Mitigating Bias in Facial Recognition Systems: Centroid Fairness Loss Optimization

Jean-Rémy Conti
LTCI, Télécom Paris
Institut Polytechnique de Paris
jean-remy.conti@telecom-paris.fr*

Stéphan Clémençon
LTCI, Télécom Paris
Institut Polytechnique de Paris
stephan.clemencon@telecom-paris.fr

Abstract

The urging societal demand for fair AI systems has put pressure on the research community to develop predictive models that are not only globally accurate but also meet new fairness criteria, reflecting the lack of disparate mistreatment with respect to sensitive attributes (*e.g.* gender, ethnicity, age). In particular, the variability of the errors made by certain Facial Recognition (FR) systems across specific segments of the population compromises the deployment of the latter, and was judged unacceptable by regulatory authorities. Designing fair FR systems is a very challenging problem, mainly due to the complex and functional nature of the performance measure used in this domain (*i.e.* ROC curves) and because of the huge heterogeneity of the face image datasets usually available for training. In this paper, we propose a novel post-processing approach to improve the fairness of pre-trained FR models by optimizing a regression loss which acts on centroid-based scores. Beyond the computational advantages of the method, we present numerical experiments providing strong empirical evidence of the gain in fairness and of the ability to preserve global accuracy.

1 Introduction

Facial Recognition (FR) systems are increasingly deployed (*e.g.* at border checkpoints), for biometric verification in particular. Although the global accuracy attained by certain FR systems is now judged satisfactory and offers clear efficiency gains (see e.g. Krizhevsky et al., 2012), their operational deployment has revealed statistically significant disparities in treatment between different segments of the population. Fairness in algorithmic decisions is now a major concern and is becoming part of the functional specifications of AI systems, and soon subject to regulation in certain application domains, including FR (Grother, 2022). Following recent scandals[2], the academic community has delved into the investigation of bias in FR systems in the last few years. This exploration extends back to early studies which examined racial bias in non-deep FR models (Phillips et al., 2003). A recent comprehensive analysis conducted by the U.S. National Institute of Standards and Technology (NIST) unveiled significant performance disparities among hundreds of academic/commercial FR algorithms, based on *e.g.* gender (Grother, 2022). In the present work, fairness is understood as the absence of (significant) disparate mistreatment, and we propose a novel methodology to reduce such bias in deep learning-based FR systems.

Related work on mitigating bias in FR. Various approaches have been explored to address bias in deep learning: pre-processing, in-processing, and post-processing methods (Caton & Haas, 2020). These strategies differ based on whether the fairness intervention occurs before, during, or after the

*Alternative correspondence: jeanremy.conti@gmail.com.

[2]See e.g. this study conducted by the American Civil Liberties Union that attracted notable media attention.

2nd Workshop on Regulatable ML at NeurIPS 2024.

training phase. Pre-processing methods are deemed unsuitable for FR purposes since balanced training datasets are actually *not enough* to mitigate bias, as illustrated by Albiero et al. (2020) for gender bias and Gwilliam et al. (2021) for racial bias. In terms of in-processing, Wang & Deng (2020) use reinforcement learning for fair decision rules, but face computational challenges. Alasadi et al. (2019) and Gong et al. (2019) employ adversarial methods to reduce bias, but these are recognized for their instability and computational needs, while Wang et al. (2019) leverage imbalanced and transfer learning techniques. Note that in-processing strategies require a complete retraining, which is notoriously costly as state-of-the-art FR systems require very large training datasets. Furthermore, these strategies lead to fairness improvements at the expense of the performance, highlighting a performance-fairness trade-off (Du et al., 2020). Concerning post-processing approaches, Dhar et al. (2021) mitigate the racial bias of a pre-trained model by enforcing the embeddings not to contain any racial information. Conti et al. (2022) reduce the gender bias using a statistical model for the embedding space, but they admit that the method is not able to tackle other types of bias. Those works change the pre-trained embeddings to improve the fairness, both in terms of false positives and false negatives. In contrast, another line of research takes a different approach, not altering the latent space but modifying the decision rule itself. Terhörst et al. (2020) intervene on the score function, while Salvador et al. (2021) rely on calibration methods. Those works focus on the bias in terms of false positives and their training set needs to have the same distribution than the test set, which may not be a realistic scenario.

Contributions. We propose a post-processing approach to mitigate the bias of a pre-trained/frozen FR model that is accessible only as a black-box, making it applicable to numerous already deployed FR systems. It is important to note that many methods fine tune state-of-the-art open-source FR models, thereby acquiring their bias, which underscores the necessity of improving their fairness properties. Our solution aims to align intra-group performance curves with those of a reference group, a functional objective that is inherently challenging. By drawing an analogy between real FR scores and centroid-based scores, we simplify the original fairness objective, enabling the use of pseudo-metrics that are easier to compute and align with modern FR loss functions. This approach thus bridges the gap between contemporary FR training and fairness mitigation. We introduce a new loss function, Centroid Fairness, which aligns these pseudo-metrics across the subgroups of the population, and use this loss to train a small model called the Fairness Module. Extensive experiments demonstrate that our Fairness Module reduces bias in pre-trained models while maintaining their performance, surpassing the traditional performance-fairness trade-off. In summary, our bias mitigation method eliminates the need for a costly retraining of a large FR model, is especially fast to train, and retains the state-of-the-art performance of the pre-trained model.

2 Background and Preliminaries

FR mainly serves two use cases: *identification*, involving the recognition of the identity of a probe image among several pre-enrolled identities, and *verification* (the primary focus of this paper), aiming at deciding whether two face images correspond to the same identity. Facial verification operates in an open-set scenario: the identities present at the test phase are often absent from the training set.

Notations. The indicator function of an event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$. Assuming that there are $K \leq +\infty$ identities within the images, a FR dataset of size N is denoted by $(\boldsymbol{x}_i, y_i)_{1 \leq i \leq N}$, where $\boldsymbol{x}_i \in \mathbb{R}^{h \times w \times c}$ is a face image of size $h \times w$, c is the color channel dimension and $y_i \in \{1, \dots, K\}$ is the identity label of $\boldsymbol{x}_i$. In face verification, the standard approach is to train an encoder function $f_\theta : \mathbb{R}^{h \times w \times c} \to \mathbb{R}^d$ (*e.g.* CNN) with learnable parameters θ to bring images from the same identity closer together and images from distinct identities far away from each other in $\mathbb{R}^d$. The latent representation of an image $\boldsymbol{x}_i$ is referred to as its *face embedding* $f_\theta(\boldsymbol{x}_i)$.

2.1 Evaluation of Face Recognition Systems

We start by explaining how the performance of a trained FR model f_θ is evaluated.

Decision rule. The *similarity score* between two face images $\boldsymbol{x}_i, \boldsymbol{x}_j$ is usually measured using the cosine similarity between their embeddings:

$$s_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \cos(f_\theta(\boldsymbol{x}_i), f_\theta(\boldsymbol{x}_j)) \in [-1, 1], \tag{1}$$

where $\cos(\boldsymbol{z}, \boldsymbol{z}') = \boldsymbol{z}^\intercal \boldsymbol{z}' / (\|\boldsymbol{z}\| \cdot \|\boldsymbol{z}'\|)$ and $||\cdot||$ is the Euclidean norm. The decision rule to decide whether both images share the same identity is obtained by applying a threshold $t \in [-1, 1]$ to this

score. If $s_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) > t$, $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is predicted to share the same identity (positive pair), while for $s_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq t$ we predict that they do not (negative pair). In this sense, face verification is a binary classification with a pair of images as input.

