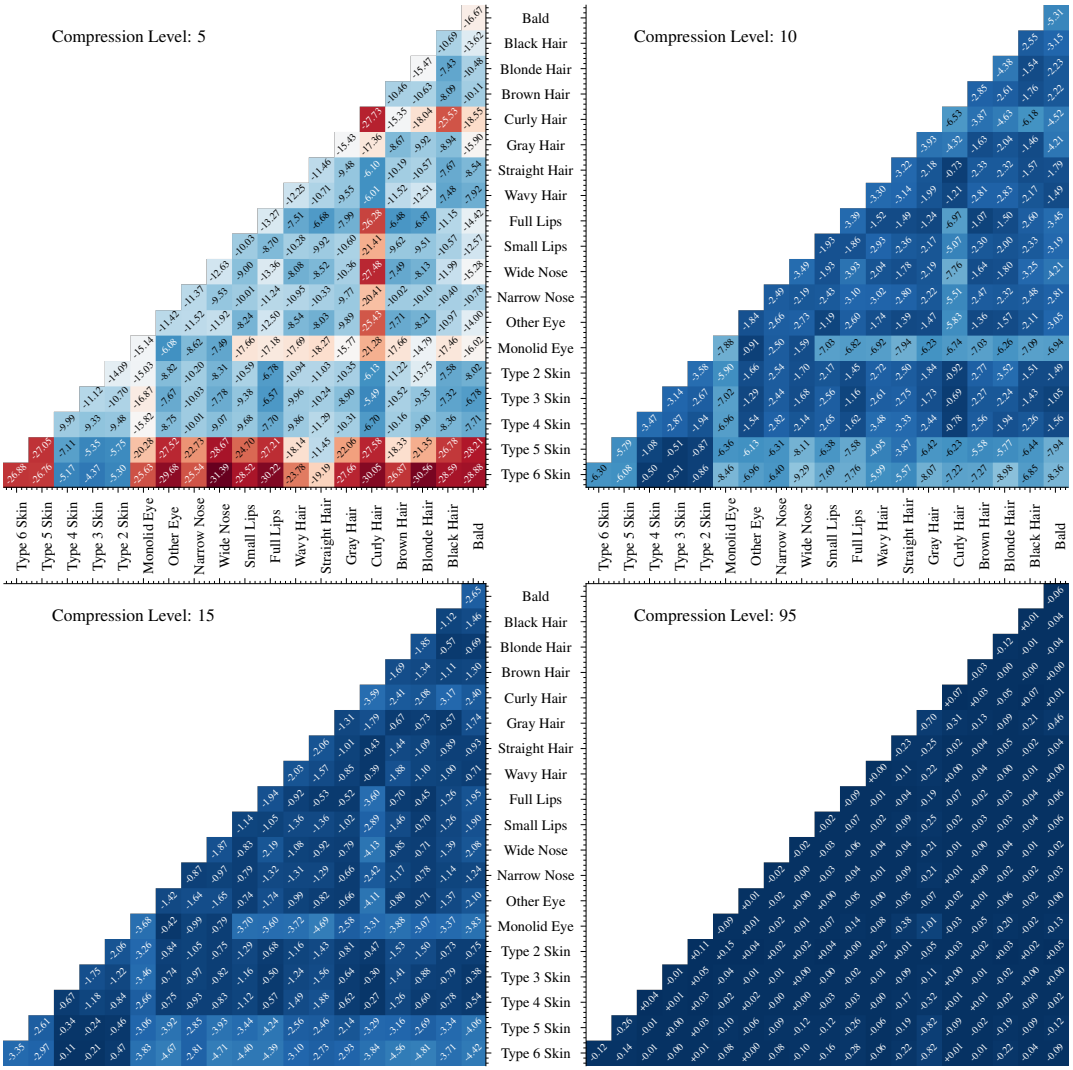


Figure 4.7: VGGFace2 original/non-compressed training imagery and compressed RFW test imagery; FMR performance differences of cross-attribute based pairings. Each cell depicts $FMR_{original} - FMR_q$.

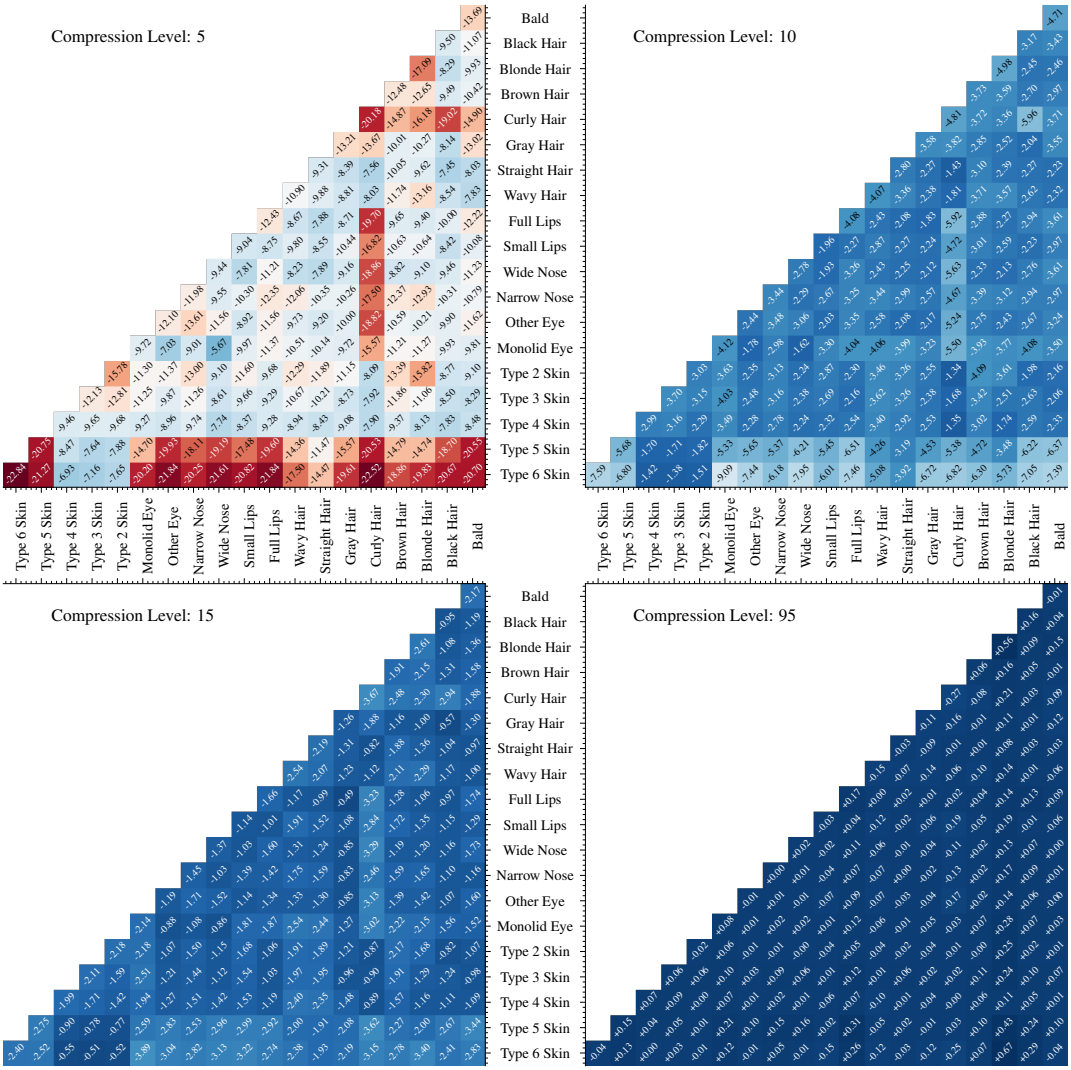


Figure 4.8: BUPT-Balanced original/non-compressed training imagery and compressed RFW test imagery FMR performance differences of cross-attribute based pairings. Each cell depicts $FMR_{original} - FMR_q$.

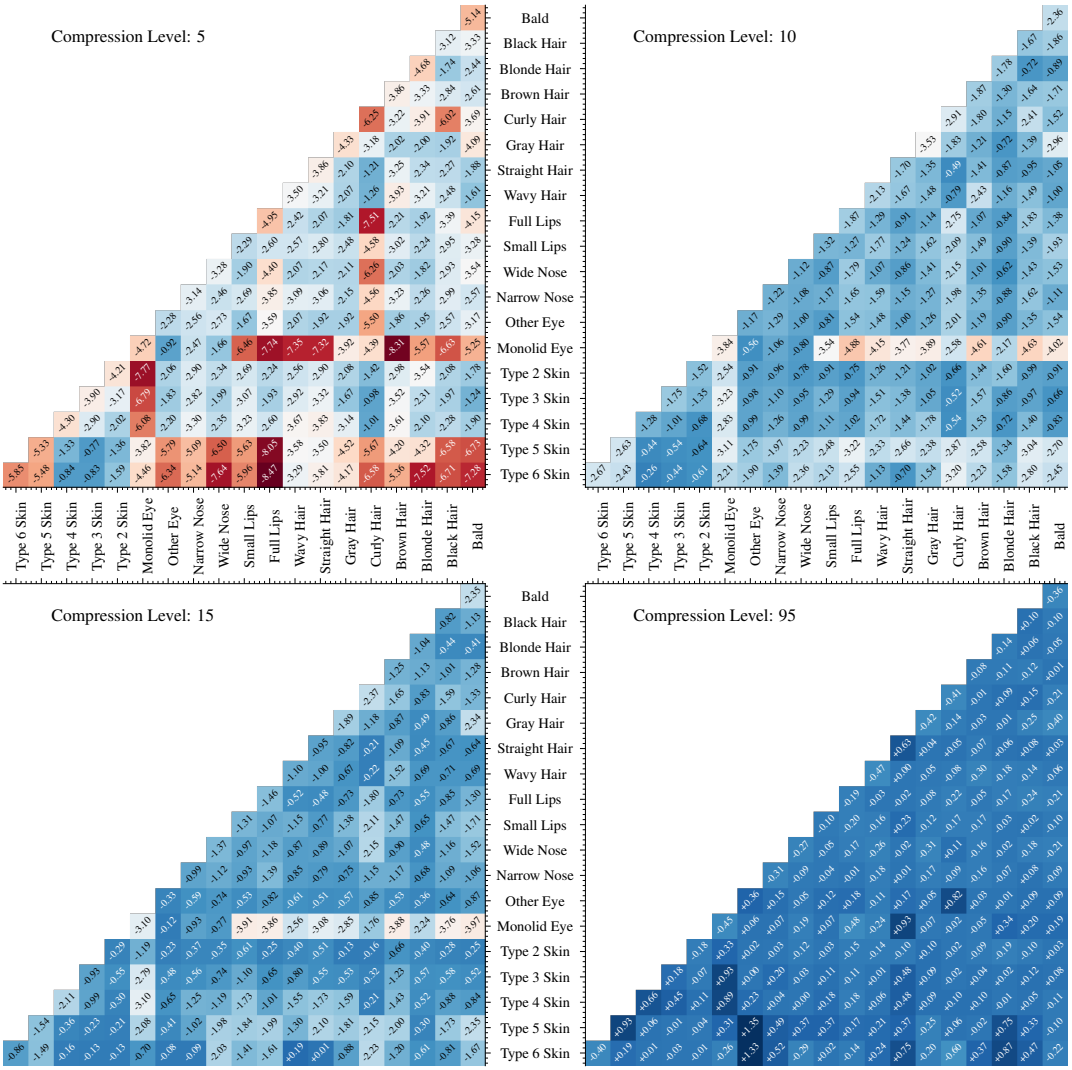


Figure 4.9: VGGFace2 compressed training imagery and compressed RFW test imagery; FMR performance differences of cross-attribute based pairings. Each cell depicts $FMR_{original} - FMR_q$.

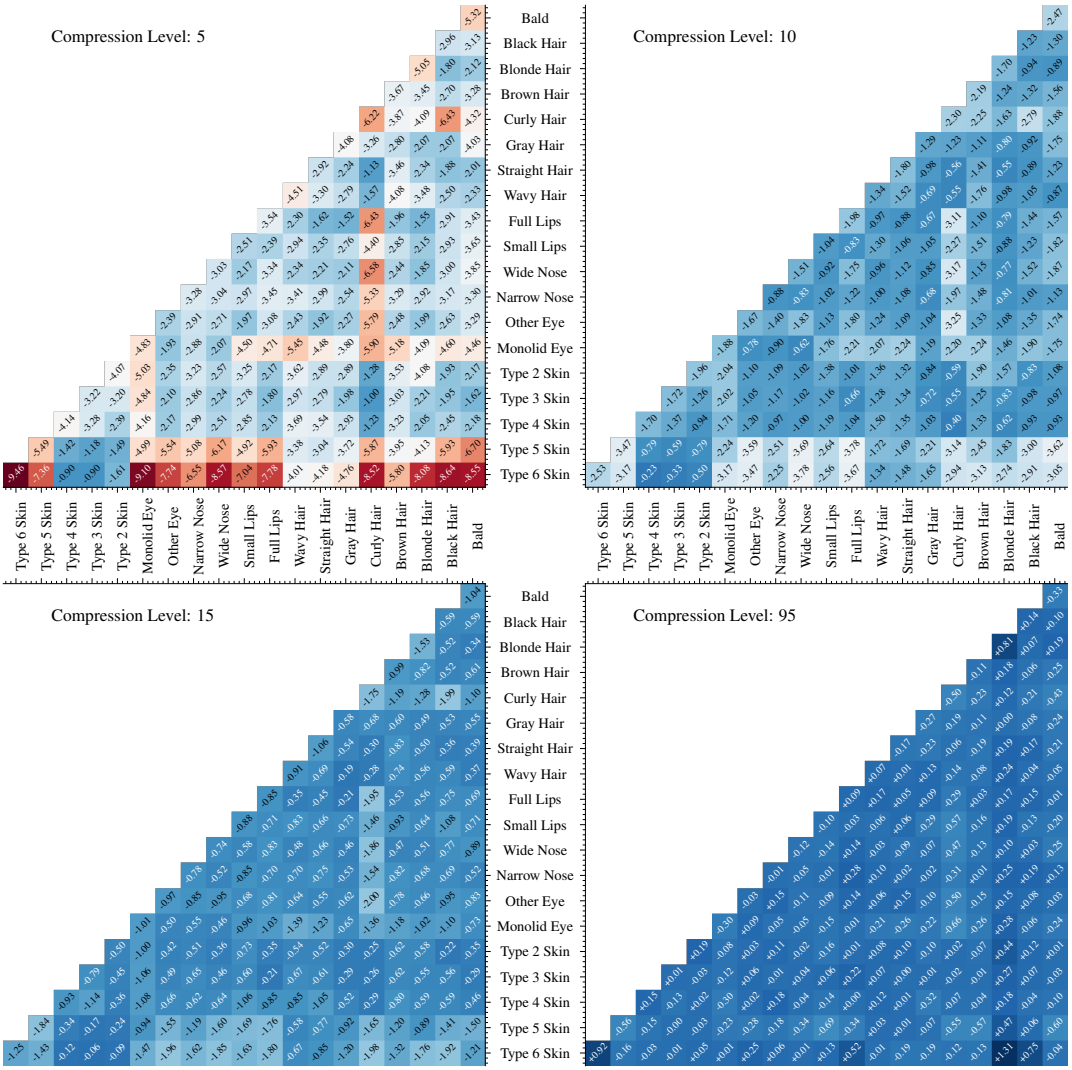


Figure 4.10: BUPT-Balanced compressed training imagery and compressed RFW test imagery; FMR performance differences of cross-attribute based pairings. Each cell depicts $FMR_{original} - FMR_q$.

4.4 Summary

This chapter examines the relationship between face verification performance for a given race-related phenotypic group under varying levels of lossy compressed sets. Overall, our evaluation finds that using lossy compressed facial image samples at inference time decreases performance more significantly on specific phenotypes, including dark skin tone, wide nose, curly hair, and monolid eye across all other phenotypic features.

Accordingly, we adopt similar training protocol in the Table 2.3, using BUPT-Balanced benchmark dataset [1], ResNet50 architecture [14], and ArcFace loss [6] and show it results in reduced accuracy and higher standard deviation as the test imagery more heavily compressed in Figure 4.6. However, the use of compressed imagery during training does make the resulting models more resilient and limits the performance degradation encountered: lower performance amongst specific racially-aligned subgroups remains. Additionally, removing chroma subsampling improves FMR for specific phenotype categories more affected by lossy compression.

CHAPTER 5

Adversarially-Enabled Data Augmentation for Racial Bias within Face Recognition

In this chapter, we propose a novel adversarial derived data augmentation methodology that aims to enable dataset balance at a per-subject level via the use of image-to-image transformation for the transfer of sensitive racial characteristic facial features. Our aim is to automatically construct a synthesised dataset by transforming facial images across varying racial domains, while still preserving identity-related features, such that racially dependant features subsequently become irrelevant within the determination of subject identity. We construct our experiments on three significant face recognition variants: Softmax [4], CosFace [5] and ArcFace [6] loss over a common convolutional neural network backbone. In comparison, we show the positive impact our proposed technique can have on the recognition performance for racial groups within an originally imbalanced training dataset by reducing the per-race variance in performance.

The material presented in this chapter of the thesis has been published in the following peer-reviewed publication:
Seyma Yucer, Samet Akcay, Noura Al Moubayed, and Toby P. Breckon., Exploring Racial Bias within Face Recognition via per-subject Adversarially-Enabled Data Augmentation.,

IEEE/CVF Computer Vision and Pattern Recognition Workshops, CVPRW, pp. 18-19, 2020.

5.1 Introduction

Recent advances in Generative Adversarial Networks (GAN), have led to realistic image generation [55] and even class generation [249]. Such advances in the field have a promising potential to overcome the bias in face recognition via realistic image generation as most of the face recognition datasets have a significantly imbalance distribution on either classes [15] or demographic groups [87].

Accordingly, in this chapter, we address the racial bias of face recognition from an adversarial augmentation point of view. As most of the datasets [1, 11, 39] consist of four major racial groups, namely African, Asian, Caucasian and Indian, we seek group-fairness among these races, in terms of facial recognition performance, by utilising generative adversarial network (GAN) [250].

Previous work [202, 210, 251] has established adversarial techniques to minimise mutual information on identity features, which reveal sensitive attributes about race, gender and age of the subject. However, such approaches [161, 210], have failed to effectively address the trade-off between suppressing the use of such sensitive attributes and the loss of key identity-related features which pertain to the overall performance of the facial recognition approach. Our solution, instead, uses an adversarial image re-synthesise technique [2], to transform sensitive attributes across a set of synthetic images comprising the full range of races being considered within the facial recognition problem. By doing so, we preserve the important identity-related features whilst making the racially dependent features of the face less prevalent due to the artificially synthesised distribution of these identity characteristics across the full range of race profiles for any given individual.

