Score Normalization for Demographic Fairness in Face Recognition

Yu Linghu[1] Tiago de Freitas Pereira[3] Christophe Ecabert[2] Sébastien Marcel[2] Manuel Günther[1]

[1]Department of Informatics, University of Zurich,
Andreasstrasse 15, CH-8050 Zurich
{yu.linghu,manuel.guenther}@uzh.ch
https://www.ifi.uzh.ch/en/aiml.html

[2]Idiap Research Institute,
Centre du Parc, Rue Marconi 19,
CH-1920 Martigny
{cecabert,marcel}@idiap.ch
https://www.idiap.ch

[3]ams OSRAM,
tiago.defreitaspereira@ams-osram.com
https://ams-osram.com/

Abstract

Fair biometric algorithms have similar verification performance across different demographic groups given a single decision threshold. Unfortunately, for state-of-the-art face recognition networks, score distributions differ between demographics. Contrary to work that tries to align those distributions by extra training or fine-tuning, we solely focus on score post-processing methods. As proved, well-known sample-centered score normalization techniques, Z-norm and T-norm, do not improve fairness for high-security operating points. Thus, we extend the standard Z/T-norm to integrate demographic information in normalization. Additionally, we investigate several possibilities to incorporate cohort similarities for both genuine and impostor pairs per demographic to improve fairness across different operating points. We run experiments on two datasets with different demographics (gender and ethnicity) and show that our techniques generally improve the overall fairness of five state-of-the-art pre-trained face recognition networks, without downgrading verification performance. We also indicate that an equal contribution of False Match Rate (FMR) and False Non-Match Rate (FNMR) in fairness evaluation is required for the highest gains. Code and protocols are available.[‡‡]

1. Introduction

The automatic identification and verification of facial images gained large attention in the last decades. With the advent of deep learning, many new methods [34, 58, 11, 38, 28] and facial image datasets [66, 6, 37, 61, 67] have been developed to train and evaluate deep learning methods. These methods have matured into being usable in security-relevant applications like automatic border control using e-gates [10].

[‡‡]https://github.com/AIML-IfI/score-norm-fairness

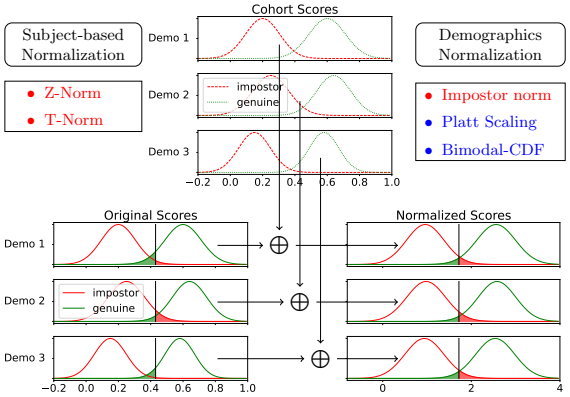

Figure 1. Improved Fairness Through Score Normalization. The original scores on the left have different False Match Rates (FMR, red area) and False Non-Match Rates (FNMR, green area) for different demographics under the same score threshold. Through modeling of score distributions from a cohort, we normalize scores such that they provide more similar FMR and FNMR across demographics, thereby improving demographic fairness. Normalization techniques in red text use cohort impostor scores only, blue ones also incorporate cohort genuine scores.

The applicability of these algorithms highly depends on the characteristics of the demographic groups in which they are employed. It was observed that *The Other Race Effect*, which is well-known in humans [35], can also be observed in Face Recognition (FR) algorithms. Since most large-scale datasets include mainly images from white people [61], and dataset biases are learned by deep learning algorithms [51, 1, 3], research has shown that algorithms perform very well on white male populations, but decrease performance on females and/or people of color [2, 32]. Consequently, most news media coverage that reports the wrong behavior of automatic FR algorithms finds the higher false negative rate in the latter demographics [50, 23]. Therefore, the

Face Recognition Vendor Test (FRVT) has a special report addressing demographic effects in FR [19], mostly observing the effect of ethnicity and gender on more than 100 Commercial-Off-The-Shelf (COTS) systems.

In face verification, a similarity value is computed between a previously enrolled gallery template (such as the face image stored in a passport) and a probe image, *e.g.* taken in an e-gate. A threshold is applied to this similarity value to decide whether the gallery template and the probe image come from the same identity. One major problem with fairness in biometrics is that the distributions of similarity scores differ between demographics, for example, the mated comparison of African people usually results in lower similarity scores than White [56]. Hence, a single threshold can have differential performance across demographics [25, 9].

Our proposed approaches to overcome this issue include score normalization techniques that have been successfully applied in FR [57, 36]. As shown in Fig. 1, the advantage of these techniques is that they can be applied to any existing FR system to improve fairness across different people and do not require any further network training. In one of these techniques, Z-norm [48], the score is normalized by a gallery-sample-specific distribution of similarities to cohort samples of different subjects. This method has been tested in [31]. In this paper, we modify these techniques to perform score normalization for a certain demographic group – instead of per sample – and show that the impostor score distribution can be normalized with this technique, but the genuine score distributions will be more disparate. Therefore, we test methods that normalize both genuine and impostor score distributions at the same time. We investigate Platt scaling [42] and propose a new method. We show that most score normalization techniques can improve demographic fairness by a good margin, by experimenting on two different datasets with a total of six evaluation protocols and five different pre-trained deep networks.

As our contributions in this paper, we:

- propose score normalization methods at post-processing stage for bias mitigation without downgrading verification performance;
- extend Z/T-norm to integrate demographics and propose three cohort-based methods, with one fitting the distribution for impostor scores and the other two also considering genuine scores;
- develop a new protocol for the RFW dataset with impostor scores selected randomly and define cohorts for the original and our new protocol;
- examine all methods on gender and ethnicity with feature extractors that perform differently;
- and analyze the relative contribution of FNMR and FMR in fairness evaluation.

