Racial Bias within Face Recognition: A Survey

SEYMA YUCER, Durham University, UK

FURKAN TEKTAS, BuboAI, UK

NOURA AL MOUBAYED, Durham University, UK

TOBY P. BRECKON, Durham University, UK

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 future potential solutions. In support, this study provides an extensive taxonomic review of research on racial bias within face recognition exploring every aspect and stage of the face recognition processing pipeline. Firstly, we discuss the problem definition of racial bias, starting with race definition, grouping strategies, and the societal implications of using race or race-related groupings. Secondly, we divide the common face recognition processing pipeline into four stages: image acquisition, face localisation, face representation, face verification and identification, and review the relevant corresponding literature associated with each stage. The overall aim is to provide 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.

CCS Concepts: • **Computing methodologies**; • **Computer Vision**; • **Face Recognition**;

Additional Key Words and Phrases: Face recognition, racial bias, race, face detection, face verification, face identification

1 INTRODUCTION

Over several decades, the objective of developing face recognition systems has gathered significant pace across research, and industry alike [3, 35, 200]. Companies, nonprofits, and governments have deployed an increasing number of face recognition systems to make autonomous decisions for millions of users [90] across various application areas, such as within employment decisions, public security, criminal justice, law enforcement surveillance, airport passenger screening, and credit reporting [4, 97]. However, such wide-scale adoption within real-world scenarios heightens public concern about their potential for abuse and the adverse effect of face recognition may have on some individuals due to the presence of bias [45, 180]. 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 [53].

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 [116] 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, [67] showed how the use of face feature descriptors varies across participants from different racial groupings. For example, it 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 tone participants use less varied descriptors than darker skin tone participants [67]. Similar

to the *own-race bias*, the conversely named *other-race effect* is also studied by a series of studies in social psychology [5, 157] to establish social implications of biased face processing and feature selection of humans in erroneous jury decisions, eyewitness identification.

Accordingly, the first technological study [144] to explore 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 face 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 [86, 141]. Subsequently, the Gender Shades study [13] 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 has emerged to understand and mitigate racial bias within face recognition [96, 117, 216]. These efforts and associated evidence of bias have forced several commercial and academical research to withdraw products, algorithms, or datasets due to the differing forms of disparities, distortions or biases [17, 118, 173].

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 face recognition 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 global research base that could readily address these issues in real-world system deployments [47, 175, 201, 218]. Moreover, face 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 relation to its processing [25]. Although previous surveys on algorithmic bias and fairness in machine learning [34, 115, 140] and face recognition in computer vision and biometrics [90, 200] exist, many aspects remain under-studied in relation to the specifics of racial bias within face recognition.

On the other hand, face recognition is a fast emerging field of research and applications alike that spans multiple more traditional fields, including machine learning, biometrics, statistics, sociology, and psychology. Therefore, we commonly find that aspects of the problem definition, in addition to the race conceptualisation and race-related performance evaluation methodologies, need to be clarified and ideally standardised. Which stages, operations and decisions in face recognition are prone to bias, and how incorrect solutions to addressing the bias issue can cause additional areas of concern that need to be highlighted in order to maximise the effectiveness of future research in this area. In this survey, we take face recognition as the central concept of our review and aim to provide coverage of all the aspects of racial bias within each stage of the face recognition processing pipeline, with additional supporting material spanning fundamental concepts from related fields.

The primary purpose of this study is to both summarise the current state of the art and to give a comprehensive critical review of prior research on the topic of racial bias within face recognition. In addition, we aim 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 aim to 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
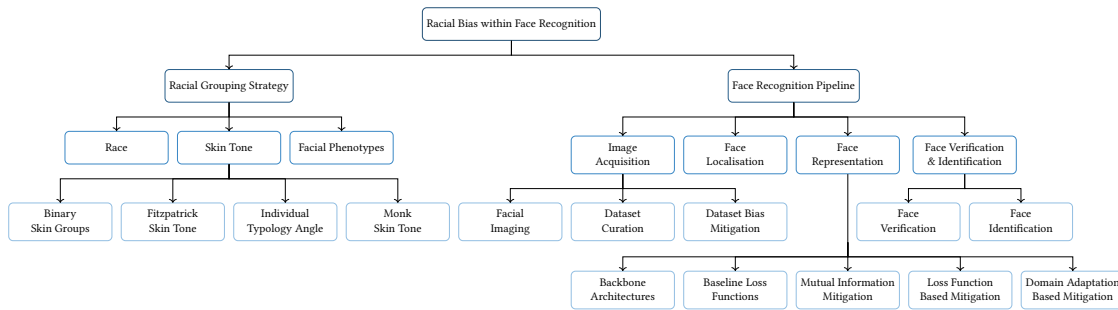
Fig. 1. Taxonomy of sections in Racial Bias within Face Recognition Survey.

recognition domain. In particular, the survey aims 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 1, right).

On this basis, we present this survey based on our taxonomy of prior work in the field and its contribution to the current state of the art (Figure 1). Subsequently, we formalise the problem definition with the corresponding evaluation and fairness criteria (Section 2). Next, we discuss standard race and race-related grouping terminology under three categories; race, skin tone and facial phenotypes (Section 3). This discussion provides an 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 4). Within this section (Section 4), 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 5).

2 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 [123] 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 [65]. In order to avoid the obvious potential for confusion,

the prior work of [7] 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 [7, 36, 95]. 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 and is smaller than 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 [28, 106, 197, 198] 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:

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

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 jth column of the weights, b_j is the jth 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,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

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:

$$\text{TMR} = \frac{TP}{TP + FN}, \quad \text{TNMR} = \frac{TN}{TP + FN}, \quad \text{FMR} = \frac{FP}{FP + TN}, \quad \text{FNMR} = \frac{FN}{FP + TN} \tag{3}$$

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 [95] 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 face 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 [36]. 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, [43] 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 [43] 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 $P(\hat{Y}|S = s) = P(\hat{Y}|S = s), s \in \{0, 1, ..r\}$ 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)* [36], 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} [59].

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 [59], 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 [206] 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, [59] 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 [95]. 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.

3 TOWARDS RACIAL GROUP FAIRNESS

Most studies on racial bias within face recognition, with a few exceptions [181, 210], 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 false match rate. 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 satisfy (*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 [9, 224]. As a result, the way race is conceptualised may perpetuate and reinforce existing forms of racial inequality [224]. Moreover, exposing or using such racial origin identifies the representation of a particular group and may lead to potential racial profiling and associated inequality [128]. 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 [13, 122]. Hanna [58] 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 [168], 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 [168]. In addition to the aforementioned points, many studies [10, 154, 168] 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 the racial bias (*group fairness criteria*) within face recognition. 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 categories and summarise recent literature along with face datasets by organising them under their grouping strategies in Table 1.

3.1 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 [151]. Race was first delineated by European naturalists and anthropologists to establish population-based research on human diversity [133]. In the seminal early scientific work of 1758, Systema Naturae [102], 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 [58, 224, 225]. 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 [224]. However, such definitions and classifications were adopted by the national census infrastructure across many jurisdictions [58]. The work of Khalid Muhammad [132] reveals how anecdotal, hereditarian and pseudo-biological race theories transformed into statistics and social surveys. Furthermore, Zuberi [224] 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 [224].

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 [53, 160] without any reference to the underlying critical theory of such categorisation and how they are defined [58, 224, 225]. 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 [120].

Previously, racial categories made an initial appearance within automated facial analysis via the task of race classification. For example, [215] propose feature extraction-based techniques for race classification using the MORPH [158], and FERET datasets [145] to predict *Caucasian, South Asian, East Asian, and African* racial classification. Later studies [158] extend the MORPH dataset for face recognition and analysis tasks (identification, 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 [1, 8, 52]. The work of [52] proposes the

Dataset Name	Year	Grouping Categories	Images	Source
Race				
ColorFERET [145]	1993	White, Asian, Black, Others	14K	Participants' photographs
MORPH [158]	2006	Caucasian, Hispanic, Asian, or African American	55K	Public Records
UTK Face [220]	2017	Asian, Black, Indian, White and Others (Hispanic, Latino, Middle Eastern)	20K	MORPH, CACD, online resources
IJB-C [113]	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 [201]	2019	African, Asian, Caucasian, Indian	45K	MS-Celeb [57]
DemogPairs [72]	2019	Asian, Black, White	10.8K	CWF, VGGFace1-2 [15, 143, 213]
BUPT-Balanced [199]	2020	African, Asian, Caucasian, Indian	1.3M	MS-Celeb [57]
VGGFace2 1200 [216]	2020	African, Asian, Caucasian, Indian	1M	VGGFace2 [15]
FairFace [96]	2021	Black, East Asian, Indian, Latino, Middle Eastern, Southeast Asian, and White	108K	Flickr, Twitter, newspapers, online resources
CASIA-Face-Africa [131]	2021	Hause (Sudan, Chad, Binin, Ivory Coast), Non-Hause	38K	Subjects from Nigeria
DiveFace [127]	2021	(Japan, China, Korea), (Europe, North America, and Latin America) (Sub-Saharan Africa, India, Bangladesh, Bhutan)	120K	MegaFace [80]
Skin Colour				
IJB-B [205]	2017	1-6 skin tones (increasing in darkness)	1K	1M FreeBase Celebrity List
PPB [13]	2018	Light, Dark skin tones (Fitzpatrick I-III,IV-VI)	68K	Gov. Official Profiles
Fair Face Challenge [175]	2020	Light, Dark skin tones (Fitzpatrick I-III,IV-VI)	152K	Flickr, Twitter, newspapers, online resources
Casual Conversations [61]	2021	Fitzpatrick Skin Tones	45K*	Vendor data
Globalface-8 [202]	2021	ITA base 8 skin tones (Tone I-VIII)	2M	1M FreeBase Celebrity List
Balancedface-8 [202]	2021	ITA base 8 skin tones (Tone I- VIII)	1.3M	1M FreeBase Celebrity List
IDS-8 [202]	2021	ITA base 8 skin tones (Tone I-VIII)	10K	1M FreeBase Celebrity List
Facial Phenotypes				
Diversity in Faces [119]	2019	ITA 6 skin tone, Craniofacial distance, area, ratio, Facial region contrast	0.97M	YFCC-100M
VGGFace2 [15] - [218]	2018	Fitzpatrick Skin Tones, Nose Shape, Eye Shape, Mouth Shape, Hair Type	3.3M	Google Image Search
RFW [201] -[218]	2019	Fitzpatrick Skin Tones, Nose Shape, Eye Shape, Mouth Shape, Hair Type	45K	MS-Celeb [57]

Table 1. Overview of most prominent face recognition datasets categorised by racial groupings, including dataset size and image sources.