Figure 5.1 illustrates how we transform the identity characteristics, and hence features, any given individual across multiple racial profiles using a CycleGAN [2]. It proposes transformation across racial domains and reconstruction to produce an identical image from a transformed image during the cyclic adversarial training. To show its robustness, we explore the performance of our approach using balanced and imbalanced

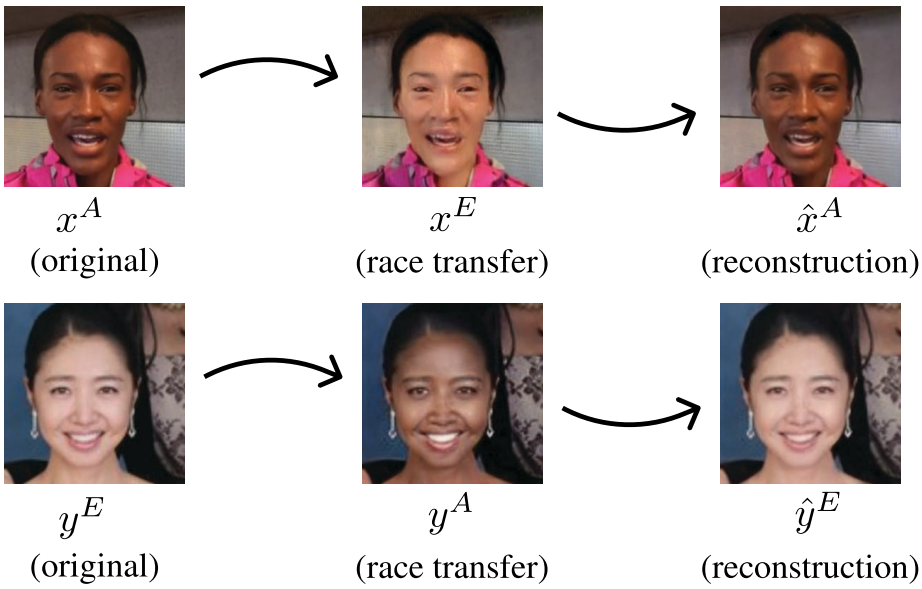


Figure 5.1: Racial transformation example using [2]. We transfer an African image x^A to Asian image y^E and obtain sythesised x^E in Asian domain and we reconstruct $\hat{x}^A$ from x^E image. Asian image y^E to African image x^A transformation follows the same procedure.

training datasets.

The main contributions of this chapter are as follows:

- We propose an adversarial image-to-image transformation technique to mitigate racial bias based on the cyclic adversarial training approach of CycleGAN [2].
- We illustrate both quantitative and qualitative performance of our proposed facial data augmentation techniques over established benchmark datasets within the face recognition domain, establishing a statistical paradigm for the presentation of recognition results on a per-race basis.
- We adopt our phenotyped-based evaluation methodology in order to show the improvement of our method on phenotype-based cross attribute pairings.

5.2 Methodology

We present our methodology in three parts: we first describe our problem definition in Section 5.2.1, explain image-to-image transfer method [2] for race transformation to mitigate face recognition bias in Section 5.2.2.

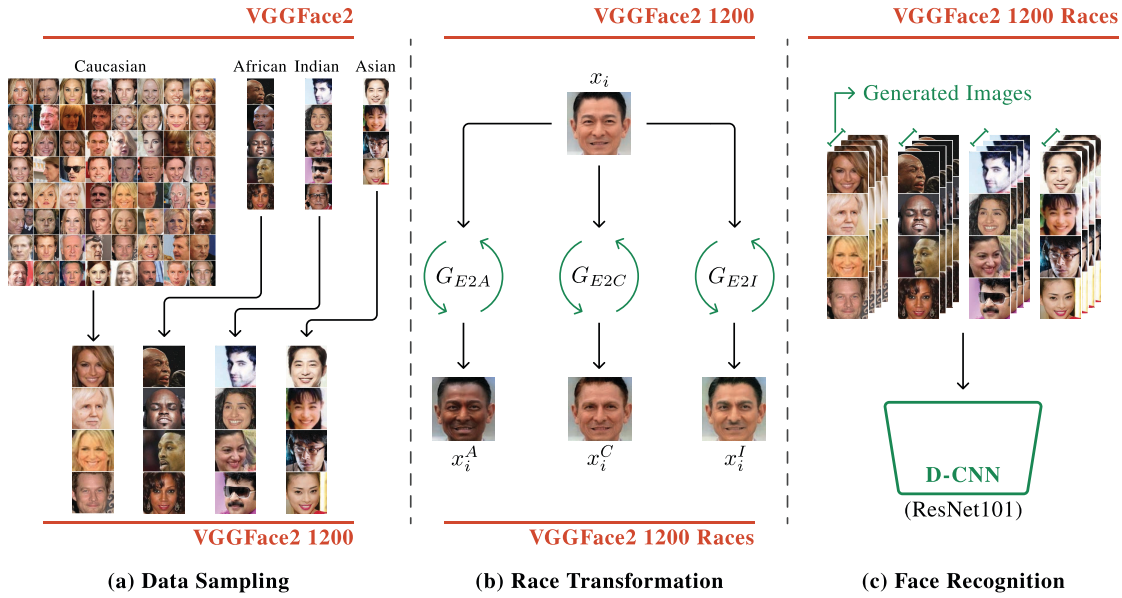


Figure 5.2: Overview of our solution in three phases: (a) describes imbalanced distribution of VGGFace2 [3] and downsampling it to VGGFace2 1200. (b) illustrates race domain transformation schema for a given image x_i (c) shows face recognition algorithms with Softmax [4], CosFace [5] and ArcFace [6] loss functions using VGGFace2 1200 Races.

5.2.1 Problem Definition

Ideally, a machine learning algorithm should require that the conditional probability P of the output given input x does not depend on any *sensitive attributes* which is demographic features in our case. This *Fairness Through Unawareness* can be formalised as $P(y \mid x) = P(y \mid x, s)$ (see Chapter 2 and Section 2.1 for more details), where x is an input, y is the corresponding subject label and s is a sensitive attribute that does not alter the outcome. However, removing dependency is highly challenging for face recognition due to high *mutual information* between facial features and sensitive attributes, such as race.

A given face image dataset, $D = [x_1, x_2, x_3, \ldots, x_N]$, provides N number of face images. A feature embedding vector of an image, $z_i = [f_1, f_2, \ldots, f_d]$, where $z_i \in \mathbb{R}^d$, is commonly statistically dependent on sensitive attributes where it causes *indirect discrimination* for particular demographic groups which potentially form overlapping, subsets of D. Although the common approach for face recognition bias is to minimise this mutual information to remove the dependency on sensitive features; it is still an extremely difficult task using face features without sacrificing any prior information for face recognition

as shown in [161, 210].

Hence, we approach the problem from a completely different perspective by transferring sensitive attributes from one domain to another whilst simultaneously preserving prior information for recognition. On the other hand, we are aware that some features are more prevalent in some demographic groups than others. The sensitive information, in this case, may improve the prior information for the recognition task. Lighter skin allows the model to learn more detailed features given characteristics of modern cameras and common scene illumination conditions. A novel input mechanism which projects different sensitive information for one image to a model makes race modelling irrelevant. As a result, we ask a question; *What if we augment and transfer sensitive information rather than removing it*? To answer this question, we present a new pre-processing based method requires augmentation of sensitive attributes of an image.

Our new inputs consist of three generated images from different domains for each image. Given the race domains $\{A, E, C, I\}$ for $\{African, Asian, Caucasian, Indian\}$ respectively, we aim to transform an image x_i from one domain as an image x_j to another domain. For instance, we transform given x_i in A to another image from different domains such as E, C, I. If we use different images belonging to these domains to transform, we can define new generated input dataset as following list $x_i^+ = \left[x_i, x_i^E, x_i^C, x_i^I\right]$ where x_i is the original image and x_i^+ is a new input list including the original image.

Transferring sensitive information while keeping prior information of the image is possible via adversarial methods, as they are capable of generating images from the training data distribution. To show that, we propose a solution of sensitive attribute transformation while keeping prior information for face recognition and present a new augmented dataset, $I_{image}^+ = \left[x_i, x_i^A, x_i^C, x_i^I, \ldots x_i, x_i^E, x_i^C, x_i^I, \ldots, x_n, x_n^A, x_n^E, x_n^C, \right]$. In the next Section 5.2.2 we present our approach to the image synthesise process to obtain D_{image}^+.

5.2.2 Adversarial Image-to-Image Transfer

Our solution transforms these sensitive attributes using a cyclic adversarial domain transfer approach, CycleGAN [2]. We assume that learning a mapping function between two different race groups domain reduces the dependency on sensitive features.

For example, given an African face image $x_i \in A$, and a Caucasian image $x_j \in C$,

we assume that the two different data distributions from these image race groups $x_i \sim p_{data}(x_i)$ and $x_j \sim p_{data}(x_j)$ can be transferable between each other. To map these two distributions between domain A and C, we introduce two mapping functions F and G, respectively from African to Caucasian domains and from Caucasian to African domains using CycleGAN [2]. Within a GAN framework, these two directional transformations need two discriminators D_A and D_C, to distinguish between x_i and $F(x_j)$, x_j and $G(x_i)$, respectively. Moreover, as an additional control on adversarial training, a cycle-consistency loss is introduced to ensure that the mapping function can transfer an individual input x_i to the desired output x_j.

$$L_{GAN}(G, D_C, A, C) = \mathbb{E}_{x_j} \sim p_d(x_j)\,[logD_C(x_j)] \quad + \mathbb{E}_{x_i} \sim p_d(x_i)\,[log(1 - D_C(G(x_i))] \tag{5.1}$$

For the first part of race transformation, an adversarial loss is used as defined in Equation 5.1 where A and C are the African and Caucasian group domains, respectively. While the generator G synthesise images using source domain A to associate to target domain C, discriminator D_C distinguishes between the real image and x_j from the synthesised image, $G(x_i)$. The same process is applied with generator F and discriminator D_A to transform domains from C to A.

The key premise of CycleGAN [2] is a controlled mechanism of adversarial training which allows us to synthesise more accurate images from the desired images in the domain. To achieve this, cycle consistency loss is introduced as defined in Equation 5.2, where $F(G(x_i))$ is reconstructed x_i from synthesised $G(x_i)$ new image. In this case, generators F and G are able to reconstruct the original images. The $L1$ norm in this loss measures the difference between the original image and reconstructed image as follows:

$$\begin{aligned} L_{cyc}(G, F) = \mathbb{E}_{x_i} \sim p_d(x_i)\,[\|\, F(G(x_i)) - x_i \,\|_1] \\ + \mathbb{E}_{x_j} \sim p_d(x_j)\,\left[\|\, G(F(x_j)) - x_j \,\|_1\right] \end{aligned} \tag{5.2}$$

The overall loss function, as defined in Equation 5.3, consists of two adversarial loss within the cycle-consistency loss where λ is a term to control the relative importance of

the cycle-consistency loss.

$$\begin{aligned} L(G, F, D_A, D_C) = & L_{GAN}(G, D_C, A, C) \\ & + L_{GAN}(F, D_A, C, A) \\ & + \lambda L_{cyc}(G, F)] \end{aligned} \tag{5.3}$$

Subsequently, overall adversarial training of this objective function aims to solve the following equation:

$$G^*, F^* = \underset{G,F}{\operatorname{argmin}} \max_{D_A, D_C} L(G, F, D_A, D_C). \tag{5.4}$$

In the intermediate step $G(x_i)$ and $F(x_j)$, the generator encodes features of inputs x_i and x_j and then $F(x_j)$ and $G(x_i)$ decodes back to obtain original images again. With reference to this set of transform Equations 5.1-5.4, we can transform both, domain A into domain C and C into A similarly for other domain pairings.

5.3 Experimental Setup

This section provides overview of our experimental evaluation in terms of the face recognition datasets used, the race classification used for racial annotation and the implementation details of our proposed approach.

5.3.1 Training Protocols

To validate our approach, we utilise BUPT-Transferface [11] for race transfer and race classification, VGGFace2 [3] for face recognition training and RFW [11] for face verification.

We train a common DCNN, ResNet [14] on proposed augmented datasets; VGGFace 2 1200, VGGFace 2 8631. We utilise ResNet100 explored by [6] with $\{BatchNorm - Dropout - FC - BatchNorm\}$ structure to get the final 512-D feature space representation after the last convolutional layer. We use same architecture for Softmax [4],

CosFace [5] and ArcFace [6] loss functions.

5.3.2 Annotation of Race

We obtain racial annotation labels for VGGFace2 [3] dataset using fine-grained classification to solely support our development of a technique to mitigate bias.

The work of [252] proposes attention-guided data augmentation to improve the spatial representation of distinctive image parts using its cropping and dropping mechanism. We adopt this solution for a race classification problem where distinctive image parts are facial attributes of eyes, nose, mouth, and forehead. Via this approach [252], we obtain racial annotations of VGGFace2 [3] and we manually check the least certain subjects according to the majority of image labels for each subject and additionally exclude some subjects who are not in the four-race set $\{Caucasian, African, Asian, Indian\}$. After this semi-automatic process, the subject distribution for training and testing sets is shown in Figure 5.3 whereby the inherent racial and gender imbalance is clearly illustrated.

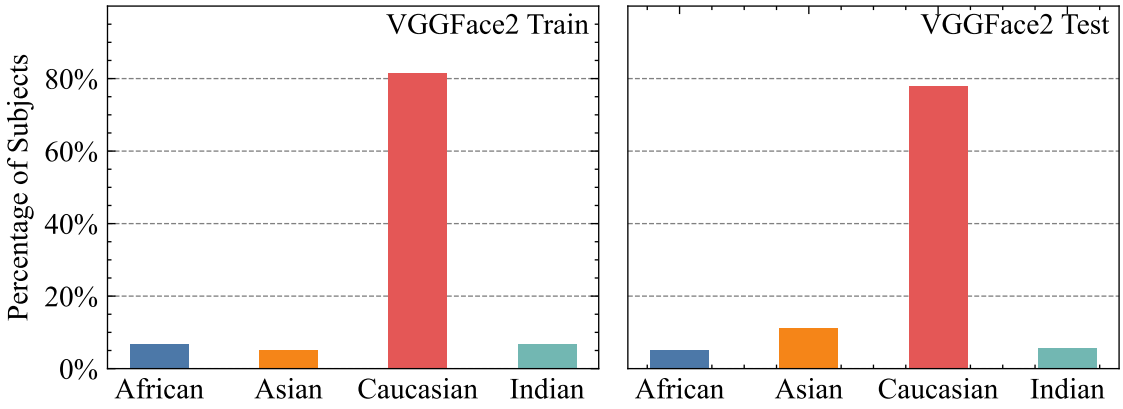


Figure 5.3: The distribution of races in the VGGFace2 dataset, both the train and test sets

5.3.3 Race Transfer

Our proposed image-to-image transformation approach creates a new dataset D^{+}_{image}, to transfer race attributes from one race group to another. To achieve that, we define separate mappings for each pair of the four different race groups. The set of 12 mappings are: {*African* $\rightarrow$ *Asian, African* $\rightarrow$ *Caucasian, African* $\rightarrow$ *Indian, Asian* $\rightarrow$ *African, Asian* $\rightarrow$ *Caucasian, Asian* $\rightarrow$ *Indian, Caucasian* $\rightarrow$ *African, Caucasian* $\rightarrow$ *Asian, Caucasian* $\rightarrow$

Method	Dataset	LFW	RFW					
			African	Asian	Caucasian	Indian	Acc	STD
Softmax	VGGFace2 1200	96.13	69.10	73.70	79.25	76.78	74.71	4.37
Softmax	VGGFace2 1200 Races	**96.27**	**70.65**	**75.68**	**80.27**	**78.28**	**76.22**	**4.16**
CosFace	VGGFace2 1200	98.16	82.78	82.68	87.53	85.41	84.60	2.33
CosFace	VGGFace2 1200 Races	**98.65**	**83.22**	83.23	**87.95**	**85.77**	**85.04**	**2.28**
Arcface	VGGFace2 1200	98.16	80.91	81.78	**86.86**	83.70	83.31	2.64
Arcface	VGGFace2 1200 Races	**98.63**	**81.28**	82.83	85.95	**84.72**	**83.69**	**2.06**

Table 5.1: Verification performance (%) of Softmax, CosFace, and ArcFace with ResNet-101 [14] on LFW [15] and RFW [11] when trained on VGGFace2 1200 and proposed VGGFace2 1200 Races datasets.

Indian, Indian $\rightarrow$ *African, Indian* $\rightarrow$ *Asian, Indian* $\rightarrow$ *Caucasian*$\}$. As our CycleGAN based approach provides two-way transformations between source and target domains, we train six models to find these two directional mappings following the approach outlined in Section 5.2.2.

For training, we generate 25K image pairs using the BUPT-Transfer [11] dataset. All face images are aligned and have a size of 256×256. To avoid gender domain differences, we only match images of the same gender as pairs. Using these six CycleGAN models, we synthesise new images and denote extended dataset as VGGFace2 1200 Races [3] which contains the original VGGFace2 1200 images and synthesised race images. Each image has three different transformed images that belong to other race domains in addition to the original. As a result, we partially absorb the downsampling effect on VGGFace2 1200. Subsequently, we synthesise all non-Caucasians images on original VGGFace2 and call the new dataset VGGFace2 8631 Races, D^{+}_{image}. We do not transform Caucasian images to other racial domains; they are already dominant in the original dataset.

5.4 Results and Discussion

In this section, we provide both racial grouping strategy and phenotype-based evaluation methodology results in order to show the improvement of our method. We also illustrate qualitative result via generated imagery and discuss the results.

5.4.1 Face Verification on Racial Groupings

To evaluate the performance of the proposed approach, we use LFW face verification protocol [15], which measures whether two images belong to the same subject or not.

We assess synthesised image quality by feeding them through a race classifier introduced in Section 5.3.2. We show examples of the correctly classified images in Figure 5.5 and the misclassified images in Figure 5.6. Each column of Figure 5.5 and 5.6 show an image transformation example where the original image is represented with green and red borders, and synthesised images are laid in the corresponding racial domain label in the y-axis. As can be seen in the Figure 5.6, image transformation is prone to fail on poor illumination and pose variations.

For face recognition, we first test our performance on **balanced datasets** VGGFace2 1200 and VGGFace2 1200 Races. We compare our results on RFW [11] using three different loss functions; Softmax, CosFace [5] and ArcFace [6] as shown in Table 5.1. Proposed facial image augmentation approach improves performance in all three methods by 0.38-1.51 %. As non-Caucasian results are improved, the standard deviation among groups is decreased. We also share LFW results in Table 5.1 to show the improvement of our solution on the imbalanced dataset. Second, we use the **imbalanced dataset** with the ArcFace as shown in Table 5.2. While LFW verification performance remains the same, RFW African and Asian performances are improved, and the standard deviation declines from 2.91 to 2.45.

