

Consistent Instance False Positive Improves Fairness in Face Recognition

Xingkun Xu[†] Yuge Huang[†] Pengcheng Shen^{†*} Shaoxin Li[†]
Jilin Li[†] Feiyue Huang[†] Yong Li[§] Zhen Cui[§]

[†]Youtu Lab, Tencent [§]Nanjing University of Science and Technology

[†]{xingkunxu, yugehuang, quantshen, darwinli, jerolinli, garyhuang}@tencent.com

[§]{yong.li, zhen.cui}@njjust.edu.cn

Abstract

Demographic bias is a significant challenge in practical face recognition systems. Existing methods heavily rely on accurate demographic annotations. However, such annotations are usually unavailable in real scenarios. Moreover, these methods are typically designed for a specific demographic group and are not general enough. In this paper, we propose a false positive rate penalty loss, which mitigates face recognition bias by increasing the consistency of instance False Positive Rate (FPR). Specifically, we first define the instance FPR as the ratio between the number of the non-target similarities above a unified threshold and the total number of the non-target similarities. The unified threshold is estimated for a given total FPR. Then, an additional penalty term, which is in proportion to the ratio of instance FPR overall FPR, is introduced into the denominator of the softmax-based loss. The larger the instance FPR, the larger the penalty. By such unequal penalties, the instance FPRs are supposed to be consistent. Compared with the previous debiasing methods, our method requires no demographic annotations. Thus, it can mitigate the bias among demographic groups divided by various attributes, and these attributes are not needed to be previously predefined during training. Extensive experimental results on popular benchmarks demonstrate the superiority of our method over state-of-the-art competitors. Code and pre-trained models are available at <https://github.com/xkx0430/FairnessFR>.

1. Introduction

With the increasing deployment of face recognition systems, fairness in face recognition has received broad interest from research communities [19, 3, 5, 15, 14, 22]. This is partially due to the enormous impact brought in our daily life by face recognition systems. For example, when auto-

*denotes the corresponding author.

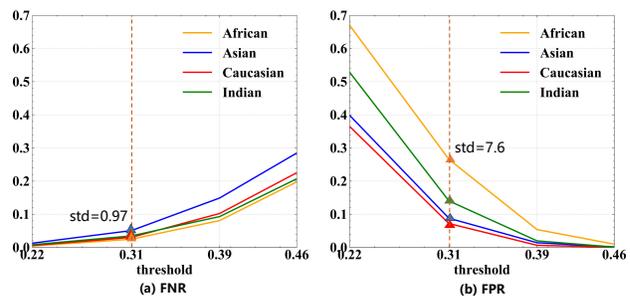


Figure 1. FNR and FPR curves of the four races in RFW [19]. FNR and FPR are calculated with a ResNet34 [2], which is trained on the public balanced dataset BUPT-Balanced [20]. Lower is better. Given a specific threshold, PPR varies significantly among different races than FNR. (e.g., the standard deviation (std) of FPR at $T_u=0.31$ is 7.6, while the std of FNR at $T_u=0.31$ is 0.97.)

matic face recognition is applied to crime prevention, unfair prediction may lead to unfair treatment of individuals across different demographic groups.

Previous studies [14, 19, 3, 4, 16] mainly improve the fairness of face recognition in two aspects, i.e., datasets and algorithms. Since the widely-used public large-scale face datasets, such as CASIA-WebFace [21], VGGFace2 [1], and MS-Celeb-1M [6] are collected from the Internet, they inevitably encode gender, ethnic, and culture biases. Thus, the works in [14, 19, 18, 9] propose some new face recognition datasets that contain relatively balanced samples in ethnicity, age, and other facial attributes. However, it is quite challenging to construct a balanced dataset in various attributes. What is more, the racial bias of the models trained with such balanced datasets cannot be eliminated completely [18]. Therefore, a novel algorithm that can mitigate the bias regardless of whether training datasets are balanced or not is imperative. Recently, several algorithms supervised by demographic attribute information are introduced to alleviate demographic bias. For example, Wang et al. [19] propose a deep information maximization adaptation network by transferring recognition knowledge from

Caucasians to other races. With similar ideas, they propose another method based on a widely-used margin-based loss function in face recognition, in which Q-learning learns the optimal margins of non-Caucasians with a manually-selected margin of Caucasians [18]. Different from the above methods take Caucasians as a reference, Gong et al. [4] present a debiasing adversarial network with four specific classifiers, in which one classifier is designed for identity and the other three are designed for demographic attributes. They further introduce a group adaptive classifier by using adaptive convolution kernels and attention mechanisms based on their demographic attributes [5]. However, all the above methods are explicitly designed to mitigate the bias in demographic groups divided by race. Thus, these methods have poor transferability and generalization. Moreover, they rely on accurate demographic attribute annotations, which are usually not available.

To address the above problem, we first evaluate the bias in face recognition from another perspective. Previous methods [4, 18, 19] mainly adopt the standard deviation of accuracy in each demographic group as the bias of a specific face recognition algorithm. In contrast, we analyze the bias in face recognition by two commonly-used evaluation metrics, *i.e.*, false positive rate (FPR), and false negative rate (FNR). As shown in Fig. 1, FPR varies significantly among different races than FNR, which shares a similar observation with [12]. Thus, it is essential to promote the consistency of FPR across each race group to mitigate the bias in face recognition. Based on this observation, we propose a false positive rate penalty loss, which mitigates face recognition bias by increasing the consistency of instance FPR. By generalizing the consistency of FPRs across each demographic group to the consistency of FPRs across each instance, our method is generic to improve the fairness of face recognition across the demographic groups divided by various attributes, such as race, gender, and age. Specifically, we first define the instance FPR as the ratio between the number of the non-target similarities above a unified threshold and the total number of the non-target similarities. Then, an additional penalty term in proportion to the ratio of instance FPR overall FPR is introduced into the denominator of the softmax-based loss. A larger ratio between each instance and the overall FPR yields a larger loss value. By such unequal penalties, the instance FPRs are supposed to be much consistent. Compared with the previous debiasing methods, our method firstly requires no demographic annotations of images; secondly can be easily embedded into the commonly used softmax-based loss function in face recognition; and finally can mitigate the bias across all demographic group divided by various kinds attributes, such as race, gender, and age.