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 1, the most commonplace face recognition datasets containing race labels [13, 201] 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 [13, 201].

Recently, the most commonly used face recognition evaluation dataset, a subset of MS-Celeb-1M [57] released as the RFW dataset [201], was constructed to measure relative face verification performance across four different racial groupings: *{African, Asian, Indian, Caucasian}*. FairFace [96] is another dataset, again drawn as a subset from the larger YFCC-100M Flickr dataset [189], 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 [220] 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 1, highlights the ambiguity and uncertainty behind the race concept upon which the 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 with social statistical science make effective performance evaluation and quantification very challenging within the racial bias problem space.

Similarly, Khan [81] identify four specific problems with the racial categories: (1) categories are not clearly defined and are often loosely associated with geographic origin, (2) 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) 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 [154] 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 [119], 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 [178]. *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 [153].

Finally, although many more studies discuss the possible negative consequences of using racial categories in face recognition datasets, Table 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 category forces racial bias literature 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
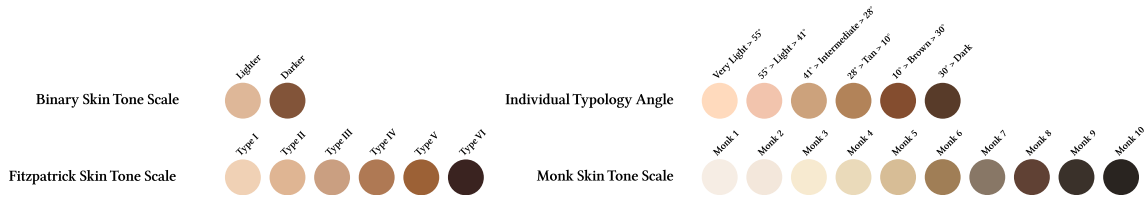
Fig. 2. Four different skin tone scales used for racial bias analysis within the context of face recognition.

racial bias. As information of racial or ethnic origin remains sensitive [66], from these observations across the face recognition field, we agree with the findings of several major studies [10, 92, 111, 122, 128, 132, 154, 224] 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 team in all cases [47].

3.2 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 [60] as skin tone variation caused by genetic differences (also exposure to the sun). In terms of its biological foundation, melanin is a group of natural pigments many organisms produce that is the fundamental driver for skin tone variance. The outermost layer of skin, the epidermis, contains melanin pigments, including red/yellow phaeomelanin and brown/black eumelanin [130]. Moreover, haemoglobin, beta carotene and bilirubin in the second layer of the skin, the dermis, absorb light and contribute to the overall pale/yellow/orange tint of human skin. It is primarily the quantity and type of melanin present in the skin that determines the skin tone. Genetic and hormonal factors control the production process of melanin, while direct exposure to ultraviolet radiation (UVR) accelerates its production [42] (hence the common phenomena of that a temporary darkening of skin tone is associated with exposure to sunlight - hence the expression *"sun tan"* where specifically the use of the term *tan* is indicative of a darkening in colour).

Over the past centuries, methods for categorising skin tone have evolved from verbal race-related descriptions (that would potentially be seen as derogatory today) with skin colour categories as *"white", "yellow", "black", "brown"*, and *"red"* [138], and colour-matching-based methods such as the Von Luschan scale [195]. The Von Luschan scale [195] uses 36 coloured glass tiles for skin colour comparison and was commonly used to racially categorise the population until the mid-20th century. Later, the Fitzpatrick Scale, established in 1975, became the most commonly used skin tone scale in dermatology and medicine. Accordingly, a plethora of work on racial bias within face recognition adopts the Fitzpatrick scale to measure and mitigate racial bias. However, Fitzpatrick's skin tone measurement was initially designed based on a subjective self-reporting or dermatology expert assessment which are often inconsistent and unreliable [69]. Subsequently, reflectance spectrophotometry and colourimetry methods [109] have now 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 [134]. Subsequently, Individual Typology Angle (ITA) [18] has been proposed to classify human skin colour using spectrophotometric measurements. More recently, a new extended skin colour scale, the Monk Skin Tone (MST) scale, was proposed with ten different tones to use in computer vision applications [125], whilst the global cosmetics

industry has traditionally used a separate set of scales [14]. The contrasting examples of these various skin tone scales are illustrated in Fig. 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 (see Section 4.1). 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, [88] 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 [120]. Specifically, [93] 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) 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) [69].

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 [92]. Accordingly, several studies suggest considering other race-related facial attributes, including lips, eye, and face shape when measuring racial bias in this context [135, 136] 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 [13, 61, 175] measure the racial bias in face recognition using either binary skin groupings, the Fitzpatrick Skin Types [41], or ITA [18] as depicted in Figure 2.

3.2.1 **Binary Skin Tone Scale.** The first research study that is based on the usage of binary skin/race groupings appeared in sociological research on race and race relations [139]. Focusing on white-black race relations in the United States brings expensive socio-economic data and analysis around such binary groupings [50]. 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 [76] 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 [13], 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 [175], which suggested researchers used a requantised (narrower) set of Fitzpatrick skin tone categories as per Gender Shades [13]. 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 [21] 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 [218].

3.2.2 **Fitzpatrick Skin Tone Scale***. 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 [40]. Later, he extended his scale to include a broader range of skin types (Type V and VI) [41] in order to offer a more granular representation across darker skin tones. The widespread adoption of this work within medical research studies [146, 179] subsequently influenced early computer vision research studies considering skin tone. Within the racial bias literature, the Gender Shades study [13] 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 [61, 175, 218]. Even recently, the extensive Casual Conversations Dataset [61] 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 [69]. 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 [93]. Consequently, [69] 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.

3.2.3 **Individual Typology Angle** *(ITA).* Another skin tone scale based on a spectrophotometric evaluation of skin colour is introduced by Chardon in 1991 [18]. 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) [93]. 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 [85, 207]. On the other hand, the utilisation of *ITA* scores and categories varies in the literature; Wang [202] 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 [119] 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 [93, 207].

3.2.4 **Monk Skin Tone (MST) Scale**. Most recently, the work of Ellis Monk [125] produced a new 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 [41], 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 [125].

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 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.

3.3 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 [55]. Such variation is most observable on faces as the face is identified as a *"biological billboard of our identity"* [19]. Subsequently, many studies [142, 172] focus on the impact of human phenotype characteristics (such as morphological attributes) on race. For example, the *Shades of Race* study [39] investigates the marginal effects of phenotypic characteristics, including skin tone, lips, nose, hair and body type on racial categorisation. Moreover, Zhuang [223] 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."*, [68]). 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 [114]. 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[163] (see Sec. 3.1).

Moreover, Maddox [110] 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 in the societal treatment and socioeconomic outcomes of disadvantaged groups [111]. 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 [75, 111, 176]. 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 [110].

Facial Coding	Description
Schema 1 [167]	Craniofacial Distances
Schema 2 [38]	Craniofacial Areas
Schema 3 [155]	Craniofacial Ratios
Schema 4 [107]	Facial Symmetry
Schema 5 [147]	Facial Regions Contrast
Schema 6 [18]	ITA-based Skin Tones
Schema 7 [162]	Age Prediction
Schema 8 [162]	Gender Prediction
Schema 9 [101]	Subjective Age & Gender Annotation
Schema 10 [84]	Pose and Resolution

(a) Summary of facial coding scheme analysis for the DiF dataset [119].

Phenotype Attribute	Categories
Skin Tone [41]	Type 1 / 2 / 3 / 4 / 5 / 6
Eyelid Type [98]	Monolid / Other
Nose Shape [223]	Wide / Narrow
Lip Shape	Full / Small
Hair Type [26]	Straight / Wavy / Curly / Bald
Hair Colour [169]	Red / Blonde / Brown / Black / Grey

(b) Facial phenotype attributes and their categorisation by [218].

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 [187] 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 [187]. 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 [6]. Subsequently, a plethora of work highlighting the shortcomings of race and skin tone-based categorisation (discussed in Sec. 3.1 and 3.2) push the current direction of research into phenotype-based categories [218]. One of the example studies, *Diversity in Faces* [119], 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 2a. 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 Sec. 3.1).

