IDIAP RESEARCH REPORT

REVIEW OF DEMOGRAPHIC BIAS IN FACE RECOGNITION

Ketan Kotwal Sébastien Marcel

Idiap-RR-01-2025

FEBRUARY 2025

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Review of Demographic Bias in Face Recognition

Ketan Kotwal, *Senior Member, IEEE,* and Sébastien Marcel, *Fellow, IEEE*

***Abstract*—Demographic bias in face recognition (FR) has emerged as a critical area of research, given its impact on fairness, equity, and reliability across diverse applications. As FR technologies are increasingly deployed globally, disparities in performance across demographic groups– such as race, ethnicity, and gender– have garnered significant attention. These biases not only compromise the credibility of FR systems but also raise ethical concerns, especially when these technologies are employed in sensitive domains. This review consolidates extensive research efforts providing a comprehensive overview of the multifaceted aspects of demographic bias in FR.**

We systematically examine the primary causes, datasets, assessment metrics, and mitigation approaches associated with demographic disparities in FR. By categorizing key contributions in these areas, this work provides a structured approach to understanding and addressing the complexity of this issue. Finally, we highlight current advancements and identify emerging challenges that need further investigation. This article aims to provide researchers with a unified perspective on the state-of-the-art while emphasizing the critical need for equitable and trustworthy FR systems.

I. Introduction

Demographic bias in face recognition (FR) systems has emerged as a critical challenge in the deployment of biometric technologies for real-world applications [1]–[4]. Bias in these systems often leads to disparities in performance across demographic groups- such as variations in recognition accuracy- based on race, gender, and age [4]–[6]. Such biases can have far-reaching consequences, especially in critical applications like border crossing, law enforcement [7], [8], security [9], and hiring processes [10]–[12], where fairness and accuracy are paramount. Fig. 1 illustrates the issue of demographic bias based on race, gender, and age. Ideally, a fair model should exhibit equitable performance across all demographic groups (*i.e.* similar error rates or comparable score distributions). The disparity between groups contributes to the persistence of demographic bias. This issue is further compounded by the growing reliance on FR technologies, making it imperative to identify, evaluate, and mitigate sources of bias effectively.

Due to its severity and widespread range of applications, demographic bias has emerged as a crucial area of research, drawing significant attention from both the biometrics and computer vision communities [3], [4], [13], [14]. This issue has been formally incorporated into the evaluation frameworks of prominent initiatives, such as the National Institute of Standards and Technology's (NIST) Face Recognition Vendor Tests (FRVT), which have included demographic effects in their reports since 2019 [15], [16]. Similarly, the Maryland Test Facility (MdTF), supported by the United States Department of Homeland Security (DHS), has conducted biometric technology rallies to evaluate demographic disparities in FR systems[1]. In Europe, organizations like the European Association for Biometrics (EAB) have hosted dedicated events on demographic fairness in biometric systems, underlining the global importance of this topic[2]. This research area is often positioned within the broader context of fairness and trustworthy biometrics, and has been receiving substantial attention– in the form of papers, workshops, or special sessions– from leading conferences such as IEEE/CVF CVPR, WACV, IEEE FG, and ICPR, and from reputable journals including IEEE Transactions on information forensics and security (TIFS), IEEE Transactions on Biometrics, Behavior, and Identity Science (TBIOM) [17] and IEEE Signal Processing Magazine [18]. Additionally, standards organizations have recognized the need for systematic approaches, with ISO/IEC 19795-10 recently publishing guidelines for quantifying demographic differentials in biometric systems emphasizing the need for addressing bias in FR technologies [19].

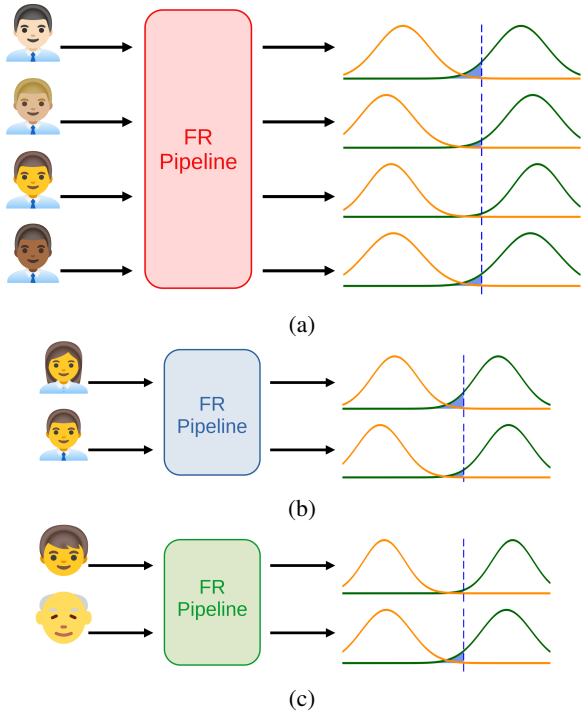


Fig. 1: Schematic illustration of issue of demographic bias in FR with reference to different demographic factors: (a) race or ethnicity, (b) gender, and (c) age.

Authors are with Idiap Research Institute, 1920-Switzerland. S Marcel is also with University of Lausanne, 1015-Switzerland.

[1] MdTF
[2] EAB event on Demographic Fairness in Biometric Systems

Scope: Fairness and bias in machine learning are expansive topics, and their application in biometrics has drawn significant attention in recent years. Several comprehensive reviews have addressed fairness and bias in machine learning broadly [5], [20], while others focus specifically on biometrics, offering insights into various modalities such as face, fingerprint, and vein, alongside applications beyond recognition, including region of interest (ROI) detection, quality assessment, and presentation attack detection [3], [4], [13], [21]. However, as face remains the most commonly used biometric trait, a substantial portion of research on demographic bias has concentrated on this modality.

While previous reviews offer broad perspectives, the extensive literature and emerging challenges specific to demographic bias in FR necessitate a dedicated review. In this article, we provide a consolidated discussion of recent advancements in the field, addressing the causes of demographic bias, available datasets for research, evaluation metrics for assessment of bias, and recent mitigation techniques. Furthermore, we explore ongoing challenges that persist in addressing demographic disparities, particularly in the light of novel use cases and emerging FR applications. Although our primary focus is on race and ethnicity, as these are dominant areas of research, we also include gender-related studies within the broader context. Age-related studies, given their distinct nature and established body of research, are referenced only where directly relevant.

Naming Conventions: In this review, we adopt the terminologies used in the referenced works while acknowledging the minor differences in naming conventions. Terms such as *race* and *ethnicity* are often used interchangeably, although they represent distinct concepts. For clarity, we retain the terms employed by the original studies. Similarly, the names of ethnic groups vary across the literature: for instance, some works use terms like Black and White, while others prefer African and Caucasian. Additionally, the South Asian group is sometimes referred to as Indian, whereas Asians often refer to East Asians. To maintain consistency and respect the source material, we adhere to the original terminologies in this review. Readers are encouraged to refer to the cited works for precise definitions and context. While fairness in related topics such as face detection, image quality, expression recognition, and attribute estimation is of interest to the research community, this review exclusively focuses on demographic bias in FR.

Contributions: This review constitutes the first comprehensive work dedicated to exploring demographic bias in FR, offering a unified and holistic perspective to researchers in the field. We systematically analyze and organize key aspects, including the causes of demographic bias, available datasets, assessment metrics, and mitigation techniques, providing a structured framework for understanding these areas. Finally, we identify emerging challenges and unresolved questions, inviting further research and innovation to advance fairness in FR systems.

The structure of this paper is as follows: Section II explores the causes of bias, analyzing factors such as distribution of demographic groups in dataset, skin-tone, image quality, and algorithmic sensitivities. We provide an overview of datasets commonly used for bias-related research in Section III, highlighting their demographic attributes and suitability for specific tasks. Section IV reviews existing metrics for bias evaluation, discussing their strengths and limitations. Section V outlines recent bias mitigation strategies across different stages of the FR pipeline. We discuss open challenges and future research directions in Section VI, and conclude the review in Section VII.

II. Causes of Demographic Bias

In this section, we consolidate findings from existing works related to causes of demographic bias in FR. Considering the wide range of research in this area, we have grouped the causes into categories for clarity and systematic review. These categories encompass factors such as imbalances in training datasets, variability in skin tones, algorithmic sensitivity, image quality and related covariates, as well as combined or intersectional demographic factors. While this categorization simplifies the organization, it is important to note that many studies attribute demographic bias to multiple, overlapping factors, making strict classification difficult. We have categorized works based on their primary focus or findings, but some studies have been referenced across multiple categories to reflect their broader relevance. This approach also highlights that causes of demographic bias are inherently multifaceted and interconnected, requiring a thorough understanding to analyze and mitigate these disparities. A quick summary of various works discussed in this section can be found in Table I.

A. Training Datasets

The issue of demographic bias in FR systems often stems from imbalanced or unrepresentative datasets, significantly influencing both training and evaluation outcomes. Research by Krishnapriya *et al.* [22] demonstrated how demographic groups, such as African-American cohorts, exhibit higher false match rates (FMR), while Caucasian cohorts face higher false non-match rates (FNMR), highlighting the interplay between ethnicity and matching thresholds. The Face Recognition Vendor Test (FRVT) conducted by NIST [15] substantiated these findings, reporting increased false positives in women, children, and the elderly, alongside higher false negatives in under-represented racial groups, emphasizing the intricate interactions between dataset characteristics and demographic attributes.

Early studies like Klare *et al.* [23] advocated for balanced datasets and the use of exclusive cohorts to enhance FR performance. Cavazos *et al.* [24] identified how dataset complexity and identification thresholds contribute to racial bias, such as the need for higher thresholds for East-Asian faces to achieve comparable false acceptance rates (FARs). Gwilliam *et al.* [25] challenged the prevailing assumptions about the necessity of balanced datasets by demonstrating that skewed distributions favoring African faces reduced racial bias more effectively than balanced datasets. Wu and Bowyer [26]

expanded this discussion, emphasizing that mere balance in identities or number of images is insufficient to address bias, highlighting additional factors like brightness and head pose during dataset assembly.

Other works delved into specific aspects of demographic balance in datasets. Wang *et al.* [27] observed that even race-balanced datasets failed to eliminate racial bias, hypothesizing that certain ethnicities are inherently more challenging to recognize. Kolla and Savadamuthu [28] highlighted the influence of facial quality and racial feature gradations on model fairness. Focusing on inter-sectional bias, Muthukumar *et al.* [29] identified structural facial features as significant contributors, particularly for dark-skinned females, over attributes like skin tone or hair length. Cook *et al.* [30] further analyzed the role of image acquisition conditions, noting how factors such as skin reflectance and environmental conditions disproportionately affect darker-skinned individuals, thus advocating for standardized acquisition protocols to mitigate bias.

Although most studies acknowledge that demographically imbalanced training data contribute to biased FR models and training with balanced datasets enhances fairness, there is a consensus that these are neither the sole causes nor complete solutions to the broader issue of demographic bias.