Evaluation metrics. The gold standard to evaluate the performance of FR models is the ROC curve, *i.e.* the False Rejection Rate (FRR) as a function of the False Acceptance Rate (FAR) as the threshold t varies[3]. In practice, FAR and FRR are computed on an evaluation dataset $(\boldsymbol{x_i}, y_i)_{1\leq i\leq N}$. The set of *genuine* (ground-truth positive) pairs is denoted as $\mathcal{G} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j), 1 \leq i < j \leq N, y_i = y_j\}$ while the set of *impostor* (ground-truth negative) pairs is $\mathcal{I} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j), 1 \leq i < j \leq N, y_i \neq y_j\}$. The metrics FAR and FRR of a FR model f_θ are then given by:

$$\mathrm{FAR}(t) = \frac{1}{|\mathcal{I}|} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j)\in\mathcal{I}} \mathbb{I}\{s_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) > t\}, \quad \mathrm{FRR}(t) = \frac{1}{|\mathcal{G}|} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j)\in\mathcal{G}} \mathbb{I}\{s_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq t\}, \quad (2)$$

where $|\cdot|$ denotes the cardinality of a set. $\mathrm{FAR}(t)$ is the proportion of *impostor scores* predicted as positive (same identity), while $\mathrm{FRR}(t)$ is the proportion of *genuine scores* predicted as negative (distinct identities). The trade-off between these two metrics play a pivotal role in assessing FR systems. For instance, in the context of airport boarding gates, achieving a very low FAR is crucial, while simultaneously maintaining a reasonable FRR to ensure a smooth and user-friendly experience.

The ROC curve, evaluated at $\alpha \in (0, 1)$, is naturally defined as $\mathrm{FRR}(t)$ with t such that $\mathrm{FAR}(t) = \alpha$. More rigorously, it is defined using the generalized inverse of a distribution function (see Appendix C.1), as $\mathrm{ROC}(\alpha) = \mathrm{FRR}\big(\ \mathrm{FAR}^{-1}(\alpha)\ \big)$. The FAR level α establishes the operational threshold of the Face Recognition system, representing the acceptable security risk. Depending on the use case, it is typically set to 10^{-i} with $i \in \{1, \ldots, 6\}$.

2.2 Fairness Metrics

To assess the fairness of a FR model, the typical approach involves examining differentials in performance across several subgroups or segments of the population. These subgroups are defined by a *sensitive attribute*, such as gender, ethnicity or age class. For a given discrete sensitive attribute that can take $A > 1$ different values in say $\mathcal{A} = \{0, 1, \ldots, A-1\}$, the attribute of identity $k \in \{1, \ldots, K\}$ is denoted by $a_k \in \mathcal{A}$. With those notations, the attribute of a face image $\boldsymbol{x}_i$ with identity y_i is thus a_{y_i}.

Intra-group metrics. For any $a \in \mathcal{A}$, the sets of genuine/impostor pairs with attribute a are:

$$\mathcal{G}_a = \{(\boldsymbol{x}_i, \boldsymbol{x}_j), i<j, y_i = y_j, a_{y_i}=a_{y_j}=a\}, \quad \mathcal{I}_a = \{(\boldsymbol{x}_i, \boldsymbol{x}_j), i<j, y_i \neq y_j, a_{y_i}=a_{y_j}=a\}. \quad (3)$$

From $\mathcal{G}_a$ and $\mathcal{I}_a$, one naturally defines the intra-group metrics $\mathrm{FAR}_a(t)$ and $\mathrm{FRR}_a(t)$ as for Eq. (2), by replacing $\mathcal{I}$ by $\mathcal{I}_a$ and $\mathcal{G}$ by $\mathcal{G}_a$. Because FR systems use a unique threshold t for their decision rule, our focus will be on comparing intra-group metrics $(\mathrm{FAR}_a(t))_{a\in\mathcal{A}}$ (and $(\mathrm{FRR}_a(t))_{a\in\mathcal{A}}$) at any fixed threshold t, as recommended by Robinson et al. (2020) and Krishnapriya et al. (2020).

Fairness metrics. We rely on fairness metrics used in previous work (Conti & Clémençon, 2024) that align with those used by the NIST in their FRVT report, analyzing the fairness of hundreds of academic/commercial FR models (Grother, 2022). As the FR performance may be more focused on the FAR metric or on FRR, depending on the use-case, we consider one fairness measure to quantify the differentials in $(\mathrm{FAR}_a(t))_{a\in\mathcal{A}}$, and another for $(\mathrm{FRR}_a(t))_{a\in\mathcal{A}}$:

$$\mathrm{BFAR}(t) = \max_{a\in\mathcal{A}} \mathrm{FAR}_a(t)/\mathrm{FAR}^\dagger(t), \quad \mathrm{BFRR}(t) = \max_{a\in\mathcal{A}} \mathrm{FRR}_a(t)/\mathrm{FRR}^\dagger(t), \quad (4)$$

where $\mathrm{FAR}^\dagger(t)$ (resp. $\mathrm{FRR}^\dagger(t)$) is the geometric mean of the values $(\mathrm{FAR}_a(t))_{a\in\mathcal{A}}$ (resp. $(\mathrm{FRR}_a(t))_{a\in\mathcal{A}}$). One can read the above acronyms as "Bias in FAR/FRR". Both metrics compare the worst FAR/FRR performance across subgroups to an aggregated performance over all subgroups, at a fixed threshold t. As for the ROC curve, this threshold t is set as $t = \mathrm{FAR}^{-1}(\alpha)$, with the FAR level $\alpha \in (0, 1)$ defining the selected appropriate security risk. The FAR level α is computed using the global population of the evaluation dataset, and not for some specific subgroup. In this sense, both fairness metrics are in fact functions of the FAR level α, instead of t, as for the ROC curve.

We note that other FR fairness metrics exist in the literature, such as the maximum difference in the values $(\mathrm{FAR}_a(t))_{a\in\mathcal{A}}$ used by Alasadi et al. (2019) and Dhar et al. (2021). However, this metric is not normalized and thus lacks interpretability.

[3]Standard ROC definitions often used $1 - \mathrm{FRR}$ (i.e., the True Positive Rate) instead of FRR. The FR community favors the use of FRR so that both metrics FAR and FRR correspond to error rates.

2.3 Training a Face Recognition System with Pseudo-Scores

We now review the state-of-the-art approach to train the deep encoder $f_\theta : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^d$ on a large training set $(\boldsymbol{x}_i, y_i)_{1\le i\le N}$ with K identities. During the training phase exclusively, a fully-connected layer is added on top of the embeddings, resulting in an output of a K-dimensional vector that predicts the identity of each image within the training set. The complete model (the encoder and the fully-connected layer) is trained as an identity classification task. The predominant form of FR loss functions is (Wang et al., 2017; Hasnat et al., 2017):

$$\mathcal{L}(\theta, \boldsymbol{\mu}) = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{e^{\kappa\ \overline{s}_\theta(\boldsymbol{x}_i, \boldsymbol{\mu}_{y_i})}}{\sum_{k=1}^{K} e^{\kappa\ \overline{s}_\theta(\boldsymbol{x}_i, \boldsymbol{\mu}_k)}}\right), \quad \text{with } \overline{s}_\theta(\boldsymbol{x}_i, \boldsymbol{\mu}_k) = \cos(f_\theta(\boldsymbol{x}_i), \boldsymbol{\mu}_k), \tag{5}$$

where the $\boldsymbol{\mu}_k$'s are the fully-connected layer's parameters ($\boldsymbol{\mu}_k \in \mathbb{R}^d$), $\kappa > 0$ is the inverse temperature of the softmax and $\boldsymbol{\mu} = (\boldsymbol{\mu}_k)_{1\le k\le K}$. Recent loss functions slightly change $\overline{s}_\theta(\boldsymbol{x}_i, \boldsymbol{\mu}_{y_i})$ in Eq. (5) by incorporating a fixed margin to penalize more the intra-class angle variations. Examples include CosFace (Wang et al., 2018) and ArcFace (Deng et al., 2019) and we apply our debiasing method on both models. Note that minimizing the loss in Eq. (5) enforces the embeddings $f_\theta(\boldsymbol{x}_i)$ of identity l to be all clustered around $\boldsymbol{\mu}_l$ on the unit hypersphere, while being far from any $\boldsymbol{\mu}_k$ such that $k \neq l$. In this sense, one might consider $\boldsymbol{\mu}_k$ as a *pseudo*-embedding which represents the identity k. The $\boldsymbol{\mu}_k$'s are often called *centroids* (Zhu et al., 2019) of a class/identity.

