Mitigating Demographic Bias in Face Recognition via Regularized Score Calibration

Ketan Kotwal[1] *and* Sébastien Marcel[1,2]

[1] Idiap Research Institute, Switzerland

[2] University of Lausanne, Switzerland

{ketan.kotwal, sebastien.marcel}@idiap.ch

Abstract

Demographic bias in deep learning-based face recognition systems has led to serious concerns. Several existing works attempt to mitigate bias by incorporating demographic-specific processing during inference, which requires knowledge or learning of demographic attribute with an additional cost. We propose to regularize training of the face recognition CNN, for demographic fairness, by imposing constraints on the distributions of matching scores. Our regularization term enforces the score distributions from different demographic groups to respect a pre-defined probability distribution, as well as it penalizes misalignment of distributions across demographic groups. The proposed method improves fairness of face recognition models without compromising the recognition accuracy, and does not require extra resources during inference. Our experiments indicate that in a cross-dataset testing, the regularized CNN can reduce the variation in accuracies (i.e., more fairness) of different demographic groups up to 25% while slightly improving recognition accuracy over baselines.

1. Introduction

Demographic bias in face recognition (FR)– which implies that certain demographic groups may experience unequal treatment or discrimination– has emereged as a serious issue in FR [12, 37, 42]. The disparity in recognition performance often leads to negative consequences for individuals from underrepresented groups, such as misidentification and limited access to important services or opportunities. [3,5,7,23,43]. This biased behavior of FR systems, is thus, not only a technical but also a social and societal concern. A detailed survey conducted by NIST FRVT on various commercial FR algorithms revealed significant differences in performance across different demographic groups, particularly concerning gender and race [18]. The prevalent nature of biased FR systems has led researchers to boost the efforts towards assessment and mitigation of bias in FR systems.

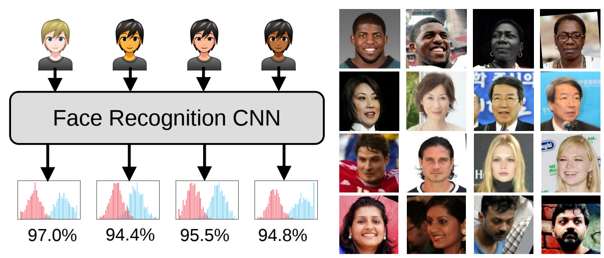


Figure 1. Non-equitable performance of recognition accuracy across different demographic groups in RFW dataset [46]. Demographic bias is apparent despite training the FR CNN on demographically balanced dataset. (The score distributions are illustrative, and do not refer to real dataset).

Several factors contribute to demographic bias in FR systems, wherein the use of imbalanced training data is one of the major factors [8, 16, 25]. Most publicly available training datasets have a skewed representation towards certain demographic groups, such as white men, while other groups (typically African and women) are underrepresented [14, 24]. Figure 1 shows a pictorial representation of issues of demographic fairness in FR along with samples of face images from different races/ ethnic groups from the RFW dataset [46]. In addition to gender and race, the age of the subject has also been shown to induce bias in FR [1, 50]. Recently, several works [9, 28, 49] have demonstrated that bias in FR might also arise from the data acquisition process.

Several works consider use of demographically balanced training data to mitigate bias in FR systems. However, recent works have shown that the use of balanced training data alone is not sufficient to mitigate bias completely [25]. For instance, an FR CNN with ResNet-50 architecture is trained with BUPT-BalancedFace dataset [45], when evaluated on RFW dataset exhibits non-equal recognition accuracies for each of the four demographics groups (numbers provided in Figure 1). Some works have proposed post-processing techniques that focus on normalizing the score distributions across different demographic groups to ensure fairness in FR outcomes [22,41]. The Z-norm and T-norm– popular score normalization techniques in biometrics– have also been considered to improve the performance of biometric systems by aligning score distributions [22, 32]. While the idea of score normalization is simple and appealing in the context of fairness, it requires knowledge of the demographic label of the data which either needs to be available a priori or has to be explicitly inferred from the input / face feaures at inference time from an additional classifier, typically a CNN. The additional resource leads to a complex system and adds computational cost and memory requirements during deployment. Training of a separate demographic classifier, too, requires significant computational and memory resources.

We incorporate the concept of score normalization as a regularization term and reformulate the objective function for training a FR CNN. We constrain the output scores of mated and non-mated pairs to follow a pre-defined distribution irrespective of the demographic groups of probe or gallery subjects. This regularization term aims to minimize the differences in score distributions across different demographic groups, thus promoting fairness in recognition score distributions as well as in (binarized) decisions. As our overall objective function consists of both the classification loss and the score regularization term, we simultaneously optimize for both the recognition accuracy and the demographic fairness in the score distributions. The training of the proposed method requires a negligible amount of extra computational and memory resources compared to training a separate demographic classifier. Since we do not modify the architecture of the FR CNN (only weights adapted to the new objective function), the overall inference pipeline remains unchanged.

The contributions of our work can be summarized as follows:

- We propose a regularization-based approach to mitigate demographic bias in FR systems by incorporating score normalization-based regularization term.
- Unlike many bias mitigation methods, our method does not require a separate classifier or additional computational resources, making it more efficient and practical for deployment.
- With the intra- and inter-demographic regularization terms, our work focuses on improving both aspects of fairness (differential performance and differential outcomes [20])– whereas many existing bias mitigation works solely focus on the latter.
- We evaluated performance of the proposed regularization method on three datasets and three backbone FR CNNs for in- and cross-dataset setups. Our experimental results demonstrate improvement in demographic fairness, without compromising recognition accuracy.

In Section 2, we present a brief review of recent works in mitigation of demographic bias in FR. We discuss the proposed regularization-based method in Section 3, followed by experimental results in Section 4. Finally, Section 5 provides conclusions.

2. Related Work

We begin by explaining how the notions of *fairness* are applied to general biometric systems. We then briefly review recent methods specifically designed to mitigate bias in FR systems.