B. Variability in Skin-tone

The influence of skin tone on the performance of FR systems has been extensively studied, revealing significant demographic disparities. In [29], Muthukumar *et al.* identified notable under-performance in recognizing dark-skinned females compared to other demographic group for commercial classifiers. Their analysis attributed these disparities to structural features such as lips, eyes, and cheeks, in addition to skin-tone itself. Similarly, Buolamwini and Gebru [31] employed the Fitzpatrick skin classification system to evaluate gender classifiers and reported the lowest accuracy for darker-skinned females. Their findings further indicated that lighter-skinned males achieved the highest performance, highlighting the intersection of skin-tone and gender as critical factors influencing recognition accuracy. In another study, Krishnapriya *et al.* [22] examined FMR and FNMR across skin-tone groups, observing higher FMR for African-American cohorts and higher FNMR for Caucasians; however, they did not find a direct causation between darker skin tones and higher error rates.

The Biometric Technology Rallies organized by MdTF have offered comprehensive insights into the role of skin-related factors in FR bias. Their 2019 report [30] emphasized skin reflectance as more significant predictor of performance disparities than race. Using systematic linear modeling, their study demonstrated that darker skin-tones were associated with longer transaction (processing overall pipeline) times and lower accuracy in biometric systems. This dependency was found to vary substantially across systems, highlighting important role of acquisition methods in determining the extent of bias. Lu *et al.* [32] provided a quantitative assessment of performance variations across five skin-tone groups, identifying light-skinned individuals as the easiest to verify and darker-skinned individuals as the most challenging. However, ambiguities in defining skin tone categories complicate direct evaluations, highlighting the need for standardized classification metrics.

C. Algorithmic Factors

In this section, we examine the sensitivity and limitations of algorithms in addressing demographic attributes. Phillips *et al.* [33] identified the "other-race effect," where algorithms developed in Western and East Asian contexts demonstrated superior performance for their respective majority racial groups. This disparity persisted even when datasets were balanced, pointing to underlying biases in algorithmic design and training processes. Klare *et al.* [23] observed recognition challenges for specific demographic groups, including females, Black individuals, and younger cohorts. They reported improved performance when models were trained exclusively on these groups, emphasizing the impact of training data composition. In [34], Nagpal *et al.* demonstrated that deep learning models encode in-group biases, mirroring human tendencies such as own-race and own-age effects. By analyzing activation maps, they showed that these biases were ingrained within the feature representations of the models. Wang *et al.* [27], using the Racial Faces in-the-Wild (RFW) dataset, validated racial bias, revealing that error rates for African faces were nearly double those for Caucasians, even with race-balanced training data. Serna *et al.* [35] similarly highlighted significant performance gaps across demographic groups, advocating for diverse training datasets and fairness-aware algorithmic designs.

Further investigations into gender-based biases by Albiero *et al.* [36] revealed skewed impostor and genuine score distributions as the primary reasons for lower accuracy in women. This bias persisted across datasets, regardless of balanced training and neutral facial expressions. Ricanek *et al.* [37] noted unique challenges in recognizing children's faces due to structural changes with age, finding that algorithms effective on adult faces performed poorly for younger subjects.

D. Image Quality

The quality of input images and associated covariates significantly influence the manifestation of demographic bias in FR systems. Numerous studies have emphasized how disparities in image quality across different demographic groups can lead to variations in system performance. For instance, Cavazos *et al.* [24] analyzed both data-driven and scenario-based factors, revealing that dataset complexity and decision thresholds have a notable impact on recognition accuracy and racial bias. Their experiments across multiple algorithms further demonstrated that East Asian faces required higher decision thresholds compared to Caucasian faces to achieve equivalent error rates, highlighting the interplay between dataset characteristics and demographic attributes.

The study conducted by MdTF highlighted skin reflectance as a critical factor influencing both the accuracy and efficiency of FR systems [38], [39]. Analyzing 158 FR systems, they found that lower skin reflectance, typically associated with darker skin-tones, correlated with reduced accuracy and higher

Reference	Year	Dataset	Attribute	Summary
Krishnapriya *et al.* [22]	2020	MORPH	ET, ST	Demonstrated demographic disparities in FMR and FNMR, with African-American cohorts having higher FMR and Caucasian cohorts higher FNMR.
NIST FRVT [15]	2019	Private	ET, GN, AG, +	Reported increased false positives in women, children, and elderly, and higher false negatives in underrepresented groups.
Klare *et al.* [23]	2012	PCSO	ET, GN, AG	Advocated for balanced datasets and exclusive cohorts to improve face recognition performance.
Cavazos *et al.* [24]	2020	GBU	ET	Highlighted how dataset complexity and thresholds affect racial bias, requiring higher thresholds for East-Asian faces.
Gwilliam *et al.* [25]	2021	BUPT, RFW	ET	Showed that skewed distributions favoring African faces can mitigate racial bias better than balanced datasets.
Wu and Bowyer [26]	2023	DemogPairs, RFW, BFW,	ET, GN, +	Emphasized that balancing identities and images alone is insufficient, stressing brightness and head pose considerations.
Wang *et al.* [27]	2019	RFW	ET	Observed that race-balanced datasets do not fully eliminate bias, suggesting inherent challenges in recognizing certain ethnicities.
Kolla and Savadamuthu [28]	2023	RFW	ET	Highlighted the influence of facial quality and racial feature gradations on fairness in face recognition models.
Muthukumar *et al.* [29]	2018	PPB	GN, ST	Identified structural facial features as primary contributors to intersectional bias for dark-skinned females.
Cook *et al.* [30]	2019	Private	ET, GN, AG, +	Analyzed image acquisition conditions, noting the impact of skin reflectance and environmental factors on darker-skinned individuals.
Muthukumar *et al.* [29]	2018	PPB	GN, ST	Identified structural facial features and skin tone as key factors for dark-skinned females' underperformance.
Buolamwini and Gebru [31]	2018	IJB-A, Adience	GN, ST	Demonstrated lowest classifier performance for darker-skinned females using the Fitzpatrick system.
Krishnapriya *et al.* [22]	2020	MORPH	ET, ST	Examined FMR and FNMR across skin tones but found no direct causation between darker skin tone and higher errors.
Cook *et al.* [30]	2019	Private	ET, GN, AG, +	Highlighted skin reflectance as a major predictor of FR disparities and emphasized acquisition methods' role.
Lu *et al.* [32]	2019	IJB-B, IJB-C	AG, GN, ST, +	Quantified performance variations across skin tone groups, noting challenges with darker skin tones.
Phillips *et al.* [33]	2011	FRVT	ET	Identified the "other-race effect," where algorithms performed better on their respective majority racial groups.
Klare *et al.* [23]	2012	PCSO	ET, GN, AG	Highlighted recognition challenges for female, Black, and younger cohorts, improved with exclusive group training.
Nagpal *et al.* [34]	2019	MORPH, RFW, CACD, +	ET, AG	Showed that deep learning models encode in-group biases, mirroring own-race and own-age human biases.
Wang *et al.* [27]	2019	RFW	ET	Demonstrated racial bias using the RFW dataset, with higher error rates for African faces compared to Caucasians.
Serna *et al.* [35]	2019	DiveFace	ET, GN	Highlighted significant performance gaps across demographic groups, calling for fairness-aware algorithm designs.
Albiero *et al.* [36]	2020	AFD, MORPH, Notre Dame	GN	Found gender-based biases in score distributions, with lower accuracy for women across balanced datasets.
Ricanek *et al.* [37]	2015	ITWCC	AG	Observed challenges in recognizing children's faces due to structural changes with age, affecting algorithm accuracy.
Cavazos *et al.* [24]	2020	GBU	ET	Highlighted dataset complexity and decision thresholds' impact on racial bias and accuracy.
MdTF [38], [39]	2023	Private	ET, GN, AG, +	Found lower skin reflectance correlated with reduced accuracy and higher transaction times.
Wu *et al.* [40]	2023	MORPH	ET, GN, +	Demonstrated how brightness inconsistencies increase FMRs and diminish similarity scores.
Krishnapriya *et al.* [41]	2019	MORPH	ET	Showed improving image quality reduces performance gaps between African-American and Caucasian cohorts.
Albiero *et al.* [36]	2020	AFD, MORPH, Notre Dame	GN	Linked gender-based disparities to score distributions and identified confounding factors like cosmetics.
Lu *et al.* [32]	2019	IJB-B, IJB-C	AG, GN, ST, +	Analyzed multiple covariates; noted lighter skin tones consistently outperformed medium-dark tones.
Vera-Rodriguez *et al.* [42]	2019	VGGFace2	GN	Highlighted gender as a covariate, with males consistently outperforming females across demographics.
Ricanek *et al.* [37]	2015	ITWCC	AG	Discussed recognition challenges due to structural changes in children's facial features over time.
Best-Rowden *et al.* [43]	2017	LEO_LS, PCSO_LS	AG, GN	Found that males generally have higher genuine scores, but their performance declines faster with age.
Sarridis *et al.* [44]	2023	RFW	ET, GN, AG	Reported high mistreatment rates for African females over 60 years, highlighting compounded biases.
El Khiyari *et al.* [45]	2016	MORPH	ET, GN, AG	Observed lower face verification accuracy in younger individuals, females, and Black racial groups.
FRVT report [15]	2021	Private	ET, GN, AG, +	Noted elevated false positives for children and elderly, especially among Asian and American Indian groups.
Cook *et al.* [39]	2023	Private	ET, GN, AG, +	Demonstrated that age and skin lightness significantly influence recognition scores, compounded by illumination.

TABLE I: Summary of works delving into various causes of demographic bias in face recognition. As several works have identified multiple causes of bias, we have categorized the works based on their primary focus or inference. For details, readers are encouraged to refer to the source materials. The demographic factors of primary interest are denoted as ET: Ethnicity or race, GN: gender or sex, AG: age; whereas + indicates study of more attributes.

transaction times. These effects varied across systems, underscoring the role of image acquisition quality as a stronger predictor of performance as mentioned earlier. Similarly, Wu *et al.* [40] explored the effects of brightness and illumination, demonstrating that under-exposed or over-exposed images result in higher FMRs, while significant brightness differences between image pairs diminish similarity scores. They recommended controlled image acquisition processes to achieve consistent brightness across demographic groups, thereby reducing accuracy disparities.

Krishnapriya *et al.* [41] further examined how variations in image quality contributed to performance gaps between African-American and Caucasian cohorts. Enhancing image quality notably reduced these disparities, particularly by minimizing low-similarity errors within the genuine distribution. Following ICAO compliance guidelines, they evaluated biometric sample quality to support these findings. Albiero *et al.* [36] investigated gender-based disparities in FR systems, linking these to differences in genuine and imposter score distributions. They also identified confounding factors such as cosmetics and image pose. Despite using neutral and balanced datasets, their study revealed that such measures alone were insufficient to fully eliminate observed disparities.

Work by Lu *et al.* provided a detailed analysis of the influence of covariates on FR performance, incorporating variables such as skin tone, age, gender, pose, facial hair, and occlusion across three datasets and five FR systems [32]. Their findings highlighted that skin tone significantly affects verification accuracy, with lighter skin tones consistently outperforming medium-dark tones. However, they also emphasized the challenges posed by ambiguities in skin tone classification, advocating for more precise methodologies for performance assessments. In alignment with earlier studies, Lu *et al.* observed that male subjects generally achieved better recognition accuracy than female subjects. They attributed this disparity to factors such as occlusion caused by longer hair and alterations in facial appearance due to makeup. These observations corroborate prior findings indicating that facial makeup can negatively impact recognition accuracy [46], [47]. Collectively, these studies underscore that demographic bias in FR systems is intrinsically tied to image quality and related covariates, necessitating focused efforts to address these issues systematically.