Pseudo-scores. As a consequence of the embedding-like nature of the centroids, we call $\overline{s}_\theta(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$ in Eq. (5) the *pseudo-score* between image $\boldsymbol{x}_i$ and centroid $\boldsymbol{\mu}_k$. Denoting the image-centroid genuine/impostor pairs by

$$\overline{\mathcal{G}} = \{(\boldsymbol{x}_i, \boldsymbol{\mu}_k),\ y_i = k\}, \qquad \overline{\mathcal{I}} = \{(\boldsymbol{x}_i, \boldsymbol{\mu}_k),\ y_i \neq k\}, \tag{6}$$

we distinguish between *genuine pseudo-scores* $\{\overline{s}_\theta(\boldsymbol{x}_i, \boldsymbol{\mu}_k), (\boldsymbol{x}_i, \boldsymbol{\mu}_k) \in \overline{\mathcal{G}}\}$ and *impostor pseudo-scores* $\{\overline{s}_\theta(\boldsymbol{x}_i, \boldsymbol{\mu}_k), (\boldsymbol{x}_i, \boldsymbol{\mu}_k) \in \overline{\mathcal{I}}\}$. These pseudo-scores effectively serve as a proxy for real scores during training. As the number of pseudo-scores is much smaller than the number of real scores, this makes the centroid-based approach much more efficient than prior strategies relying on real scores, like triplet loss-based approaches (Schroff et al., 2015). Our Fairness Module defined in the next section builds upon this strategy to preserve efficiency in the post-processing step.

3 Centroid Fairness

We now present our approach. We assume to have access to a pre-trained FR model f, trained on a training set $(\boldsymbol{x}_i, y_i)_{1\le i\le N}$. We consider here a *black-box setting*, where f is only available for inference and its training parameters are unavailable (hence we drop θ from the notation f_θ). Face images are encoded by f into embeddings, and scores (Eq. 1) are computed for all embedding pairs of the test set. Those scores define $\mathrm{FAR}_a(t)$, $\mathrm{FRR}_a(t)$ metrics for the pre-trained model f, for each subgroup $a \in \mathcal{A}$, and the $\mathrm{FAR}_a(t)$ curves are typically not aligned/equal when a varies (and similarly for $\mathrm{FRR}_a(t)$). The objective is to improve the fairness metrics of f (as defined in Eq. 4). Our post-processing method consists in training a small fair model acting on the output of f. In practice, we use the same training set as f for fair comparisons (see Appendix B.1 for a discussion).

Pseudo-scores estimation. Our approach uses the pseudo-scores of f, but assuming access to the pre-trained centroids $\boldsymbol{\mu}_k$ is not realistic as many open-source pre-trained FR models do not provide those parameters. Following Section 2.3, a natural way to estimate the centroid of identity k for f consists in taking the average of all normalized embeddings $f(\boldsymbol{x}_i)$ of identity k:

$$\boldsymbol{\mu}_k^{(0)} = \tfrac{1}{n_k}\textstyle\sum_{i=1}^{N} \mathbb{I}\{y_i = k\}\frac{f(\boldsymbol{x}_i)}{\|f(\boldsymbol{x}_i)\|} \in \mathbb{R}^d, \tag{7}$$

where n_k is the number of images $\boldsymbol{x}_i$ from the training set whose identity satisfies $y_i = k$. The (estimated) *pre-trained* pseudo-scores $\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$ can then be computed using Eq. (5).

3.1 Pseudo-Score Transformation

As seen in Section 2.2, to achieve fairness for a trained model f, the intra-group metrics $(\mathrm{FAR}_a(t))_{a\in\mathcal{A}}$ and $(\mathrm{FRR}_a(t))_{a\in\mathcal{A}}$ should be nearly constant when $a \in \mathcal{A}$ varies, at any threshold $t \in [-1, 1]$. Those intra-group metrics are computed with real scores $s(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Pseudo-metrics. Motivated by the analogy between real scores $s(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and pseudo-scores $\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$ in Section 2.3, we define centroid versions of the intra-group metrics. The real intra-group metrics involve image pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ which belong to $\mathcal{G}_a$ or to $\mathcal{I}_a$ (see Eq. 3), *i.e.* genuine/impostor pairs sharing the same attribute $a \in \mathcal{A}$. Extending the definitions of the image-centroid pairs $\overline{\mathcal{G}}$ and $\overline{\mathcal{I}}$ in Eq. (6), we define their counterparts for image-centroid pairs sharing the same attribute $a \in \mathcal{A}$:

$$\overline{\mathcal{G}}_a = \{(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}), y_i = k, a_{y_i} = a_k = a\}, \quad \overline{\mathcal{I}}_a = \{(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}), y_i \neq k, a_{y_i} = a_k = a\}. \tag{8}$$

One can then naturally define the intra-group *pseudo-metrics* for the pre-trained model f:

$$\overline{\mathrm{FAR}}_a(t) = 1/|\overline{\mathcal{I}}_a| \sum_{(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{I}}_a} \mathbb{I}\{\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) > t\}, \quad \overline{\mathrm{FRR}}_a(t) = 1/|\overline{\mathcal{G}}_a| \sum_{(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{G}}_a} \mathbb{I}\{\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \leq t\}. \tag{9}$$

Just like pseudo-scores $\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$ are surrogates for real scores $s(\boldsymbol{x}_i, \boldsymbol{x}_j)$, intra-group pseudo-metrics $\overline{\mathrm{FAR}}_a$ and $\overline{\mathrm{FRR}}_a$ are surrogates for intra-group real metrics FAR_a and FRR_a. Those pseudo-metrics are displayed in Fig. 3 and 5 for the FR model ArcFace (solid lines), evaluated on the BUPT dataset which is annotated with race labels a (see Section 4 for more details). For the same model and the same dataset, we show the real metrics FAR_a and FRR_a in Fig. 4 and 6 (solid lines). We see that pseudo and real metrics behave similarly, and the ranking of the best performance among groups is preserved. We will thus work with intra-group pseudo metrics for efficiency purposes.

Pseudo-metric curve alignment. Our objective is to align the curves $(\overline{\mathrm{FAR}}_a(t))_{a \in \mathcal{A}}$ across the values of a, for any t, and similarly for $(\overline{\mathrm{FRR}}_a(t))_{a \in \mathcal{A}}$. For simplicity, let us consider two subgroups: a and r, where r is the *reference* subgroup on which one would like to align all other subgroups. In this case, the objective is to align the curve $\overline{\mathrm{FAR}}_a(t)$ on $\overline{\mathrm{FAR}}_r(t)$, and $\overline{\mathrm{FRR}}_a(t)$ on $\overline{\mathrm{FRR}}_r(t)$.

To explain how the transformation works, we rely on Figure 1. Let $(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{G}}_a$ be an image-centroid genuine pair with attribute a, to which the pre-trained model f assigns the pseudo-score $\overline{s}_a^{(+)} := \overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$ shown in ⓐ. $\overline{s}_a^{(+)}$ is only involved in the computation of $\overline{\mathrm{FRR}}_a(t)$ of the pre-trained model f (see Eq. 9). Let $\alpha := \overline{\mathrm{FRR}}_a[\overline{s}_a^{(+)}]$ be the $\overline{\mathrm{FRR}}_a$ metric evaluated at $\overline{s}_a^{(+)}$, shown in ⓑ. By definition of $\overline{\mathrm{FRR}}_a$, $\overline{s}_a^{(+)}$ is the α-th quantile of the genuine pseudo-scores of attribute a (*i.e.* of $\{\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) : (\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{G}}_a\}$). To have $\overline{\mathrm{FRR}}_a(t)$ aligned with $\overline{\mathrm{FRR}}_r(t)$, it suffices that their respective quantiles are equal. The α-th quantile of the genuine pseudo-scores of attribute r (*i.e.* of $\{\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) : (\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{G}}_r\}$), is attained for a certain pseudo-score $\overline{s}_{\text{target}}$ satisfying $\overline{\mathrm{FRR}}_r[\overline{s}_{\text{target}}] = \alpha$ (shown in ⓒ). Using the quantile function $(\overline{\mathrm{FRR}}_r)^{-1}$, it means that $\overline{s}_{\text{target}} = (\overline{\mathrm{FRR}}_r)^{-1}(\alpha)$, as shown in ⓓ.