In parallel, [218] proposes a phenotype-based evaluation strategy for racial bias within face recognition. The study categorises representative racial characteristics on the face and explores the impact of each characteristic phenotype attribute: skin tone, eyelid type, nose shape, lips shape, hair colour and hair type. They annotate these attributes for two different publicly available face datasets: VGGFace2 (test set) [15], and RFW [201] (as presented in Table 2b). The study chooses to use Fitzpatrick Skin Types [41] for skin tones as it provides sufficient granularity, *{Type 1, Type 2, Type 3, Type 4, Type 5, Type 6}*, rather than binary skin-tone groupings, *{lighter skin-tone, darker skin-tone}*. For eye shape, [218] consider epicanthal folds and eyelid difference as a more distinctive attribute for racial bias [98]. The nose is categorised into two, wide and narrow, by examining the nasal breadth [223]. Hair texture is down-sampled from the

eight categories of [26] into three main hair texture types: straight, wavy, curly, in addition to bald. Additionally, [218] retains hair colour, as it is related to skin tone [156], with hair colour categories: red, grey, black, blonde, brown (see Table 2b).

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 a given subject's facial characteristics 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 Sec. 3.1) 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 3.1 & Sec 3.2), 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.

Overall, within this section we explore the race, skin tone and facial phenotype grouping strategies with regard to the *group fairness criteria* for racial bias within face recognition. We critically review current grouping strategies in face recognition datasets spanning race, skin tone and facial phenotypes. Whilst race remains a controversial concept that carries *historical bias*, ambiguity, ill-definition and disparity, a plethora of research identifies the possible risks and related problems of racial subject categorisation as a primary means for bias quantification within face recognition systems (Sec. 3.1). Alternatively, whilst skin colour has been utilised to both quantify and address racial bias, it remains only one trait of what is a comprehensive and multi-faceted race concept (Sec. 3.2). A broader approach, using facial phenotype as race-related facial attributes, provides a more objective and granular evaluation strategy for racial bias within face recognition (Sec. 3.3). However, whilst the overall aim is to achieve more accurate and fairer face recognition system performance across increasingly more diverse populations, we need to still ensure the race and related interpretations are not reduced to only facial phenotypes by ignoring the broader context of cultural, historical and social factors [58]. Moreover, the assessment of any grouping strategy on facial imagery creates another area of concern attributable to the often uncontrolled and inconsistent imaging conditions of facial capture that themselves lead to erroneous racial grouping annotation. As we move forward, we must address such risks, together with broader ethical considerations, within the wider development of face recognition processing pipelines.

4 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) [3, 90].

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 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 D_{train} and test D_{test} datasets for system training and evaluation (as defined in Sec. 2).

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, D_{train}, with reference to the minimisation of a loss function $\mathcal{L}$ that incites the use of a distinctive facial feature mapping (as defined in Sec. 2). Consequently, f^* provides mapping for both the curated training dataset, D_{train}, and unseen images in both test dataset, D_{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, D_{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 Sec. 2, this four stage conceptual face recognition processing pipeline is illustrated in Figure 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.
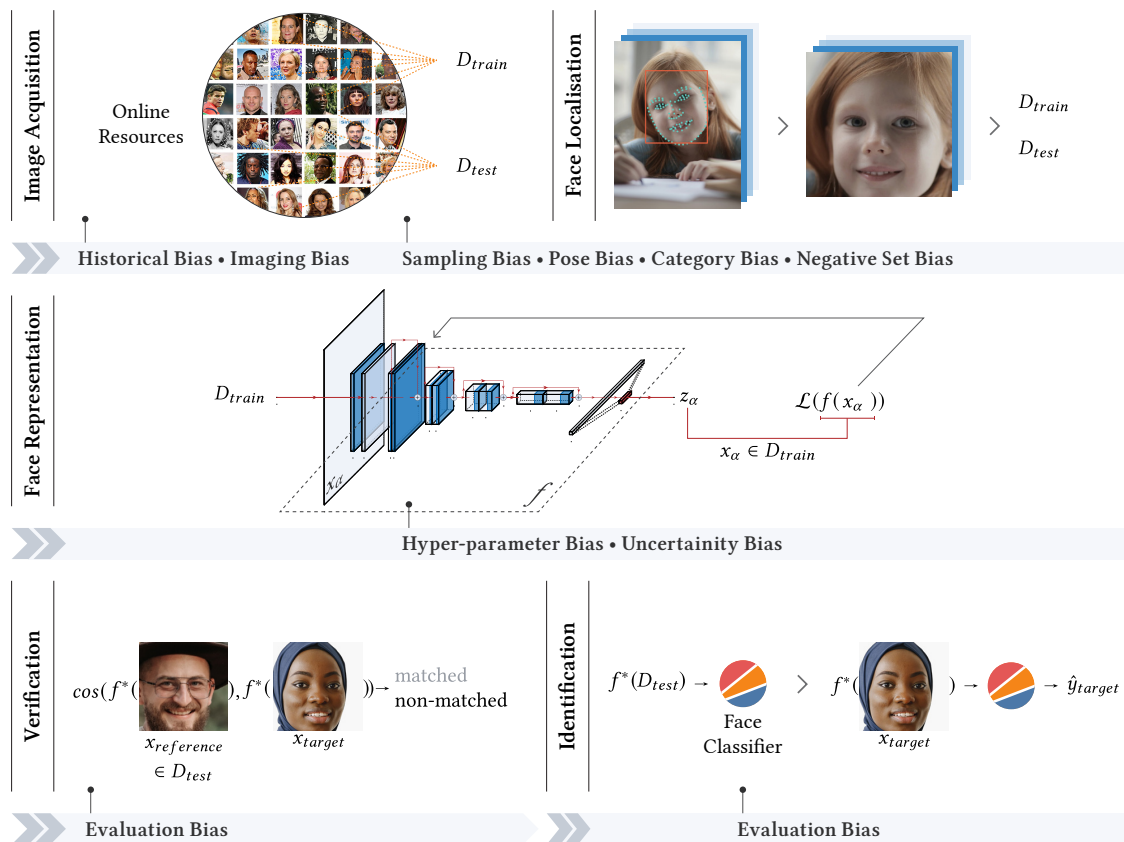
Fig. 3. Overview of the face recognition processing pipeline and bias attribution.

4.1 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*.

4.1.1 ***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 (e.g. fingerprint matching) for several decades [89]. 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 [73].

We have previously explored the *historical bias* and potential risks associated with racial categorisation and the annotation of facial images (Sec. 3). 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, [73] 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 [73]. For example, government use of such technology for racial profiling and racially-targeted restriction in some jurisdictions has been widely reported [51, 204] and investigated [24, 192]. In parallel to [73], Prabhu [12] 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 [83] 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 [12]. 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 [44] standard and ICAO 9303 guidelines [126] 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 [196] or JPEG2000 [177]; 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 [15, 57], do not conform to such requirements. Subsequently, Vangara [193] compares ICAO compliance between African and Caucasian groups in MORPH dataset [158] 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, [21] 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 [78, 112, 148, 217]. Recently, [112] 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 [112] 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, Yucer [217] finds that using lossy compressed facial test imagery decreases performance more significantly on specific phenotypes, including dark skin tone, wide nose, curly hair, and monolid eye when considered relative to a broader set of 21 phenotypic features. However, whilst the use of compressed imagery during training does make the resulting models more resilient and limits the performance degradation encountered, lower performance amongst these specific racially-aligned subgroups remains. Additionally, Yucer [217] find that removing chroma subsampling, which is itself a key lossy component of contemporary image compression schemes

FMR, improves face recognition performance for the specific phenotype categories which are otherwise more adversely affected by the use of lossy compression.

Consequently, we refer to these performance disparity effects within face recognition caused by variable imaging conditions as *imaging bias* as illustrated in Figure 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 [87] 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 [161].

4.1.2 **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 similar to selection, representation, or population bias) [20], 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 online 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 celebrities from the FreeBase listing) using public online image resources [57] (see Table 1, which then results in a dataset of millions of subjects who have/had public attention.

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 [65]. For instance, a DCNN-based face recognition model utilising certain features, such as hair colour, 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 [57] and VGGFace2 [15] - contain 10 million and 3.3 million face images respectively, and are curated from the FreeBase celebrity list as shown in Table 1. Similarly, the most common benchmark test sets for face recognition - LFW (Labeled Faces in the Wild) [70], CASIA-WebFace [213], and MegaFace [80] - are curated using online news (Yahoo), FreeBase celebrity and public online 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 [131], a large-scale African face image database or the BUPT-Balanced dataset [199], 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.

4.1.3 **Dataset Bias Mitigation***.* The prevailing assumption in machine learning is that the training set D_{train} and test set D_{test} are identically and independently distributed. However, this assumption is not valid for face recognition, and hence results in an out-of-distribution (OOD) problem. Accordingly, [190] relates the OOD problem to *"dataset bias"*. Although face recognition datasets should be representative of the real-world population in order to enable real-world face recognition model deployment, current state-of-the-art face recognition approaches remain closed-set methods, reflecting the world in a significantly biased way [190]. Subsequently, [190] groups dataset bias into four different types of bias:- 1) *selection bias* is similar to our aforementioned concept of *sampling bias*; 2) *capture bias* occurs when the dataset imagery contains targets (faces) that have minimal spatial and illumination variation and can be related with *pose bias* within face recognition context, as there is still poor pose variance (i.e. ±30° horizontal, ±15° vertical) within facial datasets; 3) *category or label bias* poses the ill-definition or mislabelling of subject identities and racial categories; 4) *negative set bias* defines bias against target appearances that are not represented in the data set (i.e. *"the rest of the world"* appearance) leading to recognition models that are overconfident and misrepresent performance by considering only a skewed subset of possible real-world data samples (i.e. the test dataset, D_{test}).