2. Background and Related Work

In recent years, deep learning has dominated and revolutionized the field of FR. Two main research directions exist in the biometrics community: developing better network topologies and implementing better-suited loss functions. Wang *et al.* [60] provide a survey of algorithms, datasets, and evaluations before 2018, and more approaches have been developed since. Most modern network architectures [26] include variations and improvements [14] of residual network architectures [22]. The latest developed loss functions, *i.e.* ArcFace [11], MagFace [38], and Adaface [28], improve the discriminability of deep features in angular space.

Also, aspects of fairness have been discussed and evaluated [19, 61]. Cavazos *et al.* [7] describe underlying factors that bias COTS FR systems with respect to ethnicity. It was observed that biases are more frequently observed in low-quality samples [19, 7]. Vangara *et al.* [56] show consistently higher False Match Rates (FMR) in African Americans compared to Caucasians using several COTS systems.

In most cases, face verification performance is evaluated via two error measures [41]. For a given score threshold, the FMR computes the number of impostor comparisons (gallery template and probe sample are from different identities) with a similarity above the threshold. Similarly, the False Non-Match Rate (FNMR) calculates the number of genuine pairs (two samples of the same identity) with a similarity below the threshold. Oftentimes, these rates are computed and averaged across demographics. Since most datasets have mainly Caucasian people [61], errors made on non-Caucasians do not influence results much.

Calculating True Match Rates (TMR) for each demographic label individually is widely used to evaluate the fairness differences across demographics [27, 29, 46, 59, 61, 62]. Though one can find differences in the performances of different algorithms [32], it is common to set a single threshold and analyze differences across demographic groups since independent thresholds are rarely applied in production. Some works in the biometrics literature advocate that the threshold should be demographic-specific [8, 44]. On the other hand, Fairness Discrepancy Rate [9], Inequality Rate [17] and Gini Aggregation Rate for Biometric Equitability [24] integrate within-demographic FMR(τ) and FNMR(τ) differences to measure the bias. Other methods consider the FNMR differences [52] and score distribution differences [30] across demographics. Here we use the metric suggested by the National Institute for Stan-

Table 1. Score Normalization. This table lists the score normalization techniques utilized in our experiments, including the data pairs used to compute the statistics.

Method	Description	Data for Statistics
M1	Z-norm subject-based	
M1.1	Z-norm subject-demo-based	Gallery × Cohort
M1.2	Z-norm demo-based	
M2	T-norm subject-based	
M2.1	T-norm subject-demo-based	Probe × Cohort
M2.2	T-norm demo-based	
M3	Impostor Norm	
M4	Platt Scaling	Cohort × Cohort
M5	Bimodal CDF	

dards and Technology (NIST) [18] which compares the Worst-case Error Rate to the geometric mean of FMR and FNMR.

There also exist methods to mitigate the bias in FR systems by improving features extracted from the networks [15, 33, 40, 16, 27, 46, 53, 63, 65, 13, 64], partially solving ethical problems by generating synthetic images for different demographics [4, 47]. With features extracted via existing FR systems, Terhorst *et al.* [55] train a classifier to replace the regular cosine similarity function which pushes score distributions of different groups to be similar. A fair score normalization (FSN) method proposed by Terhorst *et al.* [54] uses KMeans to cluster features and combine the cluster-specific thresholds and global thresholds to normalize scores. Kotwal *et al.* [31] propose a score calibration method to align the score distribution by fine-tuning a pre-trained network with additional intra- and inter-demographic loss terms. In contrast, we solely focus on boosting fairness of existing FR systems without any network training process and no need to access features. While somewhat related to our work (Kotwal *et al.* [31] investigate similar methods as ourselves, such as M1 and M2, see below), they made use of a network pre-trained on MS1MV3 [21], which overlaps identities with the RFW test set [61].

3. Approach

To apply a single score threshold τ that is suited for different demographics, it is required that each demographic follows a similar score distribution. We investigate several techniques that provide this capability, a list of these techniques is provided in Tab. 1. Most techniques collect score distributions $D = \{\text{sim}(g_1, p_1), \text{sim}(g_2, p_2), \ldots\}$ which are computed by choosing various gallery g and probe samples p. These distributions are modeled to follow a normal distribution $\mathcal{N}_{\mu,\sigma}$ with mean μ and standard deviation σ:

$$\mu = \mathbb{E}(D) \quad \sigma = \sqrt{\mathbb{E}\big((D-\mu)^2\big)} \tag{1}$$

Finally, standardization is performed for a given similarity score s between gallery g and probe sample p from the test set:

$$s' = \mathcal{S}_{\mu,\sigma}(s) = \frac{s-\mu}{\sigma} \quad \text{with} \quad s = \text{sim}(g, p) \tag{2}$$

Methods using only impostor scores to model $\mathcal{N}_{\mu,\sigma}$ are introduced in Sec. 3.1, 3.2, and 3.3, and methods including genuine scores in Sec. 3.4.

3.1. Identity-based Score Normalization

Well-known techniques such as Z-norm [48] and T-norm [5] fight differences in score distributions of single individuals that were first observed by Doddington *et al.* [12]. The identity-based impostor score distributions are computed by comparing the test gallery g or probe sample p to samples from different identities selected from a cohort dataset. When the test gallery g is compared to the cohort, this technique is called Z-norm [48], which we refer to as M1. T-norm [5], *i.e.* when the test probe sample p is compared to the cohort, is called M2 in our evaluation.

3.2. Demographics-based Score Normalization

While these identity-based approaches can make distributions more similar across demographics, no such guarantee is given and there is no restriction on the demographics of the cohort samples. In fact, the distribution contains cross-demographic comparisons, which might not reflect realistic impostor attacks. Hence, an easy extension of the Z-norm and T-norm would be to restrict the samples in the cohort to the same demographic as gallery/probe sample, respectively. We term these methods M1.1 for Z-norm and M2.1 for T-norm.