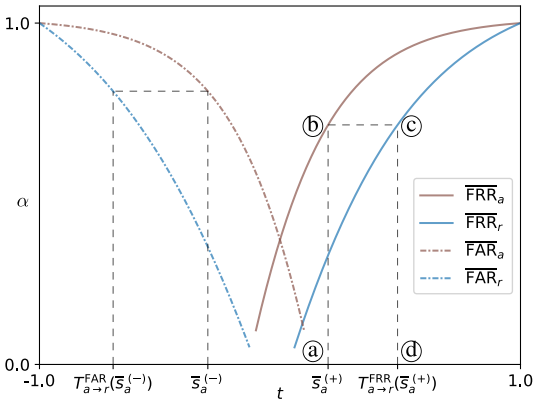

Figure 1: Pseudo-score transformation to achieve fairness. From a pseudo-score $\overline{s}_a^{(-)}$ (resp. $\overline{s}_a^{(+)}$) of an image-centroid impostor (resp. genuine) pair sharing the attribute a, the pseudo-metric $\overline{\mathrm{FAR}}_a(\overline{s}_a^{(-)}) = \alpha$ (resp. $\overline{\mathrm{FRR}}_a(\overline{s}_a^{(+)}) = \alpha$) is computed. The transformed score is the score which makes the reference pseudo-metric $\overline{\mathrm{FAR}}_r$ (resp. $\overline{\mathrm{FRR}}_r$) equal to α, among the scores from image-centroid pairs of attribute r.

Thus, for all $(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{G}}_a$, we define the pseudo-score transformation for $\overline{s}_a^{(+)} := \overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$ as

$$T^{\mathrm{FRR}}_{a\to r}(\overline{s}_a^{(+)}) = (\overline{\mathrm{FRR}}_r)^{-1} \circ \overline{\mathrm{FRR}}_a[\overline{s}_a^{(+)}].$$

Similarly, for $(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{I}}_a$, we define the pseudo-score transformation for $\overline{s}_a^{(-)} := \overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$ as

$$T^{\mathrm{FAR}}_{a\to r}(\overline{s}_a^{(-)}) = (\overline{\mathrm{FAR}}_r)^{-1} \circ \overline{\mathrm{FAR}}_a[\overline{s}_a^{(-)}].$$

The notation $\circ$ denotes the composition operator, and we refer to Appendix C.2 for rigorous definitions of the quantiles $\overline{\mathrm{FAR}}_r^{-1}$ and $\overline{\mathrm{FRR}}_r^{-1}$. The transformations $T^{\mathrm{FAR}}_{a\to r}$, $T^{\mathrm{FRR}}_{a\to r}$ are illustrated in

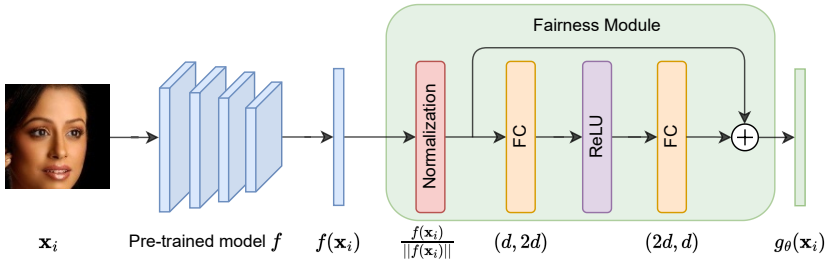

Figure 2: The proposed Fairness Module framework. A frozen pre-trained model f outputs the embedding $f(\boldsymbol{x}_i)$ for the image $\boldsymbol{x}_i$. The Fairness Module outputs a new fair embedding $g_\theta(\boldsymbol{x}_i)$.

Fig. 1. Assume that all the pseudo-scores of group a are modified as follows: all pairs from $\overline{\mathcal{I}}_a$ are transformed with $T_{a\to r}^{\mathrm{FAR}}$ while all pairs from $\overline{\mathcal{G}}_a$ are transformed with $T_{a\to r}^{\mathrm{FRR}}$. This operation changes the pseudo-metrics $\overline{\mathrm{FAR}}_a(t)$ and $\overline{\mathrm{FRR}}_a(t)$. As detailed in Appendix B.2, and proved with Proposition B.1, the newly obtained pseudo-metrics are aligned with $\overline{\mathrm{FAR}}_r$ and $\overline{\mathrm{FRR}}_r$ respectively. In other words, $T_{a\to r}^{\mathrm{FAR}}$ is an impostor pseudo-score transformation which aligns $\overline{\mathrm{FAR}}_a(t)$ on $\overline{\mathrm{FAR}}_r(t)$, and $T_{a\to r}^{\mathrm{FRR}}$ is a genuine pseudo-score transformation which aligns $\overline{\mathrm{FRR}}_a(t)$ on $\overline{\mathrm{FRR}}_r(t)$.

3.2 Fairness Module

Leveraging the pseudo-score transformations which align the group-wise pseudo-metrics with the pseudo-metrics of a reference group presented in Section 3.1, we are now ready to present the Fairness Module which will be responsible for learning these transformations.

The Fairness Module g_θ with parameters θ takes as input an embedding $f(\boldsymbol{x}_i) \in \mathbb{R}^d$ of the pre-trained model f and outputs a new embedding $g_\theta(\boldsymbol{x}_i) \in \mathbb{R}^d$ of the same dimension (see Fig. 2). To prioritize simplicity and scalability, the architecture is a shallow MLP of size (d, $2d$, d), with the first layer followed by a ReLU activation. Motivated by the normalization of FR embeddings, both for FR training and FR evaluation (see Section 2), we add a normalization step before entering the MLP. We also add a shortcut connection after the normalization step in order to allow for the model to fit the identity function easily: by setting all the MLP's weights to 0, the Fairness Module outputs the same embeddings than f. As gains in FR fairness are often accompanied by high losses in performance (Dhar et al., 2021), this shortcut connection makes it easy for the module to recover the performance of f if needed.

To learn the pseudo-score transformation introduced in Section 3.1, the Fairness Module needs to output new pseudo-scores. Thus, in addition to the new embeddings $g_\theta(\boldsymbol{x}_i)$, we learn K new centroids $\boldsymbol{\mu}_k$, as detailed in the next section. Since the shortcut connection makes it easy for the Fairness Module to recover the embedding space of the pre-trained model f, we initialize those new centroids $\boldsymbol{\mu}_k$ with the pre-trained centroids $\boldsymbol{\mu}_k^{(0)}$. Thus, the parameters used to train the Fairness Module are: the MLP's weights θ and the K new centroids $\boldsymbol{\mu}_k$. The new pseudo-scores obtained from the Fairness Module are then given by $\overline{s}_\theta(\boldsymbol{x}_i, \boldsymbol{\mu}_k) = \cos(g_\theta(\boldsymbol{x}_i), \boldsymbol{\mu}_k) \in [-1, 1]$. Note that we use the notation $\overline{s}_\theta(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$ for the pseudo-scores of the Fairness Module, while the pseudo-scores of the pre-trained model f are denoted by $\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) = \cos(f(\boldsymbol{x}_i), \boldsymbol{\mu}_k^{(0)})$.

3.3 Centroid Fairness Loss

We now present the loss function used to train the Fairness Module, following the pseudo-score transformation of Section 3.1. Recall that we have access to estimated pre-trained centroids $\boldsymbol{\mu}_k^{(0)}$, the pre-trained pseudo-scores $\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$ and thus the (pre-trained) pseudo-metrics $\overline{\mathrm{FAR}}_a$, $\overline{\mathrm{FRR}}_a$, for any $a \in \mathcal{A}$. A reference group $r \in \mathcal{A}$ is chosen, on which to align the pseudo-metrics of all groups.