To sum up, the contributions of this work are three-fold:

- To our best knowledge, it is the first work that alle-

viates the bias in face recognition by promoting the consistency of instance FPRs, which provides a new perspective to improve face recognition fairness.

- Our false positive rate penalty loss can improve the fairness across demographic groups divided by various kinds of attributes. Moreover, our method requires no demographic group annotation.
- We conduct extensive experiments on popular facial benchmarks, which demonstrate the superiority of our method over the SOTA competitors.

2. Related Work

Loss Function. Designing an effective loss function plays a vital role in deep face recognition. Many margin-based loss functions are proposed to obtain highly discriminative features for face recognition. For example, SphereFace [10], CosFace [17], ArcFace [2] are widely used margin-based loss function, which add margins in the positive logits (*i.e.*, intra-class). Recently, several works [20, 8] extend the margin-based loss function with hard sample mining strategies, which add the extra margin in the negative logits (*i.e.*, inter-class). Though the loss mentioned above functions are verified to obtain good performance, they do not consider the demographic bias. Our method can mitigate the bias in face recognition by promoting the consistency of instance FPRs, and thus improve face recognition fairness.

Bias Mitigation in Face Recognition. Firstly, we investigate the current datasets widely used for fair face recognition. The Diversity in Faces (DiF) dataset [11] provides annotations of 1 million human facial images to advance the study of fairness in facial recognition. Wang et al. [18] propose the Racial faces in-the-wild (RFW) as a testing database for studying racial bias in face recognition. In [20], Wang et al. also introduce BUPT-balanced as a balanced dataset on race, and BUPT-Globalface to reveal the real distribution of the world’s population. Both of these two public datasets are used for face recognition fairness studies. In [14], the BFW benchmark and dataset, inspired by DemogPairs [9], is introduced as a labeled data resource made available for evaluating recognition systems. BFW contains eight demographic groups for bias evaluation, and each of them consists of 200 subjects with 2.5K images. In this paper, we will evaluate our proposed algorithms on RFW and BFW. Next, we discuss the current algorithms for mitigating bias, which aims to solve unfairness of discrimination performance across groups, based on demographic information. Wang et al. [19] propose a deep information maximization adaptation network to alleviate bias, using deep unsupervised domain adaptation. Subsequently, Wang

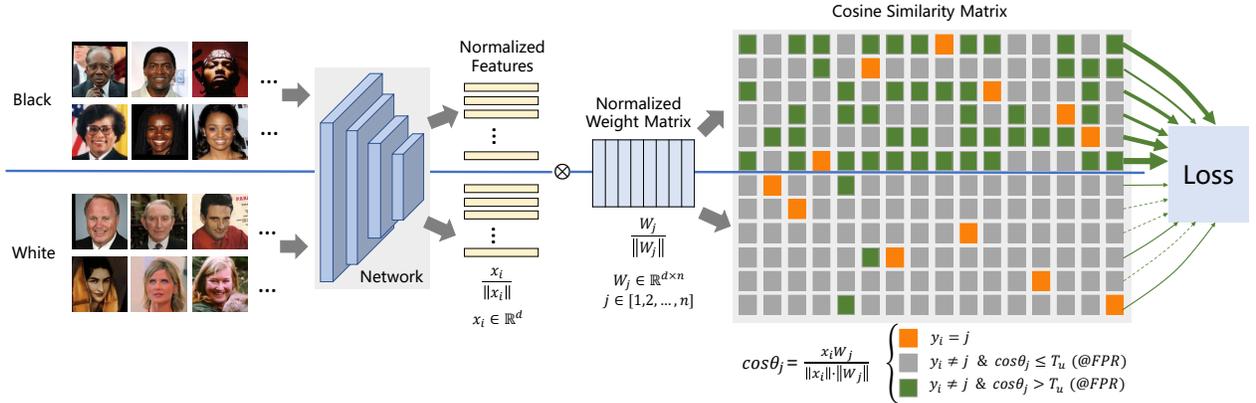


Figure 2. Illustration of instance FPR Penalty Loss. Assuming a mini-batch input X consists of samples from two races, *i.e.*, black and white, we obtain the corresponding features by an embedding network. Given a weight matrix W that each column corresponds to one identity, a cosine similarity matrix S is calculated based on the normalized feature X and weight W , and each value in this matrix is $S_{ij} = X_i W_j$. Among the matrix, the orange boxes the cosine similarities between the samples and their corresponding target (ground truth) weights. The other boxes denote the cosine similarities between the samples and the non-target weights. The green boxes indicate that their similarities are above a unified threshold T_u estimated by a preset FPR, while the grey boxes indicate equal to or less than the threshold. We take the green boxes in each row as false positives and define the instance FPR as the ratio between the number of the green boxes and the total number of the gray and green boxes. For samples from different races, the instance FPR varies significantly. Generally, the larger the instance FPR, the worse an algorithm performs at the current training stage. Thus, we introduce an additional false positive penalty term into the softmax-based losses to promote the consistency of instance FPRs.

et al. [18] introduce a reinforcement learning based race balance network, in which additive angular margin of loss functions for different races is selected by a pre-trained network module. Gong et al. [4] present a debiasing adversarial network with four specific classifiers, in which one classifier for identity and the other three for demographic attributes. They have further improved the method with a group adaptive classifier based on estimated demographic attributes recently [5].