When further splitting up the cohort into demographics as done above, the number of scores for distribution estimation is reduced, especially at the tails of very low FMRs, the estimation might not be relevant any longer and downgrade the verification performance. To increase number of scores to model normal distributions (1), we can exploit all in-demographics gallery-cohort comparisons for all enrolled gallery samples, to arrive at M1.2. Similarly, we can utilize all in-demographics probe-cohort comparisons for all probe samples, to arrive at M2.2. While combining comparisons over all gallery or probe samples can be achieved when working on specific datasets, in typical verification applications there exists no large gallery, and we have access to a single probe only. Therefore, methods M1.2 and M2.2 are more of a theoretical nature.

3.3. Pure Cohort-based Score Normalization

For a more fair and realistic evaluation, we propose to move away from subject-based normalization such

as T-norm and Z-norm. Instead, we solely rely on in-cohort in-demographics comparisons to estimate the score distributions. Particularly, we select impostor pairs of cohort samples from the same demographics (ethnicity or gender), but from different subjects, and provide mean μ_i and standard deviation σ_i for each demographic d_i via (1). We term this method M3.

This way of selecting distributions has two main advantages. First, these statistics can be pre-computed and do not require additional gallery-cohort or probe-cohort comparisons during enrollment or probing. Second, the number of scores that can be used is much larger since all different-identity same-demographics pairs are utilized. The only disadvantage is that we need to know d_i of the comparison to select correct model $\mathcal{S}_{\mu_i,\sigma_i}$ for (2). Here, we assumed d_i can be gathered during enrollment, and there is no need to know the demographic information for probe sample.

3.4. Genuine and Impostor Cohorts

Any of the above approaches have the issue that distributions are only modeled from one type of score, *i.e.*, the impostor scores. While such methods will likely be able to improve the alignment of the impostor score distributions across demographics, it is unlikely that they also normalize across genuine score distributions. Thus, they are unlikely to improve algorithmic fairness across all different operating points.

To make use of genuine score distributions, we need to compute similarities for in-cohort pairs with matching identities, which we split into different demographics. For demographic d_i, we mark the genuine score distributions $D_i^{\oplus}$, while impostor score distributions (which are the same as used in M3) are marked as $D_i^{\ominus}$. These two score distributions can now be used to provide a single monotonically increasing function to transform the raw scores into normalized scores.

There exist several techniques to incorporate the set of two distributions $D_i^{\oplus}$ and $D_i^{\ominus}$ for a given demographic d_i into one final normalization [43, 45]. Our selected representative M4 of these techniques is Platt scaling [42, 36], where logistic regression is performed to distinguish low impostor scores from large genuine scores by maximizing the weighted binary cross entropy using weights $w^{\oplus}$ and $w^{\ominus}$ for normalizing different counts of genuine and impostor scores. The final logistic function σ can be used to normalize the original test score s for a given demographic d_i:

$$s' = \sigma(s) = \frac{1}{1+e^{-\alpha s-\beta}} \qquad \text{with} \tag{3}$$

$$\alpha,\beta = \arg\max_{\alpha,\beta}\left[w^{\oplus}\sum_{s\in D_i^{\oplus}}\log\sigma(s) - w^{\ominus}\sum_{s\in D_i^{\ominus}}\log\sigma(s)\right]$$

Finally, we propose M5 which is related to M3 but incorporates both genuine and impostor score normalization. This method is inspired by the Bayesian Intrapersonal/Extrapersonal Classifier [39, 20]. We try to estimate a score that combines the probability of being a genuine and not being an impostor score: $\mathcal{P}(s) = \mathcal{P}^{\oplus}(s) - \mathcal{P}^{\ominus}(s)$. The former can be estimated by the Cumulative Distribution Function (CDF) of the genuine score distribution: $\mathcal{P}^{\oplus}(s) = \mathrm{CDF}(\mathcal{N}^{\oplus})(s)$. The latter is computed by inverting the CDF of the impostor score distribution: $\mathcal{P}^{\ominus}(s) = 1 - \mathrm{CDF}(\mathcal{N}^{\ominus})(s)$. By combining and applying them to a specific demographic d_i, we arrive at:

$$s' = \mathcal{P}(s) = \mathrm{CDF}(\mathcal{N}_i^{\oplus})(s) - 1 + \mathrm{CDF}(\mathcal{N}_i^{\ominus})(s) \tag{4}$$

4. Evaluation

For fairness evaluation, we rely on Worst-case Error Rate to geometric mean of FMR and FNMR (WERM):

$$\mathrm{WERM}_\tau = \left(\frac{\max_{d_i}\mathrm{FMR}_{d_i}(\tau)}{\left(\prod_{d_i}\mathrm{FMR}_{d_i}+\epsilon\right)^{1/n}}\right)^{\alpha} \times \left(\frac{\max_{d_i}\mathrm{FNMR}_{d_i}(\tau)}{\left(\prod_{d_i}\mathrm{FNMR}_{d_i}+\epsilon\right)^{1/n}}\right)^{(1-\alpha)} \tag{5}$$

Here we use $\epsilon = 10^{-5}$. WERM [18] ranges in $(0,\infty)$ where lower values are better. Often, the contribution of FMR and FNMR are expected to be balanced, but other weights are also possible and should be considered depending on the application needs [9, 24]. We apply equal weights: $\alpha = \frac{1}{2}$. Since the report on WERM is threshold-specific, we only focus on FMR threshold $\tau = 10^{-3}$ here, and other thresholds can be considered for different application purposes. Additionally, $\mathrm{TMR}(\tau)$ is used to measure verification performance.

4.1. New Protocols for VGGFace2

By containing over 3.31 million images of 9131 subjects, the VGGFace2 [6] dataset is one of the larger FR datasets. The training set contains 8631 identities, while the test set contains 500 identities. We utilize the given gender labels (`Male`, `Female`), and publicly available ethnicity labels (`Asian`, `Black`, `Indian`,

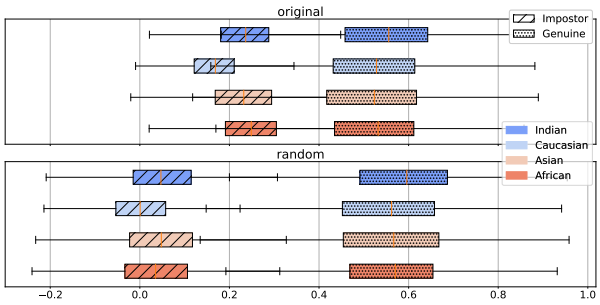

Figure 2. RFW Protocol Comparison. This figure displays distributions of baseline genuine and impostor scores of the four ethnicities on `original` and our `random` RFW protocol computed with the E2 network (cf. Tab. 2).