In order to mitigate dataset bias, many studies [77, 165] propose novel sampling methods by either down-sampling or upsampling (i.e. augmenting) the datasets in the early stage of the face recognition processing pipeline. With the latest advancements in Generative Adversarial Networks (GAN) [79, 183, 208], high-quality face image generation has become available as a potential tool to overcome the adverse effects of dataset distribution bias on subsequent real-world generalisation performance. For example, [91] addresses dataset bias within face recognition via the use of synthetic dataset augmentation. The study shows that deeper DCNN architectures generalise better to unseen facial poses, then trained using synthetically augmented datasets, and hence the impact of dataset bias can be reduced by 75%. Another work [216] aims to automatically construct a synthesised dataset by transforming facial images across varying racial domains while preserving identity-related features, such that racially dependent features subsequently become irrelevant within the determination of subject identity. Similar to [216], [46] transforms the facial images of one racial category to corresponding images of other racial categories in order to facilitate a more balanced racial category distribution via data augmentation. Moreover, [129] 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 racial groups during training. However, such generative models themselves produce samples from the underlying training set distribution upon which they are trained, meaning that they can also be impacted by dataset bias. Accordingly, [183] conduct 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 in the original GAN training data and that of the resultant distribution of the GAN output images meaning that any dataset bias present is only amplified in cases where GAN are used as a potential data augmentation strategy for face recognition.

Overall, this section outlines 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* as illustrated in Figure 3 against the corresponding stages of the face recognition pipeline.

4.2 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 [194]. Recently, face detection methods have shifted towards DCNN-based architectures and are categorised into five sub-genres by [121]: 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 [219] and Feature Pyramid Network-based RetinaFace [27], and the face detection benchmark dataset, Wider Face [211], have become widely adopted for face recognition.

The MTCNN face detector is based on a cascading multi-tasking structure [219] 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, [27] 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 [211], 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 [117], analysis the performances of five state-of-the-art face detectors; DSFD [100], Pyramid Box [184], LFD [64], RetinaFace [27], MTCNN [219] on demographic attributes including age, skin tone, gender. The study randomly samples the Casual Conversation Video Dataset [61] 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 groupings) 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 [32] investigates the robustness of three commercial online face detection capable systems: Amazon Rekognition, Microsoft Azure, and Google Cloud Platform 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 [222], YOLO5Face [149], MogFace [108]) with commercial ones [33]. They show that commercial approaches are always as biased or even more biased than non-commercial models, despite relatively larger development investment and supposed dedication to industry-level fairness commitments. More recently, [124] 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 4.1), 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.

4.3 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 [200]. 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 [191]. Following that, Gabor [104] and LBP [2] 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 [16, 99], 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 Sec.2.

4.3.1 ***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. $\sigma \in \{RELU, tanh, Softmax, ...\}$ 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 [174], the ResNet [62], and the Inception-ResNet [170]. VGG-Net [174] uses a smaller fixed number of convolutional filters compared to the AlexNet [94] to decrease the total number of trained parameters. On the other hand, ResNet [62] 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 [170] consists of multiple kernels in one layer to grasp salient features at different levels, including global and distributed features.

4.3.2 ***Baseline Loss Functions***. Contemporary, face recognition literature primarily focuses on designing novel DCNN loss functions [15, 28, 106, 198] to enhance the distinctiveness and separability of features. Mostly, such loss functions [28, 106, 198] operate on the feature embedding vectors of the last fully connected layer of the selected backbone DCNN architecture [62]. Previously, we discussed Softmax loss $\mathcal{L}_{softmax}$ (Eqn. 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 [63]. To address these problems, CosFace [198] 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.

$$\mathcal{L}_{cosface} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{\|z\|(\cos(\theta_{y_i,i})-m)}}{e^{\|z\|(\cos(\theta_{y_i,i})-m)} + \sum_{j \neq y_i}^{n} e^{\|z\|\cos(\theta_{j,i})}} \quad where \quad \cos(\theta_j, i) = W^T_j z_i \tag{4}$$

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 [28] differs from CosFace [198] 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:

$$\mathcal{L}_{arcface} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{\|z\|(\cos(\theta_{y_i,i}+m))}}{e^{\|z\|(\cos(\theta_{y_i,i}+m))} + \sum_{j \neq y_i}^{n} e^{\|z\|(\cos(\theta_{j,i}))}} \tag{5}$$

where all definitions are as per Eqn. 4. Overall the key Softmax, CosFace [198], and ArcFace [28] 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 , [191] (1991)) to above $\sim 99\%$ (ArcFace [28] (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 [123]. Moreover, [123] 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, [65] 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 [11]. 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 [182]. Another type of bias, denoted as *uncertainty bias*, is based on the probability values that are often computed 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. 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 (Sec. 4.3.3), loss function based mitigation (Sec. 4.3.4), and domain adaptation based mitigation (Sec. 4.3.5).

4.3.3 ***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

Methods	Backbone	Dataset	African	Asian	Caucasian	Indian	Avg	STD
Imbalanced Training Sets								
ArcFace [28]	ResNet-34	MegaFace	85.13	86.27	94.78	90.48	89.17	4.39
IMAN-A [201]	ResNet-34	MegaFace	91.42	91.15	94.78	94.15	92.88	1.86
ArcFace [28]	ResNet-34	VGGFace2	87.30	85.47	93.50	87.55	88.46	3.49
ARL+C [150]	ResNet-34	VGGFace2	88.57	87.65	93.48	89.35	89.76	2.57
ArcFace [28]	ResNet-50	BUPT-Global	96.28	96.03	98.22	96.77	96.83	0.98
MV-Softmax [203]	ResNet-50	BUPT-Global	95.83	95.66	99.33	95.83	96.66	1.78
DebFace-ID [48]	ResNet-50	BUPT-Global	93.67	94.33	95.95	94.78	94.68	0.96
CurricularFace [71]	ResNet-50	BUPT-Global	94.93	95.18	97.75	96.07	95.98	1.28
RamFace [212]	ResNet-50	BUPT-Global	96.73	96.17	98.28	96.77	96.99	0.90
ArcFace [28]	ResNet-101	VGGFace2	89.45	87.61	94.71	91.21	90.75	3.02
VGGF2 Races [216]	ResNet-101	VGGFace2	90.10	87.73	93.72	90.50	90.51	2.46
ArcFace [28]	ResNet-101	BUPT-Global	96.77	96.52	98.55	97.48	97.33	0.91
CurricularFace [71]	ResNet-101	BUPT-Global	96.30	95.98	97.83	96.70	96.70	0.81
RamFace [212]	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	1.38
CosFace [198]	ResNet-34	BUPT-Balanced	92.98	92.98	95.12	93.93	93.75	1.02
ArcFace [28]	ResNet-34	BUPT-Balanced	93.98	93.72	96.18	94.67	94.64	1.10
RL-RBN [199]	ResNet-34	BUPT-Balanced	95.00	94.82	96.27	94.68	95.19	0.73
RamFace [212]	ResNet-34	BUPT-Balanced	95.28	94.83	97.15	96.08	95.84	1.02
GAC-ArcFace [49]	ResNet-34	BUPT-Balanced	94.12	94.10	96.02	94.22	94.62	0.94
Fairness FR [209]	ResNet-34	BUPT-Balanced	95.95	95.17	96.78	96.38	96.07	0.69
ArcFace [28]	ResNet-50	BUPT-Balanced	96.00	95.45	97.57	96.42	96.36	0.90
CurricularFace [71]	ResNet-50	BUPT-Balanced	94.90	94.23	96.38	95.50	95.25	0.91
RamFace [212]	ResNet-50	BUPT-Balanced	96.25	95.50	97.40	96.58	96.43	0.79
GAC [49]	ResNet-50	BUPT-Balanced	94.65	94.93	96.23	95.12	95.23	0.69
Sensitive Loss [171]	ResNet-50	BUPT-Balanced	95.82	96.50	97.23	96.95	96.63	0.62
Fairness FR [209]	ResNet-50	BUPT-Balanced	96.47	95.75	97.08	96.77	96.52	0.57

Table 3. Performance of state-of-the-art face verification methods on the RFW dataset [201], with comparison based on sample standard deviation.

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 Sec. 2). A myriad of studies [22, 29, 48, 82, 127, 152, 159, 185] attempt to decrease this mutual information in order to debias the performance of face recognition approaches. For example, [185] 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 [185]). Moreover, [22] 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 [29, 48]. For example, the Protected Attribute Suppression System (PASS) [29] discourages the generator from encoding information related with sensitive attributes via discriminator. Furthermore, [82] uses a feature mapping network to unlearn biased sensitive attributes in order to disentangle the mutual information between identity and sensitive characteristics. Similarly,

[127] 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 [30, 74, 103]. Accordingly, [30] observes that the face recognition networks attend to different spatial regions in faces according to the category of an attribute 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, [103] 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 [164].