Manual annotations of demographic attribute are necessary in current studies, which are usually unavailable in practice. Auxiliary modules, such as DQN, MDP, and attribute classifier, increase the training pipelines' difficulty than standard end-to-end training methods. In contrast, our proposed approach is simple to implement in an end-to-end manner without accurate manual annotations and auxiliary network modules.

3. Proposed Approach

In this section, we introduce the details of our approach. First, we explain the relationship between false positive rate and bias in face recognition; then we deduce a new evaluation protocol for demographic bias from the corresponding FPRs. Next, we introduce our false positive rate penalty loss, which mitigates face recognition bias by increasing the consistency of instance FPR.

3.1. Demographic Bias

False Positive Rate vs. Bias In face recognition systems, a comparison of two images with the same identity generates a positive pair, while a comparison of two images with different identities generates a negative pair. In general, a unified threshold T_u should be set as the criterion for judging whether a comparison of two images is positive or negative. A negative pair with similarity above the threshold is called a false positive pair (FPP), and a positive pair with similarity below the threshold is called a false negative pair (FNP). Correspondingly, the false positive rate (FPR) is defined as the ratio of false positive pairs to all negative pairs, and the false negative rate (FNR) is defined as the ratio of false negative pairs to all positive pairs. Both FPR and FNR are the frequently used as evaluation protocols in face recognition. Given a similarity set of N^+ positive pairs $\{S^+[i]\}$, and a similarity set of N^- negative pairs $\{S^-[i]\}$, FPR and FNR, which are respectively denoted as γ^+ and γ^- , are formulated as follows:

$$\gamma^+ = \frac{\sum_{i=1}^{N^-} \mathbf{1}(S^-[i] > T_u)}{N^-}, \quad (1)$$

$$\gamma^- = \frac{\sum_{i=1}^{N^+} \mathbf{1}(S^+[i] < T_u)}{N^+}, \quad (2)$$

where $\mathbf{1}(\cdot)$ is a indicator function.

For a demographic group g of certain race, gender, or age, etc., we can further calculate its own FPR γ_g^+ and FNR

γ_g^- as follows:

$$\gamma_g^+ = \frac{\sum_{i=1}^{N_g^-} \mathbb{1}(S_g^-[i] > T_u)}{N_g^-}, \quad (3)$$

$$\gamma_g^- = \frac{\sum_{i=1}^{N_g^+} \mathbb{1}(S_g^+[i] < T_u)}{N_g^+}, \quad (4)$$

where S_g^- , S_g^+ , N_g^- and N_g^+ are the corresponding numbers and similarity sets of the g group. To analyze bias in recognition performance, we adopt the FPR and FNR protocols to exhibit the performance difference across different demographic groups. We employ the BUPT-Balanced dataset [18] to train a ResNet-34 model with ArcFace and show the FPR and FNR performance on RFW [19] in Fig. 1. By comparing the FPR and FNR of four races, we notice that the performance of different races varies greatly, and moreover, the difference in FPR is much larger than that in FNR (the standard deviation of FPR at $T_u = 0.31$ is 7.6, while the standard deviation of FNR at $T_u = 0.31$ is 0.97). We note that similar results are also reported in the NIST FRVT [12].

Considering such results, we believe that achieving higher consistency in FPR prior to FNR across demographics is essential for improving fairness in face recognition. Besides, we define the standard deviation of the ratio between the demographic and overall FPR as bias degree. Given the group set \mathcal{G} and group number $N_{\mathcal{G}}$, the bias degree δ is formulated as follows:

$$\delta = \frac{1}{N_{\mathcal{G}}} \sqrt{\sum_{g \in \mathcal{G}} \left(\frac{\gamma_g^+ - \mu}{\gamma^+} \right)^2} \quad (5)$$

where μ is the average demographic FPR. In our following experiments, this criterion is used as a fairness evaluation protocol on several benchmarks.

Consistency of Instance FPR As an extreme case, if a demographic group is consisted of one single instance, demographic group’s FPR degrades into instance’s FPR. Correspondingly, the consistency of FPRs across different demographic groups is generalized as the consistency of FPRs across different instance. During training, we choose to increase such a generalized version of consistency rather than the consistency of demographic groups’ FPRs to mitigate face recognition bias. There are two reasons: 1) Demographic groups can be divided by various kinds of attributes, such as race, gender, and age etc., which is too numerous to enumerate; 2) Evaluating the consistency of instance FPR requires no demographic annotations of image samples. In the following, we first revisit the softmax-based losses and then show our way to achieve a higher FPR consistency via introducing extra false positive penalties into the softmax-like loss.

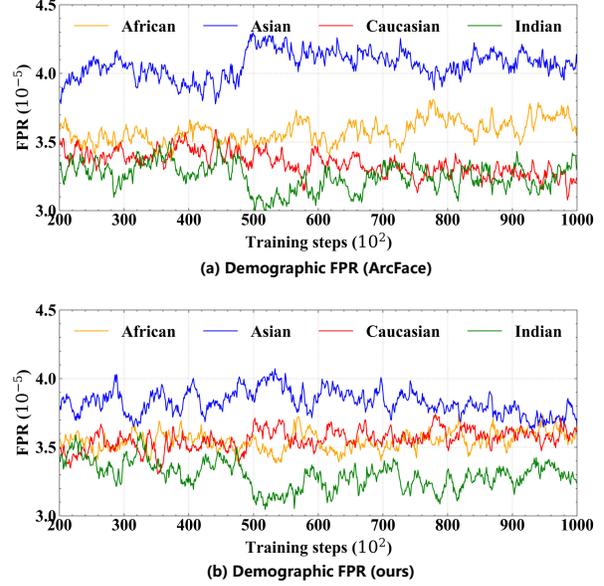


Figure 3. Comparison on demographic FPR in training. We compare the FPR trends of different races with ArcFace and our methods, respectively. In ArcFace (a), the FPRs of Caucasian and Indian are still at a low level, while the FPRs of black and Asian are at a high level, even continue to grow. In our method (b), except Indian, the FPRs of the other races trend to converge at the end of training. We notice that the FPR of Caucasian is a slightly increased. However, we evaluate this trained model on RFW, and the performance of Caucasian is not degraded.