`White`).* Less than 30 people are removed due to difficulties in determining ethnicities. Instead of using all samples for the FR experiment, we create a sub-sampled protocol. For each subject, we randomly pick five samples to compose a probe set and one sample to be registered in the gallery. We use a sample-to-sample demographic-specific all-vs-all comparison to compute baseline scores s. Since we make use of a cohort to apply the score normalization techniques described above, the same procedure is applied to obtain a cohort from the training set of VGGFace2.

VGGFace2 has a bias in `White` and `Male`, which is also reflected in the number of comparisons in the scores. Since τ is picked based on impostor score distribution, balancing for demographic groups alters τ. We want to examine the system fairness variation in the above two situations. To create balanced numbers of comparisons, for each network, we sub-sample 5,200 impostor pairs per ethnicity, and 180,000 impostor pairs per gender, such that they follow the original distributions of all scores. As a result, VGGFace2 forms four subsets: VGG2 gender, VGG2 gender-balanced, VGG2 ethnicity, and VGG2 ethnicity-balanced. The interaction of gender and ethnicity labels is not tested in our experiments, since balancing both at the same time provides splits with very few samples.

4.2. New Protocols for RFW

To overcome the issue of unbalanced evaluations with respect to ethnicity, Wang *et al.* [61] introduced the Racial Faces in the Wild (RFW) dataset. This dataset is a subset of the MS-Celeb-1M (MS1M) [21] whose identities are organized into four different ethnicities (`African`, `Asian`, `Caucasian`, `Indian`) with about 3000 individuals per ethnicity. The `original` verification protocol [61] defines around 6000 image pairs (half genuine and half impostor) per ethnicity, utilizing the most similar impostor pairs determined by a deep learning-based FR algorithm, which usually share the same gender and ethnicity. Thus, the distribution of impostor scores shown in Fig. 2 does not follow the general trend of gathering around 0. Any technique trying to learn this score distribution from cohort data is doomed to fail. Therefore, we generated a new `random` protocol for the RFW dataset to make this more comparable with other datasets' results and avoids a possible selection bias in the default protocol. It is composed of random image pairs (impostor and genuine) of same ethnicity and gender. The number of pairs is almost identical to protocol `original`.

To evaluate the difficulty, two protocols are passed into the FR experiment with network E2 (cf. Sec. 5). At $\tau = 10^{-3}$,† we reach TMR = 0.630 for `original` and 0.897 for `random`. We can also observe the distribution change of impostor scores from Fig. 2, and `random` pushes impostor distributions of all ethnicities closer to 0 without changing the bias on `Caucasian`. We suppose that impostor scores are not centered at 0 because of the performance limit of the network and the bias can be minimized if a good network is applied.

Cohort samples for RFW are taken from the BUPT-Balancedface [59] dataset, which has the same four ethnicities with 70000 subjects per ethnicity, but its image quality for `Asian` and `Indian` are not as good and stable as `African` and `Caucasian`, which have comparable quality to RFW images. We cleaned the cohort dataset by removing possible overlaps with RFW and subjects that have different labels for duplicate images. Since it is not precisely known how the impostor pairs of the default RFW protocol were selected, we rely on an IResNet100 network, which is trained on MS1M by ArcFace loss and knows RFW well,‡ to get 5000 most similar impostor pairs plus 5000 genuine pairs per ethnicity as the cohort for `original`. The cohort for `random` is selected following the same idea as the test set.

5. Experiments

In total, we evaluate five different pre-trained and publicly available FR networks, as summarized in Tab. 2. Since RFW is a subset of MS1M, any network that is trained on MS1M cannot be evaluated on RFW. Therefore, we select two Arcface networks provided§ by the RFW authors [61], *i.e.*, a ResNet34 (E1) trained on CASIA-Webfaces, and a ResNet-50 architecture (E2) trained on MS1M excluding RFW identities.

*https://gitlab.idiap.ch/bob/bob.bio.face/-/blob/master/src/bob/bio/face/database/vgg2.py

†For brevity, we write $\tau = 10^{-3}$ to refer to the threshold τ that achieves an FMR of 10^{-3} on the combined test set.

‡https://github.com/deepinsight/insightface

§http://www.whdeng.cn/RFW/model.html

Table 2. Pre-Trained Networks. This table lists the networks utilized in our experiments, including data and loss function used for training. The networks are sorted in ascending order of overall recognition performance.

Model	Network	Training Data	Loss Function
E1	ResNet34	CASIA-WebFace	ArcFace
E2	ResNet50	MS1M-w/o-RFW	ArcFace
E3	IResNet100	Webface12M	AdaFace
E4	IResNet100	MS1M	MagFace
E5	IResNet100	Webface12M	DALIFace

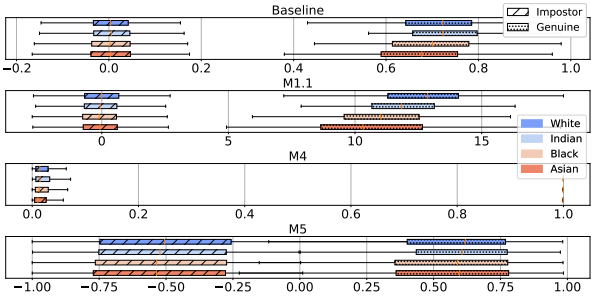

Figure 3. Impostor vs ALL. This figure compares the VGGFace2 ethnicity score distributions of baseline, impostor-based method M1.1, and impostor-genuine-based methods M4 and M5. Features are extracted by E3.