2.1. Fairness in Biometrics

The concept of fairness in the biometric community is derived from the machine learning literature, and it aims to ensure equitable treatment of individuals across different demographic groups for biometric systems using trait such as face, fingerprint, or iris [37, 41]. Broadly speaking, the demographic fairness encompasses three main notions: parity, equalized odds, and sufficiency [10, 33].

The concept of parity refers to the requirement that decision of an FR system should be unaffected by demographic attributes (such as gender or ethnicity) of the subject. Equalized odds implies that regardless of demographic attributes, rates of false negatives and false positives should be the same for all demographic groups. The notion of sufficiency indicates that the available data attributes should contain enough information to ensure accurate and fair results in FR without relying on demographic details.

2.2. Methods for Bias Mitigation in FR

The existing works on bias mitigation in FR can be categorized into three main approaches: data-processing, model-based, and post-processing.

Data processing methods aim to address bias in FR systems by modifying the training data. Kortylewski *et al.* considered synthetic data for pretraining the FR CNN and then fine-tuning it with real data to mitigate the bias (related to yaw/pose, not demographic) [26]. In [47], Wang *et al.* proposed a large-margin feature augmentation technique to balance class distributions within FR systems. In [51], a feature transfer method was discussed to enhance the feature space of under-represented individuals to address the disparity between their distribution and that of more commonly represented individuals in FR datasets.

To address bias in FR, Gong *et al.* proposed a training-based approach that utilizes adversarial techniques to extract distinct feature representations [16]. A race balance network, based on reinforcement learning, was proposed in [45] which adjusts margins for demographics to promote balanced performance across different races. This work also introduced the BUPT-GlobalFace and BUPT-Balancedface datasets to facilitate further research in this area. In [17], a group-adaptive training methodology is presented that incorporates adaptive convolution kernels and attention mechanisms into FR CNN backbones. Li *et al.* regarded debiasing as a signal-denoising problem and developed a progressive cross-transformer architecture designed specifically for fair FR by removing identity-unrelated components induced by race from identity-related ones [29]. In [48], Wang *et al.* developed a sampling strategy to address bias during training with a primary focus on gender. Gong *et al.* introduced an adversarial network for debiasing that includes one identity classifier and three demographic classifiers (gender, age, race) to achieve unbiased FR [15]. A two-stage method for adversarial mitigation of bias through disentangled representations and additive adversarial learning was proposed in [30]. Huang *et al.* proposed a cluster-based large-margin local embedding approach to reduce the effect of local data imbalance and thus, also at reducing bias coming from unbalanced training data [21]. Recent works in [35, 52] have considered contrastive loss-based training with an objective of improving the intra-class similarity and reducing the similarity between negative samples. In this context, samples with same sensitive attributes, but different target classes are considered as negative.

For mitigation of age-related bias in FR at score-level, Srinivas *et al.* used ensemble approaches for merging the scores of multiple models [40]. Terhöst *et al.* proposed the Fair Template Comparison (FTC) method which replaces the computation of the cosine similarity score by a shallow network trained using cross-entropy loss [44]. Some works in [38, 41] used score calibration or normalization to mitigate bias in FR.

3. Bias Mitigation via Regularization

In this section, we describe the proposed bias mitigation method for FR systems based on deep CNNs. Obtaining large-scale dataset with perfectly balanced demographic attributes is challenging, time-consuming, and noisy due to requirement of manual efforts. Additionally, having equally balanced demographics does not necessarily lead to unbiased FR models [25]. Therefore, it is necessary to regularize the training using some explicit criteria for fairness. First, we briefly describe procedure for training a typical FR CNN. Then we introduce our hypothesis for inducing fairness constraint, and develop the regularized loss function.

3.1. Training Regular FR CNN

Consider an FR CNN, $f(.,\theta)$ where θ represent the learnable parameters of the model. We denote $\mathbf{x}_i \in \mathbb{R}^{3hw}$, $y_i \in \mathbb{N}_0$, and $d_i \in D$ (such that $D \subset \mathbb{N}_0$) as the triplet from training data representing an RGB face image, identity label, and demographic attribute, respectively. The training procedure often considers FR as a classification problem where the subject's identity label acts as the ground truth or target, and a suitable classification loss function, $\mathcal{L}_{\text{cls}}$, is minimized via Stochastic Gradient Descend (SGD) as given by Equation 1. The left part of Figure 2 summarizes training of a typical FR CNN.

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{cls}}\big(f(x_i, \theta), y_i\big) \,. \tag{1}$$

In many FR systems, the feature representations are normalized and thus, are constrained to lie on a hyperspherical manifold. The cosine distance, as presented in Equation 2, between feature representations of two samples- $\mathbf{x}_j$ and $\mathbf{x}_k$, (usually one from pre-defined gallery, and one from probe or test sample) acts as the matching score.

$$s(f(\mathbf{x}_j), f(\mathbf{x}_k)) = \cos\Big(f(\mathbf{x}_j), f(\mathbf{x}_k)\Big). \tag{2}$$

Most state-of-the-art FR systems employ an extension of typical cross-entropy loss such as ArcFace [11], SphereFace [31], ElasticFace [4], etc. These loss functions add angular margins to the feature representations of different classes (subjects, in the present case), and have resulted in better recognition accuracy. However, it is worth noting that none of these cost functions make use of the demographic attribute, d_i.

3.2. Demographic Calibration for Fairness

For an FR system to be fair, the score distributions of mated and non-mated pairs of different demographic groups must be equally separable under single decision threshold. In ML literature, this requirement is also referred to as equalized odds estimator [19, 33]. Given an FR CNN $f(.,\theta)$, we hypothesize that the possible causes of demographic biases are: (*a*) the distribution of f(matching scores) for some demographic groups might exhibit a multi--modal behaviour (one would ideally expect a bi-modal distribution: one for mated scores and another for non-mated ones), and (*b*) non-alignment of distribution of matching scores of different demographic groups.