3.2. FPR Penalty Loss

Softmax Loss Function. The original softmax loss is formulated as follows:

$$\mathcal{L} = -\log \frac{e^{W_{y_i} x_i + b_{y_i}}}{\sum_{j=1}^n e^{W_j x_i + b_j}}, \quad (6)$$

where $x_i \in R^d$ denotes the deep feature of i -th sample which belongs to the y_i class, $W_j \in R^d$ denotes the j -th column of the weight $W \in R^{d \times n}$ and b_j is the bias term. The class number and the embedding feature size are n and d , respectively. In practice, the bias is usually set to $b_j = 0$ and the individual weight is set to $\|W_j\| = 1$ by l_2 normalization. The deep feature is also normalized and re-scaled to s . Thus, the original softmax can be modified as follows:

$$\mathcal{L} = -\log \frac{e^{s(\cos \theta_{y_i})}}{e^{s(\cos \theta_{y_i})} + \sum_{j \neq y_i}^n e^{s(\cos \theta_j)}}. \quad (7)$$

Since the learned features with the original softmax loss may not be discriminative enough for practical face recognition problem, several variants are proposed and can be formulated in a general form:

$$\mathcal{L} = -\log \frac{e^{s \cdot G(\cos \theta_{y_i})}}{e^{s \cdot G(\cos \theta_{y_i})} + \sum_{j \neq y_i}^n e^{s \cdot H(\cos \theta_j)}}, \quad (8)$$

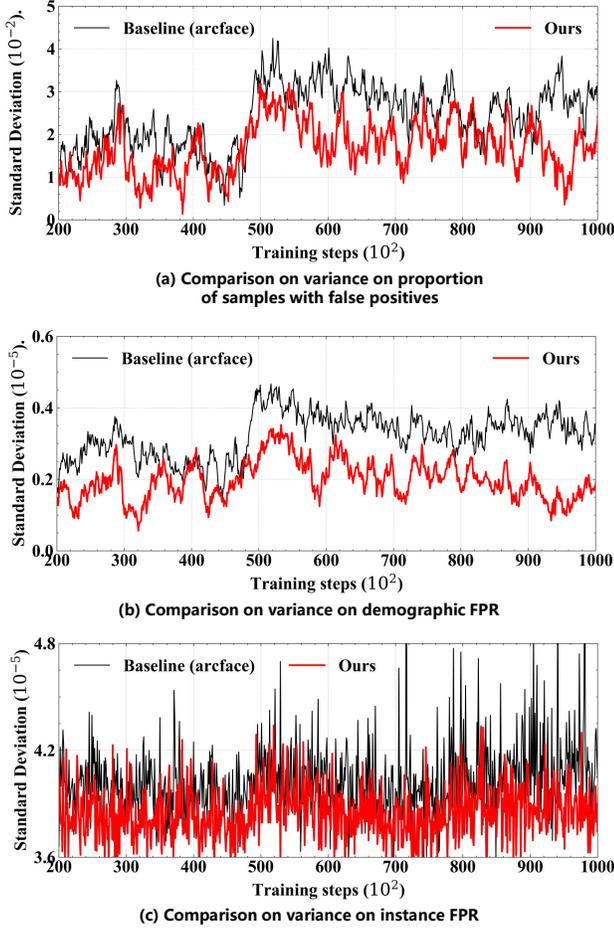


Figure 4. Comparison on variance in training. We compare the training variance in three aspects. As shown in (a), in a mini-batch, the proportion of the samples whose instance FPR is greater than 0 is compared between baseline and our method. (b) shows the standard deviations on demographics FPR, and (c) shows the standard deviations on instance FPR. Compared with arcface, our method achieves higher consistency on demographic FPR and instance FPR.

where $G(\cos \theta_{y_i})$ and $H(\cos \theta_j)$ are the functions to modulate the positive and negative cosine similarities, respectively. In margin-based loss function, such as ArcFace, $G(\cos \theta_{y_i}) = \cos(\theta_{y_i} + m)$ aims to emphasize the inter-class similarity. In mining-based loss functions, $H(\cos \theta_j)$ is designed to mining difficult negative pairs to decrease the intra-class confusion. However, the previous works focus on improving the discrimination performance on popular benchmarks, but not on enhancing the fairness of performance. Next, we introduce an extra false positive penalty term into the softmax-based losses, aiming at making the instance FPR more consistent and consequently enhancing the fairness of face recognition performance.