Since other evaluated datasets do not have this issue, we also include a more powerful IResNet-100 topology trained on MS1M using MagFace loss (E4).* Finally, we employ two IResNet-100 architectures trained on the WebFaces12M dataset using AdaFace loss (E3)† and DALIFace loss [49] (E5). For comparison, we also evaluate FSN [54] on our protocols and networks.

5.1. VGGFace2

We run tests on the VGGFace2 dataset with all feature extractors. All proposed methods are applied, independently with two different types of demographic labels, gender and ethnicity. In Fig. 3, a quantile plot, the impostor distribution for baseline of VGGFace2 ethnicity is centered around 0, but a discrepancy appears in the genuine distribution. Here the overlap of impostor and genuine scores are treated as outliers and discarded, while the overlap still exists. The impostor-based method M1.1, which is supposed to only work on one side of the scores, results in some subtle moves on the impostor side, while genuine scores are stretched to more varied distributions. FMR gaps between demographics are diminished, while FNMR gaps expand. Theoretically, a small threshold should lead to a large FNMR, but WERM drops in Tab. 4 since the change in FMR dominates. M4 and M5 attempt to harmonize the spread on both distributions simultaneously,

*https://github.com/IrvingMeng/MagFace
†https://github.com/mk-minchul/AdaFace

Table 3. Z-Norm-based Methods. This table displays WERM and TMR (%) of three Z-norm methods (M1, M1.1, M1.2) for VGGFace2 ethnicity w.r.t. feature extractors E1 - E5. The best values per column are colored in blue/red.

Network	E1		E2		E3		E4		E5	
Metrics	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓
Baseline	*93.76*	*3.8094*	*95.65*	*4.0101*	*96.80*	*1.9320*	*96.92*	*1.8356*	*96.96*	*2.3542*
M1	94.50	1.6103	95.81	3.7954	96.76	1.9046	96.92	1.8154	96.88	3.1285
M1.1	92.85	1.5870	95.73	1.3594	96.76	1.6203	96.84	2.1640	96.88	1.6728
M1.2	92.98	1.5237	95.56	2.5367	96.80	1.7845	96.88	1.9265	96.80	2.6092

and they perform differently on the task. M4 provides nicely separated score distributions, but the alignment across demographics can be worse. M5 models both distributions and achieves good alignment, but pushes the overlap of extreme values from both sides.

Tab. 4 is a compact result table. Each subtable displays TMR and WERM at $\tau = 10^{-3}$ for scores post-processed by five demographic-based methods with feature extracted by E1 - E5. An analysis for two factors of WERM is on our GitHub page.‡‡ WERM for scores before normalization (baseline) are displayed in the first row of each table. Almost all methods lead to a less biased output when normalized with respect to gender. In most cases, verification performance is not affected or even has improvements, while the drop exists with a small magnitude. The difficulty of alignment grows as the number of demographics rises to four ethnicities, especially M5 does not work well when facing three well-performing networks. M1.1, M2.1, and M4 are quite stable and outperform FSN. Comparison between pure identity- (M1) and demographic-based (M1.1/1.2) methods can be found in Tab. 3. M1 does not exhibit a notable advantage over M1.1/1.2, which proves that demographic information is more influential in mitigating bias compared to identity-only data. M1.1 and M1.2 have comparable performance, which is guaranteed when cohort size for M1.1 is large enough (Central Limit Theorem) for good estimation.

We observe that balancing ethnicity through sub-sampling worsens fairness compared to baselines, with a similar, albeit smaller, trend for gender balancing. The unexpected rise after balancing can be attributed to the limited number of samples in the minority group. Sub-sampling occurs in the majority groups and nearly all samples in the minority groups are preserved. Thus, the distribution issue in the minority groups remains, the impact of the majority groups diminishes, and ultimately, the bias is amplified. For the same reason, we observe a drop in TMR after balancing. Similar behavior on both ethnicity and gender exhibits that balancing the impostor pairs per demographic via sub-sampling and then deciding thresholds does not lead to a less biased result. Regardless of τ determination, our normalization techniques improve system fairness, though WERM magnitudes remain baseline-consistent. We prove that determining thresholds with balanced

Table 4. $\text{WERM}_{10^{-3}}$ & $\text{TMR}_{10^{-3}}$. Six tables below present results for six evaluation protocols. For each protocol, all five networks, E1 - E5 as provided in Tab. 2, are used to extract features. Five proposed score normalization methods M1 – M5, cf. Tab. 1, and FSN [54], are applied to those features and TMR (%) and WERM values at threshold $\tau = 10^{-3}$ are computed. best and runner-up TMR and best and second WERM value are highlighted. The *first row* shows the baseline.

(a) VGGFace2 Gender

Network	E1		E2		E3		E4		E5	
Metrics	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓
Baseline	*93.55*	*1.6477*	*95.48*	*1.4042*	*96.71*	*1.2908*	*96.88*	*1.1996*	*96.76*	*1.3892*
FSN	92.32	1.7575	95.36	1.3947	96.76	1.3059	96.71	1.1674	96.84	1.4596
M1.1	93.43	1.1649	95.65	1.0977	96.67	1.0997	96.88	1.0986	96.92	1.0831
M2.1	93.63	1.2033	95.77	1.0932	96.63	1.1092	96.92	1.0349	96.84	1.0926
M3	92.98	1.1779	95.32	1.1346	96.67	1.0445	96.84	1.0366	96.80	1.0476
M4	93.35	1.3064	95.36	1.1233	96.71	1.1540	96.88	1.1420	96.88	1.1289
M5	93.55	1.3906	95.44	1.1811	96.67	1.0505	96.88	1.0474	96.88	1.1871