The first factor refers to the distribution of intra-demographic scores. The correction for suppression of possible biases arising due to multimodal intra-demographic distribution can be approached by constraining $f(.,\theta)$ to respect a particular prior probability distribution function. If FR can be regarded as a binary classification, the distribution of f for a given demographic group should be bi-modal. If these distributions are constrained to be Gaussian (which is often the case for several FR datasets/ networks), for a training batch $\mathcal{B} \equiv \{f(\mathbf{x}_i, \theta) | i \in [1, B]\}$, to be mini-

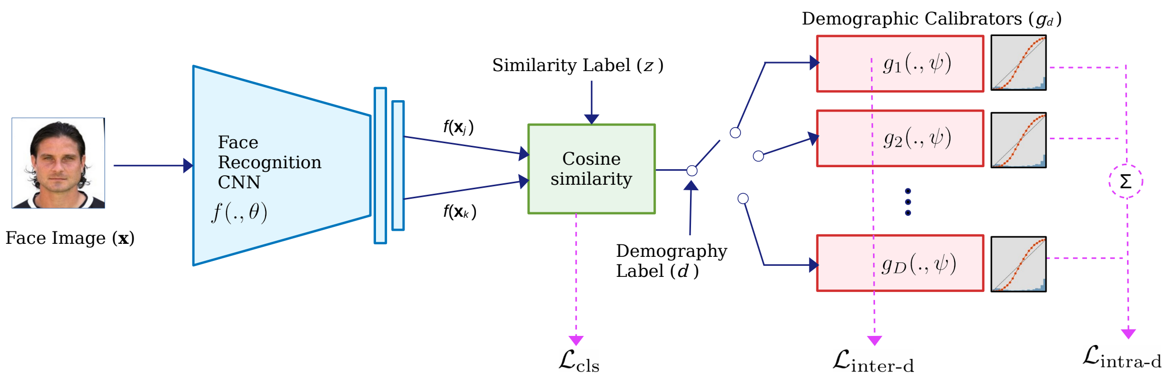


Figure 2. Training procedure for our proposed regularization method for demographic fairness in FR. The regular classification loss $\mathcal{L}_{\text{cls}}$ is regularized by intra- and inter-demographic loss terms ($\mathcal{L}_{\text{intra-d}}$ and $\mathcal{L}_{\text{inter-d}}$) that impose matching scores from each demographic group to follow specific distribution and to be aligned across demographic groups.

mized via SGD, the above constraint can be imposed using Kullback-Leibler (KL) divergence as follows:

$$\min \text{KL}[\mathcal{N}(k_c, 1) || \mathcal{N}(\mu_c, \sigma_c)], \quad c \in [\text{mated, non-mated}] \tag{3}$$

where μ_c and σ_c are respectively the estimated mean and the standard deviation of scores of mated or non-mated pairs from the batch $\mathcal{B}$.

The FR CNNs are typically trained in a contrastive learning framework. Thus, the constraint from Equation 3, too, needs to be reformulated in a contrastive framework. We accomplish this by means of Platt scaling—which is a popular method that transforms classification outputs into probability distributions [34,36,39]. Smola *et al.* [39] demonstrated the use of Platt scaling towards transforming raw matching scores into probability estimates via one-variable logistic regression. Equation 4 describes the Platt scaling function S that yields the probability of score (s) being the positive class (in biometric systems, typically mated scores are considered to be the positive class).

$$g \equiv P(\text{mated} \,|\, s) = \frac{1}{1 + e^{(\psi_a s + \psi_b)}} \tag{4}$$

Here, ψ_a and ψ_b (hereafter, ψ) are the parameters of the scaling function which are obtained by maximum likelihood (ML) estimation from scores obtained from the training set as:

$$-E\big[z_{jk} \log S(s_{jk}) + (1 - z_{jk}) \log(1 - S(s_{jk}))\big], \tag{5}$$

where s_{jk} is the matching score between two feature representations, and z_{jk} indicates the similarity label for contrastive learning. It is set to 1 if the constituent features belong to the same class (or identity), otherwise it is set to zero. While Platt scaling does not directly lead to fairness in classification models, we use this mechanism to impose mated as well as non-mated scores to follow specific distributions by computing the regression loss (as presented in Eq. 5). For a fixed (*i.e.* frozen) FR CNN, the parameters of Platt scaling are learnt for scores of a given demographic group; and subsequently the logistic regression loss, for fixed values of ψ is used to quantify the degree of mismatch in the scores of the corresponding demographic from the predefined distribution. For each constituent demographic group, we define the intra-demographic loss $\mathcal{L}_{\text{intra-d}}$ by incorporating Platt scaling function in the contrastive form as:

$$\mathcal{L}_{\text{intra-d}}(f(\mathbf{x}_j), f(\mathbf{x}_k), z_{jk}) = z_{jk} \log g(f(\mathbf{x}_j), f(\mathbf{x}_j)) + (1 - z_{jk}) \log(1 - g(f(\mathbf{x}_j), f(\mathbf{x}_k))) \tag{6}$$

Calculation of the intra-demographic loss is depicted in (right part of) Figure 2– where for a demographic calibrator is chosen based on the demographic labels of the pair of samples. The intra-demographic loss helps constraining scores to a particular distribution for each demographic. However, it does not address the issue of shifts between distributions of scores of different demographic groups. While this loss term, by clustering scores, has improved the recognition performance of each demographic group separately, we would still require different score thresholds for each demographic group for optimal classification. The use of such thresholds requires accurate knowledge of demographic label of each sample at run-time. It, thus, possibly requires a separate demographic classifier– which increases computational requirements during training and testing too. We propose to incorporate an inter-demographic loss component which penalizes large differences between intra-demographic loss values of different demographic groups.

We define the inter-demographic loss as the degree of variation between parameters ψ of each of the demographic calibrators, $g_i,\ i = 1, 2, \ldots, D$. This term, $\mathcal{L}_{\text{inter-d}}$ is defined in Equation 7.

$$\mathcal{L}_{\text{inter-d}} = \text{Var}([\psi_1, \psi_2, \ldots, \psi_D]). \tag{7}$$

The inter-demographic loss ensures that the parameters of each demographic calibrator– which in turn define the shape and location of score distributions– are aligned as much as possible.