Extra Penalty on the FPR of Instance. Since the y_i -th column of the weight W usually could be regarded as a representative of the y_i -th class, for the i -th instance belonging to class y_i , the target logit $\cos \theta_{y_i}$ could be considered as the similarity of a positive pair, while the non-target logits $\cos \theta_j, j \neq y_i$ could be considered as the similarities of negative pairs. According to Eq. 3, given these non-target similarities and a unified threshold T_u , corresponding to a overall FPR γ_u^+ , the FPR of the instance can be calculated as:

$$\gamma_i^+ = \frac{\sum_{j=1, j \neq y_i}^n \mathbb{1}(\cos \theta_j > T_u)}{n-1}, \quad (9)$$

To make instance FPRs more consistent, that is, all close to FPR γ_u^+ , we add an extra penalty term in the denominator of the softmax function which is in proportion to the ratio of instance FPR to overall FPR γ_i^+/γ_u^+ . Specially, we add the ratio (multiplied by a factor $\alpha > 0$) to the original non-target logit, leading to the loss function presented as follows:

$$\mathcal{L} = -\log \frac{e^{s \cdot G(\cos \theta_{y_i})}}{e^{s \cdot G(\cos \theta_{y_i})} + \sum_{j \neq y_i}^n e^{s \cdot (\cos \theta_j + \alpha \frac{\gamma_i^+}{\gamma_u^+})}}. \quad (10)$$

Since the extra penalty $e^{s \alpha \gamma_i^+ / \gamma_u^+}$ is always > 1 , a larger instance-overall FPR ratio yields a larger loss value. By such unequal penalties, the instance FPRs are supposed to be much consistent. Further, considering the fact that more attention should be paid to those false positive cases with higher similarity (hard samples), we introduce a weighted FPR function of instance as follows:

$$\bar{\gamma}_i^+ = \frac{\sum_{j=1, j \neq y_i}^n \mathbb{1}(\cos \theta_j > T_u) \cdot F(\cos \theta_j)}{n-1}. \quad (11)$$

Here the function $F(z)$ is supposed to give larger weights to false positive cases with higher similarities and thus should be monotone increasing. Without loss of generality, in this paper, we use the power function $F(z) = \text{sgn}(z) |z|^p$ as the weighted function, where $p \geq 1$ and $\text{sgn}(\cdot)$ is the sign function. Since T_u is usually positive, the sign function and the abs. function can be omitted, leading to $F(z) = z^p$. When $p = 1$, $F(\cdot)$ degrades into $\cos \theta_j$.

Finally, we show the effect achieved by the loss function on the training set. Fig. 3 shows the difference of the FPR trends between baseline and our method. In Fig. 3 (b), we notice that FPR of Asian in our method keeps decreasing in most time of training process, and becomes much consistent with other races at the end. In Fig. 4, we compare the training variance on proportion of samples with false positives, demographic FPR and instance FPR. As a result, both the variance in these three aspects are lower than the baseline method. In a word, our algorithm helps to mitigate the race bias effectively.

Algorithm 1: FPR Penalty Loss

Input: The deep feature of i -th sample with its label y_i , cosine similarity $\cos \theta_j$ of two vectors, last fully-connected layer parameters W , embedding network parameters Θ , class number c , sample number n , learning rate λ , and overall false positive rate γ_u^+

iteration number $k \leftarrow 0$, parameter $t \leftarrow 0$, $\gamma_u^+ \leftarrow 1e^{-4}$;

while not converged **do**

 Compute the $\lceil \gamma_u^+ n(c-1) \rceil$ -th largest value of set $\{\cos \theta_j \mid i \in [1, n], j \in [1, c], j \neq y_i\}$ as the temporary threshold T_u ;

if $\cos(\theta_j) > T_u$ **then**

$I_j = 1$;

else

$I_j = 0$;

end

 Compute the weighted FPR $\bar{\gamma}_i^+$ by Eq. 11;

 Compute the loss \mathcal{L} by Eq. 10 (replace γ_i^+ by $\bar{\gamma}_i^+$);

 Compute the gradient of W_j and x_i by Eq. 12;

 Update the parameters W and Θ by:

$$W^{(k+1)} = W^{(k)} - \lambda^{(k)} \frac{\partial \mathcal{L}_i}{\partial W},$$

$$\Theta^{(k+1)} = \Theta^{(k)} - \lambda^{(k)} \frac{\partial \mathcal{L}_i}{\partial x_i} \frac{\partial x_i}{\partial \Theta^{(k)}};$$

$k \leftarrow k + 1$;

end

Output: W, Θ .

FPR Setting and Threshold Estimation in Training.

From Eq. 9 and Eq. 11, we see that our proposed method relies on the choice of an overall γ_u^+ , which further involves the estimation of the threshold T_u . In practice, the choice of the FPR depends on the deployment scenario of the face recognition system. For example, to balance the risk and user experience, the FPR is usually set to $1e-5$ in a face access control system. And the popular public face benchmarks often focus on the FPR range $[1e-1, 1e-6]$. Note that a lower FPR means less number of false positive cases. If the overall FPR is set to be extremely small, there are rare false positive cases can get extra penalty, which may intuitively lower the performance gain of our method. Besides, it's hard to estimate a stable threshold corresponding to an extremely small FPR, because of a lack of enough negative pairs during training. For the above reasons, we choose the overall FPR range as $[1e-1, 1e-5]$ in this paper.

Given an overall FPR, usually the threshold is estimated from the quantile of the distribution of all negative pairs. Here, we utilize the non-target logits instead of the negative instance pairs in threshold estimation, for reasons that compared with the size of mini-batch, the number of class in training is often much larger, leading to more negative pairs and thus a more accurate and stable threshold estimate.