(b) VGGFace2 Gender Balanced

Network	E1		E2		E3		E4		E5	
Metrics	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓
Baseline	*93.02*	*1.7456*	*95.40*	*1.4696*	*96.63*	*1.2624*	*96.80*	*1.2490*	*96.71*	*1.3916*
FSN	91.46	1.7486	95.11	1.4429	96.67	1.2761	96.71	1.1664	96.59	1.4267
M1.1	93.35	1.1637	95.65	1.1080	96.67	1.0525	96.76	1.0811	96.92	1.1017
M2.1	93.55	1.2197	95.77	1.1014	96.63	1.0686	96.92	1.0425	96.84	1.1097
M3	92.98	1.1723	95.32	1.1426	96.67	1.0408	96.84	1.0455	96.80	1.0688
M4	93.14	1.3243	95.32	1.1330	96.63	1.1258	96.88	1.1642	96.80	1.1633
M5	93.06	1.4569	95.28	1.2022	96.67	1.0408	96.88	1.0483	96.88	1.2092

(c) VGGFace2 Ethnicity

Network	E1		E2		E3		E4		E5	
Metrics	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓
Baseline	*93.76*	*3.8094*	*95.65*	*4.0101*	*96.80*	*1.9320*	*96.92*	*1.8356*	*96.96*	*2.3542*
FSN	92.69	3.3962	95.65	4.0190	96.63	1.9067	96.92	1.7605	96.84	4.9612
M1.1	92.85	1.5870	95.73	1.3594	96.76	1.6203	96.84	2.1640	96.88	1.6728
M2.1	92.28	2.1288	95.69	1.4253	96.71	1.3468	96.96	2.9963	96.80	2.0942
M3	92.98	2.5158	95.52	2.6485	96.80	1.6644	96.88	1.8598	96.80	2.4600
M4	94.00	2.6577	95.69	3.7550	96.84	1.7912	96.92	1.6990	96.96	1.9823
M5	93.92	3.4323	95.65	3.6006	96.80	1.9328	96.88	1.9968	96.92	3.9164

(d) VGGFace2 Ethnicity Balanced

Network	E1		E2		E3		E4		E5	
Metrics	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓
Baseline	*88.62*	*6.4436*	*94.25*	*6.0192*	*96.51*	*2.9195*	*96.47*	*2.7463*	*96.39*	*3.2102*
FSN	86.57	11.1599	94.05	5.2970	96.35	3.8272	96.38	2.0060	96.35	3.9992
M1.1	93.26	2.0316	95.85	2.0109	96.71	1.9065	96.88	2.5114	96.88	1.9237
M2.1	93.18	2.5456	95.85	1.8519	96.76	1.7582	96.96	2.5224	96.88	2.4882
M3	93.43	3.3406	95.69	2.6263	96.76	2.3642	96.76	2.9416	96.80	2.5019
M4	92.65	3.6487	95.03	2.8245	96.51	1.9404	96.59	1.6599	96.67	3.0813
M5	88.91	6.1117	93.72	6.0567	96.22	3.6257	95.81	3.3750	95.73	5.6425

(e) RFW Original

Network	E1		E2		E3		E4		E5	
Metrics	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓
Baseline	*24.22*	*2.5246*	*63.05*	*2.5402*	*89.14*	*3.4611*	—	—	*89.03*	*2.7167*
FSN	2.19	6.8737	55.81	3.0454	88.60	4.5858	—	—	87.85	3.6184
M1.1	33.11	1.6128	67.13	2.0791	89.20	2.7097	—	—	88.59	2.4684
M2.1	33.38	1.4072	65.35	2.0301	90.41	7.0568	—	—	89.54	2.2140
M3	27.99	1.7419	62.17	2.2668	88.92	3.8758	—	—	89.56	2.1883
M4	26.29	2.0549	62.25	2.7727	89.60	3.5602	—	—	89.91	1.8976
M5	25.63	2.2587	62.54	2.5204	90.05	3.1712	—	—	89.50	1.6877

(f) RFW Random

Network	E1		E2		E3		E4		E5	
Metrics	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓	TMR ↑	WERM ↓
Baseline	*60.25*	*2.0418*	*89.66*	*2.7152*	*98.08*	*2.3202*	—	—	*98.04*	*4.3784*
FSN	56.15	3.4326	87.89	3.2014	98.10	3.0989	—	—	98.17	7.6021
M1.1	68.35	1.9580	90.55	1.7059	98.37	1.5468	—	—	98.66	3.2601
M2.1	63.53	1.4274	90.31	1.9231	97.97	1.7707	—	—	98.59	3.1609
M3	64.25	1.7578	89.30	2.2723	98.09	1.6354	—	—	98.42	2.0688
M4	64.11	1.7085	90.84	2.6720	98.06	2.5022	—	—	98.31	4.8942
M5	60.57	2.5452	86.83	3.4545	97.46	6.7419	—	—	97.20	5.0545

scores is unnecessary and may introduce extra bias.

5.2. RFW

All nine normalization techniques are applied to each protocol-network pair. Although `random` is proposed to mitigate the selection bias in `original`, in Tab. 4(e) and 4(f), it is not the case by checking the WERM values for the baseline scores. The bias seems not only brought by impostor score distribution and cannot be eliminated by the proposed new protocol. To ensure the results for protocol `random` are not occasionally created, we perform a simplified statistical analysis by generating four more random splits and implement all proposed methods on them, and compute TMR and WERM values for all five splits. For each network and method, we compute standard deviation (STD) across the five splits and compute the average across methods. STD ranges from 0.04 to 2.748 for TMR and from 0.022 to 0.16 for WERM, which implies that no large variation for any method, so the results discussed below are reliable and reproducible. The detailed table is available on the GitHub page.[‡‡]

Across five methods, TMR undergoes changes ranging from -0.8% to 9.2%, while FSN lowers TMR in most cases. Interestingly, normalization techniques exhibit distinct bias mitigation impacts on the two protocols. E1, E2, and E5 yield decreasing WERM across all impostor-based methods (M1.1-M3), while M5 is preferable for `original`. M4 only works well on features extracted by E1. De-biasing effects brought by M1.1 are permanent for this dataset, with the enhanced TMR performance in most cases. Consequently, the impostor-based methods, M1.1, M2.1, and M3 have steady achievement in de-biasing for both gender and ethnicity, regardless of datasets. M4 is desired only for the VGGFace2 dataset, while M5 behaves positively for gender and the hardest RFW protocol.