The loss function we propose formulates the problem as a regression task over the Fairness Module pseudo-scores $\overline{s}_\theta(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$, where the target for this regression are given by transformed pre-trained pseudo-scores. More precisely, let us consider an impostor image-centroid pair $(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{I}}_{a_k}$ shar-

ing the same attribute $a_{y_i} = a_k$. The pre-trained model assigned this pair the pseudo-score $\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$. As seen in Section 3.1, to achieve fairness, the pre-trained model f should have given the pseudo-score

$$T^{\mathrm{FAR}}_{a_k \to r}(\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})) = (\overline{\mathrm{FAR}}_r)^{-1} \circ \overline{\mathrm{FAR}}_{a_k}[\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})]$$

to this pair. Therefore, we use this as the target score for the corresponding pseudo-score $\overline{s}_\theta(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$ of the Fairness Module. In other words, we learn new impostor pseudo-scores such that the FAR pseudo-metric of the Fairness Module aligns with the pre-trained pseudo-metric $\overline{\mathrm{FAR}}_r$ of the reference group r. We use the squared error as loss function for this pair $(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$:

$$l^{(i,k)}_{\mathrm{FAR}}(\theta, \boldsymbol{\mu}) = \left[\ \overline{s}_\theta(\boldsymbol{x}_i, \boldsymbol{\mu}_k) - T^{\mathrm{FAR}}_{a_k \to r}(\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}))\ \right]^2.$$

The same idea follows for a genuine image-centroid pair $(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{G}}_{a_k}$. The pre-trained model f gave the pseudo-score $\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$ and thus the target score is:

$$T^{\mathrm{FRR}}_{a_k \to r}(\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})) = (\overline{\mathrm{FRR}}_r)^{-1} \circ \overline{\mathrm{FRR}}_{a_k}[\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})],$$

and the loss function for this pair is then:

$$l^{(i,k)}_{\mathrm{FRR}}(\theta, \boldsymbol{\mu}) = \left[\ \overline{s}_\theta(\boldsymbol{x}_i, \boldsymbol{\mu}_k) - T^{\mathrm{FRR}}_{a_k \to r}(\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}))\ \right]^2.$$

Note that the target scores of our regression task are fully determined by the pre-trained model f, and thus fixed before the training of the Fairness Module. Finally, we combine these pair-level loss functions into two *weighted* global losses, one for the FAR bias and another for the FRR bias:

$$\mathcal{L}_{\mathrm{FAR}}(\theta, \boldsymbol{\mu}) = \Bigg(\sum_{\substack{1 \le j \le N \\ 1 \le l \le K}} w^{(j,l)}_{\mathrm{FAR}} \Bigg)^{-1} \sum_{\substack{1 \le i \le N \\ 1 \le k \le K}} w^{(i,k)}_{\mathrm{FAR}}\ l^{(i,k)}_{\mathrm{FAR}}(\theta, \boldsymbol{\mu}),$$

$$\mathcal{L}_{\mathrm{FRR}}(\theta, \boldsymbol{\mu}) = \Bigg(\sum_{\substack{1 \le j \le N \\ 1 \le l \le K}} w^{(j,l)}_{\mathrm{FRR}} \Bigg)^{-1} \sum_{\substack{1 \le i \le N \\ 1 \le k \le K}} w^{(i,k)}_{\mathrm{FRR}}\ l^{(i,k)}_{\mathrm{FRR}}(\theta, \boldsymbol{\mu}).$$

We define the weights $w^{(i,k)}_{\mathrm{FAR}}$ and $w^{(i,k)}_{\mathrm{FRR}}$ given to $l^{(i,k)}_{\mathrm{FAR}}$ and $l^{(i,k)}_{\mathrm{FRR}}$ as:

$$w^{(i,k)}_{\mathrm{FAR}} = \frac{\mathbb{I}\{y_i \neq k\}\mathbb{I}\{a_{y_i} = a_k\}}{|\overline{\mathcal{I}}_{a_k}|\ \overline{\mathrm{FAR}}_{a_k}[\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})]}, \quad w^{(i,k)}_{\mathrm{FRR}} = \frac{\mathbb{I}\{y_i = k\}\mathbb{I}\{a_{y_i} = a_k\}}{|\overline{\mathcal{G}}_{a_k}|\ \overline{\mathrm{FRR}}_{a_k}[\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})]}. \tag{10}$$

These weights are carefully chosen to match our objective. First, as seen above, $l^{(i,k)}_{\mathrm{FAR}}$ is only used for pairs in $\overline{\mathcal{I}}_{a_k}$, and $l^{(i,k)}_{\mathrm{FRR}}$ for pairs in $\overline{\mathcal{G}}_{a_k}$. Thus, it is necessary to enforce $w^{(i,k)}_{\mathrm{FAR}} \propto \mathbb{I}\{y_i \neq k\}\mathbb{I}\{a_{y_i} = a_k\}$ and $w^{(i,k)}_{\mathrm{FRR}} \propto \mathbb{I}\{y_i = k\}\ \mathbb{I}\{a_{y_i} = a_k\}$. Second, the diversity of FR use-cases requires to have good performance and fairness at all FAR and FRR levels, especially at very low levels. We thus seek to give the same importance to all FAR and FRR levels. To account for the fact that the number of pairs in a given FRR or FAR interval is proportional to the length of this interval, we weight each pair $(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$ by a factor $\overline{\mathrm{FAR}}_{a_k}[\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})])^{-1}$ (if it is an impostor) or $\overline{\mathrm{FRR}}_{a_k}[\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})])^{-1}$ (if it is genuine). Without this weighting, 90% of the non-zero terms of the loss $\mathcal{L}_{\mathrm{FRR}}$ would be devoted to achieve fairness at thresholds t associated to FRR levels in $(10^{-1}, 10^0]$, by definition of $\overline{\mathrm{FRR}}_a(t)$. Third, we seek to give the same weight to all groups, irrespective of the number of images for each group. As groups are typically imbalanced in real datasets, we divide by a factor $|\overline{\mathcal{I}}_{a_k}|$ the weight of impostor pairs with attribute a_k, and by a factor $|\overline{\mathcal{G}}_{a_k}|$ the weight of genuine pairs with attribute a_k. A more detailed discussion on the choice of those weights can be found in Appendix B.3.

Finally, we define the Centroid Fairness objective as the sum of the two previous loss functions, each responsible for either FAR fairness or FRR fairness:

$$\mathcal{L}_{\mathrm{CF}}(\theta, \boldsymbol{\mu}) = \mathcal{L}_{\mathrm{FAR}}(\theta, \boldsymbol{\mu}) + \mathcal{L}_{\mathrm{FRR}}(\theta, \boldsymbol{\mu}). \tag{11}$$

Limitations. Our method is specific to post-processing as it needs some pre-trained information and it may be promising to define a new reference/target arbitrarily for in-processing. In addition, the $\mathcal{L}_{\mathrm{CF}}$ loss requires the attribute labels for the training set only. This is a small limitation as there exists such public datasets and the FR fairness literature builds on them. Moreover, current attribute predictors (Karkkainen & Joo, 2021), from face inputs, achieve $\sim 95\%$ accuracy on popular benchmarks. In terms of computation, the method is efficient as the Fairness Module is shallow and does not depend on the complexity of the pre-trained model (see A.3). The method takes ~ 2.5 less training time when dividing the train set size by 2 (see A.4).

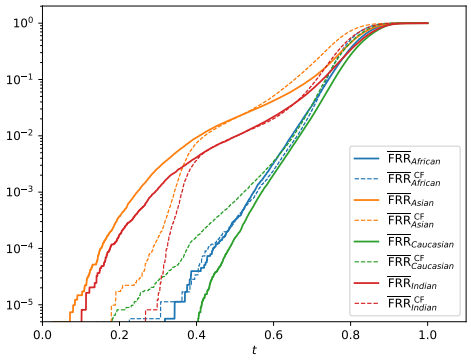

Figure 3: Pseudo-metrics $\overline{\text{FRR}}_a$ obtained with pseudo-scores, for each race. The pseudo-scores come from the pre-trained model f (solid lines) or the Fairness Module (dashed).