Optimization. We show our method 1 can be easily optimized by the conventional stochastic gradient descent. Let's denote G_i as the $sG(\cos \theta_{y_i})$ of the sample belongs to the y_i -th class, H_j as $s(\cos \theta_j + \alpha \gamma_i^+ / \gamma_u^+)$, and I_j as the mining mask. In this section, we consider the Cos-Face form of our loss function, so $G_i = \cos \theta_{y_i} - m$ and $H_i = \cos \theta_j + \alpha \bar{\gamma}_i^+ / \gamma_u^+$. In the backward propagation process, the gradients *w.r.t.* x_i and W_j are presented as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}_i}{\partial W_{y_i}} &= \frac{\partial \mathcal{L}_i}{\partial G_i} \cdot x_i, \\ \frac{\partial \mathcal{L}_i}{\partial W_j} &= \left(1 + \frac{\alpha}{\gamma_u^+} \cdot I_j \frac{\partial F}{\partial \cos \theta_j} \right) \cdot \frac{\partial \mathcal{L}_i}{\partial H_j} \cdot x_i, \\ \frac{\partial \mathcal{L}_i}{\partial x_i} &= \frac{\partial \mathcal{L}_i}{\partial G_i} \cdot W_{y_i} + \left(1 + \frac{\alpha}{\gamma_u^+} \cdot \sum_{j \neq y_i} I_j \frac{\partial F}{\partial \cos \theta_j} \right) \cdot \frac{\partial \mathcal{L}_i}{\partial H_j} \cdot W_j, \end{aligned} \quad (12)$$

Based on the above formulations, we can find the extra gradients for alleviating bias have a new composed term $\frac{\partial \mathcal{L}_i}{\partial H_j} \frac{\partial F}{\partial \cos \theta_j}$, if the j -th logit with $I_j = 1$ is chosen as a false positive case. The term $\frac{\partial \mathcal{L}_i}{\partial H_j}$ brings gradient adjustment from false positive cases above the threshold, while the term $\frac{\partial F}{\partial \cos \theta_j}$ further modulates the former adjustment by the similarity of specific false positive case.

4. Experiments

4.1. Experimental Setting

Dataset. In this study, we employ BUPT-Balancedface and BUPT-Globalface dataset [19] for training. BUPT-Balancedface dataset contains 1.3M images of 28K celebrities and is approximately race-balanced with 7K identities per race. BUPT-Globalface dataset contains 2M images of 38K celebrities, and its racial distribution is approximately the same as the real distribution of the world's population. RFW dataset [18] and BFW dataset [14] are used for fairness testing. RFW consists of faces from four race groups: African, Asian, Caucasian, and Indian. Each race group contains nearly 10K images of 3K individuals for face verification. Compared with RFW, the BFW dataset provides balanced face data with more attributes, including ID, gender, and race. There are eight demographic groups according to two genders and four ethnic groups (*i.e.* Black, White, Asian, and Indian), and each demographic group consists of 200 subjects with 2.5K images.

Training Setting. We follow [17, 2] to crop the 112×112 faces with five landmarks detected by MTCNN [23]. The RGB images are first normalized by subtracting 127.5 and divided by 128, then feeding into the embedding network. we adopt ResNet34, ResNet50 and ResNet100 as in [7, 2] as the embedding network. We conducted all the experiments on 8 NVIDIA Tesla V100 GPU with Pytorch [13]

Table 1. Verification performance (%) of different FPR parameter γ .

Methods (%)	African	Asian	Caucasian	Indian	Avg	Std
$\gamma_u^+ = 10^{-5}$	95.60	95.10	97.18	96.32	96.05	0.91
$\gamma_u^+ = 10^{-4}$	95.95	95.17	96.78	96.38	96.07	0.69
$\gamma_u^+ = 10^{-3}$	95.47	94.90	96.92	96.12	95.84	0.87
$\gamma_u^+ = 10^{-2}$	95.45	94.78	96.98	96.13	95.84	0.94
$\gamma_u^+ = 10^{-1}$	95.23	94.60	95.87	95.97	95.42	0.64

Table 2. Verification performance (%) of different exponent p in $F(z)$.

Methods (%)	African	Asian	Caucasian	Indian	Avg	Std
$p = 0.25$	95.35	95.10	96.97	96.07	95.87	0.84
$p = 0.5$	95.27	94.93	96.58	96.02	95.70	0.74
$p = 1.0$	95.18	94.92	96.90	95.83	95.71	0.88
$p = 1.5$	95.27	94.67	97.05	96.23	95.80	1.05
$p = 2.0$	95.95	95.17	96.78	96.38	96.07	0.69
$p = 2.5$	95.85	95.00	96.96	96.20	96.00	0.82
$p = 3.0$	95.60	95.18	97.17	95.98	95.98	0.85

framework. The models are trained with SGD algorithm, with momentum 0.9 and weight decay $5e - 4$. The batch size is set to be 512. On BUPT-Balancedface, the learning rate starts from 0.1 and is divided by 10 at 20, 32, 36 epochs. The training process is finished at 40 epochs. On BUPT-Globalface, we divide the learning rate at 10, 18, 22 epochs and finish at 24 epochs. We follow the common setting as [17] to set $s = 64$ and $m = 0.35$.

4.2. Ablation Study

Effect of the overall FPR γ_u^+ . We conduct experiments at five fixed FPRs from 10^{-5} to 10^{-1} , and find that nearly all the best performance of training is achieved when $\gamma_u^+ = 10^{-4}$, except that a higher accuracy of Caucasian is obtained at $\gamma_u^+ = 10^{-5}$, as shown in Tab. 1. We explain the reasons as follows: 1) When γ_u^+ is set to be 10^{-5} or even a lower value, a relatively large value of threshold is used to measure the FPR of instance and generate penalty terms. Correspondingly, the number of extra penalty would be reduced. Besides, considering that the noisy data (e.g. label flips) is ubiquitous in training dataset, with a small number of noisy false positive cases, the accuracy of estimated threshold and the extra gradient adjustment may be affected dramatically. 2) When γ_u^+ is set to be a large value, e.g. 10^{-1} or higher, we obtain a relatively small value of threshold. With such a threshold, most training instances would be forced to generate numerous false negative pairs, since the similarity of most negative pairs is around 0. As a result, the instance FPRs and its corresponding penalty would be almost equal and hard to be more consistent through optimization. Therefore, $\gamma_u^+ = 10^{-4}$ is a reasonable choice. We will use this configure in our following experiments.