Equation 8 provides the overall loss by combining Equations 1, 6, and 7.

$$\begin{aligned}\theta^*, \psi^* = \arg\min_{\theta,\psi} \frac{1}{N}\sum_{i=1}^{N}\Bigg[&\mathcal{L}_{\text{cls}}(f(\mathbf{x}_i,\theta); y_i)\\ &+ \lambda_{\text{inter-d}}\mathcal{L}_{\text{inter-d}}([\psi_1,\psi_2,...,\psi_D])\\ &+ \lambda_{\text{intra-d}}\sum_{d=1}^{D}\mathcal{L}_{\text{intra-d}}(f(\mathbf{x}_j,\theta), f(\mathbf{x}_k,\theta), z_{jk}, d_{jk})\Bigg],\end{aligned} \tag{8}$$

where $\lambda_{\text{inter-d}}$ and $\lambda_{\text{intra-d}}$ are the relative weights for corresponding loss terms. During training, we alternate between fixing the calibrators, and finetuning the FR CNN. First, the calibrators (ψ_d) are evaluated for frozen FR CNN. In the next set of training epochs, FR CNN (θ) is finetuned to improve the demographic fairness without compromising recognition accuracy as governed by the classification loss $\mathcal{L}_{\text{cls}}$. The overall training pipeline, along with computation of each loss term, is shown in Figure 2.

Both losses from equations 6 and 7 can be trivially expanded to work with any deep FR CNN architecture. On training or finetuning the FR CNN with either of the proposed regularized loss functions, the comparison of mated and non-mated scores from samples of the same demographic are inherently aligned (centered around the same value), allowing us to better set single decision thresholds for a fair behaviour without any post-processing.

4. Experimental Results

We first provide details related to experimental setup and then discuss results of proposed regularization method on different datasets.

4.1. Experimental Setup

Datasets: For our experimental analysis of the demographic regularization method, we utilized three publicly available FR datasets that provide race or ethnicity information. The first dataset is a subset of the VGGFace2 [6] dataset by Cao *et al.*, where we specifically considered 10 samples per subject identity, resulting in a training set with 86,310 samples. Our second dataset- MORPH dataset [2] comprises approximately 55,000 mugshot images of subjects from four races: Black, White Asian, and Hispanic (with few additional samples from other races). Notably, this dataset exhibits significant skewness in terms of demographic distribution; black subjects account for nearly 75% of the data while Asian subjects constitute less than 1%. Lastly, we incorporated the RFW dataset for our experiments– it has a well-balanced protocol for four demographic groups with 6,000 comparisons per group–resulting in a total of 24 k comparisons [46].

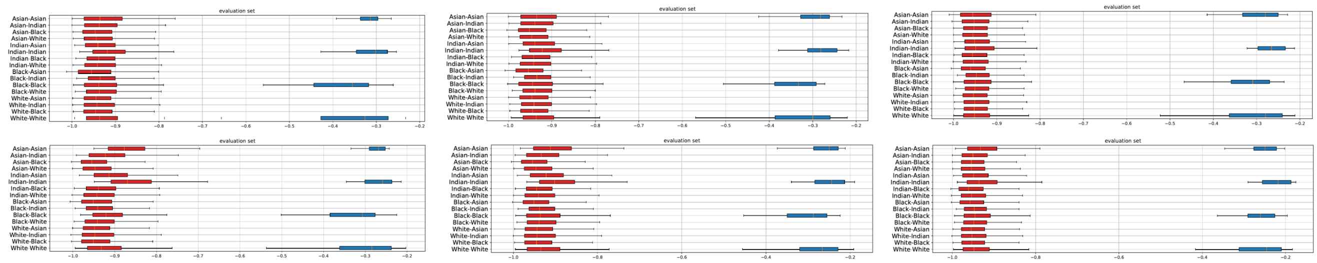
FR CNN Backbones: We worked with the FR CNN based on the iResNet architecture with either 34, 50, or 100 layers [13]. These models were trained using ArcFace loss on the MS1MV3 dataset, which is a refined version of the MS-Celeb1M dataset. The architecture and pretrained weights for our models were obtained from the InsightFace repository[1].

FR Pipeline:[2] For consistent experiments, each combination of dataset and backbone underwent a standardized preprocessing procedure. This involved using MTCNN for initial face detection and facial landmark identification. The resulting 5 landmarks were then used to align and resize the face region to meet the specified requirement of 112×112 pixels, as required by each iResNet-based FR CNN architecture. The FR CNN received the aligned, fixed size input image and produced a 512-d feature vector. The matching score was determined based on cosine similarity. To enhance the fairness of the FR CNN through fine-tuning, an SGD-based optimizer with initial learning rates ranging from 1e-4 to 1e-2 was utilized. A rate scheduler was implemented, decreasing the learning rate by a factor of 0.25 if no improvements were observed for five epochs (also known as patience). For iResNet34 and iResNet50 architectures, a batch size of 128 was employed while for iResNet100, a batch size of 64 was used during SGD-based optimization. Since contrastive setup is required by the loss functions, we predetermined that there would be ten positive samples and five negative samples per subject. Fine-tuning took place over 80 epochs with early stopping criteria; this process alternated between calibrators and FR CNN in a ratio of 1:10 epochs.

Performance Evaluation: To measure recognition accuracy, we determined the score threshold on the train set considering the Equal Error Rate (EER). This threshold was then used to convert scores from the test set into binary decisions. Alongside recognition accuracy, we also report false accept rate and false reject rate for the test set, which indicate misclassifications of imposters and genuine samples respectively. To assess demographic fairness, we evaluated the recognition accuracy for each demographic group within a cohort. We then calculated the standard deviation and skewed error ratio (SER) of recognition accuracy per demographic group. The SER is calculated as the ratio of

[1]https://github.com/deepinsight/insightface

[2]The source code for replicating our work is available at https://gitlab.idiap.ch/bob/bob.paper.wacv2024_dvpba.