4.3.4 ***Loss Function Based Mitigation****.* Another area of study [150, 202, 212] focuses on setting adaptive margins to tackle racial bias. For previous face recognition baselines [28, 198], 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, [212] proposes *Race Adaptive Margin (RAM) Loss* using a new compact margin instead of using an ArcFace-style fixed margin, m (Eqn. 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, [212]) is;

$$\mathcal{L}_{ramface} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{\|z\|(\cos(\theta_{y_i,i}+m_r))}}{e^{\|z\|(\cos(\theta_{y_i,i}+m_r))}+\sum_{j \neq y_i}^{n} e^{\|z\|(\cos(\theta_{j,i}))}}$$
$$where \quad m_r = \beta \times Z_r \times \hat{\mu}^r_{intra}, \quad \mu^r_{intra} = \frac{1}{B_r} \sum_{y=1}^{B_r} \frac{1}{M_{y_j}} \sum_{i=1}^{M_{y_j}} \cos \theta^r_{z_i^{y_j},c_{y_j}}.$$
(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. 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. 3).

Another study, [150] 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:

$$\mathcal{L}_{arl} = \mathcal{L}_L + \lambda_U \times \mathcal{L}_U + \lambda_C \times \mathcal{L}_C \quad where \quad \mathcal{L}_C = \frac{\sum_{i,j} cos(z_i, z_j)^2}{N_t}, 0 < cos(z_i, z_j) < t \qquad (7)$$

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 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 performance on well-represented groups remains unaffected when compared to other state-of-the-art approaches (Tab. 3).

4.3.5 **Domain Adaptation Based Mitigation**. Following from the discussion of Section 4.1 on the out-of-distribution problem, domain adaptation techniques have recently been introduced as a method to address racial bias issues [37, 56, 137, 201, 214]. 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, [201] prove the domain gap between racial groupings and propose a deep information maximisation adaptation network (IMAN) architecture to address this. Subsequently, [56] 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, [37] introduces Cross-Domain Triplet (CDT) loss based on the triplet loss [170] 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 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. 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 [199] (Table 3).

4.4 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, Sec. 4.4.1) and Face Identification (i.e. one-to-many facial comparison, Sec. 4.4.2).

4.4.1 **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) and matching rates (see Eqn. 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 [70, 113] where the racial diversity within these datasets is often limited, biased and overlooked [221]. Accordingly, the Labelled Faces in-the-wild Dataset (LFW) [70] 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 [28, 198] has reached over 99.5 % verification accuracy on LFW.

4.4.2   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, [80] provides two large-scale face identification benchmark datasets under various imaging conditions.

Furthermore, [182] 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 [54, 70] 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 Sec. 4.1). 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 [105]. Studies have found that a single fixed threshold often causes higher variance across demographic groups than an adaptive threshold per-group threshold [105]. Another example, [23], 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, [92] 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 following study [201] released the RFW dataset with a similar protocol to LFW [70] with the same number (6000) of pairings for each of the four racial groups African, Asian, Caucasian, Indian with separate thresholds. Another study [218] annotates RFW for face verification and VGGFace2 for face identification with facial phenotype attributes to measure racial bias. Moreover, [31] 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 [29] extend this work with PASS to deal with any sensitive attribute and proposed a novel discriminator training strategy. Subsequently, [188] (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, [160], proposes a group-specific threshold (GST) in which the sensitive attributes themselves define its calibration sets. Another study, [186] proposes the Fair Score Normalisation (FSN) method, which is essentially GST with unsupervised clusters. FSN normalises the scores by requiring the model FMRs across unsupervised clusters to be the same predefined global FMR. Salvador, [166] 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, [92], 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. Furthermore, [218] perform closed-set identification on VGGFace2 test set and show that performance difference between facial phenotypes is much smaller when compared to the face verification results. However, the study is unable to have the same proportion for each attribute, and does not measure open-set face identification. 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 3.

5   CONCLUSIONS

We provide 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.

REFERENCES

[1] Mazida A. Ahmed, Ridip Dev Choudhury, and Kishore Kashyap. 2020. Race estimation with deep networks*J. King Saud Uni. - Comput. and Info. Sci.*

[2] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. 2006. Face description with local binary patterns: Application to face recognition*IEEE Trans. Pattern Anal. Mach. Intell.* 28, 12.

[3] Waqar Ali, Wenhong Tian, Salah Ud Din, Desire Iradukunda, and Abdullah Aman Khan. 2021. Classical and Modern Face Recognition Approaches: A Complete Review*Multimedia Tools and Applications.*

[4] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. 2016. OpenFace: A General-purpose Face Recognition Library with Mobile Applications*CMU School of Computer Science.*

[5] Gizelle Anzures, Paul C Quinn, Olivier Pascalis, Alan M Slater, James W Tanaka, and Kang Lee. 2013. Developmental Origins of the other-race effect*Current Directions in Psych. Sci..*

[6] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, Duncan Wadsworth, and Hanna Wallach. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. arXiv:2103.06076

[7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness and Machine Learning*Nips tutorial.*

[8] Samir Brahim Belhaouari, Athmar N M Shamhan, and Samir Brahim Belhaouari. 2020. Fusion of Deep Learning and Handcrafted Features for Intra-Race Recognition*Advances In Natural And Applied Sciences* 14.

[9] Ruha Benjamin. 2019. Race After Technology*Social forces.*

[10] Sebastian Benthall and Bruce D Haynes. 2019. Racial categories in machine learning. In *Conf. on Fairness, Accountability, and Transparency.*

[11] Hadrien Bertrand. 2019. *Hyper-parameter optimization in deep learning and transfer learning: applications to medical imaging.* Ph. D. Dissertation. Université Paris-Saclay.

[12] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A Pyrrhic Win for Computer Vision?. In *IEEE Winter Conf. on Appl. of Comput. Vis.* IEEE.

[13] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.

[14] L Caisey, F Grangeat, A Lemasson, J Talabot, and A Voirin. 2006. Skin color and makeup strategies of women from different ethnic groups*Int. J. Cosmetic Sci..*

[15] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. 2018. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *IEEE Int. Conf. on Auto. Face Gesture Recog..*

[16] Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. 2010. Face recognition with learning-based descriptor. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[17] Davide Castelvecchi. 2020. Is facial recognition too biased to be let loose?*Nature*.

[18] Alain Chardon, Isabelle Cretois, and Colette Hourseau. 1991. Skin colour typology and suntanning pathways*International J. cosmetic science*.

[19] Peter Claes, Harold Hill, and Mark D Shriver. 2014. Toward DNA-based facial composites: Preliminary results and validation*Forensic Sci. Int.: Genetics*.

[20] William Gemmell Cochran. 1977. *Sampling Techniques: 3d Ed.* Wiley.

[21] Cynthia M Cook, John J Howard, Yevgeniy B Sirotin, Jerry L Tipton, and Arun R Vemury. 2019. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems*IEEE Trans. on Biometrics, Behavior, and Identity Science*.

[22] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly fair representation learning by disentanglement. In *Int. Conf. on Machine Learning*. PMLR.

[23] Nate Crosswhite, Jeffrey Byrne, Chris Stauffer, Omkar Parkhi, Qiong Cao, and Andrew Zisserman. 2018. Template Adaptation for Face Verification and Identification*Image and Vis. Comput.*

[24] Angela Daly. 2019. Algorithmic oppression with Chinese characteristics: AI against Xinjiang's Uyghurs. APC.

[25] Antitza Dantcheva, Petros Elia, and Arun Ross. 2016. What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics*IEEE Trans. on Inform. Forensics and Security*.

[26] Roland De La Mettrie, Didier Saint-Léger, Genevievève Loussouarn, Annelise Garcel, Crystal Porter, and André Langaney. 2007. Shape variability and classification of human hair: a worldwide approach*Human biology*.

[27] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[28] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*

[29] Prithviraj Dhar, Joshua Gleason, Aniket Roy, Carlos D Castillo, and Rama Chellappa. 2021. PASS: Protected Attribute Suppression System for Mitigating Bias in Face Recognition. In *Int. Conf. Comput. Vis.*.

[30] Prithviraj Dhar, Joshua Gleason, Aniket Roy, Carlos Domingo Castillo, P. Jonathon Phillips, and Rama Chellappa. 2021. Distill and De-bias: Mitigating Bias in Face Recognition using Knowledge Distillation*CoRR*.

[31] Prithviraj Dhar, Joshua Gleason, Hossein Souri, Carlos Domingo Castillo, and Rama Chellappa. 2020. An adversarial learning algorithm for mitigating gender bias in face recognition.

[32] Samuel Dooley, Tom Goldstein, and John P Dickerson. 2021. Robustness disparities in commercial face detection. arXiv:2108.12508

[33] Samuel Dooley, George Z Wei, Tom Goldstein, and John P Dickerson. 2022. Are Commercial Face Detection Models as Biased as Academic Models? arXiv:2201.10047

[34] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. 2020. Demographic Bias in Biometrics: A Survey on an Emerging Challenge*IEEE Trans. on Technology and Society*.

[35] Hang Du, Hailin Shi, Dan Zeng, Xiao-Ping Zhang, and Tao Mei. 2022. The Elements of End-to-End Deep Face Recognition: A Survey of Recent Advances*Comput. Surveys*.

[36] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness*Innovations in Theoretical Comput. Sci. Conf.*.

[37] Masoud Faraki, Xiang Yu, Yi-Hsuan Tsai, Yumin Suh, and Manmohan Chandraker. 2021. Cross-Domain Similarity Learning for Face Recognition in Unseen Domains. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[38] Leslie G Farkas, Marko J Katic, and Christopher R Forrest. 2005. International anthropometric study of facial morphology in various ethnic groups/races*J. Craniofacial Surgery*.

[39] Cynthia Feliciano. 2016. Shades of Race: How Phenotype and Observer Characteristics Shape Racial Classification*Amer. Behavioral Sci.*.