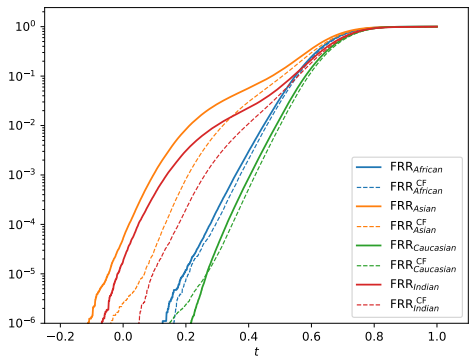

Figure 4: Real metrics FRR_a obtained with real scores, for each race. The real scores come from the pre-trained model f (solid lines) or the Fairness Module (dashed).

4 Numerical Experiments

Datasets. The training set considered in this paper is BUPT-Globalface (Wang et al., 2021). It contains 2M face images from 38k celebrities and is annotated with race attributes: $\mathcal{A} = \{\text{African, Asian, Caucasian, Indian}\}$. According to Wang et al. (2021), its racial distribution is approximately the same as the real distribution of the world's population, thus an imbalanced and realistic training set. Using the attributes in $\mathcal{A}$, we tackle the racial FR bias. Although other sensitive attributes could be relevant, we choose to mitigate the racial bias, as BUPT is the largest public FR training set, labelled with sensitive attributes (race labels only). The evaluation of all models is achieved with the test set RFW (Wang et al., 2019). RFW contains 40k images from 11k identities and is widely used to evaluate the racial bias of FR models as it is balanced in images and identities, across the four races in $\mathcal{A}$. To have good estimates of the fairness metrics BFAR and BFRR at low FAR levels α, all image pairs, sharing a same race, are used. We also employ the FairFace dataset (Sixta et al., 2020) as test set. It was introduced at the ECCV 2020 FairFace challenge and we use their predefined image pairs for evaluation. Contrary to RFW, FairFace is annotated with binary skintone attributes (Dark, Bright) and we inspect the skintone bias with BFAR and BFRR. All face images are resized to $(h, w, c) = (112, 112, 3)$ pixels with RetinaFace (Deng et al., 2020).

Pre-trained model. We take as pre-trained model f a ResNet100 (Han et al., 2017), which we train during 20 epochs on BUPT with the ArcFace (Deng et al., 2019) loss function (see B.4 for details). As many FR models, its embedding dimension is $d = 512$. To train our Fairness Module efficiently, we infer the embeddings of the whole training set and save them. Those pre-trained embeddings are the input of our Fairness Module. From those embeddings, we also compute the pre-trained centroids $\boldsymbol{\mu}_k^{(0)}$ and pseudo-scores $\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$.

Fairness models. As Caucasians are often the most represented group within FR datasets and that they benefit from better performance than other subgroups (Wang et al., 2019; Cavazos et al., 2020), we set the reference group as $r =$ Caucasian. The Fairness Module is trained with $\mathcal{L}_{\text{CF}}$ for 20 epochs on the ArcFace embeddings of BUPT, with a batch size of 4096, a learning rate equal to 10^{-3}, using the Adam (Kingma & Ba, 2014) optimizer, on 2 Tesla-V100-32GB GPUs during 40 minutes. Note that our loss function $\mathcal{L}_{\text{CF}}$ does not have any hyperparameter. The current state-of-the-art post-processing method for racial bias mitigation of FR models is achieved by PASS-s (Dhar et al., 2021), which we train on the ArcFace embeddings (see Appendix B.6 for details). As our method, PASS-s transforms the embeddings generated by a pre-trained model. However, it adopts an adversarial training paradigm, simultaneously training the model to classify identities while minimizing the encoding of race within the new embeddings. Despite the appeal of race-independent embeddings, we think that demographic characteristics are an essential part of one's identity, and their elimination may result in a notable performance loss.

Results. We show the pseudo-metrics $\overline{\text{FRR}}_a(t)$ and the metrics $\text{FRR}_a(t)$ for the pre-trained model f on BUPT respectively in Fig. 3 and 4 (solid lines). The analogy between both types of metrics is clear as they behave similarly and the ranking of the best performance among races is preserved. Fig. 3 also

Table 1: Evaluation metrics on RFW at several FAR levels. The ROC metric is expressed as a percentage (%). **Bold**=Best, Underlined=Second best.

	$\text{FAR} = 10^{-6}$			$\text{FAR} = 10^{-5}$			$\text{FAR} = 10^{-4}$		
MODEL	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR
ARCFACE	**23.15**	3.75	1.24	**13.39**	3.13	1.31	6.19	2.68	1.41
ARCFACE + PASS-S	43.80	3.42	**1.14**	30.60	2.98	**1.20**	18.62	2.77	1.28
ARCFACE + CF	23.30	**2.37**	1.16	13.44	**2.29**	1.21	**6.10**	**2.04**	**1.24**

Table 2: Evaluation metrics on FairFace at several FAR levels. The ROC metric is expressed as a percentage (%). **Bold**=Best, Underlined=Second best.

	$\text{FAR} = 10^{-4}$			$\text{FAR} = 10^{-3}$			$\text{FAR} = 10^{-2}$		
MODEL	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR
ARCFACE	**26.70**	3.15	1.08	**18.70**	1.79	1.11	**11.74**	1.26	1.11
ARCFACE + PASS-S	33.86	1.75	1.10	26.23	**1.39**	1.11	17.67	1.20	1.12
ARCFACE + CF	28.69	**1.51**	**1.07**	19.43	1.49	**1.09**	11.82	**1.06**	**1.09**

displays the pseudo-metrics $\overline{\text{FRR}}_a^{\text{CF}}(t)$ computed from the pseudo-scores $\overline{s}_\theta$ of the trained Fairness Module (dashed lines). As we set the Caucasians as the reference group, the loss $\mathcal{L}_{\text{CF}}$ enforces the pseudo-metrics $\overline{\text{FRR}}_a^{\text{CF}}(t)$ to align with $\overline{\text{FRR}}_{\text{Caucasian}}(t)$ of f. This is typically the case for the groups having the worst performance (Asians and Indians) while the counterpart is a degradation of the performance of the Caucasians, leading to pseudo-metrics $\overline{\text{FRR}}_a^{\text{CF}}(t)$ having less bias than for f. This positive impact reflects on the real metrics $\text{FRR}_a^{\text{CF}}(t)$ in Fig.4 (dashed lines) where they become closer together than for f. We postpone similar observations for the FAR pseudo-metrics and real metrics to the supplementary material (see A.1).

The pre-trained model, PASS-s and our Fairness Module (trained with Centroid Fairness loss) are evaluated on RFW in Table 1. Their performance (ROC defined in Section 2.1) as well as their fairness properties (BFAR, BFRR in Eq. (4)) are reported, at several FAR levels. Our approach succeeds in reducing the bias of ArcFace in all regimes, both in terms of FAR and FRR. PASS-s is also able to mitigate this bias, most of the time. However, note that the performance of PASS-s is significantly worse than that of the pre-trained model while our Fairness Module achieves comparable performance to ArcFace. The same evaluation metrics than for RFW are reported at several FAR levels in Table 2 for the FairFace dataset. Our approach succeeds in mitigating a good part of the skintone bias, while keeping similar performance to ArcFace. Our method is robust to a change of pre-trained model (see A.3 for three other models) and a change of training set (see A.4).

5 Conclusion

In this paper, we provide an analogy between real FR scores and centroid-based scores. This allows to define some pseudo-metrics which bridge the gap between FR training and fairness evaluation. The Centroid Fairness loss is presented and shown to align those new metrics across subgroups, allowing a small model to reduce the pre-trained racial bias without sacrificing performance. We emphasize that no sensitive attribute is needed during deployment/inference.

Reproducibility. We plan to release the code used to conduct our experiments with the Centroid Fairness loss. The training of ArcFace/PASS-s follows their official code, as specified in Appendix B.4 and Appendix B.6.

Impact statement. This paper presents a novel methodology aiming at mitigating the bias of FR systems based on deep learning, when evaluated by means of a specific fairness metric described therein. We stress that the proposed approach does not fully eliminate the bias as measured by the chosen metric, nor does it address other types of bias that may be relevant in the context of practical deployments. We also stress that the purpose of our work is not to advocate for or endorse the use of FR technologies for any application in society, nor to comment on regulatory aspects.