E. Combined or Intersectional Factors

In the preceding sections, we examined the individual factors contributing to demographic bias in FR. This section shifts focus to studies that investigate the combined effects of multiple demographic attributes, such as age, race, and gender. Vera-Rodriguez *et al.* [42] emphasized the significance of gender as a covariate in FR, observing that males consistently outperform females across various demographic groups. These findings highlight the necessity of addressing combined demographic factors to achieve equitable outcomes.

Age-related biases have been linked to structural transformations in facial features over time, particularly among children. Ricanek *et al.* [37] observed increased complexity of child aging compared to adults, attributing recognition challenges to changes in facial bone structure and the proportions of facial components. Additionally, Best-Rowden and Jain [43] reported nuanced patterns in age-related recognition performance, noting that while males generally exhibit higher genuine scores, their performance declines more rapidly with age compared to females. These observations underline the intricate interplay of demographic attributes in shaping biases in FR systems. The intersection of age, race, and gender significantly amplifies biases in FR systems. In [44], Sarridis *et al.* identified a disproportionately high mistreatment rate for African females over 60 years compared to Caucasians, illustrating the compounded effects of intersecting demographic factors. Similarly, El Khiyari *et al.* [45] demonstrated that face verification accuracy is notably lower for younger individuals (aged 18–30), females, and certain racial groups such as Black individuals, highlighting the challenges posed by such intersections of demographic factors.

Algorithmic evaluations further reinforce these findings. The FRVT report [15] observed elevated false positives among children and the elderly, particularly within Asian and American Indian groups. These disparities were intensified in low-quality imaging conditions, with younger and older demographics experiencing higher error rates. Cook *et al.* [39] extended this understanding by showing that self-reported demographic factors like age and measured skin lightness significantly impact recognition scores, often compounded by environmental factors such as illumination. Collectively, these studies underscore the intricate challenges of addressing intersecting biases in FR systems, emphasizing the necessity of age-specific and multi-faceted considerations in algorithm design and evaluation.

III. Datasets for Demographic Bias

In evaluating and mitigating demographic bias in FR systems, the selection of suitable datasets plays an important role. Although numerous FR datasets exist, those specifically intended for bias and fairness-related tasks must include demographic labels associated with each subject or identity. The datasets designed for tasks such as race, gender, ethnicity, or age estimation are particularly useful when they include demographic labels, as such factors are critical for assessing fairness. For certain tasks (related to estimation or classification of attributes), having a single image per subject may suffice, as training and testing can be conducted separately for each image. However, FR models, especially state-of-the-art systems, benefit from having multiple images (variations) of each identity to train more robust feature extractors. The testing phase (verification or identification) requires multiple images per identity, where one image serves as the gallery/ template and the others are used as test or probe samples. Occasionally, the gallery is also composed of more than one image per subject.

These requirements significantly reduce the availability of datasets suitable for assessing demographic bias, as most FR datasets do not provide adequate demographic labels or have highly skewed distributions of subjects across different

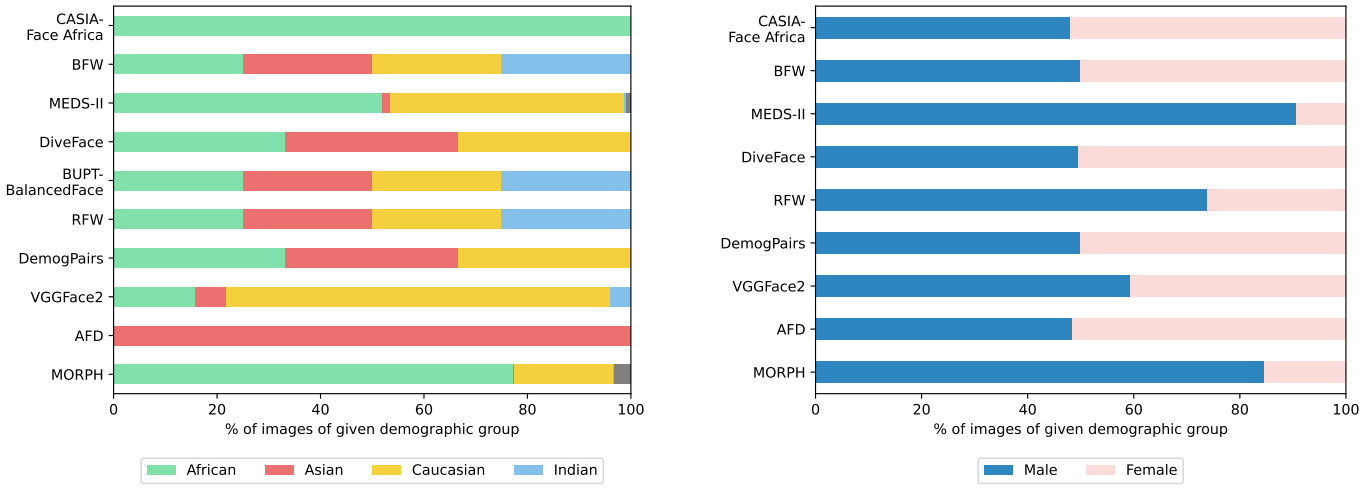
Dataset Name	Year	Number of Images / Subjects	Demographic Labels	Typical Purpose
MORPH-II [48]	2008/2016	55,000 / 13,000	(Male, Female), (Black, White, Asian, Hispanic)	Train/ Test
AFD (Curated) [49] [50]	2018	91,000+ / 1,878	(Male, Female), Asian	Train
VGGFace2 [51]	2018	3.31M / 9,131	(Asian, Black, Indian, White)	Train/ Test
DemogPairs [52]	2019	10,800 / 600	(Male, Female), (Asian, Black, White)	Test
RFW [27]	2019	24,000 pairs / -	(Asian, African, Caucasian, Indian)	Test
BUPT-BalancedFace [53]	2020	1.3M / 28,000	(Asian, African, Caucasian, Indian) (7K per race)	Train
DiveFace [54]	2020	120,000 / 24,000	(Male, Female), (Asian, African, European)	Train/ Test
MEDS-II [55]	2011	1,300+ / 518	(Male, Female), (Asian, Native American, Black, White)	Test
BFW [56] [57]	2023	20,000 / 800	(Male, Female), (Asian, Black, Indian, White)	Test
CASIA-Face-Africa [58]	2021	38,500+ / 1183	(Male, Female), African	Train/ Test
CausalFace [59]	2023	48,000 / 10,200	(Male, Female), (Asian, Black, White)	Test

TABLE II: Commonly used datasets for tasks realted to demographic bias in face recognition.

demographic groups. In this section, we outline recent and commonly used publicly available datasets that are relevant for assessing and addressing demographic bias in FR systems.

- **MORPH** [48]: The MORPH is one of the largest facial image datasets available in several variants and versions. MORPH-II is the most commonly referred academic version, comprising more than 55,000 images from more than 13,000 subjects. Despite its usefulness, it should be noted that the dataset is highly skewed in terms of gender and ethnicity, with a significant over-representation of male subjects (more than 46,000 images) and a limited number of female subjects (approximately 8,500 images). Race labels are also provided, with categories including Black, White, Asian, Hispanic, and others. However, due to the demographic imbalance, this dataset may introduce bias in FR tasks focused on fairness and equity.
- **AFD (Asian Faces Dataset - Curated)** [49]: The Asian Faces Dataset (AFD) was developed using images scraped from the web, with a focus on frontal face images [50]. The curated version, provided by Zhang *et al.* [49] includes over 42,000 images of 911 males and 49,000 images of 967 females. A gender classifier was used to filter out mislabelled images, and duplicate or near-duplicate images were removed. This curated dataset is useful for studying gender and ethnic bias in FR systems, specifically for models focused on the Asian demographic.
- **VGGFace2** [51]: The VGGFace2 is a large-scale FR dataset containing over 3.31 million images of 9,131 subjects. This dataset was annotated for gender (Male, Female) and ethnicity (Asian, Black, Indian, White) labels by Idiap Research Institute[3] making it useful for bias related tasks. The dataset has been noted to have a bias towards White and male subjects, which should be considered when using it for fairness and bias studies. The access to original download location has been removed by its creators as of 2024.
- **DemogPairs** [52]: DemogPairs is a validation set containing 10.8K images, divided into six demographic folds: Asian females, Asian males, Black females, Black males, White females, and White males. The dataset was specifically designed to evaluate the demographic bias in FR models, offering 58.3 million evaluation pairs, including cross-demographic, cross-gender, and cross-ethnicity pairs. The DemogPairs dataset was constructed with rigorous demographic annotation and is a useful resource for testing the generalization of FR systems across diverse demographic groups.
- **Racial Faces in-the-Wild (RFW)** [27]: RFW is a benchmarking dataset designed to study racial bias in FR systems. It consists of four subsets (African, Asian, Caucasian, and Indian), each containing about 3,000 individuals and 6,000 image pairs (these pairs have been defined by its creators). The dataset is specifically used for face verification tasks and includes balanced pairs of genuine (mated) and imposter (non-mated) images. The RFW dataset has been widely adopted by the research community to evaluate and compare the performance of FR algorithms across different racial groups.
- **BUPT-BalancedFace** [53]: The BUPT-BalancedFace dataset was constructed to address demographic bias by ensuring race balance across the dataset. It contains approximately 1.3 million images from 28,000 celebrities, with a balanced distribution of 7,000 identities per race. The dataset was selected from MS-Celeb-1M [60] through the FreeBase and Face++ APIs, although it has been noted that the labels may contain noise. Due to its size and balanced nature, the BUPT-BalancedFace dataset has become a popular resource for training and fine-tuning FR models while mitigating race-related biases.
- **DiveFace** [54]: DiveFace is a dataset generated from the Megaface dataset (now decommissioned) [61], containing over 120,000 images from 24,000 identities. The dataset includes two gender and three ethnicity classes, allowing for detailed demographic analysis. Annotations were made using a semi-automatic process, followed by manual inspection. This dataset is useful for studying the impact of gender and ethnicity in FR tasks, although it may exhibit bias in some groups. It should also be noted that in DiveFace dataset, the subjects of Indian and

[3] Annotations for VGGFace2 Dataset



(a) (b)

Fig. 2: Distribution of images of commonly used FR datasets considering (a) *race* and (b) *gender* as demographic factors. The details of distribution have been used from original sources (wherever available) or from other works contributing to this information; while the naming convention has been altered for unified representation aligning to the convention used by most datasets.

African ethnicities have been grouped together– which can make it difficult to use it in conjunction with other datasets that typically do not follow such grouping.