[40] Thomas B Fitzpatrick. 1975. Soleil et peau*J Med Esthet* 2.

[41] Thomas B Fitzpatrick. 1988. The Validity and Practicality of Sun-Reactive Skin Types I Through VI*Archives of dermatology*.

[42] Thomas B Fitzpatrick, Makoto Seiji, and A David McGugan. 1961. Melanin pigmentation*New England J. Medicine*.

[43] Will Fleisher. 2021. What's Fair about Individual Fairness?. In *Proc. of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

[44] Biometric Data Interchange Formats-Part. 2004. Face image data*ISO/IEC JTC1/SC37 N506, ISO/IEC IS 19794*.

[45] Raul Vicente Garcia, Lukasz Wandzik, Louisa Grabner, and Joerg Krueger. 2019. The Harms of Demographic Bias in Deep Face Recognition Research. In *IEEE Int. Conf. on Biometrics*.

[46] Jiancheng Ge, Weihong Deng, Mei Wang, and Jiani Hu. 2020. FGAN: Fan-Shaped GAN for Racial Transformation. In *IEEE Int. Joint Conf. on Biometrics*.

[47] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets*Commun. of the ACM* 64, 12.

[48] Sixue Gong, Xiaoming Liu, and Anil K. Jain. 2020. Jointly De-Biasing Face Recognition and Demographic Attribute Estimation. In *Eur. Conf. Comput. Vis.* Springer-Verlag.

[49] Sixue Gong, Xiaoming Liu, and Anil K Jain. 2021. Mitigating Face Recognition Bias via Group Adaptive Classifier. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[50] Bianca Gonzalez-Sobrino and Devon R Goss. 2019. The Mechanisms of Racialization Beyond the Black/White Binary *Ethnic and Racial Studies*.

[51] Willem H Gravett. 2020. Digital coloniser? China and artificial intelligence in Africa *Survival* 62, 6.

[52] Antonio Greco, Gennaro Percannella, Mario Vento, and Vincenzo Vigilante. 2020. Benchmarking deep network architectures for ethnicity recognition using a new large face dataset *Machine Vision and Applications* 31.

[53] Patrick Grother, Patrick Grother, Mei Ngan, and Kayee Hanaoka. 2019. *Face Recognition Vendor Test (FRVT) Part 2: Identification*. Technical Report. National Institute of Standards and Technology.

[54] Patrick Grother and Mei Ngan. 2017. The IJB-A Face Identification Challenge Performance Report *NIST Report*.

[55] Jing Guo, Jingze Tan, Yajun Yang, Hang Zhou, et al. 2014. Variation and Signatures of Selection on the Human Face *J. Human Evolution*.

[56] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. 2020. Learning meta face recognition in unseen domains. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[57] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *Eur. Conf. Comput. Vis.* Springer.

[58] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness *FAT* 2020 - Proc. of the 2020 Conf. on Fairness, Accountability, and Transparency*.

[59] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning *Advances in Neural Information Process. Syst.*

[60] Marvin Harris, Josildeth Gomes Consorte, Joseph Lang, and Bryan Byrne. 1993. Who are the Whites?: Imposed Census Categories and the Racial Demography of Brazil *Social forces*.

[61] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. 2021. Casual Conversations: A Dataset for Measuring Fairness in AI. In *IEEE Conf. Comput. Vis. Pattern Recog*.

[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog*.

[63] Lanqing He, Zhongdao Wang, Yali Li, and Shengjin Wang. 2019. Softmax Dissection: Towards Understanding Intra- and Inter-Class Objective for Embedding Learning. arXiv:1908.01281

[64] Yonghao He, Dezhong Xu, Lifang Wu, Meng Jian, Shiming Xiang, and Chunhong Pan. 2019. LFFD: A Light and Fast Face Detector for Edge Devices. arXiv:1904.10633

[65] Thomas Hellström, Virginia Dignum, and Suna Bensch. 2020. Bias in Machine Learning–What is it Good for? arXiv:2004.00686

[66] Bob Hepple. 2010. The New Single Equality Act in Britain *The Equal Rights Review*.

[67] Peter J Hills and Michael B Lewis. 2006. Short Article: Reducing the own-race bias in Face Recognition by Shifting Attention *Quart. J. Exp. Psych.*.

[68] Roos Hopman and Amade M'charek. 2020. Facing the unknown suspect: forensic DNA phenotyping and the oscillation between the individual and the collective *BioSocieties*.

[69] John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. 2021. Reliability and Validity of Image-Based and Self-Reported Skin Phenotype Metrics *IEEE Trans. on Biometrics, Behavior, and Identity Sci.*.

[70] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Technical Report. University of Massachusetts, Amherst.

[71] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. 2020. Curricularface: adaptive curriculum learning loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[72] Isabelle Hupont and Carles Fernández. 2019. DemogPairs: Quantifying the Impact of Demographic Imbalance in Deep Face Recognition. In *IEEE Int. Conf. on Auto. Face & Gesture Recog.*.

[73] Lucas D. Introna and Helen Nissenbaum. 2010. *Facial Recognition Technology A Survey of Policy and Implementation Issues*. Technical Report.

[74] S. Jung, D. Lee, T. Park, and T. Moon. 2021. Fair Feature Distillation for Visual Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog*.

[75] Kimberly Barsamian Kahn and Paul G Davies. 2011. Differentially dangerous? Phenotypic racial stereotypicality increases implicit bias among ingroup and outgroup members *Group Processes & Intergroup Relations*.

[76] Praveen Kakumanu, Sokratis Makrogiannis, and Nikolaos Bourbakis. 2007. A Survey of Skin-color Modeling and Detection methods *Pattern recognition*.

[77] Faisal Kamiran and Toon Calders. 2010. Classification with no discrimination by preferential sampling. In *Mach. Learn. Conf. Belgium and The Netherlands*.

[78] Samil Karahan, Merve Kilinc Yildirum, Kadir Kirtac, Ferhat Sukru Rende, Gultekin Butun, and Hazim Kemal Ekenel. 2016. How Image Degradations Affect Deep CNN-based Face Recognition?. In *Int. Conf. of the biometrics special interest group (BIOSIG)*. IEEE.

[79] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[80] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. 2016. The MegaFace Benchmark: 1 Million Faces for Recognition at Scale. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[81] Zaid Khan and Yun Fu. 2021. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proc. ACM Conf. on Fairness, Accountability, and transparency*.

[82] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim. 2019. Learning Not to Learn: Training DNNs with Biased Data. In *IEEE Conf. Comput. Vis. Pattern Recog*.

[83] Minjeong Kim. 2007. Creative Commons and Copyright Protection in the Digital Era *J. Comput.-mediated Commun.*.

[84] Davis E. King. 2009. *Dlib-ml: A Machine Learning Toolkit*.

[85] Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. 2019. Estimating skin tone and effects on classification performance in dermatology datasets. arXiv:1910.13268

[86] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. 2012. Face Recognition Performance: Role of Demographic Information *IEEE Trans. on Inform. Forensics and Security*.

[87] Martin Knoche, Mohamed Elkadeem, Stefan Hörmann, and Gerhard Rigoll. 2022. Octuplet Loss: Make Face Recognition Robust to Image Resolution. arXiv:2207.06726

[88] S. Kolkur, D. Kalbande, P. Shimpi, C. Bapat, and J. Jatakia. 2016. Human Skin Detection Using RGB, HSV and YCbCr Color Models. In *Proc. of the Int. Conf. on Commun. and Sign. Process.*.

[89] Peter Komarinski. 2005. *Automated Fingerprint Identification Systems (AFIS)*. Elsevier.

[90] Yassin Kortli, Maher Jridi, Ayman Al Falou, and Mohamed Atri. 2020. Face Recognition Systems: A Survey *Sensors*.

[91] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. 2019. Analyzing and Reducing the Damage of Dataset Bias to Face Recognition with Synthetic Data. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh*.

[92] KS Krishnapriya, Vítor Albiero, Kushal Vangara, Michael C King, and Kevin W Bowyer. 2020. Issues Related to Face Recognition Accuracy Varying Based on Race and Skin Tone *IEEE Trans. on Tech. and Soc*.

[93] KS Krishnapriya, Michael C King, and Kevin W Bowyer. 2021. Analysis of Manual and Automated Skin Tone Assignments for Face Recognition Applications. arXiv:2104.14685

[94] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet Classification with Deep Convolutional Neural Networks *Commun. ACM*.

[95] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness *Advances in Neural Inform. Process. Syst*.

[96] Kimmo Kärkkäinen and Jungseock Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *IEEE Winter Conf. on Appl. of Comput. Vis.*.

[97] Ruggero Donida Labati, Angelo Genovese, Enrique Muñoz, Vincenzo Piuri, Fabio Scotti, and Gianluca Sforza. 2016. Biometric Recognition in Automated Border Control: A Survey *Comput. Surveys*.

[98] Yoonho Lee, Euitae Lee, and Won Jin Park. 2000. 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*.

[99] Zhen Lei, Matti Pietikäinen, and Stan Z Li. 2013. Learning discriminant face descriptor *IEEE Trans. Pattern Anal. Mach. Intell*. 36, 2.

[100] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. 2019. DSFD: Dual Shot Face Detector. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[101] Appen Limited. 2022. Confidence to deploy AI with world-class training data. https://www.figure-eight.com/

[102] Carolus Linnaeus. 1758. *Systema Naturae*. Stockholm Holmiae (Laurentii Salvii).