Racially Balanced and Imbalanced Training Protocols: This study provides experiments on both balanced and imbalanced training datasets. Although imbalanced data may seem to be the main reason for face recognition bias, when we train algorithms on completely equally distributed data (equal number of race and gender grouping subjects.), the results still appear to exhibit performance bias. Another study experiments on a large and nearly balanced dataset and again differs on Caucasians and non-Caucasians [1]. Subsequently, we focus on a novel per-subject racial data balancing approach to understanding its impact on the face recognition bias.

Synthesised Imagery Representation for Face Recognition Models: We experimented with different image processing methods to change the input imagery. First, we attempted averaging the images belong to same subjects, which aimed to reduce translation effects

but did not yield improved results. Next, we explored concatenating the images along the y-axis and z-axis. While concatenating along the y-axis resulted in twelve channels, it significantly increased the input complexity. Ultimately, feeding the images as separate samples produced the best results.

Impact of Synthesised Imagery Quality and Transformation: We assess the quality of our synthesised images by testing them using a race classifier (Section 5.3.2). We would expect the race classifier to recognise them as the correct transformed racial label. Our overall accuracy is 49% across all transformations, but when we increase this accuracy using more pairs, and longer training, this results in an overall reduction in face recognition performance. The trade-off is complex because after transforming the main racial attributes of the face such as skin colour, eye structure and hair colour, CycleGAN proceeds to translate all facial features including those which implicitly encode unique subject identity. Other notable negatives are variations in pose and illumination on the synthesised images which could alternatively be addressed via [55] in future work.

Method	African	Asian	Caucasion	Indian	Acc	STD
ArcFace [6]	89.45	87.61	94.71	91.21	90.75	2.91
ARL+C [200]	88.57	87.65	93.48	89.35	89.76	2.57
Ours	**90.1**	**87.73**	**93.72**	**90.5**	**90.51**	**2.45**

Table 5.2: RFW [11] Verification performance comparison (%) of methods using ResNet [14] trained on VGGFace2 [3] and our proposed method is trained on VGGFace2 8631 Races with synthesised images of non-Caucasian subjects on VGGFace2.

5.4.2 Face Verification on Phenotype-based Groupings

Lastly, we apply our previously proposed racial phenotype-based bias analysis methodology (Chapter 3) to evaluate the effectiveness of our approach across various racial phenotype categories. Similarly to the previous chapter, we utilise the cross-attribute pairings provided by [13] and calculate the False Match Rate (FMR) between all attribute category combinations.

As shown in Figure 5.4, we observe a noticeable improvement in both general and specific attribute categories, including Type 3 and 4 skin tones, monolid eye, wavy hair, curly hair, as the $FMRs$ of such categories decrease. Although our proposed method is

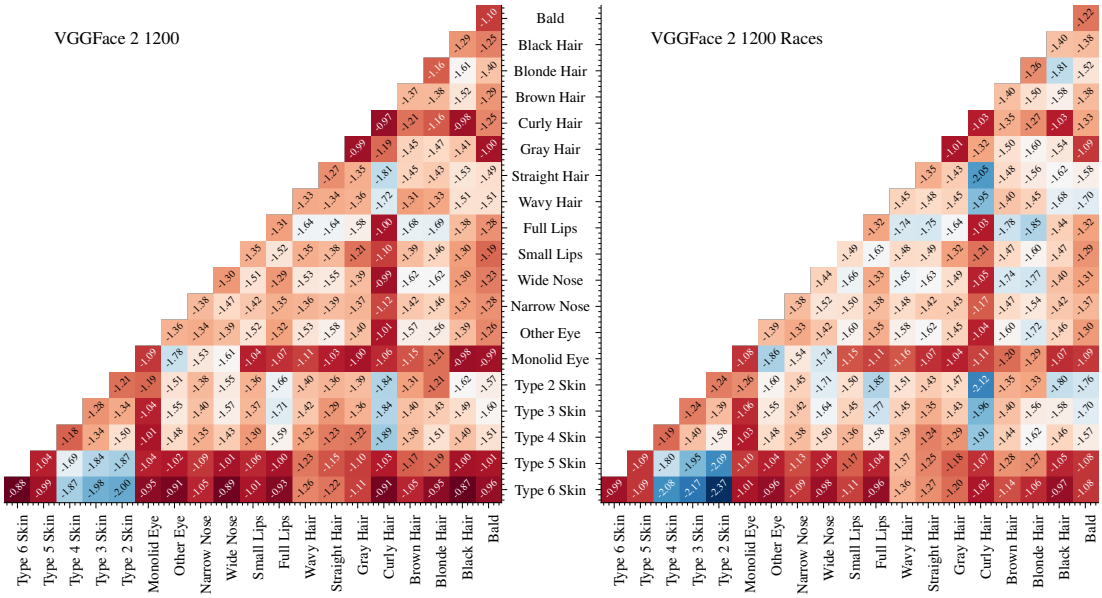


Figure 5.4: False matching rates (FMR) of cross-attribute based pairings for 21 attribute categories using VGGFace 2 1200 and augmented VGGFace 2 1200 Races training set. Each cell depicts FMR on a logarithmic scale which is $log_{10}(FMR)$ with lower negative values (close to zero) encoding superior false match rates

built on transferring racial information using racial groupings and synthesising datasets, we are able to effectively reduce racial bias across various attributes. Moreover, we observe less of a decrease in FMR when using VGGFace2 Races as the training set.

5.5 Summary

In this chapter, we explore racial bias in face recognition and present a novel adversarial derived data augmentation methodology. Transferring racial attributes of a human face whilst preserving identity features in the face recognition datasets makes face recognition algorithms more robust and less race-dependant. On our manually balanced dataset, we compare three significant face recognition variants: Softmax [4], CosFace [5] and ArcFace [6] loss functions with a common convolutional neural network backbone ResNet-101 [14].

Subsequently, using the imbalanced VGGFace2 benchmark dataset [3], ResNet-101 architecture [14], and ArcFace loss [6], we demonstrate that our proposed technique decreases the performance variations between four racial groups: {African, Asian, Caucasian, Indian} by 15.81%. Specifically, as shown in Table 2.3, the standard training

setup of VGGFace2, ResNet-101, and ArcFace loss results in a standard deviation of 2.91 and accuracy of 90.75% across the four racial groups. Our proposed approach, which involves adversarial subject-level data augmentation, achieves a similar accuracy of 90.51% but with a lower standard deviation of 2.45 in Table 5.2 across racial groups. This 15.81% reduction in standard deviation indicates our technique meaningfully reduces variability in model performance between racial groups. Although illumination, pose, and light challenge the quality of the image transformation; our technique not only improves the overall face recognition accuracy but also suppresses inter-group performance variation.

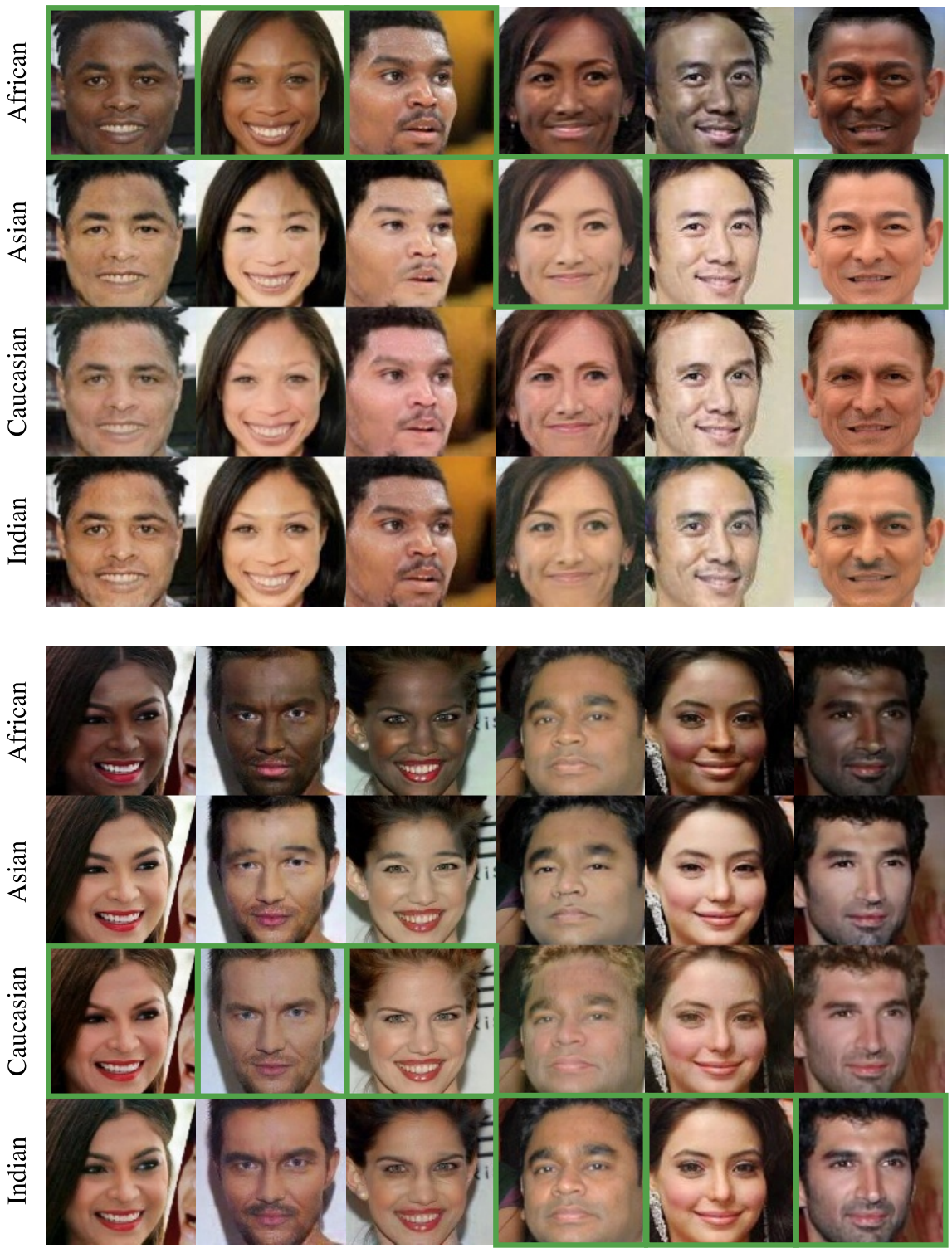


Figure 5.5: A selection of successful examples of the CycleGAN racial domain transformation of VGGFace2 dataset. Each column contains an original and sythesised face images of the same subject where the green borders indicate the original image and the corresponding race labels are laid out on the y-axis.

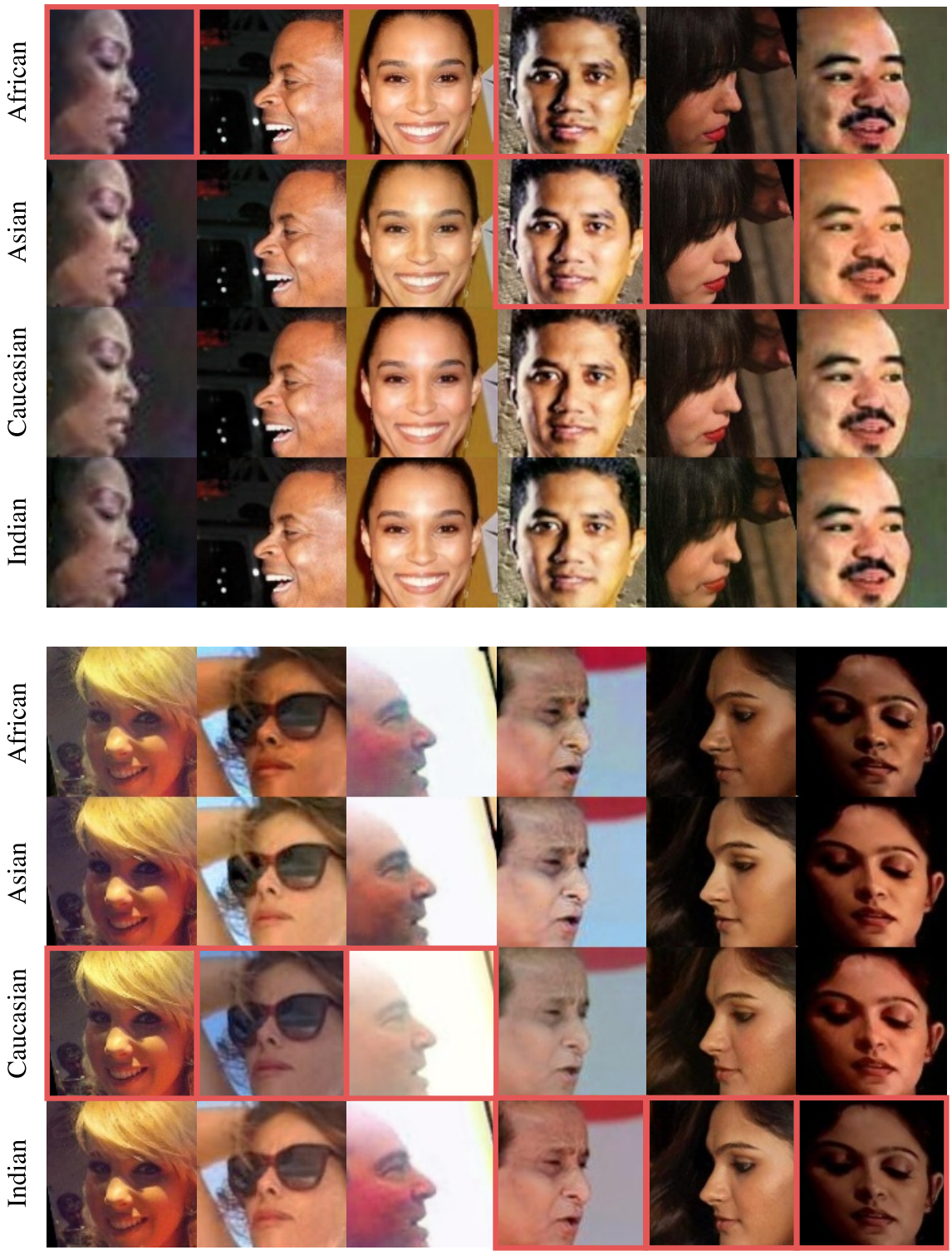


Figure 5.6: A selection of failure examples of the CycleGAN racial domain transformation of VGGFace2 dataset. Each column contains an original and sythesised face images of the same subject where the red borders indicate the original image and the corresponding race labels are laid out on the y-axis.

CHAPTER 6

Disentangling Racial Phenotypes: Fine-Grained Control of Race-related Facial Phenotype Characteristics

Achieving an effective fine-grained appearance variation over 2D facial images, whilst preserving facial identity, is a challenging task due to the high complexity and entanglement of common 2D facial feature encoding spaces. Despite these challenges, such fine-grained control, by way of disentanglement is a crucial enabler for data-driven racial bias mitigation strategies across multiple automated facial analysis tasks, as it allows to analyse, characterise and synthesise human facial diversity. In this chapter, we propose a novel GAN framework to enable fine-grained control over individual race-related phenotype attributes of the facial images. Our framework factors the latent (feature) space into elements that correspond to race-related facial phenotype representations, thereby separating phenotype aspects (e.g. skin, hair colour, nose, eye, mouth shapes), which are notoriously difficult to annotate robustly in real-world facial data. Concurrently, we also introduce a high quality augmented, diverse 2D face image dataset drawn from CelebA-HQ for GAN training. Unlike prior work, our framework only relies upon 2D imagery and related parameters to achieve state-of-the-art individual control over race-related phenotype attributes with improved photo-realistic output.

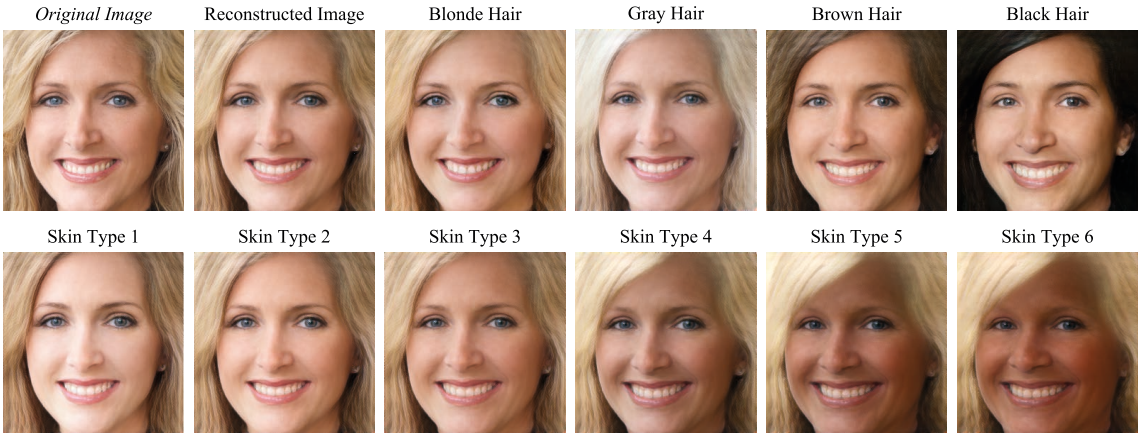


Figure 6.1: Generated images with controlled race-related phenotypes by our proposed framework.

The material presented in this chapter of the thesis has been submitted in the following peer-reviewed publication:
Yucer, Seyma, Amir Atapour, Noura Al Moubayed, Toby P. Breckon., Disentangling Racial Phenotypes: Fine-Grained Control of Race-related Facial Phenotype Characteristics, IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2024 (under review).

6.1 Introduction

Analysing and characterising human facial diversity is crucial for automated facial analysis tasks, especially as increasing research reveals the presence of racial bias causing disparate performances for racial groups [28, 238]. Moreover, we highlight the advantage of race-related facial attribute level analysis of racial bias to avoid using ill-defined racial categories and further specify the race-related facial phenotype attribute categories for racial bias evaluation in Chapter 2 and 3 respectively.

On the other hand, disentanglement learning, with its primary objective being to capture independent data variation factors, shows promise for achieving group fairness or demographic parity [253] for classification tasks and can be particularly relevant in mitigating racial bias. Earlier studies [209, 253] discuss how disentangled representation learning can enhance group fairness by isolating variations into independent components,

thereby improving interpretability, and simplifying downstream prediction tasks.

Subsequently, the latest advancements in Generative Adversarial Networks (GAN) [52, 254] not only enable high-quality face image generation but also provide control and editing capabilities within the image generation process [51, 255]. Existing literature on controllable GAN is separated into two categories following [256]: relative control [257–260] and explicit control [51, 255, 256, 261]. Relative control provides basic manipulations like changing illumination or facial rotation, whilst explicit control enable precise manipulations, such as setting the illumination to a lighter shade or rotating the face by exact angles (e.g. 30° to the left).