(a) FR CNN: iResNet34 (b) FR CNN: iResNet50 (c) FR CNN: iResNet100

Figure 3. Score distributions for pairs of subjects from different demographics of the VGGFace2 dataset for different FR CNN backbones. Red and Blue boxes represent boxplots of non-mated (imposter) and mated (genuine) scores respectively. For each plot, top row: baseline; bottom row: regularized FR CNN.

highest error rate to the lowest error rate among all demographic groups (SER $= \max \text{error}_d / \min \text{error}_d,\ d \in D$). A lower standard deviation (std) and SER value, coupled with higher recognition accuracy, indicates a more accurate and equitable FR system.

We establish a baseline by measuring accuracy and fairness without any additional processing or regularization. We could not implement some of the comparative methods (from Sec. 2.2) due to either non-availability of code or lack of information related to protocols (dataset splits, score thresholds, etc.). Thus, we evaluate the performance of two score normalization techniques: Z- and T-normalizations. These normalization techniques are used to center the distribution of impostor scores for each demographic group around zero, which can help improve fair behavior of FR CNNs by allowing for a single decision threshold. Finally, we present the matching scores obtained from the regularized FR CNNs.

4.2. Results of Regularization Experiments

Results on VGGFace2: In the initial experiment, we trained and evaluated the FR CNN using different partitions of the VGGFace2 dataset. The calibration loss within demographic groups was unbalanced (*i.e.*, the loss value from each demographic group was equally weighted irrespective of group's share in training partition). Table 1 shows the recognition accuracy and demographic fairness results for both the baseline (non-calibrated) and score-calibrated FR CNNs, as well as accuracy values specifically for each of the four demographic groups. We also provide the corresponding metrics for Z- and T-normalizations as these methods are effective and, in some sense, close to the underlying principle of our method. Overall recognition accuracy increased by 0.63%, 0.12%, and 0.33% respectively for FR CNNs with 34, 50, and 100 layers after proposed regularization. There were slight improvements in accuracy for almost each demographic group for the proposed method. At the same time, the variation in recognition accuracy among different demographics decreased as indicated by the reduced standard deviation (std) and skewed error rate (SER) metrics as listed in the table. The decrease in variation suggests that the FR CNN treats samples from different races/ethnic groups more fairly after calibrated regularization. Figure 3 provides the boxplots of score distributions for pairwise demographics, with colored boxes representing scores within the first and third quartiles (*i.e.*, $Q3$-$Q1$). It can be observed that mode (shown in colored boxes) for each demographic's scores are better aligned in regularized cases, especially for 50- and 100-layer FR backbones.

Method	FMR (↓)	FNMR (↓)	Avg Acc (↑)	A_acc	I_acc	B_acc	W_acc	std (↓)	SER (↓)
baseline	1.97	3.20	98.02	98.10	98.08	98.62	97.92	0.26	1.50
Z-norm	1.82	2.88	98.17	98.27	98.30	98.68	98.06	0.22	1.47
T-norm	1.93	3.28	98.07	98.19	98.30	98.69	97.93	0.27	1.58
Proposed	1.40	1.20	98.65	98.65	98.34	99.07	98.57	0.25	1.48

Method	FMR (↓)	FNMR (↓)	Avg Acc (↑)	A_acc	I_acc	B_acc	W_acc	std (↓)	SER (↓)
baseline	1.74	2.56	98.26	98.42	98.29	98.69	98.17	0.20	1.40
Z-norm	1.76	2.48	98.24	98.44	98.32	98.68	98.12	0.20	1.42
T-norm	1.70	2.56	98.30	98.41	98.53	98.74	98.19	0.20	1.44
Proposed	1.62	1.28	98.38	98.55	98.42	98.81	98.38	0.18	1.36

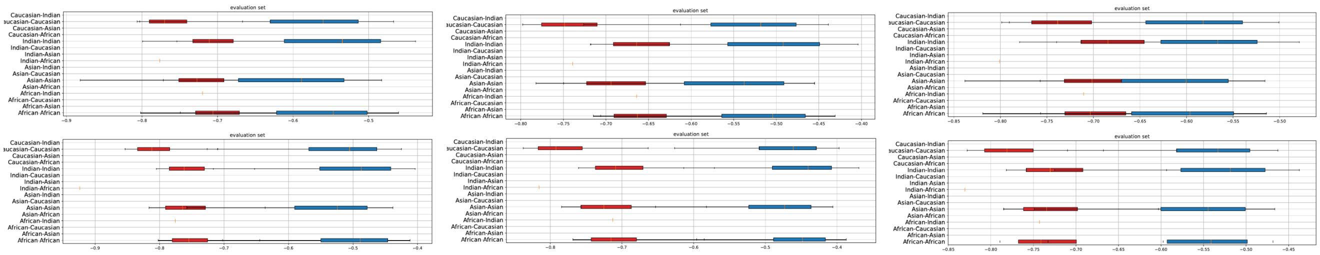
Method	FMR (↓)	FNMR (↓)	Avg Acc (↑)	A_acc	I_acc	B_acc	W_acc	std (↓)	SER (↓)
baseline	1.79	2.80	98.21	98.19	98.13	98.56	98.19	0.17	1.30
Z-norm	1.72	2.80	98.27	98.34	98.22	98.56	98.23	0.14	1.24
T-norm	1.88	2.80	98.12	98.15	98.26	98.53	98.05	0.18	1.32
Proposed	1.46	0.48	98.54	98.50	98.61	98.84	98.51	0.14	1.29

Table 1. Performance evaluation of the proposed method on VGGFace2 dataset. top: iResNet34, middle: iResNet50, and bottom: iResNet100 FR backbones. The FMR, FNMR, and all accuracy values are indicated as percentages.