- **Multiple Encounter Dataset (MEDS-II)** [55]: MEDS-II is an extension of the MEDS-I dataset and was created to assist with the NIST Multiple Biometric Evaluation. The dataset includes over 1,300 images of 518 subjects, with many subjects having only a single image, limiting its usefulness for verification tasks. The MEDS-II is dominated by male subjects of White and Black ethnicities. Despite its limitations in demographic diversity, it remains a useful resource for testing FR systems in real-world scenarios, especially where multiple encounters of a subject are available.
- **BFW (Balanced Faces in the Wild)** [56] [57]: The BFW dataset was designed to provide a more balanced evaluation of FR systems by creating subgroups that are evenly split across gender and ethnicity. The dataset is compiled from VGGFace2 [51] and offers a refined approach to subgroup analysis with less overlap between training and testing data. The corresponding demographic labels were generated using ethnicity [62] and gender [63] classifiers, followed by manual validation. Additionally, the BFW dataset is also balanced with respect to the number of images, subjects, and the (ratio of) images per subject; making it particularly useful for evaluating the demographic fairness of FR models, offering a more balanced alternative to other datasets.
- **CASIA-Face-Africa** [58]: The CASIA-Face-Africa dataset is the first large-scale face dataset of African subjects– comprising 38,546 images from 1,183 individuals, captured under varying illumination conditions using multi-spectral cameras. It includes detailed demographic attributes and facial expressions along with manually annotated with facial key points. The dataset exhibits a well-distributed age representation, with a significant portion belonging to the subjects upto 40 years, aligning with the majority workforce demographics. Additionally, it maintains an almost balanced gender ratio (48% male, 52% female), making it useful for geneder-based analysis as well. In terms of ethnic variations, the dataset includes multiple African ethnic groups, with a notable dominance of the Hausa ethnic group.
- **CausalFace** [59]: It is a large-scale dataset of synthetically generated faces comprising 48,000 synthetic face image pairs generated from 10,200 unique identities, along with 555,000 human annotations covering individual attributes and pairwise identity comparisons. The dataset is constructed using EG3D [64], a state-of-the-art GAN framework that allows explicit control over geometry and pose. In terms of demographic attributes, the CausalFace dataset includes six demographic groups created by combining three ethnicities (White, Black, and East Asian) and two genders (male and female). Using a GAN-based generator, face images were systematically modified across unprotected attributes such as pose, age, expression, and lighting while maintaining identity consistency. The dataset images were selected on the basis of perceived matching scores provided by human annotators.

The datasets are summarized in Table II, while Fig. 2 illustrates an overview of the demographic distribution, with race and gender as demographic variables, for the datasets reviewed in this section.

IV. Assessment of Demographic Bias in FR

For an FR system, the FMR represents the proportion of matching scores where a non-mated pair of biometric samples

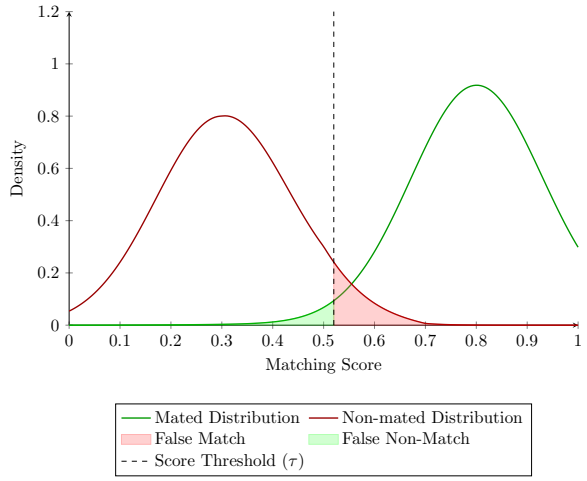


Fig. 3: Illustration of false matches and false non-matches arising from distributions of mated and non-mated scores along with the score threshold.

(*i.e.* samples from different subjects) is incorrectly classified as a match, while FNMR corresponds to the scenarios where a mated pair (*i.e.* samples of the same subject) is incorrectly classified as non-matching. These metrics are essential for evaluating the performance of FR algorithms– where lower values of error rates are preferred. These errors are determined by the score threshold (often denoted as τ) that binarizes the similarity score into match or no-match decision. As illustrated in Figs. 1a–1c, the overlap between distributions of mated scores (shown in green) and non-mated scores (shown in red) defines the regions where classification errors (*i.e.* false matches and false non-matches) occur.

The choice of score threshold plays significant role in FMR and FNMR. A higher threshold reduces FMR by limiting the likelihood of imposter pairs being classified as matches but increases FNMR by rejecting more genuine pairs, and vice-versa. However, it is essential to distinguish between score overlap and binary decision-making. The overlap reflects the inherent ambiguity in the score distributions, while binarized decisions result from applying the threshold to assign matches or non-matches. For instance, two demographic groups with identical overlap may exhibit different FMR and FNMR due to variations in their score distribution shapes or population characteristics. Similarly, the distributions (mated, non-mated, or both) for two demographic groups can be significantly different albeit exhibiting same error rates.

The demographic bias in FR system leads to variations in score distributions and their overlaps across different demographic groups. Such disparities inherently result in different FMR and FNMR values for each group when a single (global) threshold is used. Fig. 3 illustrates score distributions of different groups may differ, leading to unequal error rates. In this section we briefly review performance measures designed for demographic-aware assessments of bias in FR systems. Given the overlap and distinct characteristics of scores and decisions, establishing a well-defined evaluation framework is crucial for accurate analysis. Howard *et al.* [6] introduced the concepts of differential performance and differential outcomes, providing two key terms that aid in achieving a precise understanding and categorization of assessment metrics.

- **Differential Performance:** Variations in genuine (mated) or imposter (non-mated) distributions across demographic groups, independent of thresholds.
- **Differential Outcome:** Differences in FMRs or FNMRs between groups, based on decision thresholds.

Quantifying demographic bias, in terms of both- demographic performance and outcomes, is critical for developing fair and reliable FR systems. By analyzing the overlap of distributions and error disparities, one can identify specific areas requiring intervention to ensure equity across diverse user groups.

Since the 2019 Face Recognition Vendor Test (FRVT)[4] report, NIST has included demographic effects in FR algorithms [15]. This assessment involved comparing non-mated pairs within the same demographic group, setting thresholds for algorithms to achieve an FMR of 0.001 for white males (since this demographic typically associated with the lowest FMR). The report [15] offered a comprehensive analysis of recognition processes and identified areas where demographic effects might occur. To quantify demographic disparity, NIST initially employed Inequity Ratios (IR), calculating the ratio of maximum to minimum of FMR and FNMR across demographic groups. However, considering potentially large range of the error rates, these ratios can become numerically unstable, especially in extreme cases. To alleviate this shortcoming, NIST has considered few modified versions of Inequity Ratios such as adjusting the score threshold (τ) score, incorporating fixed constants in the denominator, or expressing worst-case error rates relative to arithmetic or geometric means [65]. Using the geometric mean is particularly advantageous due to its extended range over FMR/FNMR values. Another possible approach involves referencing a standard FMR or FNMR in the denominator, which inherently resolves stability issues while providing more robust evaluations of demographic bias.

The Fairness Discrepancy Rate (FDR) is one of the initial efforts of quantifying bias in FR [66]. For an FR system using a single decision threshold, the FDR combines FMR and FNMR using weighted sums, and evaluates fairness through a unified measure that captures the trade-offs between both error rates. The FDR requires two hyper-parameters: one for the score threshold, and another for defining relative importance of FMR over FNMR. Thus, it offers flexibility of assessing the fair nature of the model at pre-defined score threshold and application-dependent weighing of false matches to false non-matches. In a similar vein, to define the balance between false matches and false non-matches, NIST proposed an approach to calculate the Inequity Measure by raising the terms representing demographic disparities to specific exponents (serving as weights) and then multiplying them.

Schuckers *et al.* highlighted the importance of accounting for statistical variation when evaluating fairness in FR sys-

[4]Since 2023, the FRVT initiative has been restructured into the Face Recognition Technology Evaluation (FRTE) and Face Analysis Technology Evaluation (FATE) programs.

Bias Assessment Metric	Ref Publication	Year	Description
Inequity Rate (IR)	NIST / Grother *et al.* [16]	2021	This metric involves the analysis of the minimum and maximum ratios of False Match Rate (FMR) and False Non-Match Rate (FNMR), with the application of weighted exponents.
Statistical Approaches	Schuckers *et al.* [67]	2022	Utilizes a bootstrap-based hypothesis testing approach to assess bias.
Separation/Compactness Metrics	Kotwal & Marcel [68]	2022	Investigates the distributions of genuine and impostor scores, focusing on shape and compactness characteristics.
Fairness Discrepancy Rate (FDR)	Pereira & Marcel [66]	2021	A weighted combination of the maximum differential values of FMR and FNMR.
GARBE (Gini Coefficient Based Metric)	MdTF / Howard *et al.* [69]	2022	Measures statistical dispersion in FMR and FNMR, integrating these through linear weighting.
Sum of Group Error Differences	Elobaid *et al.* [70]	2024	Examines relative deviations in group-level FMR and FNMR compared to global scores, providing insight into disparities.
Comprehensive Equity Index (CEI)	Solano *et al.* [71]	2024	A weighted combination that assesses disparities in both tail and central distributions of performance metrics.
Mean Absolute Percentage Error (MAPE)	Villalobos *et al.* [72]	2022	Calculates the relative deviation of FMR from a pre-established benchmark.
Standard Deviation	-	-	The standard deviation of FMR, FNMR, and True Match Rate (TMR) is assessed to gauge variability.
Skewed Error Ratio (SER)	-	-	A ratio that compares the worst-case error rates across different demographic groups.
Trade-Off (TO)	-	-	Evaluates the trade-off between performance metrics, particularly the difference in average accuracy and standard deviation.

TABLE III: Summary of bias assessment metrics employed in demographic analysis of face recognition systems.

tems [67]. They noted that the differences among demographic groups can arise either from actual performance disparities or by chance due to sampling variability, leading to potential Type-I errors. To address this, they proposed two statistical methodologies: a bootstrap-based hypothesis test and a simpler test methodology tailored for non-statistical audiences. Their study also conducted simulations to explore the relationship between margin of error and factors such as the number of subjects, attempts, correlation between attempts, underlying FNMRs, and the number of demographic groups. In [69], the researchers at MdTF proposed new metric for assessment of bias based on demographic outcomes. Their metric, GARBE (Gini Aggregation Rate for Biometric Equitability), is inspired by the Gini coefficient– which has a long history of use as a dispersion measure in socio-economic context. The GARBE evaluates statistical dispersion in error rates and emphasizes equitable treatment across demographic groups. Similar to the FDR and IR, this metric combines weighted contributions of FMR and FNMR to produce a single fairness score for a given fixed score threshold.

Villalobos *et al.* proposed the Mean Absolute Percentage Error (MAPE) as a metric to quantify differences in error rates across demographic groups [72]. MAPE measures the relative deviation of FMRs from a policy-defined FMR, ensuring that low error rates for one group do not mask higher error rates for another. High deviations in error rates, particularly towards lower values of FMR, can negatively impact the system by increasing FNMR. A MAPE score of zero indicates that all demographic groups achieve the desired FMR, making it an effective metric for fairness evaluation. The Sum of Group Error Differences (SED_G) was introduced as the fairness assessment metric address disparities in biometric verification systems in [70]. The SED_G calculates relative deviations in FMR and FNMR across demographic groups from the FMR/ FNMR of global scores. They consider the Equal Error Rate (EER) threshold as a reference to compute the error rates. In other words, it adapts a relative difference formula to quantify demographic bias by comparing individual group performances to a global standard. Authors argue that by incorporating both within-demographic (WDI) and cross-demographic (CDI) interactions, SED_G is able to provides provide better understanding of the magnitude and type of bias making it a versatile measure.