Acknowledgments and Disclosure of Funding

This research was partially supported by the French National Research Agency (ANR), under grant ANR-20-CE23-0028 (LIMPID project). The authors would like to thank Aurélien Bellet for his helpful comments.

References

Alasadi, J., Al Hilli, A., and Singh, V. K. Toward fairness in face matching algorithms. In *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*, pp. 19–25, 2019. 2, 3

Albiero, V., Zhang, K., and Bowyer, K. W. How does gender balance in training data affect face recognition accuracy? *arXiv preprint arXiv:2002.02934*, 2020. 2

Caton, S. and Haas, C. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020. 1

Cavazos, J. G., Phillips, P. J., Castillo, C. D., and O'Toole, A. J. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE transactions on biometrics, behavior, and identity science*, 3(1):101–111, 2020. 8

Chen, S., Liu, Y., Gao, X., and Han, Z. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices, 2018. 13

Conti, J.-R. and Clémençon, S. Assessing uncertainty in similarity scoring: Performance & fairness in face recognition. In *International Conference on Learning Representations*, 2024. 3

Conti, J.-R., Noiry, N., Clémençon, S., Despiegel, V., and Gentric, S. Mitigating gender bias in face recognition using the von mises-fisher mixture model. In *International Conference on Machine Learning*, pp. 4344–4369. PMLR, 2022. 2

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019. 4, 8, 16

Deng, J., Guo, J., Ververas, E., Kotsia, I., and Zafeiriou, S. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5203–5212, 2020. 8

Dhar, P., Gleason, J., Roy, A., Castillo, C. D., and Chellappa, R. Pass: Protected attribute suppression system for mitigating bias in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15087–15096, 2021. 2, 3, 6, 8, 12, 17

Du, M., Yang, F., Zou, N., and Hu, X. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4):25–34, 2020. 2

Gong, S., Liu, X., and Jain, A. K. Jointly de-biasing face recognition and demographic attribute estimation. *arXiv preprint arXiv:1911.08080*, 2019. 2

Grother, P. Face recognition vendor test (frvt) part 8: Summarizing demographic differentials. 2022. 1, 3

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

Han, D., Kim, J., and Kim, J. Deep pyramidal residual networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. doi: 10.1109/cvpr.2017.668. URL `http://dx.doi.org/10.1109/CVPR.2017.668`. 8

Hasnat, M., Bohné, J., Milgram, J., Gentric, S., Chen, L., et al. von mises-fisher mixture model-based deep learning: Application to face verification. *arXiv preprint arXiv:1706.04264*, 2017. 4

Hsieh, F. and Turnbull, B. W. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, 24(1):25–40, 1996. 17

Karkkainen, K. and Joo, J. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1548–1558, 2021. 7

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 8, 16

Krishnapriya, K., Albiero, V., Vangara, K., King, M. C., and Bowyer, K. W. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020. 3

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

Norwood, K. J. *Color matters: Skin tone bias and the myth of a postracial America*. Routledge, 2013. 12

Phillips, P. J., Grother, P., Micheals, R., Blackburn, D. M., Tabassi, E., and Bone, M. Face recognition vendor test 2002. In *2003 IEEE International SOI Conference. Proceedings (Cat. No. 03CH37443)*, pp. 44. IEEE, 2003. 1

Robinson, J. P., Livitz, G., Henon, Y., Qin, C., Fu, Y., and Timoner, S. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–1, 2020. 3

Salvador, T., Cairns, S., Voleti, V., Marshall, N., and Oberman, A. M. Faircal: Fairness calibration for face verification. In *International Conference on Learning Representations*, 2021. 2

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015. 4

Sixta, T., Jacques Junior, J. C., Buch-Cardona, P., Vazquez, E., and Escalera, S. Fairface challenge at eccv 2020: Analyzing bias in face recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pp. 463–481. Springer, 2020. 8, 12

Terhörst, P., Kolf, J. N., Damer, N., Kirchbuchner, F., and Kuijper, A. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognition Letters*, 140:332–338, 2020. 2

Wang, F., Xiang, X., Cheng, J., and Yuille, A. L. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1041–1049, 2017. 4

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018. 4, 12

Wang, M. and Deng, W. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9322–9331, 2020. 2

Wang, M., Deng, W., Hu, J., Tao, X., and Huang, Y. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 692–702, 2019. 2, 8

Wang, M., Zhang, Y., and Deng, W. Meta balanced network for fair face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):8433–8448, 2021. 8, 13

Zhu, Q., Zhang, P., Wang, Z., and Ye, X. A new loss function for cnn classifier based on predefined evenly-distributed class centroids. *IEEE Access*, 8:10888–10895, 2019. 4

A Additional experiments

A.1 FAR pseudo-metrics and real metrics on BUPT

The equivalents of Fig. 3 and 4 are displayed in Fig. 5 and 6. The same analogy between pseudo-metrics and real metrics can be observed (similar behavior, preserved ranking among races), for the pre-trained model f and for the Fairness Module. However, our Fairness Module seems to focus its efforts on aligning the pseudo-metrics $\overline{\mathrm{FAR}}_a^{\mathrm{CF}}$ onto the reference curve $\overline{\mathrm{FAR}}_{\mathrm{Caucasian}}$, at low FAR levels only. This might be due to our choice of weights $w_{\mathrm{FAR}}^{(i,k)}$ appearing in $\mathcal{L}_{\mathrm{FAR}}$. We observed empirically that this choice of weights tends to give too much importance at aligning $\overline{\mathrm{FAR}}_a^{\mathrm{CF}}$ at low FAR levels. Choosing these weights is a difficult question and we tried to provide some rational arguments about them. However, one benefit of our choices is that the gap between the dashed curves is reduced, for both pseudo-metrics and real metrics of our Fairness Module, at low FAR levels, where it is generally the most difficult task.

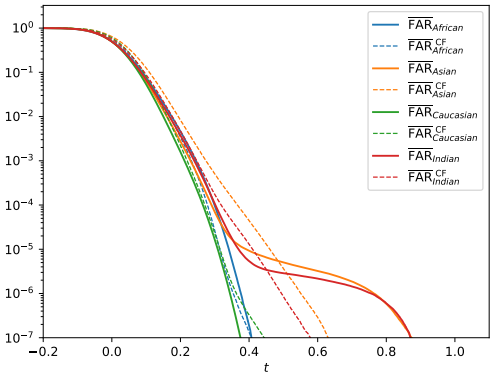

Figure 5: Pseudo-metrics $\overline{\mathrm{FAR}}_a$ obtained with pseudo-scores, for each race. The pseudo-scores come either from the pre-trained model f (solid lines), or from the Fairness Module (dashed lines).

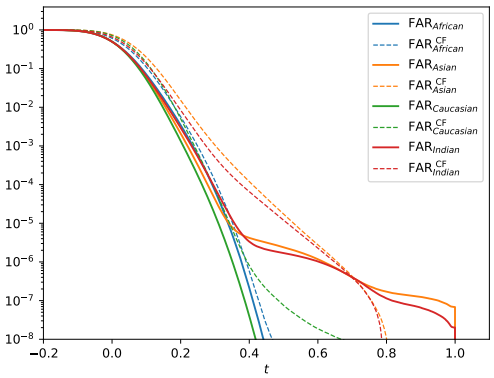

Figure 6: Real metrics FAR_a obtained with real scores, for each race. The real scores come either from the pre-trained model f (solid lines), or from the Fairness Module (dashed lines).

A.2 Evaluation on FairFace

We test the pre-trained model ArcFace, PASS-s and our Fairness Module trained with Centroid Fairness loss on another dataset: the evaluation is performed with the FairFace dataset Sixta et al. (2020). It was introduced at the ECCV 2020 FairFace challenge and we use their predefined image pairs for evaluation. Contrary to RFW, FairFace is annotated with binary skintone attributes: $\mathcal{A}_{\mathrm{FF}} = \{\text{Dark, Bright}\}$. The skintone bias is related to the racial bias as there is some correlation between race and skintone Norwood (2013); Dhar et al. (2021). Note that we consider $\mathcal{A} = \{\text{African, Asian, Caucasian, Indian}\}$ for training and for evaluating on RFW, and $\mathcal{A}_{\mathrm{FF}}$ for the FairFace evaluation.