Method	FMR (↓)	FNMR (↓)	Avg Acc (↑)	African_acc	Asian_acc	Caucasian_acc	Indian_acc	std (↓)	SER (↓)
baseline	8.69	8.68	91.32	89.58	89.60	95.27	90.82	2.34	2.20
Z-norm	11.08	11.07	88.92	86.21	87.53	93.00	88.96	2.54	1.97
T-norm	10.84	10.84	89.16	86.50	87.65	93.52	88.97	2.66	2.08
Proposed	7.46	7.46	92.54	91.48	91.71	95.85	91.12	1.92	2.13

Method	FMR (↓)	FNMR (↓)	Avg Acc (↑)	African_acc	Asian_acc	Caucasian_acc	Indian_acc	std (↓)	SER (↓)
baseline	3.29	3.29	96.71	96.44	95.48	98.17	96.75	0.96	2.46
Z-norm	4.63	4.63	95.37	95.14	93.63	97.32	95.38	1.31	2.37
T-norm	4.66	4.66	95.34	94.77	94.10	97.23	95.26	1.17	2.13
Proposed	2.24	2.24	97.76	97.79	97.17	98.72	97.36	0.60	2.20

Table 2. Performance evaluation of the proposed method on RFW dataset using VGGFace2 dataset to regularize FR backbone. top: iResNet50, bottom: iResNet100 FR backbone. The FMR, FNMR, and all accuracy values are indicated as percentages.



(a) Baseline

(b) Regularized with VGGFace2

(c) Regularized with MORPH

Figure 4. Score distributions for pairs of subjects from different demographics of the RFW dataset for different FR CNN backbones. Red and Blue boxes represent boxplots of imposter and genuine scores respectively. For each plot, top row: iResNet50 backbone; bottom row: iResNet100 backbone.

During the evaluation of the same FR CNN on the RFW dataset, we noticed a substantial decrease in accuracy and fairness metrics compared to the baseline performance. This deterioration can largely be attributed to the fact that, in the regularized model, weights were adjusted based on intra-demographic loss, which was influenced by the imbalanced demographic distribution of VGGFace2. Subsequently, we fine-tuned the FR CNNs on VGGFace2 with balanced demographic weights.

The performance of three FR CNNs improved when trained with balanced settings on the RFW dataset. For both backbones, 50- and 100-layered iResNets, the overall accuracy increased by nearly 1%, while reducing the standard deviation by 25% compared to the respective baseline numbers. Although there were not consistent improvements in recognition accuracy for individual demographic groups, regularizing the FR CNNs resulted in improving overall performance in terms of both accuracy and fairness. Table 2 shows evaluation of the performance on RFW dataset using VGGFace2 regularization with balanced settings. The left and middle columns of Figure 4 show the score distributions of pairwise demographics (the RFW protocol does not have cross-demographic pairs) for baseline and for FR CNN regularized with VGGFace2 dataset. While the tails of score distributions mated and non-mated pairs still overlap, the extent of overlap has reduced by the use of regularized FR CNN.

Results on MORPH: In Table 3, we present the results regularized FR CNN using different partitions of the MORPH dataset– which is highly imbalanced for ethnic demographics. Since the baseline CNNs already provide near-perfect recognition, this experiment does not shed much light in terms of qualitative performance metrics. However, it should be noted that the aspect of fairness is not only limited to disparity in differential outcome (classification decisions), but also to the differential performance (distributions of mated/ non-mated scores) [20, 27]. Hence, in addition to improved accuracy/ reduced std, a bias mitigation technique should also attempt to improve the score distributions towards specific desired properties [27]. Figure 5 shows the boxplots depicting the distribution of scores for different demographic groups. A comparison between score distributions of the baseline (top row in Figure) and those of the regularized FR CNN (bottom row) demonstrates improvements in this regard: a better alignment across demographic groups (for mated scores) and more compact distributions (shorter whisks) can be observed in most cases.

By regularizing the FR CNNs at learning rate of $1e$-4 with a balanced intra-demographic term, we observed im-

Method	FMR (↓)	FNMR (↓)	Avg Acc (↑)	A_acc	H_acc	B_acc	W_acc	std (↓)	SER (↓)
baseline	0.06	0.00	99.94	99.99	99.98	99.92	99.99	0.03	7.99
Z-norm	0.06	0.00	99.94	99.99	99.99	99.92	99.98	0.03	7.99
T-norm	0.06	0.02	99.94	100.00	99.99	99.92	99.98	0.03	-
Proposed	0.09	0.00	99.91	99.98	99.95	99.89	99.97	0.03	5.50

Method	FMR (↓)	FNMR (↓)	Avg Acc (↑)	A_acc	H_acc	B_acc	W_acc	std (↓)	SER (↓)
baseline	0.06	0.00	99.94	99.99	99.99	99.92	99.99	0.03	7.99
Z-norm	0.09	0.00	99.91	99.99	99.98	99.88	99.98	0.04	11.98
T-norm	0.09	0.00	99.91	99.99	99.99	99.88	99.97	0.04	11.98
Proposed	0.11	0.03	99.89	99.98	99.94	99.87	99.97	0.04	6.50

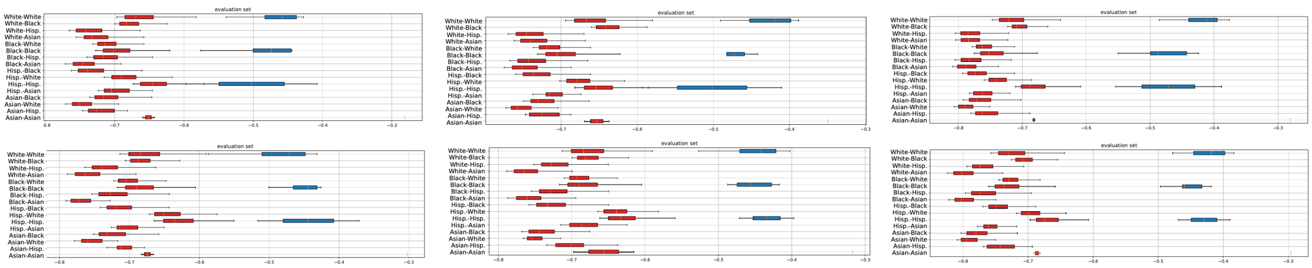
Method	FMR (↓)	FNMR (↓)	Avg Acc (↑)	A_acc	B_acc	H_acc	W_acc	std (↓)	SER (↓)
baseline	0.05	0.00	99.95	99.99	99.94	99.98	100.00	0.02	5.99
Z-norm	0.05	0.00	99.95	99.99	99.94	99.99	99.99	0.02	5.99
T-norm	0.05	0.00	99.95	100.00	99.94	99.99	99.99	0.02	-
Proposed	0.05	0.00	99.95	99.99	99.94	99.97	99.99	0.02	5.99

Table 3. Performance evaluation of the proposed method on MORPH dataset. top: iResNet34, middle: iResNet50, and bottom: iResNet100 FR backbones. The FMR, FNMR, and all accuracy values are indicated as percentages.