Effect of exponent p in $F(z)$. With the fixed FPR as 10^{-4} , we further investigate the effect of exponent p in $F(z) =$

Table 3. Verification performance (%) of protocol on RFW with SOTA methods ([BUPT-Balancedface]).

Methods (%)	African	Asian	Caucasian	Indian	Avg	Std
ArcFace-R34 [18]	93.98	93.72	96.18	94.67	94.64	1.11
CosFace-R34 [18]	92.93	92.98	95.12	93.93	93.74	1.03
DebFace-R34 (ECCV'20)	93.67	94.33	95.95	94.78	94.68	0.83
PfE-R34 [5]	95.17	94.27	96.38	94.60	95.11	0.93
GAC-R34 [5]	94.65	94.93	96.23	95.12	95.23	0.60
RL-RBN-R34(cos) (CVPR'20)	95.27	94.52	95.47	95.15	95.10	0.41
RL-RBN-R34(arc) (CVPR'20)	95.00	94.82	96.27	94.68	95.19	0.93
Ours-R34	95.95	95.17	96.78	96.38	96.07	0.69
ArcFace-R50	95.55	94.95	96.68	95.47	95.66	0.73
Ours-R50	96.47	95.75	97.08	96.77	96.52	0.57
ArcFace-R100	96.43	94.98	97.37	96.17	96.24	0.98
Ours-R100	97.03	95.65	97.6	96.82	96.78	0.82

Table 4. Verification accuracy (%) of protocol on RFW with SOTA methods ([BUPT-Globalface]).

Methods (%)	African	Asian	Caucasian	Indian	Avg	Std
ArcFace-R34 [18]	93.87	94.55	97.37	95.86	95.37	1.53
CosFace-R34 [18]	92.17	93.50	96.63	94.68	94.25	1.90
RL-RBN-R34(cos) (CVPR'20)	94.27	94.58	96.03	95.15	95.01	0.77
RL-RBN-R34(arc) (CVPR'20)	94.87	95.57	97.08	95.63	95.79	0.93
Ours-R34	95.77	95.85	97.92	96.70	96.56	0.75
ArcFace-R50	96.23	96.43	97.98	96.92	96.89	0.78
Ours-R50	96.85	96.75	98.30	96.95	97.21	0.73
ArcFace-R100	96.68	96.10	98.17	97.32	97.07	0.89
Ours-R100	97.37	96.48	98.57	97.4	97.45	0.85

Table 5. Bias degree of protocol on RFW with SOTA methods.

overall FPR	10^{-5}	10^{-4}	10^{-3}	10^{-2}
RL-RBN-R34(arc)	351.98	208.44	92.18	16.70
Ours-R34	257.53	185.91	59.25	10.33

Table 6. Bias degree of protocol on BFW with SOTA methods.

overall FPR	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}
RL-RBN-R34(arc)	2.44	2.01	2.49	2.91	2.43
Ours-R34	1.18	1.08	1.18	1.67	1.80

z^p . Here, we set the value varies from 0.25 and 3.0. Tab. 2 shows that the performance of our method decreases as n increasing from 0.25 to 1.0, and then gradually increases until n reaches near 2.0, and then the performance begins to decrease again. Note that when $p > 1$, the gradient is convex with regard to z , and with a moderate value of p , e.g. $p = 2$, we can give a proper but not too much penalty on false positive cases with larger similarity. Based on these reasons, we choose $p = 2$ in our following experiments.

4.3. Comparisons with SOTA methods

Accuracy on RFW. We train a ResNet34 model on BUPT-Balancedface with our method, and report the results of the competitors following the RFW protocol, shown in Tab. 3. Our method shows superiority over the competitors with the balanced dataset. Compared with the SOTA results, it achieves about 0.77% gains for average accuracy, and its standard deviation is decreased to 0.69 which is slightly higher than SOTA. Though the standard deviation of GAC and RL-RBN(cos) is much lower, its per-

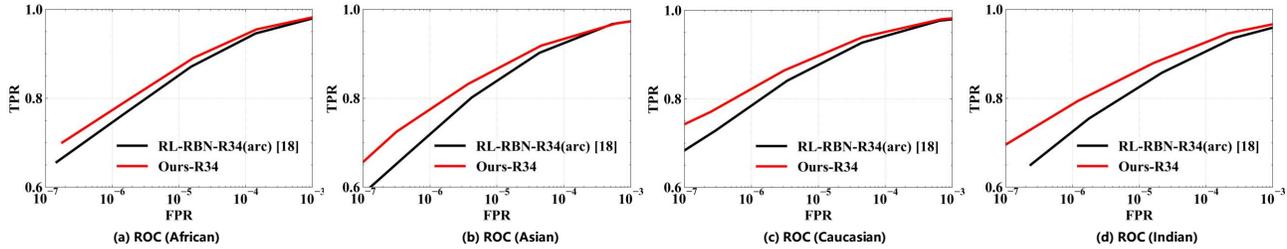


Figure 5. ROC for RFW.