Relatively fewer attempts have been made to assess the demographic bias at score-level (ie, based on differential performance). Kotwal and Marcel introduced three fairness evaluation measures that emphasize the separation, compactness, and distribution of genuine and impostor scores [68]. Unlike conventional approaches that depend on system accuracy, these measures focus on assessing differential performance without requiring external parameters such as score thresholds. By examining how well the match is, rather than merely determining a match, this approach provides a more nuanced evaluation of demographic fairness. Additionally, they also discussed a weighted fusion strategy to improve relative contributions from under-represented groups, addressing the challenges posed by imbalanced datasets. Building on the work from [68], Solano *et al.* developed the Comprehensive Equity Index (CEI)– combining error rate differences and recognition score distribution disparities [71]. The CEI enhances bias quantification methods by considering both the distribution tails and overall shapes of score distributions, enabling the detection of subtle biases across demographic groups. They also conducted experiments on high performing FR systems (as per NIST evaluations)

using real challenging datasets. Their experiments showed that CEI was able to effectively capture the demographic bias on several challenging datasets with several covariates.

Several studies have evaluated bias and fairness in FR systems using the standard deviation of performance metrics calculated across demographic groups. These metrics include FMR, FNMR, and True Match Rate (TMR)– where higher value in standard deviation corresponds to greater demographic disparities [32], [53], [57], [72]–[74]. Another significant metric is the Skewed Error Ratio (SER), which specifically focuses on worst-case error ratios, providing insights into the performance imbalance across groups [53], [72], [75]. Recent competitions [76]–[78] exploring the use of synthetic data for FR and bias mitigation have adopted a trade-off performance metric: the mean accuracy adjusted by the standard deviation. This metric aims to ensure that efforts to mitigate bias do not come at the expense of recognition performance. This metric emphasizes the development of FR models that achieve both high recognition performance and fairness across constituent demographic groups.

Table III provides a brief summary of the bias assessment metrics discussed in this section.

V. Bias Mitigation in Face Recognition Systems

In an FR system, bias mitigation can be applied at different stages of the recognition pipeline, providing a structured approach to addressing demographic disparities. These stages align with general categories of bias mitigation strategies commonly used in machine learning: pre-processing, in-processing, and post-processing [5], [20], [79], [80].

Pre-processing methods aim to address bias at the level of training data by modifying or augmenting it to reduce discriminatory patterns or imbalances before the data is used for training. This approach is particularly relevant when representational disparities in the data contribute to demographic bias. The in-processing methods focus on modifications to the FR model during training or fine-tuning phase, often by incorporating constraints or objectives that optimize fairness without significantly compromising recognition accuracy. Finally, post-processing techniques involve adjustments to the output of trained models to ensure fairness across demographic groups. These techniques modify the results, such as classification scores or decision thresholds, to achieve equitable performance without altering the underlying model. Fig. 4 summarizes the categorization of bias mitigation methods– that correlates with the corresponding stages of FR pipeline.

A. Pre-Processing Methods

Pre-processing methods, also known as data-based methods, focus on modifying the biometric samples before feeding into the FR system. These techniques aim to normalize the characteristics of the data to make them robust for subsequent feature extraction. Most of the methods in this category can also be regarded as special case of data augmentation, specifically designed to reduce demographic biases in the training data.

In [23], Klare *et al.* showed that FR algorithms were performing worse for female, Black, and younger individuals. To address this concern, they proposed two mitigation strategies based on selection of training data. First, training models on specific demographic cohorts to enhance recognition for those groups, and second, implementing a dynamic face matcher selection approach, where different algorithms, each trained on distinct demographic groups, would be chosen based on probe information. A similar approach was followed by Deb *et al.* in the context of longitudinal study of FR [81]. Using the data of more than 900 children captured over time, they demonstrated that recognition accuracy decreases as the time gap between image captures grows. To address this, they fine-tuned FaceNet [82] on a separate child face dataset, showing improved accuracy. The authors advocated the need for tailored training and evaluation of FR systems for different age groups. Lu *et al.* [32] investigated the influence of covariates—such as age, gender, pose, and skin tone— toward the performance of face verification. Their findings showed that using gender information to curate training data improves performance, particularly at low FMRs. In [83], Kortylewski *et al.* demonstrated the effectiveness of using synthetic face images to mitigate the bias arising from real-world datasets. By utilizing synthetic data for pre-training the FR model, they showed that the negative effects of dataset bias, particularly with regard to pose variations, could be significantly reduced. Their experiments revealed that pre-training the CNNs with synthetic data, the need for real-world data can be reduced by up to 75%. Although this study did not directly address demographic bias, we include it in the review as it highlights the potential of synthetic data in improving the generalization performance of FR systems and reducing (non-demographic) dataset bias.

Wang *et al.* explored the issue of highly-skewed class distributions in FR datasets [84]. They proposed Large Margin Feature Augmentation (LMFA) and Transferable Domain Normalization (TDN) as methods to balance class distributions by augmenting and normalizing the feature space. These methods were shown to enhance the performance of underlying FR models by mitigating issues arising from class imbalance, which often correlates with demographic bias in under-represented groups. In [85], Yin *et al.* introduced a center-based feature transfer framework to address the under-representation of certain demographic groups in FR datasets. By transferring feature distributions from well-represented classes to under-represented ones, they augmented the feature space for these groups, reducing bias and improving recognition performance for under-represented subjects. Recently, Kotwal and Marcel proposed an Image-to-Image transformation module called Demographic Fairness Transformer (DeFT), which enhances image representations before passing them to pretrained CNNs [86]. The DeFT uses multi-head encoders and soft-attention mechanisms to selectively enhance images based on inferred demographic information. The demographic labels of race or ethnicity are often non-discrete– but this concern has rarely been addressed. In [86], they replaced hard labels with probabilistic weights which are implicitly inferred at run-time. Their experiments show that DeFT reduces bias and improves model fairness, with some models also achieving slightly better accuracy compared to

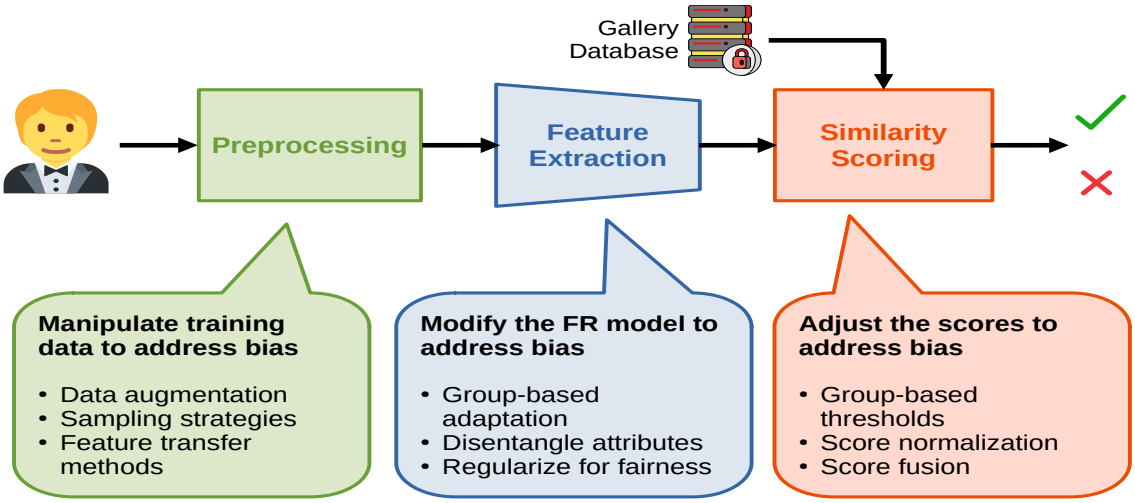


Fig. 4: Illustration of categories of methods of bias mitigation in FR.

baseline systems.

B. In-Processing Methods

This category of works, also known as model-based methods, are applied during the feature extraction stage by modifying the weights of the FR model. The goal is to learn weights that generate features or embeddings that are less sensitive to demographic differences.

Amini *et al.* [87] proposed a debiasing algorithm that adjusts the sampling probabilities of data points in large datasets to reduce hidden biases. When applied for face detection use-case, their algorithm led to decrease in race and gender bias while improving classification accuracy. To our knowledge, similar approaches have not been tested for recognition or verification applications. In [53], Wang and Deng introduced a reinforcement learning-based race balance network (RL-RBN) where they applied adaptive margins through deep Q-learning. Their method aimed to reduce the skewness of feature scatter between racial groups, leading to more balanced performance across different demographics. As a part of this work, they also released two datasets- BUPT-GlobalFace and BUPT-BalancedFace datasets- that were specifically designed to study racial bias in FR systems.

Gong *et al.* proposed a group-adaptive classifier (GAC) that uses adaptive convolution kernels and attention mechanisms tailored to different demographic groups [88]. By applying kernel masks and attention maps specific to each group, their method activates facial regions that are more discriminative for each demographic, thereby improving recognition accuracy and fairness across demographic groups. Another approach by the same authors [74] introduced an adversarial network, called DebFace, which utilizes a multi-task learning framework to simultaneously learn identity and demographic attributes. Their method employed adversarial training to disentangle identity features from demographic attributes such as gender, age, and race to effectively reducing bias in the recognition process. Their experiments demonstrated that, for DebFace, not only recognition but also demographic attribute estimation tasks were less biased.

Another approach for domain-specific bias mitigation using disentangled representation learning was proposed by Liang *et al.* [89]. They introduced a two-stage method combining modules for disentangled representation learning with additive adversarial learning (AAL). While this work does not directly address demographic bias, it provides useful insights into how domain-specific biases can be mitigated by learning disentangled representations. The effectiveness of this method in reducing bias across various domains suggests its potential applicability in the context of demographic bias in FR. In [90], a Progressive Cross Transformer (PCT) was proposed to mitigate racial bias by decoupling face representations into identity-related and race-induced components. Using dual cross-transformers, the PCT refines identity features and suppresses racial noise, demonstrating lower racial bias without compromising recognition accuracy. The concept of score normalization was incorporated as a regularization term into the training objective enabling simultaneous optimization of recognition accuracy and demographic fairness [75]. This was facilitated by constraining the output scores of mated and non-mated pairs to adhere to a pre-defined distribution, and followed by minimizing differences in score distributions across demographic groups. During inference, the overall pipeline did not require modifications as the FR CNN architecture remained unaltered while only the weights were fine-tuned to the new objective.

Finally, we discuss a couple of works dealing with fairness in facial attribute recognition. Both of these are based on contrastive learning– which can be useful mechanism to address the demographic bias in FR as well. Park *et al.* addressed fairness issues in attribute classification using a contrastive learning framework [91]. They constructed a Fair Supervised Contrastive Loss (FSCL) which reduces bias by normalizing intra-group compactness and inter-group separability, penalizing sensitive attribute information in representations.