Method	FMR (↓)	FNMR (↓)	Avg Acc (↑)	African_acc	Asian_acc	Caucasian_acc	Indian_acc	std (↓)	SER (↓)
baseline	8.69	8.68	91.32	89.58	89.60	95.27	90.82	2.34	2.20
Z-norm	11.08	11.07	88.92	86.21	87.53	93.00	88.96	2.55	1.97
T-norm	10.84	10.84	89.16	86.50	87.65	93.52	88.97	2.67	2.08
Proposed	10.87	10.87	89.13	88.31	87.58	92.75	87.89	2.11	1.71

Method	FMR (↓)	FNMR (↓)	Avg Acc (↑)	African_acc	Asian_acc	Caucasian_acc	Indian_acc	std (↓)	SER (↓)
baseline	3.29	3.29	96.71	96.44	95.48	98.17	96.75	0.96	2.46
Z-norm	4.63	4.63	95.37	95.14	93.63	97.32	95.38	1.31	2.37
T-norm	4.66	4.66	95.34	94.77	94.10	97.23	95.26	1.17	2.12
Proposed	4.32	4.32	95.67	95.77	94.75	97.27	94.91	0.99	1.92

Table 4. Performance evaluation of the proposed method on RFW dataset using MORPH dataset to regularize FR backbones. top: iResNet50, bottom: iResNet100 FR backbones. The FMR, FNMR, and all accuracy values are indicated as percentages.



(a) FR CNN: iResNet34 (b) FR CNN: iResNet50 (c) FR CNN: iResNet100

Figure 5. Score distributions for pairs of subjects from different demographics of the MORPH dataset for different FR CNN backbones. Red and Blue boxes represent boxplots of imposter and genuine scores respectively. For each plot, top row: baseline; bottom row: regularized FR CNN.

provements in demographic fairness (decrease in std and SER) on RFW dataset as provided in Table 4. However, the regularization resulted in lowering the overall recognition accuracy by around 1%. We believe that using a larger learning rate may have been beneficial, but the corresponding experiment did not converge on training set. From Figure 4(c), it may be observed that the regularized FR CNN with iResNet100 backbone was able to improve the score distributions (*i.e.*, demographic fairness), however, other FR CNN was not capable of producing fair models. A better procedure, possibly with different learning rates for calibration and classification, may be required to work with datasets that have near-perfect recognition accuracy baselines.

5. Conclusion

In this work, we have developed a regularization-based approach to improve demographic fairness, primarily related to ethnicity or race, of an FR CNN without compromising its recognition accuracy. For this finetuning, we use score-calibrators for each demographic groups as a means to quantify the disparity in matching scores of demographic samples– which in turn acts as regularization term. This regularization or disparity consists of two components: one penalizes the scores of each demographic for not adhering to specific distribution, and another one related to misalignment of score distributions of different demographic groups. Our work, possibly for the fist time for FR, demonstrates how a popular concept of score calibration (typically a post-processing method) can be transformed into training-time regularization. Since the regularized FR CNN does not modify the interfaces or architecture of the baseline CNN, the inference pipeline does not require any changes from the baseline one. Also, the generic nature of regularization loss terms (Eqs.6, 7) implies that the proposed bias mitigation method can be easily extended to different FR CNNs. We have demonstrated the efficacy of the proposed method in in- and cross-dataset testing. Additionally, we have also demonstrated that the proposed regularization improves, not just classification accuracy, but also score distributions of mated and non-mated pairs of different demographic groups.

The initial success of the proposed method is encouraging, however, several factors influencing demographic fairness (and recognition accuracy) of an FR system are required to examined further. We would like to examine different aspects of calibration namely: impact of weighing / balancing demographic groups across different terms of loss function. It was observed that for FR CNNs that are highly accurate for datasets, the learning rate was a crucial factor. Further ablation studies in this regard (learning rate and scheduler) can throw light on jointly improving fairness and recognition accuracy of an FR CNN.

Acknowledgement

Authors would like to thank the Hasler foundation for their support through the SAFER project.

References

[1] Vítor Albiero, Kevin W. Bowyer, Kushal Vangara, and Michael C. King. Does face recognition accuracy get better with age? deep face matchers say no. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, volume 1, pages 250–258, 2020. 1

[2] G Bingham, B Yip, M Ferguson, and C Nansalo. MORPH-II: Inconsistencies and Cleaning. *University of North Carolina Wilmington NSF REU*, 2017. 5

[3] Melissa Bosque. Facial recognition bias frustrates Black asylum applicants to US, advocates say. https://www.theguardian.com/us-news/2023/feb/08/us-immigration-cbp-one-app-facial-recognition-bias, 2023. [Online; accessed 15-Oct-2023]. 1

[4] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on*

computer vision and pattern recognition, pages 1578–1587, 2022. 3

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, Feb 2018. 1

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Proceedings of the IEEE international conference on automatic face & gesture recognition*, pages 67–74. IEEE, 2018. 5

[7] Davide Castelvecchi. Is facial recognition too biased to be let loose? *Nature*, 587(7834):347–350, 2020. 1

[8] Valeriia Cherepanova, Steven Reich, Samuel Dooley, Hossein Souri, John Dickerson, Micah Goldblum, and Tom Goldstein. A deep dive into dataset imbalance and bias in face identification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 229–247, 2023. 1

[9] Cynthia Cook, John Howard, Yevgeniy B. Sirotin, Jerry Tipton, and Arun 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*, 1(1):32–41, 2019. 1