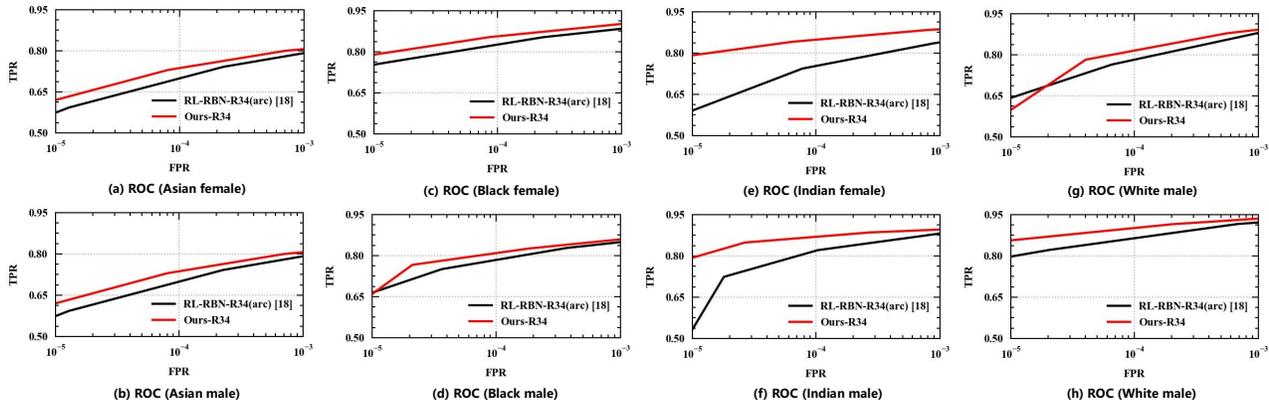


Figure 6. ROC for BFW.

formance on Caucasian is actually worse than that of the CosFace baseline. In contrast, for our method, the reduction in bias is obtained along with the accuracy improvement of all four races. We also train a ResNet34 model on BUPT-Globalface with our method and arcface. In Tab. 4, it shows that the average accuracy of our method is still much better than other competitors, while our standard deviation is also lower than others. Besides, we train Arcface and our method with ResNet50 and ResNet100 as in [2]. As shown in Tab. 3 and Tab. 4, our method also performs better than the common baseline. The above results show that our method can achieve competitive performances on both race balanced and unbalanced datasets, with regard to the mean and standard deviation of accuracy.

FPR on RFW. According to the FPR and TPR evaluation protocols discussed in Sec.2, we compare the performance between baseline and our method. Fig. 5 (a) shows the African ROC curves of our method and the SOTA competitor, and it is clear that our method performs best. Besides, Fig. 5 (b)(c)(d) respectively show the other groups' ROC curves. This experiment on RFW proves that our loss leads to the face recognition model with more discriminative features than RL-RBN(arc). According to the evaluation protocol defined in Eq. 5, we also compare the bias degree with the SOTA method in Tab. 5. The lower bias degree at each threshold corresponding with the overall FPR

demonstrates our method can achieve better performance on fairness recognition than that of RL-RBN(arc).

FPR on BFW. Fig. 6 shows the ROC curves on all 8 demographic groups in BFW. Across all ethnicity, our method achieves better performance on the female group and the male group. Across all gender, the TPR on each ethnicity in our method is much better than RL-RBN(arc). As defined in Eq. 5, we calculate the bias degree on BFW shown in Tab. 6, which proves that our algorithm also can alleviate both gender and race bias across demographics effectively.

5. Conclusions

In this paper, We develop a novel penalty term into the softmax loss function to alleviate bias and improve the fairness performance in face recognition. We propose the concept of instance FPR as an extreme case of demographic FPR, and convert consistency of instance FPR as a penalty item of softmax-based loss. Extensive experiments on popular facial benchmarks demonstrate the effectiveness of our method compared to the SOTA competitors. Following the main idea of this work, future research can be expanded in various aspects, including designing a better weight function $F(\cdot)$ for inconsistency penalty, and investigating the effects of noise samples that might be mistakenly optimized as false positive cases.

References

- [1] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE Int. Conf. Automatic Face and Gesture Recog.*, 2018. [1](#)
- [2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4690–4699, 2019. [1](#), [2](#), [6](#), [8](#)
- [3] 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*, 2020. [1](#)
- [4] Sixue Gong, Xiaoming Liu, and Anil K Jain. Jointly debiasing face recognition and demographic attribute estimation. In *Eur. Conf. Comput. Vis.*, pages 330–347. Springer, 2020. [1](#), [2](#), [3](#)
- [5] Sixue Gong, Xiaoming Liu, and Anil K Jain. Mitigating face recognition bias via group adaptive classifier. *arXiv preprint arXiv:2006.07576*, 2020. [1](#), [2](#), [3](#), [7](#)
- [6] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Eur. Conf. Comput. Vis.*, 2016. [1](#)
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [6](#)
- [8] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5901–5910, 2020. [2](#)
- [9] Isabelle Hupont and Carles Fernández. Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019. [1](#), [2](#)
- [10] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 212–220, 2017. [2](#)
- [11] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019. [2](#)
- [12] Mei Ngan, Patrick J Grother, and Mei Ngan. *Face recognition vendor test (FRVT) performance of automated gender classification algorithms*. US Department of Commerce, National Institute of Standards and Technology, 2015. [2](#), [4](#)
- [13] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *Adv. Neural Inform. Process. Syst. Worksh.*, 2017. [6](#)
- [14] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 0–1, 2020. [1](#), [2](#), [6](#)
- [15] Tomáš Sixta, Julio Junior, CS Jacques, Pau Buch-Cardona, Eduard Vazquez, and Sergio Escalera. Fairface challenge at eccv 2020: Analyzing bias in face recognition. *arXiv preprint arXiv:2009.07838*, 2020. [1](#)
- [16] 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 Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. [1](#)
- [17] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. [2](#), [6](#), [7](#)
- [18] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9322–9331, 2020. [1](#), [2](#), [3](#), [4](#), [6](#), [7](#)
- [19] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 692–702, 2019. [1](#), [2](#), [4](#), [6](#)
- [20] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. Mis-classified vector guided softmax loss for face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12241–12248, 2020. [1](#), [2](#)
- [21] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014. [1](#)
- [22] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020. [1](#)
- [23] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Sign. Process. Letters*, 23(10):1499–1503, 2016. [6](#)