[103] Boxiao Liu, Shenghan Zhang, Guanglu Song, Haihang You, and Yu Liu. 2021. Rectifying the Data Bias in Knowledge Distillation. In *Int. Conf. Comput. Vis. Worksh*.

[104] Chengjun Liu and Harry Wechsler. 2002. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition *IEEE Trans. on Image Process*. 11, 4.

[105] Jiaheng Liu, Zhipeng Yu, Haoyu Qin, Yichao Wu, Ding Liang, Gangming Zhao, and Ke Xu. 2022. OneFace: One Threshold for All. In *Eur. Conf. Comput. Vis.* Springer.

[106] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep Hypersphere Embedding for Face Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog*.

[107] Yanxi Liu, Karen L Schmidt, Jeffrey F Cohn, and Sinjini Mitra. 2003. Facial asymmetry quantification for expression invariant human identification *Comput. Vis. and Image Understanding*.

[108] Yang Liu, Fei Wang, Jiankang Deng, Zhipeng Zhou, Baigui Sun, and Hao Li. 2022. MogFace: Towards a Deeper Appreciation on Face Detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[109] Bao Chau K Ly, Ethan B Dyer, Jessica L Feig, Anna L Chien, and Sandra Del Bino. 2020. Research Techniques Made Simple: Cutaneous Colorimetry: A Reliable Technique for Objective Skin Color Measurement *J. Investigative Dermatology*.

[110] Keith B Maddox. 2004. Perspectives on Racial Phenotypicality Bias *Personality and Social Psych. Review*.

[111] Keith B Maddox and Jennifer M Perry. 2018. Racial appearance bias: Improving evidence-based policies to address racial disparities *Policy Insights from the Behavioral and Brain Sciences*.

[112] Puspita Majumdar, Surbhi Mittal, Richa Singh, and Mayank Vatsa. 2021. Unravelling the Effect of Image Distortions for Biased Prediction of Pre-trained Face Recognition Models. In *Proc. IEEE/CVF Int. Conf. on Comput. Vis.*.

[113] Brianna Maze, Jocelyn Adams, James A. Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. 2018. IARPA Janus Benchmark - C: Face Dataset and Protocol. In *Int. Conf. on Biometrics (ICB)*. IEEE.

[114] Amade M'charek. 2020. Tentacular faces: Race and the return of the phenotype in forensic identification *American anthropologist*.

[115] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning *Comput. Surveys*.

[116] Christian A Meissner and John C Brigham. 2001. Thirty Years of Investigating the own-race bias in Memory for Faces: A Meta-Analytic Review. *Psych., Public Policy, and Law*.

[117] Hanna F Menezes, Arthur SC Ferreira, Eanes T Pereira, and Herman M Gomes. 2021. Bias and Fairness in Face Detection. In *IEEE Conf. on Graphics, Patterns and Images*.

[118] Joseph Menn. 2019. Microsoft turned down facial-recognition sales on human rights concerns. Reuters.

[119] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. 2019. Diversity in faces. arXiv:1901.10436

[120] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision*Proc. ACM on Human-Comput. Interac.*.

[121] Shervin Minaee, Ping Luo, Zhe Lin, and Kevin Bowyer. 2021. Going Deeper Into Face Detection: A Survey*arXiv:2103.14983*.

[122] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and Inclusion Metrics in Subset Selection. In *Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society*. ACM.

[123] Tom M Mitchell. 1980. *The Need for Biases in Learning Generalizations*. Dep. of Comput. Science, Laboratory for Comput. Science Research.

[124] Surbhi Mittal, Kartik Thakral, Puspita Majumdar, Mayank Vatsa, and Richa Singh. 2023. Are Face Detection Models Biased?. In *IEEE Int. Conf. on Auto. Face and Gesture Recog.*.

[125] Ellis Monk. 2022. *Skin tone research* Google.

[126] Jean Monnerat, Serge Vaudenay, and Martin Vuagnoux. 2007. *Machine-readable Travel Documents*. Technical Report. Springer.

[127] Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. 2021. SensitiveNets: Learning Agnostic Representations with Application to Face Images*IEEE Trans. Pattern Anal. Mach. Intell.*.

[128] Paul Mozur. 2019. One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority*The New York Times*.

[129] Youssef Mroueh et al. 2021. Fair Mixup: Fairness via Interpolation. In *Int. Conf. Learn. Represent.*.

[130] Michael P Muehlenbein. 2010. *Human Evolutionary Biology*. Cambridge University Press.

[131] Jawad Muhammad, Yunlong Wang, Caiyong Wang, Kunbo Zhang, and Zhenan Sun. 2021. CASIA-Face-Africa: A Large-scale African Face Image Database*IEEE Trans. on Inform. Forensics and Security*.

[132] Khalil Gibran Muhammad. 2019. *The Condemnation of Blackness*. Harvard University Press.

[133] Staffan Müller-Wille. 2014. Race and History: Comments from an epistemological point of view*Sci., Tech., & Human Values*.

[134] Deepani Munidasa, Gerrit Schlippe, and Sharnika Abeyakirthi. 2018. *Measurements of Skin Colour*. Springer Int. Publishing, Cham. 55–61 pages.

[135] Vidya Muthukumar. 2019. Color-Theoretic Experiments to Understand Unequal Gender Classification Accuracy From Face Images. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*.

[136] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R Varshney. 2018. Understanding Unequal Gender Classification Accuracy from Face Images. arXiv:1812.00099

[137] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo. 2021. Reducing Domain Gap by Reducing Style Bias. In *IEEE Conf. Comput. Vis. Pattern Recog*.

[138] Pamela E Oliver. 2017. Race Names.

[139] Michael Omi and Howard Winant. 2014. *Racial Formation in the United States*. Routledge.

[140] Kalia Orphanou, Jahna Otterbacher, Styliani Kleanthous, Khuyagbaatar Batsuren, Fausto Giunchiglia, Veronika Bogina, Avital Shulner Tal, Alan Hartman, and Tsvi Kuflik. 2022. Mitigating Bias in Algorithmic Systems-A Fish-Eye View*Comput. Surveys*.

[141] Alice J O'Toole, P Jonathon Phillips, Xiaobo An, and Joseph Dunlop. 2012. Demographic Effects on Estimates of Automatic Face Recognition Performance*Image and Vis. Comput.*.

[142] Stephen Ousley, Richard Jantz, and Donna Freid. 2009. Understanding Race and Human Variation: Why Forensic Anthropologists are Good at Identifying Race*Amer. J. of Physical Anthropology*.

[143] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition.

[144] P Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J O'Toole. 2011. An other-race effect for Face Recognition Algorithms*ACM Trans. on Appl. Perception*.

[145] P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. 2000. The FERET evaluation methodology for face-recognition algorithms*IEEE Trans. Pattern Anal. Mach. Intell.*.

[146] Latrice C Pichon, Hope Landrine, Irma Corral, Yongping Hao, Joni A Mayer, and Katherine D Hoerster. 2010. Measuring skin cancer risk in African Americans: is the Fitzpatrick Skin Type Classification Scale culturally sensitive?*Ethnicity & Disease*.

[147] Aurélie Porcheron, Emmanuelle Mauger, et al. 2017. Facial Contrast Is a Cross-Cultural Cue for Perceiving Age*Frontiers in Psych.*.

[148] Matt Poyser, Amir Atapour-Abarghouei, and Toby P Breckon. 2021. On the Impact of Lossy Image and Video Compression on the Performance of DCNNs. In *IEEE Int. Conf. on Pattern Recog*.

[149] Delong Qi, Weijun Tan, Qi Yao, and Jingfeng Liu. 2021. YOLO5Face: why reinventing a face detector. arXiv:2105.12931

[150] Haoyu Qin. 2020. Asymmetric Rejection Loss for Fairer Face Recognition. arXiv:2002.03276

[151] Willard V Quine. 1953. Three grades of modal involvment. In *Int. Congress of Philosophy*.

[152] Ruggero Ragonesi, Riccardo Volpi, Jacopo Cavazza, and Vittorio Murino. 2020. Learning unbiased representations via mutual information backpropagation. arXiv:2003.06430

[153] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proc. of the 2019 AAAI/ACM Conf. on AI, Ethics, and Society*.

[154] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *AAAI/ACM Conf. on AI, Ethics, and Society*.

[155] Narayanan Ramanathan and Rama Chellappa. 2006. Modeling Age Progression in Young Faces. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[156] Jonathan L Rees. 2003. Genetics of Hair and Skin color *Annual Review of Genetics*.

[157] Gillian Rhodes, Vance Locke, Louise Ewing, and Emma Evangelista. 2009. Race Coding and the other-race Effect in Face Recognition *Perception*.

[158] Karl Ricanek and Tamirat Tesafaye. 2006. MORPH: A Longitudinal Image Database of Normal Adult Age-Progression. In *IEEE Int. Conf. on Auto. Face and Gesture Recog.*

[159] Joseph Robinson, Yun Fu, Samson Timoner, Yann Henon, and Can Qin. 2021. Balancing Biases and Preserving Privacy on Balanced Faces in-the-wild.

[160] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. 2020. Face Recognition: Too Bias, or Not Too Bias?. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*

[161] Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. 2021. Sample Selection for Fair and Robust Training *Advances in Neural Inform. Process. Sys.*.