A widely adopted approach for both relative and explicit control of images within generative process is based on identifying disentanglement properties in the latent space corresponding to image attributes [51, 255, 256, 261]. Numerous studies [262–264] have identified such facial attribute properties, such as head pose, lighting, facial expressions, facial accessories, gender, and age, aiming to effectively disentangle such attributes from the facial identity. Such facial attributes can be categorised as either identity-relevant or identity-irrelevant [262]. Identity-relevant attributes, such as racial features such as nose and eye shapes, define distinctive facial characteristics that remain same under different expressions and poses. Conversely, identity-irrelevant attributes such as smiling or head pose are non-distinctive, as any alterations to them do not impact the overall identity. Consequently, disentangling identity-relevant attributes is more complex task due to their higher mutual information with facial identity, compared to identity-irrelevant attributes. Yet, much of the existing disentanglement literature primarily addresses identity-irrelevant attributes including head pose, expressions, mouth openness, smiling, and makeup [264–266].

For example, StyleRig [261] provides fine-grained control over facial images generated by StyleGAN2, integrating an additional layer that captures 3D pose and expression variations. Another work, MOST-GAN [267] proposes using explicit 3D parameters extracted by 3D Morphable Models [268], to train StyleGAN2 for expression, lighting and pose manipulation. More recently, [269] proposes a novel self-supervised disentanglement framework to decouple pose and expression without using 3D Morphable Face Models (3DMM) [270] and paired data. Whilst alternative approaches including domain

translation [264, 271] and latent space interpolation [272, 273] offer ways to control facial attributes, they often lead to entanglement where modifying one attribute can inadvertently affect others.

However, despite this progress in GAN, achieving explicit control on identity-relevant facial attributes over the generative process remains a challenge. Such explicit control requires not only keeping photo-realism and facial identity but also changing the single individual attribute in a desired way. Consequently, 3D face representations in generative models, such as 3DMM or equivalent 3D meshes, provide a deeper level of control in the latent space [274–276]. While it can facilitate disentanglement by leveraging depth and shape information, obtaining an accurate and detailed 3D imagery and supervision (attribute labels and representations) is challenging and furthermore such high-fidelity 3D imagery makes GAN training even more complex and computationally expensive [276].

Consequently, in this chapter, we aim to explicitly control race-related facial attributes, setting the foundation for creating controlled face image variations for future potential solutions to mitigate racial bias within automated facial analysis tasks. Most pertinent to our research, ConfigNet [51] provides a framework for parametric rendering over 2D facial images by incorporating 3D parameters from synthetic data. The objective of ConfigNet [51] is to generate realistic and controllable face images via modelling and generating of intricate attribute parameters (not present in the 2D dataset) within a 3D synthetic image dataset, bridging the gap between neural rendering and traditional rendering pipeline. Our aim of is specifically related with its ability to render both complex, multiple identity-relevant and -irrelevant factors into the latent space. Yet, instead of utilising 3D synthetic data, we derive the parameters in a 2D image space, which is significantly more challenging but yet has greater real-world applicability. We aim to have realistic image generation with controllable identity-relevant attributes in a factorised latent space.

To this end, inspired by ConfigNet [51] and StyleGAN2 [52], we develop an enhanced framework, solely grounded on 2D imagery and its metric-based parameters, for controlling specific race-related facial phenotypes such as skin and hair colour, and shapes of nose, eyes, and mouth. Our approach emphasises explicit control over these facial parameters, which are delineated and quantified using 2D image evaluations. Initially, we define these race-related phenotype parameters through 2D metric-based evaluations, subse-

quently factorised them into the latent space. We then improve the ConfigNet framework by adopting the generator-discriminator architecture of StyleGAN2, replace the synthetic data and its 3D parameters in favour of 2D high-resolution training data for which we curate an augmented, diverse dataset derived from CelebHQ.

In this chapter, our key contributions are as follows:

- We propose a framework that achieves explicit control over identity-relevant race-related facial phenotypes via a single factorised and disentangled latent space.
- Our framework relies on simple hand-crafted 2D metrics parameters obtained by public face dataset, eliminating the need for 3D render data or manual auditing.
- We introduce the CelebA-HQ-Augmented-Cleaned dataset, which is the first semi-synthesised, manually-cleaned, high-quality dataset encompassing over 26,500 images with an improved racially diverse distribution.
- We demonstrate that our proposed framework achieves both higher image quality and improved controllability on race-related facial phenotype attributes when compared to contemporary state of the art approaches [51].

6.2 Methodology

Our method employs two 2D face image datasets: a supervised set I_C sampled from CelebA-HQ [7] and an unsupervised set I_F from FFHQ [55]. The primary distinction between I_C and I_F is their intended use. I_C introduces race-related facial phenotype attributes into the factorised latent space, while I_F is used without any paired supervision (facial phenotype attribute). Our framework does not require any supervision during the test phase. We detail the process of acquiring race-related facial phenotype attributes of I_C to factorise in latent space in Section 6.2.1 and further explain our framework in Section 6.2.2.

6.2.1 Race-related Facial Phenotypes in Factorised Latent Space

Previously we identify a set of observable race-related facial phenotype characteristics that are specific to face and correlated to the racial profile of the subject. These rep-

resentative race-related facial attributes encompass skin, hair colour, eye, nose, and lip shape. We use the same attribute categories within the factorised latent space and denote each of them with θ corresponding naming as in Table 6.1. As a result, each facial image within the supervised dataset contains various predetermined facial phenotype: skin colour θ_{skin}, hair colour θ_{hair}, nose θ_{nose}, eye θ_{left_eye} and θ_{right_eye}, and mouth θ_{mouth} features. We derive hand-crafted metric-driven representations for these specific phenotype attributes, avoiding subjective annotations. Following this, akin to the methodology in ConfigNet [51], each phenotype attribute is factorised into k components θ_1 to θ_k, as follows:

$$\theta \in \mathbb{R}^m = \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \cdots \times \mathbb{R}^{m_k} \tag{6.1}$$

Each θ_i corresponds to a semantically meaningful facial phenotype attribute to generate I_C. The supervised data encoder E_C maps each θ_i to z_i, a part of z, which thus factorises z into k parts. The factorised latent space enables manipulation of pre-defined attributes in generated images by swapping specific attributes such as skin colour of the part represented by $z_i = E_{C_i}(\theta_i)$. We also present such attributes and descriptions in Table 6.1.

Figure 6.2: Metric-based parameters for race-related facial phenotypes: (a) Top column images are sourced from CelebA-HQ [7], (b) Mask images provided by MaskGAN [8]. (c) The facial skin area used for skin colour and (d) the hair area used for hair colour. (e-h) The specific face patch inputs applied for feature extraction.

Skin and Hair Colour: We utilise skin and hair segmentation masks on face images in order to quantify skin and hair colour. MaskGAN [8] provides hand-annotated mask images (as shown in the second column (b) of Figure 6.2.) for CelebA-HQ [7] dataset with

Phenotype	**Representation**	**Description**	$Input \rightarrow Output$
Skin Colour	$\theta_{skin} = \{V_{mean}, S_{mean}, Cr_{mean}\}$	Melanin, Greyness, Redness	$\mathbb{R}^3 \rightarrow \mathbb{R}^3$
Hair Colour	$\theta_{hair} = \{V_{mean}, S_{mean}, Cr_{mean}\}$	Melanin, Greyness, Redness	$\mathbb{R}^3 \rightarrow \mathbb{R}^3$
Left Eye	$\theta_{lefteye} = \{q_1, q_2..., q_{125}\}$	Left eye feature vector	$\mathbb{R}^{125} \rightarrow \mathbb{R}^{125}$
Right Eye	$\theta_{righteye} = \{q_1, q_2..., q_{125}\}$	Right eye feature vector	$\mathbb{R}^{125} \rightarrow \mathbb{R}^{125}$
Nose	$\theta_{nose} = \{q_1, q_2..., q_{128}\}$	Nose feature vector	$\mathbb{R}^{128} \rightarrow \mathbb{R}^{125}$
Mouth (Lips)	$\theta_{mouth} = \{q_1, q_2..., q_{128}\}$	Mouth feature vector	$\mathbb{R}^{128} \rightarrow \mathbb{R}^{125}$

Table 6.1: Dimensions and descriptions of race-related facial phenotype attributes in factorised latent space.

19 classes including all facial components and accessories. We restrict the skin region on the skin segments via facial landmark points, considering the potential overlap of beard and eyeglasses on the face. Subsequently, we measure the melanin, greyness, and redness values within the selected skin region and the hair region (column (c) for skin and (d) for hair in Figure 6.2). As a baseline for our work, ConfigNet [51] employs these values for hair colour analysis using a 3D image rendering software. Instead, we estimate the 2D colour spaces of the skin and hair regions to capture the *melanin, greyness, and redness* values within these regions. Specifically, for the *melanin* representation, we convert the skin and hair pixels (separately) from the RGB colour space to the HSV colour space and measure the mean value of the (V) channel describing the intensity of the colour. Increased (V) corresponds to a lighter skin tone due to decreased melanin levels, with reverse correlation providing skin colour representation. Similarly, to assess the *greyness* representation, we estimate the mean saturation value (S) from the HSV space, which represent the degree of greyness. Lastly, we convert the RGB colour space to the YCrCb colour space and extract the (Cr) channel mean value within the selected skin and hair regions to capture the redness component.

Nose, Lip, Eye Shape Feature: We extract representations of the eyes, nose, and mouth from images, and produce 64×64 pixel patch images, as shown in Figure 6.2 columns (e-g) using facial landmarks. For each facial region (left eye, right eye, lips, and mouth), we train individual MobilenetV2 networks [277] using the original CelebA dataset and its facial attribute categories excluding CelebA-HQ [7] samples to be later utilised as I_C. Features are then extracted from the final layer of corresponding model. As prior work [13] also categorises the eyes, nose, and mouth into two groups, we utilise the ground truth labels from CelebA attributes: *"Big Nose"* for the nose patch, *"Big Lips"*

for the mouth patch, and ‘‘*Narrow Eyes*” for both left and right eye patch images.

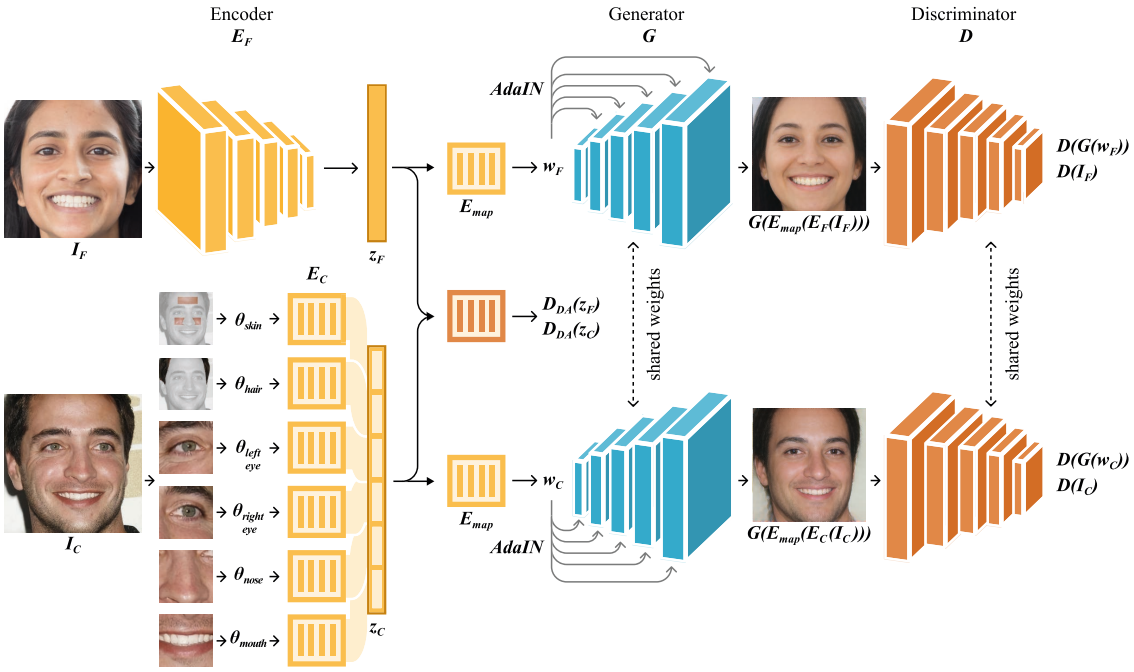


Figure 6.3: ConfigNet employs two encoders E_F and E_C that encode face images I_F and I_C in latent space vectors z_F and z_C, respectively. These vectors are further transformed into w_F and w_C using E_{map}, which are then fed into the shared decoder G for image generation. A domain discriminator D_{DA} ensures the similarity of latent distributions generated by E_F and E_C.

6.2.2 Proposed Framework

Building on the structure of the baseline [51], our method incorporates a decoder G and two encoders, E_F and E_C and a discriminator D as can be seen in Figure 6.3. E_F is a ResNet-50 backbone architecture [14] pre-trained on ImageNet [278]. E_C is a set of separate multi-layer perceptrons (MLPs) E_{C_i} for each of the corresponding θi in Table 6.1. These encoders E_C and E_F embed both I_F and I_C into a unified factorised latent space z_F and z_C respectively.Unsupervised set I_F is provided to its encoder as images from the set I_F, whereas supervised data is represented as vectors $\theta \in R^m$, which thoroughly delineate the content of the associated image in I_C (as explained in Section 6.2.1). Subsequently, both z_F and z_C are transformed into w_F and w_C using the StyleGAN2 mapping network E_{map}, which comprises eight fully-connected layers. The vector size of z_F, z_C and w are all 512.

(a) (b) (c)

Figure 6.4: The impact of one-shot learning through fine-tuning. (a) Original image. (b) Reconstructed image after second-stage training. (c) Reconstructed image after fine-tuning.

Whilst the baseline work [51] employs separate discriminator networks, D_F and D_C, for both real and synthetic data to enhance image realism, we implement a shared discriminator D in the second stage, given our sole dependence on 2D image sets, negating the need to close the realism gap caused by the use of synthetic data in [51]. Similar to [55], we apply a two-stage training strategy.

In the first stage, we train a shared StyleGAN2 generator G with its mapping encoder E_{map} [55], and separate discriminators D_F and D_C and encoder E_C. z_F is sampled from the normal distribution and encoder E_F is not included in this stage. With the combined StyleGAN2 architecture [55], the first stage loss is:

$$\begin{aligned} L_1 = {} & L_{GAN_G}(D_F, G(w_F)) + L_{GAN_G}(D_{DA}, z_C) \\ & + L_{GAN_G}(D_C, G(w_C)) + \lambda_{perc} L_{perc}(G(w_C), I_C) \end{aligned} \tag{6.2}$$

where $L_{GAN_G}(D, x) = -\log(D(x))$. As StyleGAN2 maps the input latent vector z to an intermediate latent space w, we first map factorised latent space z_C to w_C and then control the generator through adaptive instance normalisation (AdaIN) at each convolution layer of G. We remove eye loss and identity loss as we do not observe any improvement after adopting StyleGAN2. Following [51], we set the same loss weights as follows: domain adversarial loss weight $\lambda_{DA} = 5$, gradient penalty loss weight $\lambda_{R1} = 10$, perceptual loss weight in the first stage $\lambda_{perc} = 0.00005$. The adversarial losses on the images including

the style generator and discriminator losses are equally weighted.

In the second stage, we introduce E_F and a single shared discriminator D, where the pre-trained weights of D_F are utilised for training D. The second stage loss is:

$$L_2 = L_1 + \lambda_{perc} L_{perc}(G(w_F), I_F) + \log\left(1 - D_{DA}(z_F)\right) \tag{6.3}$$

where the aim of $\log(1 - D_{DA}(z_F))$ is to align the output distribution of E_F with that of E_C. We set perceptual loss weight $\lambda_{perc} = 10$ in this stage. In our experiments, the two-stage training enhanced both controllability and image quality, while attempts to single-stage training process (training all encoders, the generator, discriminator collectively in one iteration) result in unsatisfactory image generation.

One-shot learning by fine-tuning: Following the approach in [51], we employ a one-shot learning procedure to reduce the identity gap by fine-tuning the generator using individual images. This identity gap between the original and reconstructed images as well as improved reconstruction achieved in this stage are presented in Figure 6.4. In a similar vein, we fine-tune our generator on I_F by minimising the subsequent loss:

$$L_{ft} = L_{GAN_G}(D, G(\hat{w_F})) + \log\left(1 - D_{DA}(\hat{z_F})\right) + L_{perc}(G(\hat{w_F}), I_F) + L_{face}(G(\hat{w_F}), I_F)] \tag{6.4}$$

where L_{face} is a perceptual loss with VGGFace [89] as the pre-trained network. We optimise over G as well as z_F which is initialised with $E_F(I_F)$. The addition of a L_{face} improves the perceptual quality of the generated face images, whilst it is not noticeable during the main training phase, since fine-tuning lacks the regularisation achieved through training on a large number of images.

Fine-grained Phenotype Control: To have fine-grained control over the latent space generated by E_F, we adopt the gradient descent-based minimisation algorithm presented by [51]. This enables targeted modifications, such as adjusting skin colour or hair colour darkness level, while ensuring the rest of the facial attributes remain the same (for a detailed description, see [51]).

6.3 Experimental Results

In this section, we explain our training setup and experimental results to evaluate photo-realism and controllibility.

6.3.1 Datasets

We utilise the FFHQ [55] and CelebA-HQ datasets for training of our framework. FFHQ dataset [55] contains 60,000 high-resolution images of size 1024×1024 pixels. We utilise 50,000 samples from FFHQ for our training set as our primary source of unsupervised images I_F and the same 10,000 samples for the validation set (I_{test}) for a consistent comparison of results with ConfigNet [51]. CelebA-HQ, a subset of CelebA, offers 30,000 high-resolution images, each at a resolution of 1024×1024 [7] and is the source of CelebA-HQ-Clean-Augmented (supervised set, I_C).

These datasets consist of an imbalanced racial distribution. For instance, [279] reveals that the FFHQ dataset consists of 69% White, 4% African, and 27% individuals who are neither African nor white. Similarly, [280] indicates that CelebA-HQ contains over 70% White individuals and fewer than 10% of African. To address this, we introduce CelebA-HQ-Clean-Augmented which is a semi-augmented high-quality image set. We align all the face images from those datasets to a standard reference frame using landmarks from OpenFace [281] and reduce the resolution to 256×256 pixels.

Figure 6.5: A selection of images from CelebA-HQ-Clean-Augmented. While some images are augmented using the method proposed by [9], others, both original and augmented, are removed due to low imaging conditions and pose discrepancies.

CelebA-HQ-Clean-Augmented: To address the lack of diversity within the GAN training dataset, we apply our previous adversarial data augmentation technique to facilitate the transfer of race-specific facial features (Chapter 5). From the original 30,000 CelebA-

HQ images, we augmented another 30,000 images by transferring all the images from the Caucasian to the African domain (Figure 6.5). However, both the original and synthesised images exhibit poor imaging conditions and not all of the original images actually belong to Caucasian subjects, which may cause faulty or erroneous parameter estimation. Moreover, as skin colour estimation relies on colour spaces, we prioritised images without prominent shading or lighting that may mislead the skin colour evaluation. Accordingly, we manually clean and select a refined dataset containing 26,513 images; 17,861 original and 8,652 augmented. Figure 6.5 shows exemplar images from the curated CelebA-HQ-Clean-Augmented dataset.