Reference	Year	Type of Method	Test Dataset	Method Description
Klare *et al.* [23]	2012	Data-Processing	PCSO	Training models on specific demographic cohorts
Deb *et al.* [81]	2018	Data-Processing	CLF	Finetuning models on specific cohort (age, in this case)
Lu *et al.* [32]	2019	Data-Processing	IJB-B, IJB-C	Curation of training data
Kortylewski *et al.* [83]	2019	Data-Processing	Multi-PIE, LFW, IJB-A	Synthetic data pretraining followed by real data fine-tuning
Wang *et al.* [84]	2019	Data-Processing	RFW	Large-margin feature augmentation to balance class distributions
Yin *et al.* [85]	2019	Data-Processing	LFW, IJB-A, MS-Celeb-1M	Feature transfer to enhance under-represented groups
Kotwal & Marcel [86]	2024	Data-Processing	RFW	Demographic-dependent transformation of input image
Amini *et al.* [87]	2019	In-Processing	PPB	Sampling data probabilities for face detection
Wang & Deng [53]	2020	In-Processing	RFW	Reinforcement learning-based race balance network
Gong *et al.* [88]	2021	In-Processing	RFW	Group-adaptive training with adaptive convolution kernels and attention mechanisms
Gong *et al.* [74]	2020	In-Processing	RFW	Adversarial debiasing network with identity and demographic classifiers
Liang *et al.* [89]	2019	In-Processing	CelebA	Two-stage adversarial bias mitigation through disentangled representations and additive adversarial learning
Li *et al.* [90]	2021	In-Processing	RFW, BFW	Progressive cross-transformer to remove race-induced identity-unrelated components
Kotwal & Marcel [75]	2024	In-Processing	VGGFace2, MORPH, RFW	Regularization constraints based on score calibrations for demographic groups
Park *et al.* [91]	2022	In-Processing	CelebA, UTK-Face	Contrastive setups to enhance intra-class similarity and diminish similarity between negative samples
Zhang *et al.* [92]	2023	In-Processing	CelebA, UTK-Face	Generating contrastive sample pairs with visual similarity and unsupervised feature reweighting
Michalski *et al.* [93]	2018	Post-Processing	Private	Dynamic thresholds based on age differences
Srinivas *et al.* [94]	2019	Post-Processing	ITWCC-D1	Score fusion and ensemble strategies to address age-related bias
Robinson *et al.* [57]	2020	Post-Processing	BFW	Demographic-specific thresholds
Terhörst *et al.* [95]	2020	Post-Processing	Color-Feret, LFW	Fairness-driven NN classifier
Terhörst *et al.* [96]	2020	Post-Processing	Adience, Color-Feret, MORPH	Fair score normalization to mitigate demographic bias
Linghu *et al.* [97]	2024	Post-Processing	VGGFace2, RFW	Integrating demographic information in Z/T score normalization

TABLE IV: Summary of recent works on mitigation of demographic bias in face recognition.

Although predominantly an in-processing method, the authors also incorporate a loss function to address the imbalance in training data where majority groups are constrained to have a better intra- group compactness and inter-class separability compared to the under-represented ones. Similarly, Zhang *et al.* proposed Fairness-aware Contrastive Learning (FairCL) for unsupervised representation learning and demonstrated the use-case of facial attribute recognition [92]. In addition to fair contrastive learning of feature representations, they also attempted to address the dataset bias by specifically generating contrastive sample pairs that share the same visual information apart from sensitive attributes. They also suggested unsupervised feature reweighting to strike balance the utility and fairness of learned representations.

C. Post-Processing Methods

These techniques are applied after the feature extraction and matching processes to adjust the final decision scores and reduce bias. While these methods are less commonly used compared to pre- and in-processing techniques, they can still play a role in ensuring equitable outcomes. Additionally, these approaches are relatively easy to integrate into existing systems.

Michalski *et al.* investigated the impact of age variation on FR, particularly for children [93]. They showed that dynamic thresholding improves performance. To address age-related

bias, they adjusted thresholds based on age differences as opposed to a fixed threshold. In another work on age-related bias, Srinivas *et al.* experimented with score-level fusion strategies to improve recognition accuracy for the children (age as demographic) [94]. They considered six fusion schemes that combined different score-normalization techniques and fusion rules. For normalizing, they considered z-norm and min-max strategies; while fusion was conducted using min, max, or sum rules. Robinson *et al.* [57] showed that applying a single threshold across different demographic groups leads to significant variations in the FMR. They addressed this issue by using per-subgroup thresholds to balance the FMRs across ethnic and gender groups, improving both recognition fairness and performance.

A typical FR pipeline employs similarity functions to obtain a matching score. Terhorst *et al.* replaced the conventional similarity function by a fairness-driven neural network classifier [95]. By adding a penalization term in the loss function, their method was able to equalizes score distributions across ethnic groups, reducing intra-ethnic bias while maintaining high recognition performance. In another work [96], Terhorst *et al.* introduced an unsupervised fair score normalization method based on individual fairness principles, which treats similar individuals similarly. During training, they partitioned the identities in finite number of groups using K-means clustering on face embeddings. At inference, they computed the cluster-specific thresholds for both samples contributing to the score, and these threshold were combined with a global threshold to yield normalized scores. In a recent work, Linghu *et al.* extended traditional score normalization methods, such as Z and T normalization, by incorporating demographic information to enhance fairness in FR systems [97]. Furthermore, they evaluated three cohort-based approaches based on imposter scores, Platt scaling, and bi-modal cumulative distribution functions (CDF). Their findings demonstrated that the proposed method improved fairness across both race and gender demographic groups, particularly at low FMRs.

The works on mitigation are summarized in Table IV.

VI. Future Directions

In recent years, with advanced architectures and increase in the number of layers and parameters, FR models have gained a substantial improvement in their capacity to learn complex facial representations. These deeper architectures, often comprising over a hundred layers with millions of parameters, have significantly enhanced the ability of FR systems to generalize across challenging scenarios, resulting in higher overall recognition performance (as well as reduced bias). Additionally, the availability of larger, more diverse datasets has contributed to better learning outcomes. These datasets, which incorporate substantial variations demographics and covariates such as pose, illumination, and expression (PIE), have facilitated measurable progress in improving both accuracy and fairness FR.

However, despite these advancements, several key challenges in this area continue to exist. Most of the existing efforts primarily center on extremely deep models, which demand extensive computational resources and memory footprint for both training and deployment. This emphasis on high-capacity architectures does not adequately address quality issues in data or labels nor does it cater to the requirements of resource-constrained environments. Thus, while the combination of deeper models and diverse data has been pivotal, future research must explore avenues to address residual biases and expand fairness to a broader range of applications.

In this section, we examine some of the emerging challenges associated with demographic bias. These challenges highlight the need for ongoing research to adapt bias mitigation strategies to align with the advancements in FR applications.

Bias in Lightweight Models: The lightweight FR models, often used in handheld devices and resource-constrained environments, encounter significant challenges concerning demographic bias. These systems, crucial for privacy-sensitive applications, often inherit limitations in capacity and architecture, leading to non-equitable performance across demographic groups. Bias in lightweight models has garnered limited research attention, despite their widespread deployment in mobile and IoT devices with varying sensor qualities. Techniques like knowledge distillation (KD) and pruning, while essential for model compression, introduce or amplify bias. For instance, Liu *et al.* highlighted that KD inherits biases from larger teacher models [98], while pruning has been shown to disproportionately impacts underrepresented groups [99], [100]. Incorporating fairness-aware techniques is crucial for mitigating these issues. Lin *et al.* [101] introduced FairGRAPE, a pruning method that evaluates network connections with demographic considerations, reducing performance disparities. Achieving demographic fairness in lightweight models requires targeted compression strategies and ethical evaluations of demographic-specific impacts. By integrating fairness principles into compression techniques, lightweight FR systems can achieve more equitable outcomes while maintaining efficiency and accuracy.

Bias in Quantization: Quantization, a model compression technique, retains the original architecture (as opposed to KD and pruning– which often modify the structure) while reducing parameter precision, producing smaller and faster models. However, it can lead to higher demographic bias by prioritizing global performance over the accurate classification of under-represented groups. Such bias underscores the need for rigorous fairness evaluation across demographic subgroups when deploying compressed models. Quantization converts floating-point (FP) models into lower-precision formats like 8-bit, balancing efficiency and accuracy [102]. It typically consists of two approaches: post-training quantization (PTQ) and quantization-aware training (QAT). Some studies, such as Stoychev *et al.* demonstrated that 8-bit PTQ maintained fairness and accuracy in gender bias for face expression recognition [102]. However, reducing precision to 6 bits significantly degraded fairness, indicating a trade-off between compression and bias mitigation. Although similar investigations for FR systems remain limited, these findings

highlight the need to evaluate and ensure demographic fairness in quantized models.

Low Resolution: The research on demographic bias in FR has predominantly focused on high-resolution images, often overlooking the challenges posed by low-resolution images typically captured by surveillance cameras or from significant distances. One of the primary impediment to research this issue is the scarcity of low-resolution datasets that include demographic attributes. Consequently, demographic disparities in low-resolution FR remain under-explored, despite their importance in real-world applications. A recent work from Atzori *et al.* attempted to address this gap by designing a novel framework to investigate demographic bias in low-resolution FR [103]. They trained state-of-the-art FR models on various combinations of high- and low-resolution images. Testing on degraded images from five datasets revealed significant disparities across gender and ethnic groups, underscoring the need for timely interventions in low-resolution FR. It may be noted that their approach involved use of a generative model to convert high-resolution face images into realistic low-resolution counterparts. The importance of low-resolution FR is evident in programs like BRIAR[5], which aim to enhance recognition technologies for challenging scenarios, such as long-distance identification and low-quality image acquisition. To tackle demographic bias in low-resolution FR, there is a need to develop both datasets and models tailored to these unique use-cases.

Training Datasets: Demographically balanced datasets, while not entirely eliminating bias, significantly contribute to reducing non-equitable performance in FR. Thus, large, diverse, and balanced datasets are pivotal for achieving fairer and accurate models. However, acquiring such datasets is increasingly challenging due to cost and ethical and privacy concerns surrounding biometric data collection. The commercial sector, in particular, faces difficulties as most available datasets are collected locally from consenting individuals— which are often in limited size and demographic representation. These constraints necessitate innovative approaches to address demographic bias using smaller or synthetic datasets.

The use of synthetic data in FR has recently gained traction as a potential solution to privacy and data-sharing concerns. Competitions like the FRSyn series have encouraged advancements in synthetic data usage [76]–[78]. Despite these efforts, FR models trained exclusively on synthetic datasets continue to underperform compared to those trained on real datasets of similar size [76], [104]. This gap is evident in both recognition accuracy (measured by metrics like FMR and FNMR) and demographic fairness (assessed by standard deviation of performance across groups). The analysis of synthetic datasets by Huber *et al.* revealed that demographic bias might worsen compared to the (real) training dataset [105]. Enhancing the quality and utility of synthetic datasets, beyond the aspect of bias, remains an open problem, requiring further exploration.

Use-Cases of Remote Checking: Last few years have witnessed tremendous surge in online activities: financial Transactions, banking, user-onboarding, etc. These activities have driven widespread adoption of remote identity verification (RIdV) technologies. These systems authenticate individuals by comparing real-time images or selfies, captured via smart devices, against official identity documents, such as work permits or driver's licenses. Such solutions are integral to online Know Your Customer (KYC) processes, which are now standard for banks and financial institutions. While RIdV systems enhance convenience and scalability, it is essential to ensure their fairness across demographic groups as they become more prevalent.