6. Discussion

In WERM (5), $\alpha = \frac{1}{2}$ is set to balance the contribution of FNMR and FMR, but balancing is not guaranteed by α alone. We compute the relative contribution of FMR and FNMR (R_{FMR}, R_{FNMR}, respectively) in (5) by dividing scaled worst-case FMR and FNMR by WERM, take $\delta = R_{\text{FMR}} - R_{\text{FNMR}}$, and then analyze the distribution of δ with respect to each method. A higher δ implies a lower R_{FNMR}. Our hypothesis is confirmed by the distribution plot in Fig. 4, where genuine-impostor-based methods (M4, M5) and identity-based methods (M1, M2) more frequently result in a large δ than the other methods. Most methods have δ centered lower than baselines. Protocols like VGGFace2 gender-balanced with all decreases on WERM locates mostly

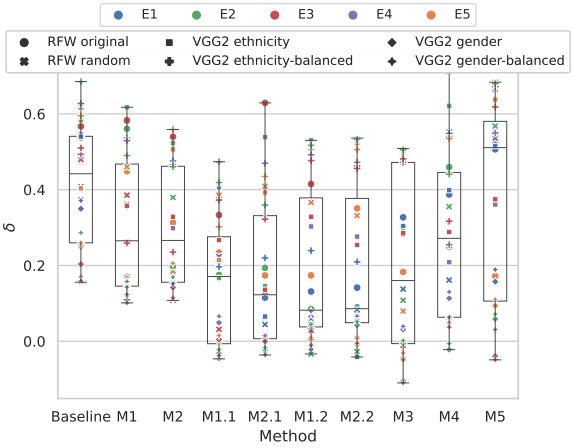

Figure 4. Distribution of δ. This figure exhibits the distribution of FMR and FNMR contribution difference δ with respect to the baseline and each method.

at the lower tail, and method like M5 has a large portion of δ clustered above 0.5 which is consistent with Tab. 4. Impostor-based methods (M1-M3), which focus on aligning impostor scores, lead to a smaller FMR and higher FNMR differences so that δ decreases. However, for genuine-impostor-based methods, taking care of both sides simultaneously does not ensure the change in two error rates will be the same. Some unpredictable effects occur depending on the dominant side. For example, better alignment in the genuine side results in higher FMR and smaller FNMR differences, leading to a larger δ, or vice versa. In general, impostor-based methods are more stable in de-biasing at scoring time.

7. Conclusion

We propose nine score normalization techniques, two are well-known identity-based methods (M1, M2), each followed by two extensions that integrate demographic information (M1.1, M1.2, M2.1, M2.2), and three pure cohort-based methods (M3, M4, M5). All techniques improve demographic fairness for high-security applications, *i.e.*, at low FMR, by working solely on scores without requiring network or feature adaptation. Importantly, in opposition to many feature-based fairness improvement techniques, none of our methods decreases verification performance, even small improvements can be observed. Experiments on six protocols from two datasets and five pre-trained feature extractors demonstrate the consistency of impostor-based methods (M1.1, M1.2, M3) with different verification performances. Analysis of the WERM value reveals the unequal contribution of FNMR and FMR in fairness evaluation, which is planned to be improved next. Also, other distributions

than Normal will be explored to capture tail behavior.

Acknowledgement

The authors thank the Hasler foundation for their support through the SAFER project.

References

[1] V. Albiero and K. W. Bowyer. Is face recognition sexist? no, gendered hairstyles and biology are. In *British Machine Vision Virtual Conference BMVC*. BMVA Press, 2020.

[2] V. Albiero, K. KS, K. Vangara, K. Zhang, M. C. King, and K. W. Bowyer. Analysis of gender inequality in face recognition accuracy. In *Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 81–89. IEEE/CVF, 2020.

[3] V. Albiero, K. Zhang, and K. W. Bowyer. How does gender balance in training data affect face recognition accuracy? In *International Joint Conference on Biometrics (IJCB)*, 2020.

[4] 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*, 2023.

[5] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1), 2000.

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Automatic Face & Gesture Recognition (FG)*. IEEE, 2018.

[7] 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? *Transactions on Biometrics, Behavior, and Identity Science (TBIOM)*, 3(1):101–111, 2020.

[8] 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. *Transactions on Biometrics, Behavior, and Identity Science (TBIOM)*, 1(1):32–41, 2019.

[9] T. de Freitas Pereira and S. Marcel. Fairness in biometrics: A figure of merit to assess biometric verification systems. *Transactions on Biometrics, Behavior, and Identity Science (TBIOM)*, 4(1):19–29, 2021.

[10] J. S. del Rio, D. Moctezuma, C. Conde, I. M. de Diego, and E. Cabello. Automated border control e-gates and facial recognition systems. *Computers & Security*, 62, 2016.

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] G. R. Doddington, W. Liggett, A. F. Martin, M. A. Przybocki, and D. A. Reynolds. SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *International Conference on Spoken Language Processing (ICSPL)*, 1998.

[13] S. Dooley, R. S. Sukthanker, J. P. Dickerson, C. White, F. Hutter, and M. Goldblum. On the importance of architectures and hyperparameters for fairness in face recognition. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022.

[14] I. C. Duta, L. Liu, F. Zhu, and L. Shao. Improved residual networks for image and video recognition. In *International Conference on Pattern Recognition (ICPR)*, pages 9415–9422. IEEE, 2021.

[15] S. Gong, X. Liu, and A. K. Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *European Conference on Computer Vision (ECCV)*. Springer, 2020.

[16] S. Gong, X. Liu, and A. K. Jain. Mitigating face recognition bias via group adaptive classifier. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[17] P. Grother. Demographic differentials in face recognition algorithms. *Virtual Events Series–Demo-Graphic Fairness in Biometric Systems*, 2021.

[18] P. Grother. Face recognition vendor test (FRVT) part 8: Summarizing demographic differentials. Technical report, National Institute of Standards and Technology (NIST), 2022.