6.3.2 Image Quality - Photorealism

In Table 6.2, we measure the photorealism of our generated images using the Frechet Inception Distance (FID) [282] and compare our results with ConfigNet [51]. First, we examine the FID score between the FFHQ and our CelebA-HQ-Clean-Augmented dataset. Since ConfigNet [51] utilises raw synthetic images, the SynthFace dataset, there is a noticeable feature distance when compared to FFHQ. By replacing SynthFace dataset with CelebA-HQ-Clean-Augmented face dataset, we not only eliminate the need for synthetic data but also significantly improve the distribution difference of training sets by lowering FID score by 12 points (from 52.19 to 40.81 $\downarrow$ compared to [51]). In the subsequent evaluation, we test the FID performance of the first stage by generating random images from the first-stage trained generator G. Notably, our framework achieves a lower perceptual distance score, indicating higher image quality and more realistic image generation (Table 6.2). Subsequently, we show our second-stage trained model reconstruction quality using E_F, we re-generate FFHQ evaluation set, I_{test}, and calculate FID score between $G(E_{map}(E_F(I_{test}))$ and I_{test}. Our approach consistently produces more realistic images 6.2 compared to [51] as illustarted by the quantitive results of Table 6.2 and the qualitative results of Figure 6.1 and 6.6.

Additionally, we modify the relevant attribute index location of the latent space vector $z_F = E_F(I_{test})$ to control the skin and hair colour of the generated image while preserving the other features. As a result, we present further qualitative results for our generated images, encompassing both reconstructed and manipulated images with focused attribute

variations in Figure 6.6.

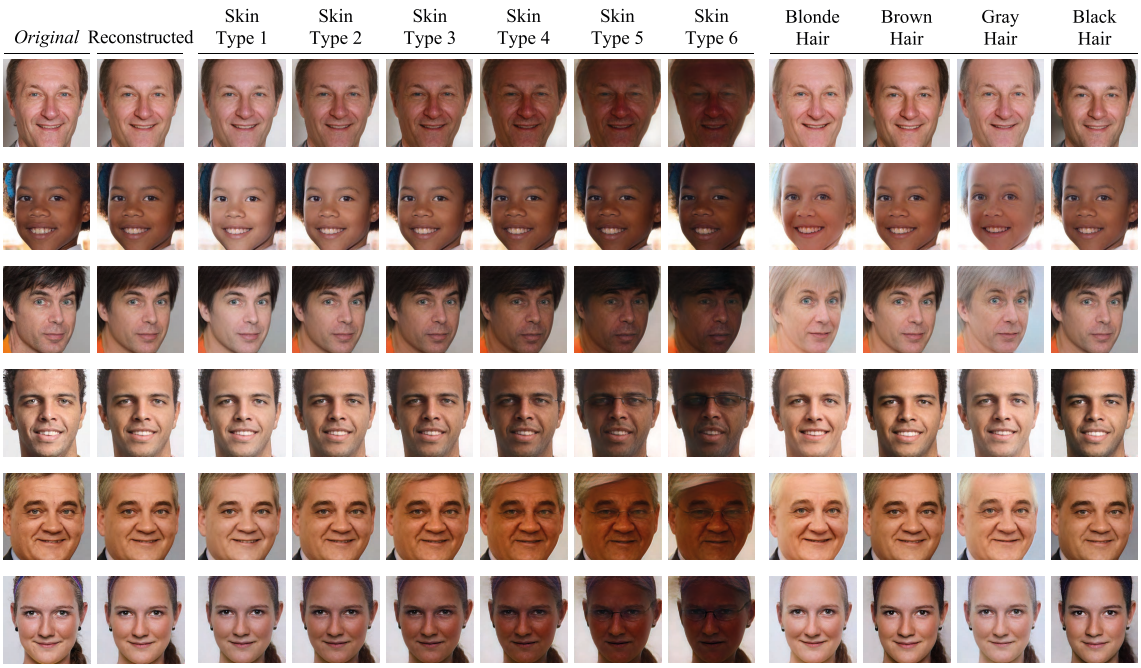


Figure 6.6: Generated and controlled images from $G(E_{map}(E_F(I_{test}))$. From the top row to the following rows, the sequence respectively shows original and reconstructed images, followed by generated images with associated attribute changes. We modify the corresponding index of $z_{test} = (E_F(I_{test}))$ to synthesise attribute-modified images.

6.3.3 Controllability

We adopt the ConfigNet controllability experiment to evaluate the effects of modifying specific attributes, such as skin colour or hair colour. Our generator successfully alters the hair and skin colour of faces within its latent space, and achieves higher control over hair colour than [51] on the generated images. Figures 6.1 and 6.6 show the qualitative results of controllability for these attributes.

Method	**ConfigNet** [51]	**Ours**
I_C	52.19	40.81
$G(z), z \approx N(0,(I))$ no 2nd stage	43.05	39.55
$G(E_F(I_F))$	33.41	28.64

Table 6.2: FID score for FFHQ, CelebA-HQ-Clean-Augmented, and images obtained with our decoder G and latent vectors z_F from the real-image encoder E_F.

To quantitatively assess the controllability of our framework, we follow [51] and randomly select 1000 images I_{test} from the FFHQ validation set, encode them into the latent space $z = E_F(I_{test})$, and then exchange the latent factor z_i associated with a specific attribute v (such as hair colour) with a factor obtained from E_C. For each attribute v, we generate two images: I^+ where the attribute is set to a value v^+ (e.g., blonde hair), and I^- where the attribute takes a semantically opposite value v^- (e.g., black hair). This results in pairs of images (I^+, I^-) that should be nearly identical except for the selected attribute v, highlighting the differences. To measure these differences, we employ an attribute predictor denoted as C_{pred}. We train a MobileNet v2 architecture on skin and hair colour, leveraging attribute labels and images from [13], and validate it on I_{test}. In an ideal scenario, $C_{\text{pred}}(I+)$ should be 1, $C_{\text{pred}}(I-)$, and the Mean Absolute Difference (MD) for other facial attributes should converge to 0.

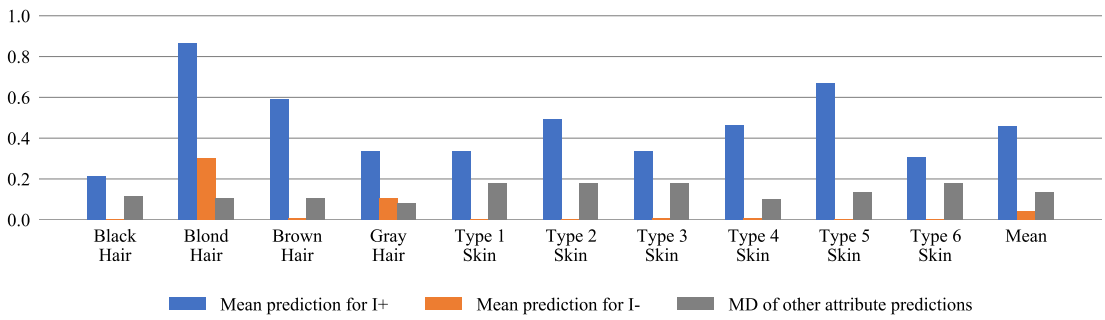


Figure 6.7: Evaluation of control and disentanglement ability of our proposed framework. Blue and orange bars represent attribute values for images with the respective attribute ($I+$ for higher values, $I-$ for lower values). Gray bars indicate differences in other attributes (MD and C_{diff} for lower values).

Figure 6.7 illustrates that $C_{\text{pred}}(I+)$ is generally greater than $C_{\text{pred}}(I-)$, while the MD for other attributes remains near 0. The highest controllability is observed for skin type 5 and blond and brown hair attributes, where $C_{\text{pred}}(I+)$ approximates the ideal value of 1. In contrast, the lowest level of control is observed for skin type 1 and black hair attributes. These substantial discrepancies arise from the attribute prediction model capacity on such attributes, as it is trained on VGGFace2 dataset [3], which contains a notably low count of Type 1 instances (as indicated by the distribution in Chapter 3). Consequently, we achieve superior control over hair colour attributes in comparison to [51], the only possible identical attributes available for comparison.

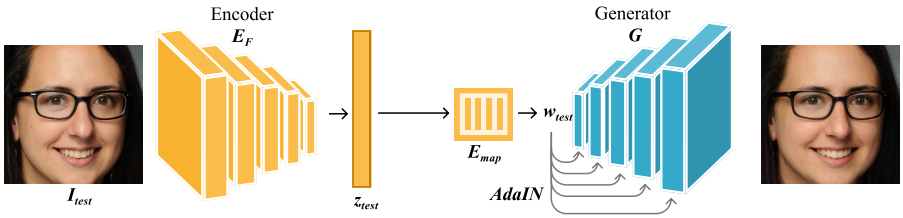


Figure 6.8: Inference of our proposed framework

Conversely, our framework encounters challenges in disentangling nose, eye, and mouth shapes. For instance, interchanging left-right eyes leads to alterations in the shape of both eyes. Moreover, altering the nose or lips causes changes in the facial pose and shape. The failure modes of these shape-related attribute changes are presented in Figure 6.9.

6.3.4 Inference

We present the inference pipeline of our framework in Figure 6.8. Importantly, our approach achieves disentanglement of race-related facial phenotypes without requiring additional attribute labels or representations. This is achieved through the training of E_F, which encodes these phenotypes within a factorised latent vector space utilised by the Generator G. For any given 2D image I_{test}, it is encoded by E_F and E_{map} in sequence, and then reconstructed by G. Simultaneously, the control of the generated image is enabled by modifying specific components of z_{test} Figure 6.8.

Furthermore, we present additional results obtained from randomly selected examples within the FFHQ validation dataset in Figure 6.10. These results demonstrate the effectiveness of our framework in manipulating the pre-defined race-related attributes. While it excels in generating variations in skin colour and hair colour, it encounters challenges in controlling nose and lip attributes.

6.3.5 Failure Modes

Figure 6.9 show failure modes of various subjects associated with changes in shape-related parameters. In the left or right narrow eye control, our framework exhibits two common issues: firstly, it tends to simultaneously alter both eyes or neither, and secondly, it misinterprets narrow eyes as closed eyes in some cases, as seen in the middle row of

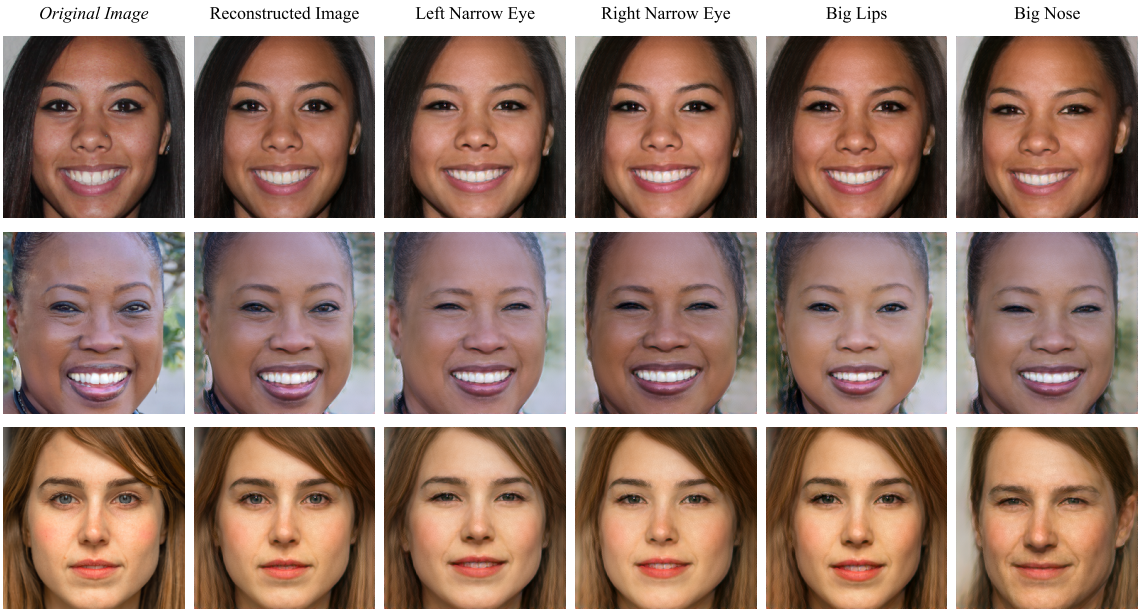


Figure 6.9: Failure modes. Eye Shape Control: leads to a slight appearance shift, affecting both eyes simultaneously. Nose and Lips Control: results in change of unrelated attributes such as pose and mouth openness.

Figure 6.9.

Similarly, for controlling the nose and lips attributes, we observe entanglement with unrelated factors such as pose and mouth openness, as presented in Figure 6.9. We hypothesise that adopting an enhanced feature representation models, such as visual transformers [283] applied to manually generated patch imagery, could lead to substantial improvements in our ability to disentangle these facial features effectively.

6.3.6 Race-related Facial Phenotypes

We utilise the phenotypes presented in Chapter 3 in two ways: firstly, we use such phenotypes as parts in our factorised latent space and secondly, to train our attribute prediction model using VGGFace2 [3] dataset, with skin and hair tone labels. While we attempted to automatically generate hand-crafted skin tone labels, the change in imaging conditions, ranging from low to high quality, lightning and shades make automatic skin tone assignment unreliable and as such this was not pursued further.

6.4 Discussion

Importance of Training Distribution of Generative Models: Race-related phenotype disentanglement through generative processes can address racial bias and provide deeper insights into the underlying reasons for disparate performances within racial groupings. However, GAN [279] reflect the discrepancies of the training data in the synthesised outputs. Despite our efforts with the CelebA-HQ-Clean-Augmented dataset to reduce the influence of imbalanced distribution of training data on GAN, some unintended correlations still appear. Specifically, when our model was fine-tuned to modify skin colour, it displayed an unintended correlation: associating darker skin tones with eyeglasses (likely due to numerous eyeglass samples within FFHQ) and blonde hair with femininity (17% of the CelebHQ samples were women with blonde hair).

Additionally, we noted challenges in controlling darker skin tones compared to lighter skin tone ones, possibly due to the symmetric algorithmic bias arises when the imbalances in the training data are magnified in the generated data [279].

Comparison of Entanglement for Shape and Colour Parameters: Achieving explicit control over shape-related parameters is more challenging than colour-related ones. This difficulty could arise from inadequate representation of shape features or the greater entanglement of shape with identity, or limitations of StyleGAN2 in handling shape information. Failure modes of such attribute parameter change are illustrated in the Figure 6.9

6.5 Summary

In this chapter, we introduce a framework, building upon ConfigNet [51], that disentangles race-related facial phenotypes in a latent space. We, first, introduce the CelebA-HQ-Augmented-Cleaned dataset, which is the first semi-synthesised, manually-cleaned, high-quality dataset encompassing over 26,500 images with an improved racially diverse distribution. Furthermore, our approach leverages such 2D publicly available FFHQ dataset and CelebA-HQ-Augmented-Cleaned dataset and further employs straightforward 2D handcrafted metrics for latent space factorisation. Our 2D handcrafted metrics does not require manual annotation or 3D rendering.

Consequently, our approach achieves fine-grained control over racial phenotypes with improves photorealism and controllability compared to ConfigNet [51] without requiring any 3D rendered synthetic data. Although the disentanglement of certain identity-relevant attributes was not entirely controllable, we believe improved and more representative feature metrics will address this in the future.

To the best of our knowledge, our study is the first to attempt disentangling and exerting explicit control over such crucial race-related facial phenotype, paving new avenues for evaluating racial bias in automated facial analysis tasks. Unlike prior work [51], our framework only relies upon 2D imagery and related parameters to achieve both higher image quality and improved controllability on race-related facial phenotype attributes.

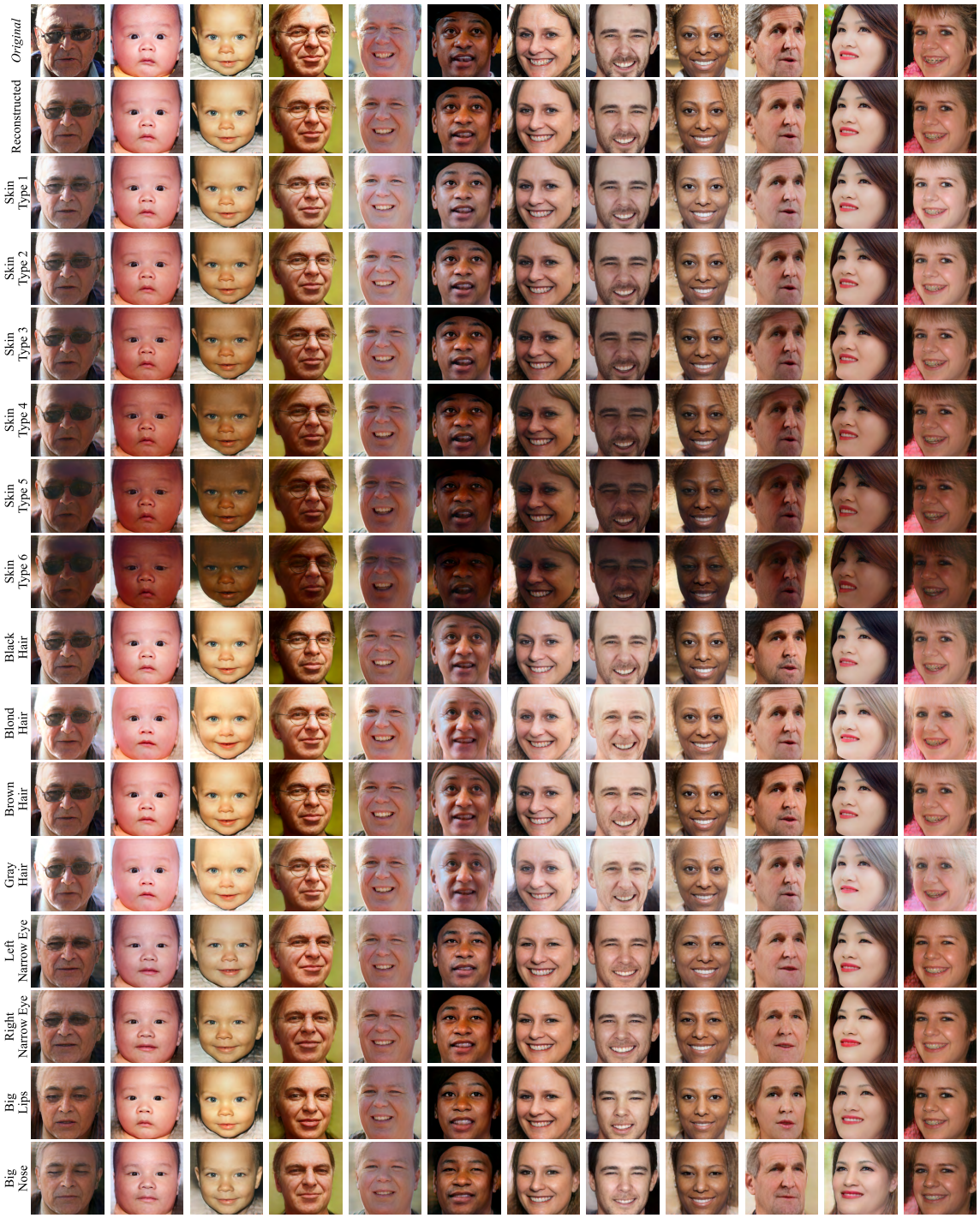


Figure 6.10: Additional examples from the FFHQ validation set, with both reconstructed and controlled images with associated attribute change.