Recognizing the increasing reliance on remote verification, the MdTF and DHS S&T introduced the Remote Identity Validation Technology Demonstration (RIVTD) initiative[6]. In addition to security, accuracy, and liveness detection requirements, this program also places particular emphasis on ensuring demographic fairness in such technologies. A recent study by Fatima *et al.* [106] investigated demographic fairness in RIdV technologies using statistical methods to analyze performance disparities. Their analysis of five commercial RIdV systems revealed that only two achieved equitable outcomes across demographic groups. Notably, higher FNMRs were observed among African cohorts and individuals with darker skin-tones. Such findings highlight the necessity of evaluating RIdV technologies across demographic groups to ensure equitable and unbiased performance.

Complex Bias Factors (Intersectionality): The majority of research on mitigation of bias in FR, as discussed in Sec V has focused on single demographic attribute, such as race, age, or gender. However, several studies have identified that combination or intersection of various demographic factors causes (or amplifies) bias in FR models (cf. Sec II). Existing bias mitigation techniques typically target one demographic attribute at a time, achieving measurable improvements in fairness for that specific attribute. However, it remains unclear whether such processing inadvertently introduces imbalances in other demographic attributes. For instance, enhancing fairness for gender-related bias may increase disparities linked to ethnicity and vice-versa. This highlights the need for systematic evaluations of the intersectionality of demographic factors, such as race and gender combined. Consequently, developing mitigation methods capable of addressing multiple demographic attributes simultaneously remains an open challenge.

Noisy Labels: The assignment of demographic attributes such as race, ethnicity, and skin-tone in FR datasets typically involves discrete labeling into finite categories. Some attributes, such as race, are often self-reported. In many cases, race and ethnicity annotations may be derived from automatic classifiers or manual efforts. Automatic classifiers, predominantly based

[5]BRIAR Programme by IARPA.

[6]Remote Identity Validation Technology Demonstration (RIVTD)

on deep learning, are likely to be susceptible to bias too; while manual annotations are prone to human judgment errors. Similarly, skin-tone is frequently categorized using scales like Fitzpatrick's, which discretizes it into specific values, ignoring its continuous spectrum. This reliance on discrete labels introduces noise into the training data, as samples near category boundaries are often inaccurately labeled. Recent work [86], addressed this issue by using probabilistic weights (soft labels) for demographic information instead of utilizing rigif (categorical) labels. However, most existing methods overlook the issue of errors in training data.

Noisy labels in training datasets pose a significant challenge, especially considering massive scale of FR datasets, often in few hundreds of thousands of images. Manually verifying or curating such datasets is labor-intensive, impractical, and still prone to errors. Furthermore, removing samples with ambiguous labels can lead to reducing dataset diversity and robustness. Thus, developing robust mitigation strategies capable of handling noisy labels without compromising the effectiveness of training processes or dataset diversity is essential for improving fairness and accuracy in FR systems.

VII. Conclusion

In this work, we have systematically explored the issue of demographic bias in FR through different yet interrelated sections: causes, datasets, assessment metrics, and mitigation strategies. The section on causes delves into the underlying factors, categorizing them into aspects such as imbalances in training datasets, variability of skin-tones, algorithmic sensitivities, image quality and covariates, and combined demographic factors. The discussion on datasets presents an overview of existing resources and their demographic distributions, Furthermore, we reviewed state-of-the-art metrics for evaluating bias, ranging from traditional statistical measures to advanced fairness indicators, highlighting differences in demographic performance and outcomes across demographic groups. The mitigation section highlights diverse approaches, including preprocessing, in-processing, and post-processing techniques, offering a comprehensive overview of existing work in this area.

In addition to summarizing existing research, we identified emerging challenges including the fairness implications of lightweight models, the role of synthetic datasets, the complexities of remote identity verification systems, and intersectional bias. These challenges reflect the evolving nature of FR technologies and underscore the need for innovative strategies to ensure equitable and reliable outcomes in real-world applications.

Acknowledgment

The work has been supported by the Hasler foundation (through the SAFER project) and the Swiss Center for Biometrics Research and Testing.

References

[1] C. Rathgeb, P. Drozdowski, D. Frings, N. Damer, and C. Busch, "Demographic fairness in biometric systems: What do the experts say?" *IEEE Technology and Society Magazine*, vol. 41, no. 4, pp. 71–82, 2022. 1

[2] T. Sixta, J. C. Jacques Junior, P. Buch-Cardona, E. Vazquez, and S. Escalera, "Fairface challenge at eccv 2020: Analyzing bias in face recognition," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 463–481. 1

[3] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020. 1, 2

[4] A. Jain, D. Deb, and J. Engelsma, "Biometrics: Trust, but verify," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 3, pp. 303–323, 2021. 1, 2

[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021. 1, 2, 10

[6] J. J. Howard, Y. B. Sirotin, and A. R. Vemury, "The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance," in *2019 ieee 10th international conference on biometrics theory, applications and systems (btas)*. IEEE, 2019, pp. 1–8. 1, 8

[7] A. Limantė, "Bias in facial recognition technologies used by law enforcement: Understanding the causes and searching for a way out," *Nordic Journal of Human Rights*, vol. 42, no. 2, pp. 115–134, 2024. 1

[8] C. Jones, "Law enforcement use of facial recognition: bias, disparate impacts on people of color, and the need for federal legislation," *NCJL & Tech.*, vol. 22, p. 777, 2020. 1

[9] D. Leslie, "Understanding bias in facial recognition technologies," *arXiv preprint arXiv:2010.07023*, 2020. 1

[10] B. Burgess, A. Ginsberg, E. W. Felten, and S. Cohney, "Watching the watchers: bias and vulnerability in remote proctoring software," in *31st USENIX security symposium (USENIX security 22)*, 2022, pp. 571–588. 1

[11] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, "Saving face: Investigating the ethical concerns of facial recognition auditing," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 145–151. 1

[12] A. Pena, I. Serna, A. Morales, and J. Fierrez, "Bias in multimodal ai: Testbed for fair automatic recruitment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 28–29. 1

[13] A. Ross, S. Banerjee, C. Chen, A. Chowdhury, V. Mirjalili, R. Sharma, T. Swearingen, and S. Yadav, "Some research problems in biometrics: The future beckons," in *2019 international conference on biometrics (ICB)*. IEEE, 2019, pp. 1–8. 1, 2

[14] C. Busch, "Challenges for automated face recognition systems," *Nature Reviews Electrical Engineering*, pp. 1–10, 2024. 1

[15] P. Grother, M. Ngan, and K. Hanaoka, "Face recognition vendor test part 3: Demographic effects," 12 2019. 1, 2, 4, 5, 8

[16] P. Grother, "Demographic differentials in face recognition algorithms," *EAB Virtual Event Series-Demographic Fairness in Biometric Systems*, 2021. 1, 9

[17] W. Deng, T. Hassner, X. Liu, and M. Pantic, "Tbiom special issue on trustworthy biometrics-editorial," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 3, pp. 301–302, 2022. 1

[18] J. Cheong, S. Kalkan, and H. Gunes, "The hitchhiker's guide to bias and fairness in facial affective signal processing: Overview and techniques," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 39–49, 2021. 1

[19] I. J. S. Biometrics, "Iso/iec dis 19795-10. information technology – biometric performance testing and reporting – part 10: Quantifying biometric system performance variation across demographic group," 2023. 1

[20] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–44, 2022. 2, 10

[21] S. Yucer, F. Tektas, N. Al Moubayed, and T. Breckon, "Racial bias within face recognition: A survey," *ACM Computing Surveys*, 2023. 2

[22] K. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer, "Issues related to face recognition accuracy varying based on race and skin tone," *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 8–20, 2020. 2, 3, 4

[23] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on information forensics and security*, vol. 7, no. 6, pp. 1789–1801, 2012. 2, 3, 4, 10, 12

[24] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on

measuring race bias?" *IEEE transactions on biometrics, behavior, and identity science*, vol. 3, no. 1, pp. 101–111, 2020. 2, 3, 4

[25] M. Gwilliam, S. Hegde, L. Tinubu, and A. Hanson, "Rethinking common assumptions to mitigate racial bias in face recognition datasets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4123–4132. 2, 4

[26] H. Wu and K. W. Bowyer, "What should be balanced in a "balanced" face recognition dataset?" *arXiv preprint arXiv:2304.09818*, 2023. 2, 4

[27] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2019, pp. 692–702. 3, 4, 6

[28] M. Kolla and A. Savadamuthu, "The impact of racial distribution in training data on face recognition bias: A closer look," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 313–322. 3, 4

[29] V. Muthukumar, T. Pedapati, N. Ratha, P. Sattigeri, C.-W. Wu, B. Kingsbury, A. Kumar, S. Thomas, A. Mojsilovic, and K. R. Varshney, "Understanding unequal gender classification accuracy from face images," *arXiv preprint arXiv:1812.00099*, 2018. 3, 4

[30] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 32–41, 2019. 3, 4

[31] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81, Feb 2018, pp. 77–91. 3, 4

[32] B. Lu, J.-C. Chen, C. D. Castillo, and R. Chellappa, "An experimental evaluation of covariates effects on unconstrained face verification," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 42–55, 2019. 3, 4, 5, 10, 12

[33] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole, "An other-race effect for face recognition algorithms," *ACM Transactions on Applied Perception (TAP)*, vol. 8, no. 2, pp. 1–11, 2011. 3, 4

[34] S. Nagpal, M. Singh, R. Singh, and M. Vatsa, "Deep learning for face recognition: Pride or prejudiced?" *arXiv preprint arXiv:1904.01219*, 2019. 3, 4

[35] I. Serna, A. Morales, J. Fierrez, M. Cebrian, N. Obradovich, and I. Rahwan, "Algorithmic discrimination: Formulation and exploration in deep learning-based face biometrics," *arXiv preprint arXiv:1912.01842*, 2019. 3, 4

[36] V. Albiero, K. Ks, K. Vangara, K. Zhang, M. C. King, and K. W. Bowyer, "Analysis of gender inequality in face recognition accuracy," in *Proceedings of the ieee/cvf winter conference on applications of computer vision workshops*, 2020, pp. 81–89. 3, 4, 5

[37] K. Ricanek, S. Bhardwaj, and M. Sodomsky, "A review of face recognition against longitudinal child faces," *BIOSIG 2015*, pp. 15–26, 2015. 3, 4, 5

[38] C. Cook, J. Howard, Y. B. Sirotin, J. Tipton, and A. Vemury, "Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 32–41, 2019. 3, 4

[39] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Demographic effects across 158 facial recognition systems," Technical report, DHS, Tech. Rep., 2023. 3, 4, 5

[40] H. Wu, V. Albiero, K. Krishnapriya, M. King, and K. Bowyer, "Face recognition accuracy across demographics: Shining a light into the problem," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1041–1050. 4, 5

[41] K. Krishnapriya, K. Vangara, M. C. King, V. Albiero, K. Bowyer *et al.*, "Characterizing the variability in face recognition accuracy relative to race," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0. 4, 5

[42] R. Vera-Rodriguez, M. Blazquez, A. Morales, E. Gonzalez-Sosa, J. C. Neves, and H. Proença, "Facegenderid: Exploiting gender information in dcnns face recognition systems," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0. 4, 5

[43] L. Best-Rowden and A. K. Jain, "Longitudinal study of automatic face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 1, pp. 148–162, 2017. 4, 5