[10] Tiago de Freitas Pereira and Sébastien Marcel. Fairness in biometrics: a figure of merit to assess biometric verification systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(1):19–29, 2021. 2

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 3

[12] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020. 1

[13] Ionut Cosmin Duta, Li Liu, Fan Zhu, and Ling Shao. Improved residual networks for image and video recognition. In *Proceedings of the International Conference on Pattern Recognition*, pages 9415–9422. IEEE, 2021. 5

[14] Markos Georgopoulos, James Oldfield, Mihalis A Nicolaou, Yannis Panagakis, and Maja Pantic. Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. *International Journal of Computer Vision*, 129(7):2288–2307, 2021. 1

[15] Sixue Gong, Xiaoming Liu, and Anil K Jain. Debface: Debiasing face recognition. *arXiv preprint arXiv:1911.08080*, 2019. 3

[16] Sixue Gong, Xiaoming Liu, and Anil K Jain. Jointly debiasing face recognition and demographic attribute estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 330–347. Springer, 2020. 1, 2

[17] Sixue Gong, Xiaoming Liu, and Anil K Jain. Mitigating face recognition bias via group adaptive classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3414–3424, 2021. 3

[18] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019. 1

[19] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. 3

[20] John J. Howard, Yevgeniy B. Sirotin, and Arun Vemury. The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In *Proceedings of the International Conference on Biometrics Theory, Applications and Systems*, pages 1–8, 2019. 2, 7

[21] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2781–2794, 2019. 3

[22] Anil Jain, Karthik Nandakumar, and Arun Ross. Score normalization in multimodal biometric systems. *Pattern recognition*, 38(12):2270–2285, 2005. 2

[23] Thaddeus Johnson and Natasha Johnson. Police Facial Recognition Technology Can't Tell Black People Apart. https://www.scientificamerican.com/article/police-facial-recognition-technology-cant-tell-black-people-apart, 2023. [Online; accessed 15-Oct-2023]. 1

[24] Rie Kamikubo, Lining Wang, Crystal Marte, Amnah Mahmood, and Hernisa Kacorri. Data representativeness in accessibility datasets: A meta-analysis. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–15, 2022. 1

[25] Manideep Kolla and Aravinth 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*, pages 313–322, 2023. 1, 3

[26] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas 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*, pages 0–0, 2019. 2

[27] Ketan Kotwal and Sébastien Marcel. Fairness Index Measures to Evaluate Bias in Biometric Recognition. In *Proceedings of the International Conference on Pattern Recognition*, pages 479–493. Springer, 2022. 7

[28] K. Krishnapriya, Vítor Albiero, Kushal Vangara, Michael King, and Kevin Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020. 1

[29] Yong Li, Yufei Sun, Zhen Cui, Shiguang Shan, and Jian Yang. Learning fair face representation with progressive cross transformer. *arXiv preprint arXiv:2108.04983*, 2021. 3

[30] Jian Liang, Yuren Cao, Chenbin Zhang, Shiyu Chang, Kun Bai, and Zenglin Xu. Additive adversarial learning for unbiased authentication. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11428–11437, 2019. 3

[31] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 3

[32] Miranti Indar Mandasari, Manuel Günther, Roy Wallace, Rahim Saeidi, Sébastien Marcel, and David A van Leeuwen. Score calibration in face recognition. *IET Biometrics*, 3(4):246–256, 2014. 2

[33] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021. 2, 3

[34] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the International Conference on Machine learning*, pages 625–632, 2005. 4

[35] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10398, 2022. 3

[36] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 4

[37] Christian Rathgeb, Pawel Drozdowski, Dinusha Frings, Naser Damer, and Christoph Busch. Demographic fairness in biometric systems: What do the experts say? *IEEE Technology and Society Magazine*, 41(4):71–82, 2022. 1, 2

[38] Tiago Salvador, Stephanie Cairns, Vikram Voleti, Noah Marshall, and Adam Oberman. Bias mitigation of face recognition models through calibration. *arXiv preprint arXiv:2106.03761*, 2021. 3

[39] Alexander J Smola, Peter J Bartlett, Dale Schuurmans, and Bernhard Schölkopf. *Advances in large margin classifiers*. MIT press, 2000. 4

[40] Nisha Srinivas, Karl Ricanek, Dana Michalski, David S Bolme, and Michael 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*, pages 0–0, 2019. 3

[41] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognition Letters*, 140:332–338, 2020. 2, 3

[42] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, 3(1):16–30, 2021. 1

[43] Philipp Terhörst, Kevin Riehl, Naser Damer, Peter Rot, Blaz Bortolato, Florian Kirchbuchner, Vitomir Struc, and Arjan Kuijper. Pe-miu: A training-free privacy-enhancing face recognition approach based on minimum information units. *IEEE Access*, 8:93635–93647, 2020. 1

[44] Philipp Terhörst, Mai Ly Tran, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Comparison-level mitigation of ethnic bias in face recognition. In *Proceedings of the International Workshop on Biometrics and Forensics*, pages 1–6. IEEE, 2020. 3

[45] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9322–9331, 2020. 1, 2

[46] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai 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*, pages 692–702, 2019. 1, 5

[47] Pingyu Wang, Fei Su, Zhicheng Zhao, Yandong Guo, Yanyun Zhao, and Bojin Zhuang. Deep class-skewed learning for face recognition. *Neurocomputing*, 363:35–45, 2019. 2

[48] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020. 3

[49] Haiyu Wu, Vítor Albiero, K Krishnapriya, Michael King, and Kevin 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*, pages 1041–1050, 2023. 1

[50] Wang Yao, Muhammad Ali Farooq, Joseph Lemley, and Peter Corcoran. A study on the effect of ageing in facial authentication and the utility of data augmentation to reduce performance bias across age groups. *IEEE Access*, 2023. 1

[51] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5704–5713, 2019. 2

[52] Fengda Zhang, Kun Kuang, Long Chen, Yuxuan Liu, Chao Wu, and Jun Xiao. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *Proceedings of the International Conference on Learning Representations*, 2022. 3