CHAPTER 7

Conclusion

The widespread adoption of face recognition in various real-world applications has brought a rise in the occurrences of disparate face recognition performance across inter-sectional racial groupings. Despite the growing interest in the academic research and industrial endeavours, prior work proposing dataset, and evaluation strategies has raised another set of concerns regarding the social implications of racial groupings and their definition in dataset and evaluation. Furthermore, such racial bias mitigation work and their associated interpretation of racial groupings often remain limited or misguiding against specific face datasets, their annotations, and proposed methodologies.

Accordingly, this thesis aims to directly address issues of racial bias within face recognition domain and review the literature to cover the broader context of historical and social factors with the goal of achieving more accurate and fairer face recognition performance across increasingly more diverse populations. Subsequently, it provides a comprehensive overview of existing methods while drawing upon a general concept of the face recognition processing pipeline. Each phase of the processing pipeline is examined in relation to racial bias to provide a broader and deeper understanding of the current advancements and challenges within the field.

Consequently, the thesis addresses the ambiguity of racial bias problem definition and

ill-defined grouping strategies by proposing a new evaluation methodology using facial phenotype attributes. The thesis introduces a more objective and granular strategy to evaluate and ultimately address racial bias within face recognition whilst avoiding exposing potentially protected or ill-defined attributes (Chapter 2 and 3).

Subsequently, one of the key operations in the initial phase of the face recognition processing pipeline, namely lossy image compression, is investigated to uncover its impact on face recognition performance. The relationship between face verification performance for a given race-related phenotypic group under varying levels of lossy compressed sets revealed more significant decrease in performance for specific phenotypic groups. Removing chroma subsampling and the use of compressed imagery during training do make models more resilient and improves FMR for specific phenotype categories more affected by lossy compression (Chapter 4).

Moreover, this thesis proposes transferring racial information over the facial subjects while keeping prior identity information for face recognition. The methodology lies on augmenting and transferring sensitive racial information rather than removing it to make racially dependant features subsequently irrelevant within the determination of subject identity (Chapter 5).

Consequently, this thesis advances the domain transfer concept by addressing race-related facial phenotypes discussed in the Chapter 5. Our proposed methodology enable the factorisation of these phenotype attributes within the latent space through a generative process. This explicit control over facial phenotypes within the latent space not only facilitates the modification of specific phenotypes but also enables the analysis of their impact on face recognition performance. Consequently, we delve into future prospects, examining potential solutions and limitations in mitigating racial bias within face recognition (Chapter 6).

7.1 Contributions

We exhaustively trace the evolving field of racial bias within face recognition providing a thorough examination of both the technical aspects of face recognition, such as algorithmic and dataset biases, and the socio-cultural constructs of race concept that have an

impact on the field. Our literature review in Chapter 2 establishes a strong foundation for understanding the complexities surrounding racial bias in face recognition, setting the stage for further exploration and solutions in subsequent Chapters 3-6. Presenting the literature on annotation processes for grouping categories and summarising recent works and face datasets organised by grouping strategies in Table 2.1 reveals a lack of consensus that significantly hinders collaborative efforts to address bias due to inconsistent problem definition across the field. As one of the biggest highlights within this work, we address such the usage of ill-defined racial groupings and introduce of a race-related based facial phenotypes (Chapter 3). Our phenotype-based grouping methodology uncover more considerable performance disparities among phenotype attributes than racial groups and hence a more resolute measure of performance bias.

Furthermore, we contribute to the understanding of face recognition performance via use of race-related phenotype attributes subjected to varying levels of lossy compression (Chapter 4). Our evaluation reveals that the utilisation of lossy compressed facial image samples during inference significantly decreases performance, particularly impacting phenotypic features such as dark skin tone, wide nose, curly hair, and monolid eyes, while affecting other features to a lesser extent. Interestingly, employing compressed imagery during model training increases resilience and decreases performance difference, yet disparities persist among racially-aligned subgroups. Furthermore, we observe that the removal of chroma subsampling notably reduces False Match Rates (FMR) for certain phenotype categories, demonstrating the potential for targeted improvements in the face verification process (Chapter 4).

Subsequently, we address racial bias in face recognition through the introduction of an innovative adversarial-derived data augmentation method (Chapter 5). Our approach focuses on transferring racial attributes within facial images while preserving essential identity features within face recognition datasets. This methodology enhances the robustness of face recognition algorithms, reducing their dependency on race as a distinguishing factor. Subsequently, our study, has showed measurable improvements in face recognition performance disparities across racial groups, as detailed in Table 5.1 and 5.2. Although these improvements may not reach statistical significance (via Statistical Hypothesis Tests), complicating the interpretation of their practical impact, machine learning

algorithms requires much more complicated evaluations for statistical significance [284]. Moreover, it is crucial to recognise that face recognition systems are frequently deployed on a large scale. Therefore, even modest improvements can yield significant real-world effects, enhancing both the performance and fairness of these systems across various applications. Moreover, the methods by which we measure performance, along with how we design benchmark test sets and focus on evaluating racial bias, significantly influence the statistical significance of the results.

Consequently, our results demonstrate a remarkable 15.81% reduction in performance variations among four distinct racial groups (African, Asian, Caucasian, Indian). Furthermore, we undertake a comprehensive evaluation of our approach using our phenotype-based evaluation methodology, demonstrating notable improvements in performance. Despite challenges posed by illumination, pose, and lighting variations, our technique not only enhances overall face recognition accuracy but also effectively mitigates inter-group performance disparities (Chapter 5).

Consequently, we propose a novel framework that extends ConfigNet [51] to disentangle race-related facial phenotypes within a latent space (Chapter 6). Leveraging 2D datasets and straightforward handcrafted metrics, our approach provides fine-grained control over race-related phenotypes, enhancing photorealism and controllability without the need for synthetic data. While some aspects of identity-related attribute disentanglement present challenges, our work paves the way for future research with the goal of addressing these issues through improved feature metrics. With our primary future aim of advancing research on racial bias by facilitating the generation of race-related facial appearance variations, our proposed method, enabling a disentangled feature space, may open new avenues for evaluating racial bias in automated facial analysis task (Chapter 4).

7.2 Limitations and Future Work

Given the inherent complexity of computer vision problems, particularly within the context of the socially controversial topic of racial bias, our contributions within this thesis entail certain limitations that open the way for future research work. In this section, we discuss these limitations and potential avenues for future work.

7.2.1 Face Imagery and Face Recognition Datasets

As discussed in Chapter 2, facial imagery represents one of the most distinctive forms of data within the computer vision, encapsulating not only the biometric characteristics of human subjects but also other valuable information. However, this significant biometric data is often not sensitively curated or thoughtfully structured within facial datasets, despite serving as a foundational component for the evaluation of algorithms and methodologies.

Contemporary face recognition datasets [3,15] aim to challenge face recognition models by introducing challenging imaging conditions while often disregarding demographic statistics and distributions within such datasets. In Chapters 3 and 5, we address these imbalances by either proposing a generative model for subject-level augmentation or curating them to mitigate the skewed distribution that can amplify dataset bias and hence the results. However, generated imagery can introduce distortions and lower image quality, potentially exacerbating bias towards certain attributes. While recent facial dataset efforts including RFW [11], BuptBalanced datasets [1] have aimed to improve data distribution, they often inherit socially constructed racial groupings, which introduce a new set of concerns as discussed in Chapters 2 and 3.

Future work, with a focus on acquiring and releasing diverse, ethically sourced, and challenging datasets, while providing comprehensive details about the data collection process, may contribute to standardising benchmarks for evaluation of racial bias within face recognition.

7.2.2 Dataset Annotation and Grouping Strategies

The issue of racial bias within face recognition is defined as a supervised learning problem, requiring the race or race-related labels alongside subject identity labels. However, given the relatively recent emergence of this topic, there is a few limited datasets that provide such racial labels [1, 11]. Consequently, numerous studies have resorted to different grouping strategies on contemporary face recognition datasets (already constructed with racially imbalance distributions) without offering comprehensive explanations for these grouping decisions or their annotation processes [1, 11, 73]. The diversity and absence of

a consensus approach to this challenge are discussed in Chapters 2 and 3.

In these chapters, we highlight the importance of reaching a consensus and elucidate the drawbacks associated with the usage of socially ill-defined racial categories. We also introduce a novel evaluation methodology that leverages facial phenotypes. However, the scope of such phenotypes can be extended both in depth by introducing more detailed categories and in breadth by including additional phenotypes such as face shape and nose length, as suggested by [130]. Nevertheless, this expansion adds complexity to the evaluation process which is already a significantly more complex evaluation strategy due to the significant number of phenotype categories.

Although race-related facial phenotypes are effective for evaluation of racial bias in face recognition, it is essential to recognise that the concept of race extends beyond facial attributes meaning that degrading racial bias to these categories may divert attention from other aspects of the racial bias issue [71].

Crucially, adopting an unsupervised learning approach for racial bias within face recognition, rather than overly categorising every attribute of the face, has the potential to address issues related to group fairness criteria, dataset annotation, and may led the way for more robust and effective solutions in future research.

7.2.3 Image Generation for Fairness-Racial Bias

Generative neural networks have gained significant pace in computer vision, offering a different types of utilities for various downstream tasks. More generally, the generative process enhances our comprehension of data distributions, enables control over data generation, and facilitates the new data generation (whether random, manipulated, or conditioned) to address imbalanced learning problem.

In Chapters 5 and 6, we leverage these capabilities to mitigate and analyse racial bias within face recognition. Despite limitations in our generation process, particularly in challenging imaging conditions including pose and lighting changes, our image-to-image generation methodology in Chapter 5 have shown significant improvements in mitigating performance disparities in face recognition. Furthermore, the methodology presented in Chapter 6 introduces new avenues for examination of racial bias, encompassing both feature latent space and 2D image space, through its explicit control over race-related

facial phenotypes (presented in Chapter 3).

The limitations inherent to our methodology in Chapter 6, specifically the inability to control more intricately entangled features like the nose and mouth, may be mitigated through the utilisation of enhanced representations beyond those introduced in Chapter 6 or by employing more sophisticated generative models including the integration of diffusion models and transformers [285–287], designed for the precise analysis of racial bias.

Future work in this area will build upon our existing efforts, with a continued focus on achieving a fully controllable latent space that encompasses race-related facial phenotypes. This advancement holds significant promise for deepening our understanding of the attribute-level impact of race-related features on face recognition performance.

It is worth emphasizing that the generative process, while offering great potential, can introduce various forms of bias. This potential for bias amplification necessitates careful consideration, especially regarding its impact on contemporary face datasets [279]. As their applicability to real-world scenarios continues to expand, future research work should focus on the assessment of the bias of these models.

Finally, given the limited body of research on the bias of generative networks, we must remain vigilant about the potential consequences of amplifying specific data aspects. These consequences could significantly affect the development and evaluation of face recognition. The impact of such amplification may not be evenly distributed among all individuals globally, which raises again another concerns about the exacerbation of social biases and inequalities, particularly for groups already experiencing disparities and injustices [68].

Bibliography

[1] M. Wang and W. Deng, "Mitigating bias in face recognition using skewness-aware reinforcement learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, 2020. xii, 7, 18, 36, 43, 62, 68, 70, 72, 78, 83, 89, 91, 99, 128

[2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision*, 2017. xiii, 91, 92, 94, 95

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *IEEE International Conference on Auto. Face Gesture Recognition*, p. 6774, 2018. xiii, xvii, 7, 18, 36, 41, 54, 60, 69, 78, 93, 96, 97, 98, 100, 101, 118, 120, 128

[4] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, 2017. xiii, 12, 41, 62, 70, 90, 93, 96, 101

[5] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, 2018. xiii, 12, 41, 42, 43, 45, 48, 90, 93, 97, 99, 101

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, 2019. xiii, 12, 41, 42, 43, 45, 48, 62, 69, 74, 78, 89, 90, 93, 96, 97, 99, 100, 101

[7] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018. xiv, 7, 109, 110, 111, 115

[8] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, 2020. xiv, 110

[9] S. Yucer, S. Akçay, N. Al-Moubayed, and T. P. Breckon, "Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition Workshop*, 2020. xiv, 3, 18, 72, 115

[10] M. Merler, N. Ratha, R. S. Feris, and J. R. Smith, "Diversity in faces," *arXiv preprint arXiv:1901.10436*, 2019. xvi, 18, 20, 24, 28, 29

[11] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in-the-wild: Reducing racial bias by information maximization adaptation network," in *International Conference Computer Vision*, 2019. xvi, xvii, 3, 7, 18, 19, 43, 47, 49, 60, 77, 78, 91, 96, 98, 99, 100, 128

[12] C. Feliciano, "Shades of race: How phenotype and observer characteristics shape racial classification," *American Behavioral Science*, pp. 390–419, 2016. xvi, 7, 27, 56, 58

[13] S. Yucer, F. Tektas, N. A. Moubayed, and T. P. Breckon, "Measuring hidden bias within face recognition via racial phenotypes," in *Proceedings IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3202–3211, 2022. xvii, 3, 18, 23, 25, 72, 74, 80, 82, 83, 84, 100, 111, 118

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, 2016. xvii, 41, 62, 70, 78, 89, 96, 98, 100, 101, 112

[15] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," tech. rep., University of Massachusetts, Amherst, 2007. xvii, 36, 48, 49, 50, 65, 91, 98, 99, 128

[16] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *Conference on Graphics, Patterns and Images*, IEEE, 2018. 1

[17] S. Bashbaghi, E. Granger, R. Sabourin, and M. Parchami, "Deep learning architectures for face recognition in video surveillance," in *Deep Learning in Object Detection and Recognition*, Springer, 2019. 1

[18] L. Hemamou, G. Felhi, V. Vandenbussche, J.-C. Martin, and C. Clavel, "Hirenet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews," in *AAAI Conference on Artificial Intelligence*, 2019. 1

[19] M. A. Uddin, J. B. Joolee, and Y.-K. Lee, "Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm," *IEEE Transactions on Affective Computing*, 2020. 1

[20] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435, 2019. 1, 21

[21] H. Du, H. Shi, D. Zeng, X.-P. Zhang, and T. Mei, "The elements of end-to-end deep face recognition: A survey of recent advances," *ACM Computing Surveys*, 2022. 2

[22] W. Ali, W. Tian, S. U. Din, D. Iradukunda, and A. A. Khan, "Classical and modern face recognition approaches: A complete review," *Multimedia Tools and Applications*, pp. 4825–4880, 2021. 2, 30

[23] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, pp. 215–244, 2021. 2, 4, 40

[24] Y. Kortli, M. Jridi, A. A. Falou, and M. Atri, "Face recognition systems: A survey," *Sensors*, p. 342, 2020. 2, 4, 30, 54

[25] R. D. Labati, A. Genovese, E. Muñoz, V. Piuri, F. Scotti, and G. Sforza, "Biometric recognition in automated border control: A survey," *ACM Computing Surveys*, 2016. 2

[26] A. K. Jain and S. Z. Li, *Handbook of Face Recognition*. Springer, 2011. 2

[27] B. Amos, B. Ludwiczuk, M. Satyanarayanan, *et al.*, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, p. 20, 2016. 2

[28] R. V. Garcia, L. Wandzik, L. Grabner, and J. Krueger, "The harms of demographic bias in deep face recognition research," in *IEEE International Conference on Biometrics*, 2019. 2, 54, 106

[29] N. Srinivas, K. Ricanek, D. Michalski, D. S. Bolme, and M. King, "Face recognition algorithm bias: Performance differences on images of children and adults," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition Workshop*, 2019. 2, 54

[30] P. Grother, M. Ngan, and K. Hanaoka, *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019. 2, 18, 54, 64, 69

[31] C. A. Meissner and J. C. Brigham, "Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review.," *Psychology, Public Policy, and Law*, p. 3, 2001. 2

[32] P. J. Hills and M. B. Lewis, "Short article: Reducing the own-race bias in face recognition by shifting attention," *Quartarly Journal Experimental Psychology*, pp. 996–1002, 2006. 2

[33] G. Anzures, P. C. Quinn, O. Pascalis, A. M. Slater, J. W. Tanaka, and K. Lee, "Developmental origins of the other-race effect," *Current Directions in Psychology Science*, pp. 173–178, 2013. 3

[34] G. Rhodes, V. Locke, L. Ewing, and E. Evangelista, "Race coding and the other-race effect in face recognition," *Perception*, pp. 232–241, 2009. 3

[35] G. L. Wells and E. A. Olson, "The other-race effect in eyewitness identification: What do we do about it?," *Psychology, Public Policy, and Law*, p. 230, 2001. 3

[36] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole, "An other-race effect for face recognition algorithms," *ACM Transactions on Appl. Perception*, pp. 1–11, 2011. 3

[37] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on Inform. Forensics and Security*, pp. 1789–1801, 2012. 3

[38] A. J. O'Toole, P. J. Phillips, X. An, and J. Dunlop, "Demographic effects on estimates of automatic face recognition performance," *Image and Vision Computer*, pp. 169–176, 2012. 3

[39] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings Conference on Fairness, Accountability and Transparency* (S. A. Friedler and C. Wilson, eds.), pp. 77–91, PMLR, 2018. 3, 16, 18, 19, 22, 24, 25, 26, 55, 91

[40] K. Krkkinen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proceedings IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1547–1557, 2021. 3, 18, 19

[41] H. F. Menezes, A. S. Ferreira, E. T. Pereira, and H. M. Gomes, "Bias and fairness in face detection," in *IEEE Conference on Graphics, Patterns and Images*, pp. 247–254, 2021. 3, 38

[42] S. Shiaeles, "Facebook will drop its facial recognition system - but heres why we should be sceptical," *The Conversation*, 2021. 3

[43] J. Menn, "Microsoft turned down facial-recognition sales on human rights concerns," *UK Reuters*, vol. 17, 2019. 3

[44] D. Castelvecchi, "Is facial recognition too biased to be let loose?," *Nature*, 2020. 3

[45] T. Sixta, J. C. S. Jacques Junior, P. Buch-Cardona, E. Vazquez, and S. Escalera, "Fairface challenge at eccv 2020: Analyzing bias in face recognition," in *Computer Vision – ECCV 2020 Workshops* (A. Bartoli and A. Fusiello, eds.), pp. 463–481, Springer International Publishing, 2020. 3, 18, 23, 25, 26, 72