[44] I. Sarridis, C. Koutlis, S. Papadopoulos, and C. Diou, "Towards fair face verification: An in-depth analysis of demographic biases," *arXiv preprint arXiv:2307.10011*, 2023. 4, 5

[45] H. El Khiyari and H. Wechsler, "Face verification subject to varying (age, ethnicity, and gender) demographics using deep learning," *Journal of Biometrics and Biostatistics*, vol. 7, no. 323, p. 11, 2016. 4, 5

[46] A. Dantcheva, C. Chen, and A. Ross, "Can facial cosmetics affect the matching accuracy of face recognition systems?" in *2012 IEEE Fifth international conference on biometrics: theory, applications and systems (BTAS)*. IEEE, 2012, pp. 391–398. 5

[47] K. Kotwal, Z. Mostaani, and S. Marcel, "Detection of age-induced makeup attacks on face recognition systems using multi-layer deep features," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 1, pp. 15–25, 2019. 5

[48] G. Bingham, B. Yip, M. Ferguson, and C. Nansalo, "MORPH-II: Inconsistencies and Cleaning," *University of North Carolina Wilmington NSF REU*, 2017. 6

[49] V. A. Kai Zhang and K. W. Bowyer, "A method for curation of web-scraped face image datasets," in *International Workshop on Biometrics and Forensics (IWBF)*, 2020. 6

[50] Z. Xiong, Z. Wang, C. Du, R. Zhu, J. Xiao, and T. Lu, "An asian face dataset and how race influences face recognition," in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 372–383. 6

[51] Q. Cao, L. Shen, W. Xie, O. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *Proceedings of the IEEE international conference on automatic face & gesture recognition*. IEEE, 2018, pp. 67–74. 6, 7

[52] I. Hupont and C. Fernández, "Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition," in *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*. IEEE, 2019, pp. 1–7. 6

[53] M. Wang and W. Deng, "Mitigating bias in face recognition using skewness-aware reinforcement learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9322–9331. 6, 10, 11, 12

[54] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana, "Sensitivenets: Learning agnostic representations with application to face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2158–2164, 2020. 6

[55] A. P. Founds, N. Orlans, W. Genevieve, and C. I. Watson, "Nist special databse 32-multiple encounter dataset ii (meds-ii)," 2011. 6, 7

[56] J. P. Robinson, C. Qin, Y. Henon, S. Timoner, and Y. Fu, "Balancing biases and preserving privacy on balanced faces in the wild," *IEEE Transactions on Image Processing*, 2023. 6, 7

[57] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, "Face recognition: too bias, or not too bias?" in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, 2020, pp. 0–1. 6, 7, 10, 12, 13

[58] J. Muhammad, Y. Wang, C. Wang, K. Zhang, and Z. Sun, "Casia-face-africa: A large-scale african face image database," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3634–3646, 2021. 6, 7

[59] H. Liang, P. Perona, and G. Balakrishnan, "Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4977–4987. 6, 7

[60] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Proceedings of 14th European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 87–102. 6

[61] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4873–4882. 6

[62] S. Fu, H. He, and Z.-G. Hou, "Learning race from face: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 12, pp. 2483–2509, 2014. 7

[63] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 34–42. 7

[64] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. de Mello, O. Gallo, L. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, "Efficient geometry-aware 3d generative adversarial networks," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 102–16 112. 7

[65] P. Grother, "Face recognition vendor test (frvt) part 8: Summarizing demographic differentials," *National Institute of Standards and Technology (NIST)*, vol. 8429, 2022. 8

[66] T. de Freitas Pereira and S. Marcel, "Fairness in biometrics: a figure of merit to assess biometric verification systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 1, pp. 19–29, 2021. 8, 9

[67] M. Schuckers, S. Purnapatra, K. Fatima, D. Hou, and S. Schuckers, "Statistical methods for assessing differences in false non-match rates across demographic groups," in *International Conference on Pattern Recognition*. Springer, 2022, pp. 570–581. 9

[68] K. Kotwal and S. Marcel, "Fairness Index Measures to Evaluate Bias in Biometric Recognition," in *Proceedings of the International Conference on Pattern Recognition*. Springer, 2022, pp. 479–493. 9

[69] J. J. Howard, E. J. Laird, R. E. Rubin, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Evaluating proposed fairness models for face recognition algorithms," in *International Conference on Pattern Recognition*. Springer, 2022, pp. 431–447. 9

[70] A. Elobaid, N. Ramoly, L. Younes, S. Papadopoulos, E. Ntoutsi, and I. Kompatsiaris, "Sum of group error differences: A critical examination of bias evaluation in biometric verification and a dual-metric measure," *arXiv preprint arXiv:2404.15385*, 2024. 9

[71] I. Solano, A. Peña, A. Morales, J. Fierrez, R. Tolosana, F. Zamora-Martinez, and J. S. Agustin, "Comprehensive equity index (cei): Definition and application to bias evaluation in biometrics," *arXiv preprint arXiv:2409.01928*, 2024. 9

[72] E. Villalobos, D. Mery, and K. Bowyer, "Fair face verification by using non-sensitive soft-biometric attributes," *IEEE Access*, vol. 10, pp. 30 168–30 179, 2022. 9, 10

[73] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. M. Moreno, J. Fierrez, and A. Kuijper, "A comprehensive study on face recognition biases beyond demographics," *IEEE Transactions on Technology and Society*, vol. 3, no. 1, pp. 16–30, 2021. 10

[74] S. Gong, X. Liu, and A. Jain, "Jointly de-biasing face recognition and demographic attribute estimation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*. Springer, 2020, pp. 330–347. 10, 11, 12

[75] K. Kotwal and S. Marcel, "Mitigating demographic bias in face recognition via regularized score calibration," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, January 2024, pp. 1150–1159. 10, 11, 12

[76] P. Melzi, R. Tolosana, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, I. DeAndres-Tame, A. Morales, J. Fierrez, J. Ortega-Garcia *et al.*, "Frcsyn-ongoing: Benchmarking and comprehensive evaluation of real and synthetic data to improve face recognition systems," *Information Fusion*, vol. 107, p. 102322, 2024. 10, 14

[77] ——, "Frcsyn challenge at wacv 2024: Face recognition challenge in the era of synthetic data," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 892–901. 10, 14

[78] I. DeAndres-Tame, R. Tolosana, P. Melzi, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, A. Morales, J. Fierrez, J. Ortega-Garcia *et al.*, "Frcsyn challenge at cvpr 2024: Face recognition challenge in the era of synthetic data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3173–3183. 10, 14

[79] R. Singh, P. Majumdar, S. Mittal, and M. Vatsa, "Anatomizing bias in facial analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 12 351–12 358. 10

[80] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro, "Bias mitigation for machine learning classifiers: A comprehensive survey," *ACM Journal on Responsible Computing*, vol. 1, no. 2, pp. 1–52, 2024. 10

[81] D. Deb, N. Nain, and A. K. Jain, "Longitudinal study of child face recognition," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 225–232. 10, 12

[82] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823. 10

[83] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0. 10, 12

[84] P. Wang, F. Su, Z. Zhao, Y. Guo, Y. Zhao, and B. Zhuang, "Deep class-skewed learning for face recognition," *Neurocomputing*, vol. 363, pp. 35–45, 2019. 10, 12

[85] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Feature transfer learning for face recognition with under-represented data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5704–5713. 10, 12

[86] K. Kotwal and S. Marcel, "Demographic fairness transformer for bias mitigation in face recognition," in *2024 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2024, pp. 1–10. 10, 12, 15

[87] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus, "Uncovering and mitigating algorithmic bias through learned latent structure," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 289–295. 11, 12

[88] S. Gong, X. Liu, and A. Jain, "Mitigating face recognition bias via group adaptive classifier," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3414–3424. 11, 12

[89] J. Liang, Y. Cao, C. Zhang, S. Chang, K. Bai, and Z. Xu, "Additive adversarial learning for unbiased authentication," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 428–11 437. 11, 12

[90] Y. Li, Y. Sun, Z. Cui, S. Shan, and J. Yang, "Learning fair face representation with progressive cross transformer," *arXiv preprint arXiv:2108.04983*, 2021. 11, 12

[91] S. Park, J. Lee, P. Lee, S. Hwang, D. Kim, and H. Byun, "Fair contrastive learning for facial attribute classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 389–10 398. 11, 12

[92] F. Zhang, K. Kuang, L. Chen, Y. Liu, C. Wu, and J. Xiao, "Fairness-aware contrastive learning with partially annotated sensitive attributes," in *Proceedings of the International Conference on Learning Representations*, 2023. 12

[93] D. Michalski, S. Y. Yiu, and C. Malec, "The impact of age and threshold variation on facial recognition algorithm performance using images of children," in *2018 international conference on biometrics (icb)*. IEEE, 2018, pp. 217–224. 12

[94] N. Srinivas, K. Ricanek, D. Michalski, D. S. Bolme, and M. King, "Face recognition algorithm bias: Performance differences on images of children and adults," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0. 12, 13

[95] P. Terhörst, M. L. Tran, N. Damer, F. Kirchbuchner, and A. Kuijper, "Comparison-level mitigation of ethnic bias in face recognition," in *Proceedings of the International Workshop on Biometrics and Forensics*. IEEE, 2020, pp. 1–6. 12, 13

[96] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Post-comparison mitigation of demographic bias in face recognition using fair score normalization," *Pattern Recognition Letters*, vol. 140, pp. 332–338, 2020. 12, 13

[97] Y. Linghu, T. de Freitas Pereira, C. Ecabert, S. Marcel, and M. Günther, "Score normalization for demographic fairness in face recognition," in *2024 IEEE International Joint Conference on Biometrics (IJCB)*, 2024, pp. 1–11. 12, 13

[98] B. Liu, S. Zhang, G. Song, H. You, and Y. Liu, "Rectifying the data bias in knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1477–1486. 13

[99] M. Paganini, "Prune responsibly," *arXiv preprint arXiv:2009.09936*, 2020. 13

[100] E. Iofinova, A. Peste, and D. Alistarh, "Bias in pruned vision models: In-depth analysis and countermeasures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 364–24 373. 13

[101] X. Lin, S. Kim, and J. Joo, "Fairgrape: Fairness-aware gradient pruning method for face attribute classification," in *European Conference on Computer Vision*. Springer, 2022, pp. 414–432. 13

[102] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2704–2713. 13

[103] A. Atzori, G. Fenu, and M. Marras, "Demographic bias in low-resolution deep face recognition in the wild," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 3, pp. 599–611, 2023. 14

[104] A. George and S. Marcel, "Digi2real: Bridging the realism gap in synthetic data face recognition via foundation models," *arXiv preprint arXiv:2411.02188*, 2024. 14

[105] M. Huber, A. T. Luu, F. Boutros, A. Kuijper, and N. Damer, "Bias and diversity in synthetic-based face recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 6215–6226. 14

[106] K. Fatima, M. Schuckers, G. Cruz-Ortiz, D. Hou, S. Purnapatra, T. Andrews, A. Neupane, B. Marshall, and S. Schuckers, "A large-scale study of performance and equity of commercial remote identity verification technologies across demographics," in *2024 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2024, pp. 1–8. 14