[19] P. Grother, M. Ngan, and K. Hanaoka. Face recognition vendor test (FRVT) part 3: Demographic effects. Technical report, National Institute of Standards and Technology (NIST), 2018.

[20] M. Günther and R. P. Würtz. Face detection and recognition using maximum likelihood classifiers on Gabor graphs. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 23(03):433–461, 2009.

[21] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision (ECCV)*. Springer, 2016.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[23] K. Hill. Wrongfully accused by an algorithm. *New York Times*, 6 2020. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

[24] 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 (ICPR)*, pages 431–447. Springer, 2022.

[25] 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 *International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2019.

[26] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[27] C. Huang, Y. Li, C. C. Loy, and X. Tang. Deep imbalanced learning for face recognition and attribute prediction. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[28] M. Kim, A. K. Jain, and X. Liu. AdaFace: Quality adaptive margin for face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[29] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] K. Kotwal and S. Marcel. Fairness index measures to evaluate bias in biometric recognition. In *International Conference on Pattern Recognition Workshops (ICPRW)*, 2022.

[31] K. Kotwal and S. Marcel. Mitigating demographic bias in face recognition via regularized score calibration. In *Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[32] K. S. 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. *Transactions on Technology and Society (TTS)*, 1(1):8–20, 2020.

[33] J. Liang, Y. Cao, C. Zhang, S. Chang, K. Bai, and Z. Xu. Additive adversarial learning for unbiased authentication. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[34] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. SphereFace: Deep hypersphere embedding for face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[35] R. S. Malpass and J. Kravitz. Recognition for faces of own and other race. *Journal of Personality and Social Psychology*, 1969.

[36] M. I. Mandasari, M. Günther, R. Wallace, R. Saeidi, S. Marcel, and D. A. van Leeuwen. Score calibration in face recognition. *IET Biometrics*, 3(4):246–256, 2014.

[37] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. IARPA Janus Benchmark - C: Face dataset and protocol. In *International Conference on Biometrics (ICB)*, 2018.

[38] Q. Meng, S. Zhao, Z. Huang, and F. Zhou. MagFace: A universal representation for face recognition and quality assessment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[39] B. Moghaddam, W. Wahid, and A. Pentland. Beyond eigenfaces: Probabilistic matching for face recognition. In *International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1998.

[40] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana. SensitiveNets: Learning agnostic representations with application to face images. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[41] P. J. Phillips, P. Grother, and R. Micheals. *Handbook of Face Recognition*, chapter Evaluation Methods in Face Recognition. Springer, 2nd edition, 2011.

[42] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

[43] N. Poh and S. Bengio. F-ratio client dependent normalisation for biometric authentication tasks. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2005.

[44] N. Poh, J. Kittler, A. Rattani, and M. Tistarelli. Group-specific score normalization for biometric systems. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 38–45. IEEE, 2010.

[45] N. Poh, A. Merati, and J. Kittler. Adaptive client-impostor centric score normalization: A case study in fingerprint verification. In *International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*. IEEE, 2009.

[46] H. Qin. Asymmetric rejection loss for fairer face recognition. *arXiv*, 2020.

[47] P. Rahimi, C. Ecabert, and S. Marce. Toward responsible face datasets: modeling the distribution of a disentangled latent space for sampling face images from demographic groups. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2023.

[48] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19–41, 2000.

[49] W. Robbins, G. Bertocco, and T. E. Boult. DaliID: Distortion-adaptive learned invariance for identification – a robust technique for face recognition and person re-identification. *IEEE Access*, 2024.

[50] T. Romm. Amazon's facial-recognition tool misidentified 28 lawmakers as people arrested for a crime, study finds. *Washington Post*, July 2018. Retrieved from http://www.washingtonpost.com.

[51] E. M. Rudd, M. Günther, and T. E. Boult. MOON: A mixed objective optimization network for the recognition of facial attributes. In *European Conference on Computer Vision (ECCV)*, pages 19–35. Springer, 2016.

[52] 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 (ICPR)*, pages 570–581. Springer, 2022.

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

[54] 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*, 140:332–338, 2020.

[55] P. Terhörst, M. L. Tran, N. Damer, F. Kirchbuchner, and A. Kuijper. Comparison-level mitigation of ethnic bias in face recognition. In *International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2020.

[56] K. Vangara, M. C. King, V. Albiero, K. Bowyer, et al. Characterizing the variability in face recognition accuracy relative to race. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[57] R. Wallace, M. McLaren, C. McCool, and S. Marcel. Cross-pollination of normalisation techniques from speaker to face authentication using gaussian mixture models. *Transactions on Information Forensics and Security (TIFS)*, 7(2), 2012.

[58] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. CosFace: Large margin cosine loss for deep face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[59] M. Wang and W. Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[60] M. Wang and W. Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.

[61] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *International Conference on Computer Vision (ICCV)*, pages 692–702. IEEE, 2019.

[62] P. Wang, F. Su, Z. Zhao, Y. Guo, Y. Zhao, and B. Zhuang. Deep class-skewed learning for face recognition. *Neurocomputing*, 2019.

[63] X. Xu, Y. Huang, P. Shen, S. Li, J. Li, F. Huang, Y. Li, and Z. Cui. Consistent instance false positive improves fairness in face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 578–586, 2021.

[64] Y. Yang, A. Gupta, J. Feng, P. Singhal, V. Yadav, Y. Wu, P. Natarajan, V. Hedau, and J. Joo. Enhancing fairness in face detection in computer vision systems by demographic bias mitigation. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2022.

[65] Z. Yang, X. Zhu, C. Jiang, W. Liu, and L. Shen. RamFace: race adaptive margin based face recognition for racial bias mitigation. In *International Joint Conference on Biometrics (IJCB)*. IEEE, 2021.

[66] T. Zheng and W. Deng. Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments. Technical report, Beijing University of Posts and Telecommunications, 2018.

[67] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, et al. WebFace260M: A benchmark unveiling the power of million-scale deep face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.