[46] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Commun. of the ACM*, pp. 86–92, 2021. 3, 21

[47] A. Dantcheva, P. Elia, and A. Ross, "What else does your biometric data reveal? a survey on soft biometrics," *IEEE Transactions on Inform. Forensics and Security*, pp. 441–467, 2016. 4

[48] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Transactions on Technology and Society*, 2020. 4

[49] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, 2021. 4, 72

[50] K. Orphanou, J. Otterbacher, S. Kleanthous, K. Batsuren, F. Giunchiglia, V. Bogina, A. S. Tal, A. Hartman, and T. Kuflik, "Mitigating bias in algorithmic systems-a fish-eye view," *ACM Computing Surveys*, 2022. 4

[51] M. Kowalski, S. J. Garbin, V. Estellers, T. Baltruaitis, M. Johnson, and J. Shotton, "Config: Controllable neural face image generation," in *European Conference on Computer Vision*, 2020. 7, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 121, 122, 127

[52] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 8110–8119, 2020. 7, 37, 107, 108

[53] A. Fakhro, H. W. Yim, Y. K. Kim, and A. H. Nguyen, "The evolution of looks and expectations of asian eyelid and eye appearance," in *Seminars in plastic surgery*, Thieme Medical Publishers, 2015. 7

[54] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3676–3684, 2015. 7

[55] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019. 7, 91, 100, 109, 113, 115

[56] T. M. Mitchell, *The Need for Biases in Learning Generalizations*. Dep. of Computer Science, Laboratory for Computer Science Research, 1980. 11, 42

[57] T. Hellström, V. Dignum, and S. Bensch, "Bias in machine learning–what is it good for?," *arXiv preprint arXiv:2004.00686*, 2020. 11, 36, 42

[58] S. Barocas, M. Hardt, and A. Narayanan, "Fairness and machine learning," *Nips tutorial*, p. 2, 2017. 11

[59] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," *Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012. 11, 14

[60] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in Neural Inform. Process. Syst.*, 2017. 11, 14, 15

[61] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Sign. Process. Letters*, 2018. 12

[62] W. Fleisher, "What's fair about individual fairness?," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, (New York, NY, USA), p. 480490, ACM, 2021. 14

[63] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in Neural Information Process. Syst.*, 2016. 15

[64] B. Woodworth, S. Gunasekar, M. Ohannessian, and N. Srebro, "Learning non-discriminatory predictors," in *Conference on Learning Theory*, pp. 1920–53, PMLR, 2017. 15

[65] Y. Sun, Y. Li, and Z. Cui, "Nfw: Towards national and individual fairness in face recognition," in *Pattern Recognition*, pp. 540–553, 2021. 15

[66] S. Xue, M. Yurochkin, and Y. Sun, "Auditing ml models for individual bias and unfairness," in *International Conference on Artificial Intelligence and Statist.*, pp. 4552–4562, PMLR, 2020. 15

[67] R. Benjamin, "Race after technology," *Social Forces*, 2019. 16

[68] T. Zuberi, *Thicker than Blood: How Racial Statistics Lie*. U of Minnesota Press, 2001. 16, 17, 18, 21, 130

[69] P. Mozur, "One month, 500,000 face scans: How china is using a.i. to profile a minority," *The New York Times*, 2019. 16, 21, 55

[70] M. Mitchell, D. Baker, N. Moorosi, E. Denton, B. Hutchinson, A. Hanna, T. Gebru, and J. Morgenstern, "Diversity and inclusion metrics in subset selection," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, (New York, NY, USA), p. 117123, ACM, 2020. 16, 21, 55

[71] A. Hanna, E. Denton, A. Smart, and J. Smith-Loud, "Towards a critical race methodology in algorithmic fairness," *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 501–512, 2020. 16, 17, 18, 129

[72] M. K. Scheuerman, K. Wade, C. Lustig, and J. R. Brubaker, "How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis," *ACM on Human-Computer Interac.*, 2020. 16, 54

[73] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, "Saving face: Investigating the ethical concerns of facial recognition auditing," in *AAAI/ACM Conference on AI, Ethics, and Society*, 2020. 16, 20, 21, 55, 128

[74] S. Benthall and B. D. Haynes, "Racial categories in machine learning," in *Conference on Fairness, Accountability, and Transparency*, 2019. 16, 21

[75] W. V. Quine, "Three grades of modal involvment," in *International Congress of Philosophy*, 1953. 17, 56

[76] S. Müller-Wille, "Race and history: Comments from an epistemological point of view," *Science, Technology, & Human Values*, pp. 597–606, 2014. 17

[77] C. Linnaeus, *Systema Naturae*. Stockholm Holmiae (Laurentii Salvii), 1758. 17

[78] T. Zuberi, E. Bonilla-Silva, *et al.*, *White Logic, White Methods: Racism and Methodology*. Rowman & Littlefield Publishers, 2008. 17, 18

[79] K. G. Muhammad, *The Condemnation of Blackness*. Harvard University Press, 2019. 17, 21

[80] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, "Face recognition: Too bias, or not too bias?," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition Workshop*, 2020. 18, 50

[81] M. Miceli, M. Schuessler, and T. Yang, "Between subjectivity and imposition: Power dynamics in data annotation for computer vision," *Proceedings ACM on Human-Computer Interac.*, pp. 1–25, 2020. 18, 26

[82] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions Pattern Anal. Mach. Intell.*, pp. 1090–1104, 2000. 18

[83] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *IEEE International Conference on Auto. Face and Gesture Recognition*, 2006. 18, 34

[84] Z. Zhang, Y. Song, and H. Qi, "Age progress./regress. by conditional adversarial autoencoder," *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, 2017. 18, 19

[85] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother, "Iarpa janus benchmark - c: Face dataset and protocol," in *International Conference on Biometrics (ICB)*, pp. 158–165, IEEE, 2018. 18, 48

[86] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European Conference Computer Vision*, Springer, 2016. 18, 19, 35, 36, 54

[87] I. Hupont and C. Fernández, "Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition," in *IEEE International Conference on Auto. Face & Gesture Recognition*, pp. 1–7, 2019. 18, 91

[88] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014. 18, 36

[89] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 41.1–41.12, BMVA Press, 2015. 18, 114

[90] J. Muhammad, Y. Wang, C. Wang, K. Zhang, and Z. Sun, "Casia-face-africa: A large-scale african face image database," *IEEE Transactions on Inform. Forensics and Security*, pp. 3634–3646, 2021. 18, 36

[91] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana, "Sensitivenets: Learning agnostic representations with application to face images," *IEEE Transactions Pattern Anal. Mach. Intell.*, 2021. 18, 44

[92] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 4873–4882, 2016. 18, 36, 49

[93] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, *et al.*, "Iarpa janus benchmark-b face dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition Workshop*, pp. 90–98, 2017. 18

[94] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer, "Casual conversations: A dataset for measuring fairness in ai," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, 2021. 18, 23, 26, 38

[95] M. Wang, Y. Zhang, and W. Deng, "Meta balanced network for fair face recognition," *IEEE Transactions Pattern Anal. Mach. Intell.*, 2021. 18, 24, 45, 72

[96] O. Yongsheng, W. Xinyu, Q. Huihuan, and X. Yangsheng, "A real time race classification system," *Proceedings IEEE International Conference on Inform. Acquisition*, pp. 378–383, 2005. 18

[97] A. Greco, G. Percannella, M. Vento, and V. Vigilante, "Benchmarking deep network architectures for ethnicity recognition using a new large face dataset," *Machine Vision and Applications*, 2020. 19

[98] S. B. Belhaouari, A. N. M. Shamhan, and S. B. Belhaouari, "Fusion of deep learning and handcrafted features for intra-race recognition," *Advances In Natural And Applied Sciences*, pp. 76–83, 2020. 19

[99] M. A. Ahmed, R. D. Choudhury, and K. Kashyap, "Race estimation with deep networks," *J. King Saud Uni. - Computer and Info. Science*, 2020. 19

[100] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, pp. 64–73, 2016. 19

[101] Z. Khan and Y. Fu, "One label, one billion faces: Usage and consistency of racial categories in computer vision," in *Proceedings ACM Conference on Fairness, Accountability, and transparency*, pp. 587–597, 2021. 20

[102] O. Solon, "Facial recognition's 'dirty little secret': Millions of online photos scraped without consent," *NBC News*, 2019. 20

[103] B. Hepple, "The new single equality act in britain," *The Equal Rights Review*, 2010. 21, 55

[104] K. B. Maddox and J. M. Perry, "Racial appearance bias: Improving evidence-based policies to address racial disparities," *Policy Insights from the Behavioral and Brain Sciences*, 2018. 21, 28, 55

[105] K. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer, "Issues related to face recognition accuracy varying based on race and skin tone," *IEEE Transactions on Tech. and Soc.*, 2020. 21, 26, 49, 50

[106] M. Harris, J. G. Consorte, J. Lang, and B. Byrne, "Who are the whites?: Imposed census categories and the racial demography of brazil," *Social forces*, pp. 451–462, 1993. 21

[107] P. Oliver, "Race names," 2017. 22

[108] F. von Luschan, *Beiträge zur Völkerkunde Der Deutschen Schutzgebiete*. D. Reimer, 1897. 22

[109] T. B. Fitzpatrick, "Soleil et peau," *J Med Esthet*, pp. 33–34, 1975. 22

[110] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types i through vi," *Archives of dermatology*, pp. 869–871, 1988. 22, 24, 26, 57

[111] M. S. Sommers, J. D. Fargo, Y. Regueira, K. M. Brown, B. L. Beacham, A. R. Perfetti, J. S. Everett, and D. J. Margolis, "Are the fitzpatrick skin phototypes valid for cancer risk assessment in a racially and ethnically diverse sample of women?," *Ethnicity & Disease*, p. 505, 2019. 22

[112] L. C. Pichon, H. Landrine, I. Corral, Y. Hao, J. A. Mayer, and K. D. Hoerster, "Measuring skin cancer risk in african americans: is the fitzpatrick skin type classification scale culturally sensitive?," *Ethnicity & Disease*, pp. 174–179, 2010. 22

[113] J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Reliability and validity of image-based and self-reported skin phenotype metrics," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pp. 550–560, 2021. 23, 26

[114] K. Krishnapriya, G. Pangelinan, M. C. King, and K. W. Bowyer, "Analysis of manual and automated skin tone assignments," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 429–438, 2022. 23, 24, 26

[115] B. C. K. Ly, E. B. Dyer, J. L. Feig, A. L. Chien, and S. Del Bino, "Research techniques made simple: Cutaneous colorimetry: A reliable technique for objective skin color measurement," *Journal Investigative Dermatology*, pp. 3–12, 2020. 23

[116] D. Munidasa, G. Schlippe, and S. Abeyakirthi, *Measurements of Skin Colour*. Cham: Springer International Publishing, 2018. 23

[117] A. Chardon, I. Cretois, and C. Hourseau, "Skin colour typology and suntanning pathways," *International Journal cosmetic science*, pp. 191–208, 1991. 23, 26, 28

[118] Y. Wu, T. Tanaka, and M. Akimoto, "Utilization of ita and hue angle in the measurement of skin color on images," *Bioimages*, pp. 1–8, 2020. 24

[119] N. M. Kinyanjui, T. Odonga, C. Cintas, N. C. Codella, R. Panda, P. Sattigeri, and K. R. Varshney, "Estimating skin tone and effects on classification performance in dermatology datasets," *arXiv preprint arXiv:1910.13268*, 2019. 24

[120] E. Monk, "Skin tone research google," 2022. 24

[121] M. Omi and H. Winant, *Racial Formation in the United States*. Routledge, 2014. 24

[122] B. Gonzalez-Sobrino and D. R. Goss, "The mechanisms of racialization beyond the black/white binary," *Ethnic and Racial Studies*, pp. 505–510, 2019. 24

[123] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern recognition*, pp. 1106–1122, 2007. 24

[124] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2019. 25, 34

[125] S. Kolkur, D. Kalbande, P. Shimpi, C. Bapat, and J. Jatakia, "Human skin detection using rgb, hsv and ycbcr color models," in *Proceedings of the International Conference on Commun. and Sign. Process.*, pp. 324–332, 2016. 25

[126] V. Muthukumar, T. Pedapati, N. Ratha, P. Sattigeri, C.-W. Wu, B. Kingsbury, A. Kumar, S. Thomas, A. Mojsilovic, and K. R. Varshney, "Understanding unequal gender classification accuracy from face images," *arXiv preprint arXiv:1812.00099*, 2018. 26

[127] V. Muthukumar, "Color-theoretic experiments to understand unequal gender classification accuracy from face images," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition Workshop*, 2019. 26

[128] J. Guo, J. Tan, Y. Yang, H. Zhou, *et al.*, "Variation and signatures of selection on the human face," *Journal of Human Evolution*, pp. 143–152, 2014. 27

[129] P. Claes, H. Hill, and M. D. Shriver, "Toward dna-based facial composites: Preliminary results and validation," *Forensic Science International: Genetics*, pp. 208–216, 2014. 27

[130] N. Sesardic, "Race: a social destruction of a biological concept," *Biology & Philosophy*, 2010. 27, 56, 129

[131] S. Ousley, R. Jantz, and D. Freid, "Understanding race and human variation: Why forensic anthropologists are good at identifying race," *American Journal of Physical Anthropology*, 2009. 27, 56

[132] Z. Zhuang, D. Landsittel, S. Benson, R. Roberge, and R. Shaffer, "Facial anthropometric differences among gender, ethnicity, and age groups," *The Annals of Occupational Hygiene*, 2010. 27, 56, 57, 58, 81

[133] R. Hopman and A. Mcharek, "Facing the unknown suspect: forensic dna phenotyping and the oscillation between the individual and the collective," *BioSocieties*, pp. 438–462, 2020. 27

[134] A. M'charek, "Tentacular faces: Race and the return of the phenotype in forensic identification," *American anthropologist*, pp. 369–380, 2020. 28

[135] B. K. Rothman, *Genetic Maps and Human Imaginations: The Limits of Science in Understanding Who We Are*. WW Norton & Company, 1998. 28

[136] K. B. Maddox, "Perspectives on racial phenotypicality bias," *Personality and Social Psychology Review*, 2004. 28, 55

[137] A. Skinner and G. Nicolas, "Looking black or looking back? using phenotype and ancestry to make racial categorizations," *Journal Exper. Social Psychology*, 2015. 28, 55

[138] K. B. Kahn and P. G. Davies, "Differentially dangerous? phenotypic racial stereotypicality increases implicit bias among ingroup and outgroup members," *Group Processes & Intergroup Relations*, 2011. 28, 55

[139] S. A. Schendel, "Anthropometry of the head and face," *Plastic and Reconstructive Surgery*, p. 480, 1995. 28

[140] L. G. Farkas, M. J. Katic, and C. R. Forrest, "International anthropometric study of facial morphology in various ethnic groups/races," *Journal Craniofacial Surgery*, pp. 615–646, 2005. 28

[141] N. Ramanathan and R. Chellappa, "Modeling age progression in young faces," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 387–394, 2006. 28

[142] Y. Liu, K. L. Schmidt, J. F. Cohn, and S. Mitra, "Facial asymmetry quantification for expression invariant human identification," *Computer Vision and Image Understanding*, pp. 138–159, 2003. 28

[143] A. Porcheron, E. Mauger, *et al.*, "Facial contrast is a cross-cultural cue for perceiving age," *Frontiers in Psychology*, p. 1208, 2017. 28

[144] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *IEEE International Conference on Computer Vision Workshop*, pp. 252–257, 2015. 28

[145] Appen Limited, "Confidence to deploy ai with world-class training data," tech. rep., Appen, 2022. 28

[146] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal Machine Learning Research*, pp. 1755–1758, 2009. 28

[147] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. M. Moreno, J. Fierrez, and A. Kuijper, "A comprehensive study on face recognition biases beyond demographics," *IEEE Transactions on Technology and Society*, pp. 16–30, 2021. 28, 29

[148] S. Barocas, A. Guo, E. Kamar, J. Krones, M. R. Morris, J. W. Vaughan, W. D. Wadsworth, and H. Wallach, "Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 368–378, 2021. 29

[149] P. Komarinski, *Automated Fingerprint Identification Systems (AFIS)*. Elsevier, 2005. 33

[150] L. D. Introna and H. F. Nissenbaum, "Facial recognition technology: A survey of policy and implementation issues," tech. rep., Center for Catastrophe Preparedness and Response, New York University, 2009. 33

[151] W. H. Gravett, "Digital coloniser? china and artificial intelligence in africa," *Survival*, pp. 153–178, 2020. 33

[152] S.-L. Wee, "China uses dna to track its people, with the help of american expertise," *The New York Times*, p. 2019, 2019. 33

[153] A. Daly, "Algorithmic oppression with chinese characteristics: Ai against xinjiang's uyghurs," 2019. 33

[154] R. Van Noorden, "The ethical questions that haunt facial-recognition research," *Nature*, pp. 354–359, 2020. 33

[155] A. Birhane and V. U. Prabhu, "Large image datasets: A pyrrhic win for computer vision?," in *Proceedings IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1536–1546, IEEE, 2021. 33

[156] M. Kim, "Creative commons and copyright protection in the digital era," *Journal Computer-mediated Commun.*, pp. 187–209, 2007. 33

[157] B. D. I. Formats-Part, "Face image data," *ISO/IEC JTC1/SC37 N506, ISO/IEC IS 19794*, 2004. 34, 72

[158] J. Monnerat, S. Vaudenay, and M. Vuagnoux, "Machine-readable travel documents," tech. rep., Springer, 2007. 34, 72

[159] G. K. Wallace, "The JPEG still picture compression standard," *Commun. ACM*, 1991. 34, 72, 73, 74, 75

[160] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *IEEE Sign. Process. Magazine*, 2001. 34, 72, 73

[161] K. Vangara, M. King, V. Albiero, *et al.*, "Characterizing the variability in face recognition accuracy relative to race," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition Workshop*, 2019. 34, 91, 94

[162] S. Karahan, M. K. Yildirum, K. Kirtac, F. S. Rende, G. Butun, and H. K. Ekenel, "How image degradations affect deep cnn-based face recognition?," in *International Conference of the biometrics special interest group (BIOSIG)*, IEEE, 2016. 34

[163] M. Poyser, A. Atapour-Abarghouei, and T. P. Breckon, "On the impact of lossy image and video compression on the performance of dcnns," in *IEEE International Conference on Pattern Recognition*, 2021. 34, 73

[164] P. Majumdar, S. Mittal, R. Singh, and M. Vatsa, "Unravelling the effect of image distortions for biased prediction of pre-trained face recognition models," in *Proceedings IEEE/CVF International Conference on Computer Vision*, pp. 3786–3795, 2021. 34, 71, 73