[162] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2015. DEX: Deep Expectation of Apparent Age from a Single Image. In *IEEE Int. Conf. on Comput. Vis. Worksh.*

[163] Barbara Katz Rothman. 1998. *Genetic Maps and Human Imaginations: The Limits of Science in Understanding Who We Are.* WW Norton & Company.

[164] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. 2017. InclusiveFaceNet: Improving Face Attribute Detection with Race and Gender Diversity. arXiv:1712.00193

[165] Payel Sadhukhan. 2019. Learning Minority Class prior to Minority Oversampling. In *Int. Joint Conf. on Neural Networks.* IEEE.

[166] Tiago Salvador, Stephanie Cairns, Vikram Voleti, Noah Marshall, and Adam M Oberman. 2022. FairCal: Fairness Calibration for Face Verification. In *Int. Conf. Learn. Represent.*

[167] Stephen A Schendel. 1995. Anthropometry of the Head and Face *Plastic and Reconstructive Surgery*.

[168] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis *ACM on Human-Comput. Interac.*

[169] Peter M Schneider, Barbara Prainsack, and Manfred Kayser. 2019. The Use of Forensic DNA Phenotyping in Predicting Appearance and Biogeographic Ancestry *Deutsches Ärzteblatt International*.

[170] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[171] Ignacio Serna, Aythami Morales, Julian Fierrez, and Nick Obradovich. 2022. Sensitive Loss: Improving Accuracy and Fairness of Face Representations with Discrimination-Aware Deep Learning *Artificial Intelligence*.

[172] Neven Sesardic. 2010. Race: a social destruction of a biological concept *Biology & Philosophy*.

[173] Stavros Shiaeles. 2021. Facebook will drop its facial recognition system - but heres why we should be sceptical *The Conversation*.

[174] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition *arXiv:1409.1556*.

[175] Tomáš Sixta, Julio Junior, CS Jacques, Pau Buch-Cardona, Eduard Vazquez, and Sergio Escalera. 2020. FairFace Challenge at ECCV 2020: Analyzing Bias in Face Recognition. arXiv:2009.07838

[176] Allison Skinner and Gandalf Nicolas. 2015. Looking Black or looking back? Using Phenotype and Ancestry to Make Racial Categorizations *J. Exper. Social Psych.*

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

[178] Olivia Solon. 2019. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent *NBC News*.

[179] Marilyn S Sommers, Jamison D Fargo, Yadira Regueira, Kathleen M Brown, Barbara L Beacham, Angela R Perfetti, Janine S Everett, and David J Margolis. 2019. Are the Fitzpatrick skin phototypes valid for cancer risk assessment in a racially and ethnically diverse sample of women? *Ethnicity & Disease*.

[180] Nisha Srinivas, Karl Ricanek, Dana Michalski, David S Bolme, and Michael King. 2019. Face Recognition Algorithm Bias: Performance differences on images of children and adults. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*

[181] Yufei Sun, Yong Li, and Zhen Cui. 2021. NFW: Towards National and Individual Fairness in Face Recognition. In *Pattern Recognition*.

[182] Harini Suresh and John V Guttag. 2019. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. arXiv:1901.10002

[183] Shuhan Tan, Yujun Shen, and Bolei Zhou. 2020. Improving the Fairness of Deep Generative Models without Retraining. arXiv:2012.04842

[184] Xu Tang, Daniel Du, Zeqiang He, and Jingtuo Liu. 2018. PyramidBox:A Context-assisted Single Shot Face Detector. In *Eur. Conf. Comput. Vis.*.

[185] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. 2021. EnD: Entangling and Disentangling deep representations for bias correction. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[186] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. 2020. Post-comparison mitigation of demographic bias in face recognition using fair score normalization *Pattern Recognition Letters*.

[187] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales, Julian Fierrez, and Arjan Kuijper. 2021. A Comprehensive Study on Face Recognition Biases Beyond Demographics. arXiv:2103.01592

[188] Philipp Terhörst, Mai Ly Tran, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. 2020. Comparison-Level Mitigation of Ethnic Bias in Face Recognition. In *Int. Worksh. on Biometrics and Forensics.* https://doi.org/10.1109/IWBF49977.2020.9107956

[189] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The New Data in Multimedia Research*Commun. ACM.*

[190] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[191] Matthew Turk and Alex Pentland. 1991. Eigenfaces for Recognition*J. Cognitive Neuroscience.*

[192] Richard Van Noorden. 2020. The ethical questions that haunt facial-recognition research*Nature* 587, 7834.

[193] Kushal Vangara, Michael King, Vitor Albiero, et al. 2019. Characterizing the Variability in Face Recognition Accuracy Relative to Race. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*

[194] Paul Viola and Michael J Jones. 2004. Robust Real-time Face Detection*Int. J. of Comput. Vis.*

[195] Felix von Luschan. 1897. *Beiträge zur Völkerkunde Der Deutschen Schutzgebiete.* D. Reimer.

[196] Gregory K. Wallace. 1991. The JPEG Still Picture Compression Standard*Commun. ACM.*

[197] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. 2018. Additive margin softmax for face verification*IEEE Sign. Process. Letters* 25, 7.

[198] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*

[199] Mei Wang and Weihong Deng. 2020. Mitigating Bias in Face Recognition Using Skewness-Aware Reinforcement Learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*

[200] Mei Wang and Weihong Deng. 2021. Deep Face Recognition: A Survey*Neurocomputing.*

[201] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. 2019. Racial Faces in-the-Wild: Reducing Racial Bias by Information Maximization Adaptation Network. In *Int. Conf. Comput. Vis.*

[202] Mei Wang, Yaobin Zhang, and Weihong Deng. 2021. Meta Balanced Network for Fair Face Recognition*IEEE Trans. Pattern Anal. Mach. Intell.*

[203] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. 2020. Mis-classified vector guided softmax loss for face recognition. In *Proc. of the AAAI Conf. on Artificial Intelligence*, Vol. 34.

[204] Sui-Lee Wee. 2019. China Uses DNA to Track Its People, With the Help of American Expertise*The New York Times* 21.

[205] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. 2017. IARPA Janus Benchmark-B Face Dataset. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*.

[206] Blake Woodworth, Suriya Gunasekar, Mesrob Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. In *Conf. on Learning Theory.*

[207] Yue Wu, Toshiyuki Tanaka, and Makio Akimoto. 2020. Utilization of ITA and Hue Angle in the Measurement of Skin Color on Images*Bioimages.*

[208] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. FairGAN: Fairness-aware Generative Adversarial Networks. In *IEEE Int. Conf. on Big Data.*

[209] Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. 2021. Consistent Instance False Positive Improves Fairness in Face Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[210] Songkai Xue, Mikhail Yurochkin, and Yuekai Sun. 2020. Auditing ML Models for Individual Bias and Unfairness. In *Int. Conf. on Artificial Intelligence and Statist.*.

[211] Shuo Yang, Ping Luo, Chen Loy, and Xiaoou Tang. 2016. WIDER FACE: A Face Detection Benchmark. In *IEEE Conf. Comput. Vis. Pattern Recog.*

[212] Zhanjia Yang, Xiangping Zhu, Changyuan Jiang, Wenshuang Liu, and Linlin Shen. 2021. RamFace: Race Adaptive Margin Based Face Recognition for Racial Bias Mitigation. In *IEEE Int. Joint Conf. on Biometrics (IJCB)*. IEEE.

[213] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Learning Face Representation from Scratch. arXiv:1411.7923

[214] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. 2019. Feature transfer learning for face recognition with under-represented data. In *IEEE Conf. Comput. Vis. Pattern Recog.*.

[215] Ou Yongsheng, Wu Xinyu, Qian Huihuan, and Xu Yangsheng. 2005. A real time race classification system*Proc. IEEE Int. Conf. on Inform. Acquisition.*

[216] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. 2020. Exploring Racial Bias within Face Recognition via per-subject Adversarially-Enabled Data Augmentation. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*

[217] Seyma Yucer, Matt Poyser, Noura Al-Moubayed, and Toby P Breckon. 2022. Does lossy image compression affect racial bias within face recognition?. In *IEEE Int. Joint Conf. on Biometrics.*

[218] Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby P. Breckon. 2022. Measuring Hidden Bias within Face Recognition via Racial Phenotypes. In *IEEE Winter Conf. on Applications of Comput. Vis.*.

[219] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment using Multi-task Cascaded CNNs*IEEE Sig. Process. Letters.*

[220] Zhifei Zhang, Yang Song, and H. Qi. 2017. Age Progress./Regress. by Conditional Adversarial Autoencoder*IEEE Conf. Comput. Vis. Pattern Recog.*

[221] Erjin Zhou, Zhimin Cao, and Qi Yin. 2015. Naive-deep face recognition: Touching the limit of LFW benchmark or not? arXiv:1501.04690

[222] Yanjia Zhu, Hongxiang Cai, Shuhan Zhang, Chenhao Wang, and Yichao Xiong. 2020. Tinaface: Strong but simple baseline for face detection. arXiv:2011.13183

[223] Ziqing Zhuang, Douglas Landsittel, Stacey Benson, Raymond Roberge, and Ronald Shaffer. 2010. Facial Anthropometric Differences among Gender, Ethnicity, and Age Groups*The Annals of Occupational Hygiene.*

[224] Tukufu Zuberi. 2001. *Thicker than Blood: How Racial Statistics Lie.* U of Minnesota Press.

[225] Tukufu Zuberi, Eduardo Bonilla-Silva, et al. 2008. *White Logic, White Methods: Racism and Methodology.* Rowman & Littlefield Publishers.