[165] M. Knoche, M. Elkadeem, S. Hörmann, and G. Rigoll, "Octuplet loss: Make face recognition robust to image resolution," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, IEEE, 2023. 35

[166] Y. Roh, K. Lee, S. Whang, and C. Suh, "Sample selection for fair and robust training," *Advances in Neural Inform. Process. Sys.*, pp. 815–827, 2021. 35

[167] W. G. Cochran, *Sampling Techniques: 3d Ed.* Wiley, 1977. 35

[168] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 1521–1528, 2011. 36

[169] B. Zhang, S. Gu, B. Zhang, J. Bao, D. Chen, F. Wen, Y. Wang, and B. Guo, "Styleswin: Transformer-based gan for high-resolution image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11304–11314, 2022. 37

[170] D. Xu, S. Yuan, L. Zhang, and X. Wu, "Fairgan: Fairness-aware generative adversarial networks," in *IEEE International Conference on Big Data*, pp. 570–75, 2018. 37

[171] S. Tan, Y. Shen, and B. Zhou, "Improving the fairness of deep generative models without retraining," *arXiv preprint arXiv:2012.04842*, 2020. 37

[172] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition Workshop*, 2019. 37

[173] J. Ge, W. Deng, M. Wang, and J. Hu, "Fgan: Fan-shaped gan for racial transformation," in *IEEE International Joint Conference on Biometrics*, pp. 1–7, 2020. 37

[174] Y. Mroueh *et al.*, "Fair mixup: Fairness via interpolation," in *International Conference Learn. Represent.*, 2021. 37

[175] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, pp. 137–154, 2004. 38

[176] S. Minaee, P. Luo, Z. Lin, and K. Bowyer, "Going deeper into face detection: A survey," *arXiv:2103.14983*, 2021. 38

[177] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded cnns," *IEEE Sig. Process. Letters*, pp. 1499–1503, 2016. 38

[178] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 5203–5212, 2020. 38

[179] S. Yang, P. Luo, C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 5525–5533, 2016. 38

[180] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "Dsfd: Dual shot face detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 5060–5069, 2019. 38

[181] X. Tang, D. Du, Z. He, and J. Liu, "Pyramidbox:a context-assisted single shot face detector," in *European Conference Computer Vision*, pp. 797–813, 2018. 38

[182] Y. He, D. Xu, L. Wu, M. Jian, S. Xiang, and C. Pan, "Lffd: A light and fast face detector for edge devices," in *arXiv:1904.10633*, 2019. 38

[183] S. Dooley, T. Goldstein, and J. P. Dickerson, "Robustness disparities in commercial face detection," *arXiv preprint arXiv:2108.12508*, 2021. 39

[184] Y. Zhu, H. Cai, S. Zhang, C. Wang, and Y. Xiong, "Tinaface: Strong but simple baseline for face detection," *arXiv preprint arXiv:2011.13183*, 2020. 39

[185] D. Qi, W. Tan, Q. Yao, and J. Liu, "Yolo5face: Why reinventing a face detector," in *Computer Vision ECCV 2022 Workshops: Tel Aviv, Israel, October 2327, 2022, Proceedings, Part V*, (Berlin, Heidelberg), p. 228244, Springer-Verlag, 2023. 39

[186] Y. Liu, F. Wang, J. Deng, Z. Zhou, B. Sun, and H. Li, "Mogface: Towards a deeper appreciation on face detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 4093–4102, 2022. 39

[187] S. Dooley, G. Z. Wei, T. Goldstein, and J. P. Dickerson, "Are commercial face detection models as biased as academic models?," *arXiv preprint arXiv:2201.10047*, 2022. 39

[188] S. Mittal, K. Thakral, P. Majumdar, M. Vatsa, and R. Singh, "Are face detection models biased?," in *IEEE International Conference on Auto. Face and Gesture Recognition*, pp. 1–7, 2023. 39

[189] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal Cognitive Neuroscience*, pp. 71–86, 1991. 40, 42

[190] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image Process.*, pp. 467–476, 2002. 40

[191] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions Pattern Anal. Mach. Intell.*, pp. 2037–2041, 2006. 40

[192] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 2707–2714, 2010. 40

[193] Z. Lei, M. Pietikäinen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Transactions Pattern Anal. Mach. Intell.*, pp. 289–302, 2013. 40

[194] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 41

[195] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 815–823, 2015. 41, 47

[196] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, pp. 84–90, 2017. 41

[197] L. He, Z. Wang, Y. Li, and S. Wang, "Softmax dissection: Towards understanding intra-and inter-class objective for embedding learning," in *Proceedings of the AAAI conference on artificial intelligence*, pp. 10957–10964, 2020. 41

[198] H. Bertrand, *Hyper-parameter optimization in deep learning and transfer learning: applications to medical imaging*. PhD thesis, Université Paris-Saclay, 2019. 42

[199] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Equity and access in algorithms, mechanisms, and optimization*, pp. 1–9, 2021. 42, 49

[200] H. Qin, "Asymmetric rejection loss for fairer face recognition," *arXiv preprint arXiv:2002.03276*, 2020. 43, 45, 46, 100

[201] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei, "Mis-classified vector guided softmax loss for face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12241–12248, 2020. 43

[202] S. Gong, X. Liu, and A. K. Jain, "Jointly de-biasing face recognition and demographic attribute estimation," in *European Conference Computer Vision*, Springer-Verlag, 2020. 43, 44, 91

[203] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: adaptive curriculum learning loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 5901–5910, 2020. 43

[204] Z. Yang, X. Zhu, C. Jiang, W. Liu, and L. Shen, "Ramface: Race adaptive margin based face recognition for racial bias mitigation," in *IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, 2021. 43, 45

[205] S. Gong, X. Liu, and A. K. Jain, "Mitigating face recognition bias via group adaptive classifier," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, 2021. 43, 72

[206] X. Xu, Y. Huang, P. Shen, S. Li, J. Li, F. Huang, Y. Li, and Z. Cui, "Consistent instance false positive improves fairness in face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 578–586, 2021. 43

[207] I. Serna, A. Morales, J. Fierrez, and N. Obradovich, "Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning," *Artificial Intelligence*, 2022. 43, 72

[208] E. Tartaglione, C. A. Barbano, and M. Grangetto, "End: Entangling and disentangling deep representations for bias correction," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 13503–13512, 2021. 44, 72

[209] E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel, "Flexibly fair representation learning by disentanglement," in *International Conference on Machine Learning*, PMLR, 2019. 44, 106

[210] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning not to learn: Training dnns with biased data," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, (Los Alamitos, CA, USA), pp. 9004–9012, IEEE Computer Society, 2019. 44, 91, 94

[211] J. P. Robinson, C. Qin, Y. Henon, S. Timoner, and Y. Fu, "Balancing biases and preserving privacy on balanced faces in the wild," *IEEE Transactions on Image Processing*, 2023. 44

[212] R. Ragonesi, R. Volpi, J. Cavazza, and V. Murino, "Learning unbiased representations via mutual information backpropagation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2729–2738, 2021. 44

[213] P. Dhar, J. Gleason, A. Roy, C. D. Castillo, and R. Chellappa, "Pass: Protected attribute suppression system for mitigating bias in face recognition," in *International Conference Computer Vision*, pp. 15087–15096, 2021. 44, 50

[214] S. Jung, D. Lee, T. Park, and T. Moon, "Fair feature distillation for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, (Los Alamitos, CA, USA), IEEE Computer Society, 2021. 44

[215] P. Dhar, J. Gleason, A. Roy, C. D. Castillo, P. J. Phillips, and R. Chellappa, "Distill and de-bias: Mitigating bias in face recognition using knowledge distillation," *CoRR*, 2021. 44

[216] B. Liu, S. Zhang, G. Song, H. You, and Y. Liu, "Rectifying the data bias in knowledge distillation," in *International Conference Computer Vision Workshop*, 2021. 44, 45

[217] H. J. Ryu, H. Adam, and M. Mitchell, "Inclusivefacenet: Improving face attribute detection with race and gender diversity," *arXiv preprint arXiv:1712.00193*, 2017. 45

[218] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Feature transfer learning for face recognition with under-represented data," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 5704–5713, 2019. 47

[219] J. Guo, X. Zhu, C. Zhao, D. Cao, Z. Lei, and S. Z. Li, "Learning meta face recognition in unseen domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 6163–6172, 2020. 47

[220] M. Faraki, X. Yu, Y.-H. Tsai, Y. Suh, and M. Chandraker, "Cross-domain similarity learning for face recognition in unseen domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 15292–15301, 2021. 47

[221] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo, "Reducing domain gap by reducing style bias," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 8686–8695, IEEE Computer Society, 2021. 47

[222] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of lfw benchmark or not?," *arXiv preprint arXiv:1501.04690*, 2015. 48

[223] P. Grother and M. Ngan, "The ijb-a face identification challenge performance report," *NIST Report*, 2017. 49

[224] J. Liu, Z. Yu, H. Qin, Y. Wu, D. Liang, G. Zhao, and K. Xu, "Oneface: One threshold for all," in *European Conference Computer Vision*, pp. 545–561, Springer, 2022. 49

[225] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman, "Template adaptation for face verification and identification," *Image and Vision Computer*, pp. 35–48, 2018. 49

[226] P. Dhar, J. Gleason, H. Souri, C. D. Castillo, and R. Chellappa, "Towards gender-neutral face descriptors for mitigating bias in face recognition," *arXiv preprint arXiv:2006.07845*, 2020. 50

[227] P. Terhrst, M. L. Tran, N. Damer, F. Kirchbuchner, and A. Kuijper, "Comparison-level mitigation of ethnic bias in face recognition," in *International Workshop on Biometrics and Forensics*, pp. 1–6, 2020. 50

[228] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Post-comparison mitigation of demographic bias in face recognition using fair score normalization," *Pattern Recognition Letters*, pp. 332–338, 2020. 50

[229] T. Salvador, S. Cairns, V. Voleti, N. Marshall, and A. M. Oberman, "Faircal: Fairness calibration for face verification," in *International Conference Learn. Represent.*, 2022. 50

[230] J. Chong-Soon Lee, "Navigating the topology of race," *Stan. L. Rev.*, vol. 46, p. 747, 1993. 54

[231] W. D. Roth, "The multiple dimensions of race," *Ethnic and Racial Studies*, 2016. 55

[232] T. Alzahrani, W. Al-Nuaimy, and B. Al-Bander, "Integrated multi-model face shape and eye attributes identification for hair style and eyelashes recommendation," *Computation*, 2021. 57

[233] Y. Lee, E. Lee, and W. J. Park, "Anchor epicanthoplasty combined with outfold type double eyelidplasty for asians: Do we have to make an additional scar to correct the asian epicanthal fold?," *Plastic and Reconstructive Surgery*, 2000. 57

[234] L. B. Kumar, V. Jayaraman, P. Mathew, S. Ramasamy, and R. D. Austin, "Reliability of lip prints in personal identification: An inter-racial pilot study," *Journal of forensic dental sciences*, p. 178, 2016. 58

[235] R. De La Mettrie, D. Saint-Léger, G. Loussouarn, A. Garcel, C. Porter, and A. Langaney, "Shape variability and classification of human hair: a worldwide approach," *Human biology*, 2007. 58

[236] J. L. Rees, "Genetics of hair and skin color," *Annual Review of Genetics*, 2003. 58

[237] J. Hernandez-Ortega, J. Galbally, J. Fierrez, and L. Beslay, "Biometric quality: Review and application to face recognition with faceqnet," *arXiv preprint arXiv:2006.03298*, 2020. 71, 73

[238] M. Georgopoulos, J. Oldfield, M. A. Nicolaou, Y. Panagakis, and M. Pantic, "Mitigating demographic bias in facial datasets with style-based multi-attribute transfer," *International Journal Computer Vision*, 2021. 72, 106

[239] A. R. Joshi, X. Suau Cuadros, N. Sivakumar, L. Zappella, and N. Apostoloff, "Fair SA: Sensitivity analysis for fairness in face recognition," in *Proc, of The Algorithmic Fairness through the Lens of Causality and Robustness*, Proceedings of Machine Learning Research, pp. 40–58, PMLR, 2022. 72

[240] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. OToole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020. 72

[241] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *International Conference on Quality of Multimedia Experience*, IEEE, 2016. 73

[242] I. Vasiljevic, A. Chakrabarti, and G. Shakhnarovich, "Examining the impact of blur on recognition by convolutional networks," *arXiv preprint arXiv:1611.05760*, 2016. 73

[243] M. Koziarski and B. Cyganek, "Impact of low resolution on image recognition with deep neural networks: An experimental study," *International Journal Applied Mathematics and Computer Science*, 2018. 73

[244] S. Dodge and L. Karam, "A study and comparison of human and deep learning recognition performance under visual distortions," in *International Conference on Computer Communication and Networks*, IEEE, 2017. 73

[245] R. Torfason, F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, "Towards image understanding from deep compression without decoding," *arXiv preprint arXiv:1803.06131*, 2018. 73

[246] F. G. Zanjani, S. Zinger, B. Piepers, S. Mahmoudpour, P. Schelkens, and P. H. N. de With, "Impact of jpeg 2000 compression on deep convolutional neural networks for metastatic cancer detection in histopathological images," *Journal of Medical Imaging*, p. 027501, 2019. 73

[247] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image Process.*, 2012. 77

[248] P. Grother, P. Grother, M. Ngan, and K. Hanaoka, "Face recognition vendor test (frvt) part 1: Verification," tech. rep., National Institute of Standards and Technology, 2022. 80

[249] A. Ali-Gombe and E. Elyan, "Mfc-gan: class-imbalanced dataset classification using multiple fake class generative adversarial network," *Neurocomputing*, 2019. 91

[250] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014. 91

[251] X. Wang and H. Huang, "Approaching machine learning fairness through adversarial network," *arXiv preprint arXiv:1909.03013*, 2019. 91

[252] T. Hu and H. Qi, "See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification," *arXiv preprint arXiv:1901.09891*, 2019. 97

[253] F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem, "On the fairness of disentangled representations," *Advances in Neural Information Processing Systems*, 2019. 106

[254] B. Zhang, S. Gu, B. Zhang, J. Bao, D. Chen, F. Wen, Y. Wang, and B. Guo, "Styleswin: Transformer-based gan for high-resolution image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11304–11314, 2022. 107

[255] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong, "Disentangled and controllable face image generation via 3d imitative-contrastive learning," in *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, pp. 5154–5163, 2020. 107

[256] A. Shoshan, N. Bhonker, I. Kviatkovsky, and G. Medioni, "Gan-control: Explicitly controllable gans," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14083–14093, 2021. 107

[257] Y. Shen, P. Luo, J. Yan, X. Wang, and X. Tang, "Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, pp. 821–830, 2018. 107

[258] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable gan controls," *Advances in Neural Information Processing Systems*, pp. 9841–9850, 2020. 107

[259] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9243–9252, 2020. 107

[260] G. Balakrishnan, Y. Xiong, W. Xia, and P. Perona, "Towards causal benchmarking of bias in face analysis algorithms," in *Deep Learning-Based Face Analytics*, pp. 327–359, Springer, 2021. 107

[261] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt, "Stylerig: Rigging stylegan for 3d control over portrait images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6142–6151, 2020. 107

[262] A. Nickabadi, M. S. Fard, N. M. Farid, and N. Mohammadbagheri, "A comprehensive survey on semantic facial attribute editing using generative adversarial networks," *ArXiv*, 2022. 107

[263] S.-Z. Xu, H.-Z. Huang, F.-L. Zhang, and S.-H. Zhang, "Faceshapegene: A disentangled shape representation for flexible face image editing," *Graphics and Visual Computing*, p. 200023, 2021. 107

[264] Y.-H. Lee and S.-H. Lai, "Byeglassesgan: Identity preserving eyeglasses removal for face images," in *European Conference on Computer Vision*, pp. 243–258, Springer, 2020. 107, 108

[265] Q. Gu, G. Wang, M. T. Chiu, Y.-W. Tai, and C.-K. Tang, "Ladn: Local adversarial disentangling network for facial makeup and de-makeup," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 10481–10490, 2019. 107

[266] J. Wang, J. Zhang, Z. Lu, and S. Shan, "Dft-net: Disentanglement of face deformation and texture synthesis for expression editing," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3881–3885, IEEE, 2019. 107

[267] S. C. Medin, B. Egger, A. Cherian, Y. Wang, J. B. Tenenbaum, X. Liu, and T. K. Marks, "Most-gan: 3d morphable stylegan for disentangled face image manipulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1962–1971, 2022. 107

[268] V. Blanz and T. Vetter, "Face recognition based on fitting a 3d morphable model," *IEEE Transactions on pattern analysis and machine intelligence*, pp. 1063–1074, 2003. 107

[269] Y. Pang, Y. Zhang, W. Quan, Y. Fan, X. Cun, Y. Shan, and D.-m. Yan, "Dpe: Disentanglement of pose and expression for general video portrait editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2023. 107

[270] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, *et al.*, "3d morphable face modelspast, present, and future," *ACM Transactions on Graphics (ToG)*, pp. 1–38, 2020. 107

[271] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8188–8197, 2020. 108

[272] Z. He *et al.*, "Facial attribute editing by only changing what you want," *IEEE Transactions on Image Processing*, pp. 5464–5478, 2018. 108

[273] R. Abdal, Y. Qin, and P. Wonka, "Image2stylegan++: How to edit the embedded images?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8296–8305, 2020. 108

[274] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y.-L. Yang, "Hologan: Unsupervised learning of 3d representations from natural images," in *IEEE International Conference on Computer Vision*, 2019. 108

[275] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5799–5809, 2021. 108

[276] W. Xia and J.-H. Xue, "A survey on 3d-aware image synthesis," *arXiv preprint arXiv:2210.14267*, 2022. 108

[277] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE/CVF Conference on Computer Vision Pattern Recognition*, 2018. 111

[278] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Ieee, 2009. 112

[279] V. H. Maluleke, N. Thakkar, T. Brooks, E. Weber, T. Darrell, A. A. Efros, A. Kanazawa, and D. Guillory, "Studying bias in gans through the lens of race," in *European Conference on Computer Vision*, pp. 344–360, Springer, 2022. 115, 121, 130

[280] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, and C. C. Loy, "Celebv-hq: A large-scale video facial attributes dataset," in *European Conference on Computer Vision*, pp. 650–667, Springer, 2022. 115

[281] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 59–66, IEEE, 2018. 115

[282] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), Curran Associates, Inc., 2017. 116

[283] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. 120

[284] J. Brownlee, "Statistical significance tests for comparing machine learning algorithms," 2018. 127

[285] Y. Zhong and W. Deng, "Face transformer for recognition," *arXiv preprint arXiv:2103.14803*, 2021. 130

[286] Z. Sun and G. Tzimiropoulos, "Part-based face recognition with vision transformers," *arXiv preprint arXiv:2212.00057*, 2022. 130

[287] M. Kim, F. Liu, A. Jain, and X. Liu, "Dcface: Synthetic face generation with dual condition diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12715–12725, 2023. 130