The performance (ROC) of the three models, as well as their fairness properties (BFAR, BFRR), are reported at several FAR levels in Table 3. The same conclusions as for RFW (Table 1) can be drawn. Our approach succeeds in mitigating a good part of the racial bias, while keeping similar performance to ArcFace.

A.3 Evaluations for Other Pre-Trained Models

To go beyond the pre-trained model ArcFace used in Section 4, a ResNet100 is trained on BUPT during 20 epochs, with the CosFace Wang et al. (2018) loss. Then, we train our Fairness Module with the Centroid Fairness (CF) loss on top of the CosFace embeddings of BUPT. The hyperparameters used to train CosFace and the Fairness Module are the same as for ArcFace (Section 4). Both models are evaluated on the RFW dataset in Table 4, and on FairFace (see A.2) in Table 5.

Table 3: Evaluation metrics on FairFace at several FAR levels. The ROC metric is expressed as a percentage (%). **Bold**=Best, Underlined=Second best.

	$\text{FAR} = 10^{-4}$			$\text{FAR} = 10^{-3}$			$\text{FAR} = 10^{-2}$		
MODEL	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR
ARCFACE	**26.70**	3.15	1.08	**18.70**	1.79	1.11	**11.74**	1.26	1.11
ARCFACE + PASS-S	33.86	1.75	1.10	26.23	**1.39**	1.11	17.67	1.20	1.12
ARCFACE + CF	28.69	**1.51**	**1.07**	19.43	1.49	**1.09**	11.82	**1.06**	**1.09**

Table 4: Evaluation metrics on RFW at several FAR levels, for the pre-trained model CosFace. The ROC metric is expressed as a percentage (%). **Bold**=Best.

	$\text{FAR} = 10^{-6}$			$\text{FAR} = 10^{-5}$			$\text{FAR} = 10^{-4}$		
MODEL	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR
COSFACE	**21.46**	4.14	1.15	12.03	3.81	1.20	5.30	3.01	1.28
COSFACE + CF	21.95	**3.21**	**1.10**	**11.94**	**2.82**	**1.12**	**5.18**	**2.37**	**1.15**

Table 4 and Table 5 show the robustness of our method when varying the loss function used to train the pre-trained model, but keeping the same architecture (ResNet100). We now inspect the robustness when varying the architecture of the pre-trained model.

Instead of considering a ResNet100, a ResNet34 is trained on BUPT during 20 epochs with the ArcFace loss. The hyperparameters used to train this pre-trained model ArcFace-R34 and its Fairness Module are the same as for ArcFace ResNet100 (Section 4). Both models are evaluated on the RFW dataset in Table 6, and on FairFace (see A.2) in Table 7.

Finally, a MobileFaceNet Chen et al. (2018) is trained on BUPT during 40 epochs with the ArcFace loss. This architecture is much smaller than ResNet architectures. The Fairness Module is trained on BUPT during 7 epochs using the Centroid Fairness loss. The other hyperparameters used to train this pre-trained model ArcFace-MBF and its Fairness Module are the same as for ArcFace ResNet100 (Section 4). Both models are evaluated on the RFW dataset in Table 8, and on FairFace (see A.2) in Table 9.

All the results from this section confirm the robustness of our Fairness Module to a change of pre-trained model.

A.4 Evaluations for Another Training Set

We now investigate the robustness of our method when varying the training set used to train the pre-trained model and its Fairness Module. Note that very few open-source FR datasets have ethnicity labels. The only large-scale FR datasets that satisfy this property are BUPT-Globalface (used in Section 4) and BUPT-Balancedface Wang et al. (2021). BUPT-Balancedface contains 1.3M face images from 28k celebrities and is also annotated with race attributes. Contrary to BUPT-Globalface, it is approximately race-balanced with 7k identities for each of the four available ethnicities.

In the following, BUPT-Balancedface is employed as the training set of the pre-trained model ArcFace ResNet100 and its Fairness Module. The Fairness Module is trained during 16 epochs with the Centroid Fairness loss, on 2 Tesla-V100-32GB GPUs during 15 minutes.. The framework and the remaining hyperparameters used for training are the same than for BUPT-Globalface (see Section 4). Both models are evaluated on the RFW dataset in Table 10. Those results underline the robustness of our method when varying the training set (and the distribution of its identities).

Table 5: Evaluation metrics on FairFace at several FAR levels, for the pre-trained model CosFace. The ROC metric is expressed as a percentage (%). **Bold**=Best.

	FAR = 10^{-4}			FAR = 10^{-3}			FAR = 10^{-2}		
MODEL	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR
COSFACE	**24.42**	1.77	1.08	**17.77**	1.49	1.08	11.13	1.25	1.09
COSFACE + CF	25.23	**1.60**	**1.07**	17.78	**1.36**	**1.07**	**10.95**	**1.14**	**1.08**

Table 6: Evaluation metrics on RFW at several FAR levels, for the pre-trained model ArcFace ResNet34. The ROC metric is expressed as a percentage (%). **Bold**=Best.

	FAR = 10^{-6}			FAR = 10^{-5}			FAR = 10^{-4}		
MODEL	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR
ARCFACE-R34	**29.10**	4.35	1.14	17.89	4.02	1.21	8.80	3.14	1.29
ARCFACE-R34 + CF	29.81	**3.58**	**1.08**	**17.80**	**3.06**	**1.11**	**8.63**	**2.44**	**1.14**

B Further Remarks

B.1 Reason for Using the Same Training Set as f

Note that the debiasing approach presented in this paper could work by training the fair model on another training set than the dataset used to train f. It would imply the inference of $f(\boldsymbol{x}_i)$ and the computation of the centroids $\boldsymbol{\mu}_k^{(0)}$ on a different training set. The whole procedure to train the fair model starts with those inputs and can be applied on any training set. However, in order to fairly compare f and its post-processing fair model, the same training set is used for both models, without adding data.

B.2 Effect of the Pseudo-Score Transformations $T_{a\to r}^{\mathrm{FAR}}$ and $T_{a\to r}^{\mathrm{FRR}}$

For simplicity, let us consider a population with two subgroups: a and r. r stands for the *reference* subgroup, on which one would like to align all other subgroups. In this case, the objective is to align the curve $\overline{\mathrm{FAR}}_a(t)$ on $\overline{\mathrm{FAR}}_r(t)$, and $\overline{\mathrm{FRR}}_a(t)$ on $\overline{\mathrm{FRR}}_r(t)$.

For all $(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{I}}_a$, we consider the following pseudo-score transformation for $\overline{s}_a^{(-)} := \overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$:

$$T_{a\to r}^{\mathrm{FAR}}(\overline{s}_a^{(-)}) = (\overline{\mathrm{FAR}}_r)^{-1} \circ \overline{\mathrm{FAR}}_a[\overline{s}_a^{(-)}].$$

Assume that all pseudo-scores $\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$ for $(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{I}}_a$ were transformed with this mapping. Then, $\overline{\mathrm{FAR}}_a(t)$ would be equal to:

$$\overline{\mathrm{FAR}}_{a\to r}(t) = \frac{1}{|\overline{\mathcal{I}}_a|} \sum_{(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{I}}_a} \mathbb{I}\{T_{a\to r}^{\mathrm{FAR}}(\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})) > t\}.$$

Similarly, for all $(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{G}}_a$, we set the following pseudo-score transformation for $\overline{s}_a^{(+)} := \overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$:

$$T_{a\to r}^{\mathrm{FRR}}(\overline{s}_a^{(+)}) = (\overline{\mathrm{FRR}}_r)^{-1} \circ \overline{\mathrm{FRR}}_a[\overline{s}_a^{(+)}].$$

Assume that all pseudo-scores $\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$ for $(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{G}}_a$ were transformed with this mapping. The pseudo-metric $\overline{\mathrm{FRR}}_a(t)$ would become:

$$\overline{\mathrm{FRR}}_{a\to r}(t) = \frac{1}{|\overline{\mathcal{G}}_a|} \sum_{(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{G}}_a} \mathbb{I}\{T_{a\to r}^{\mathrm{FRR}}(\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})) \leq t\}.$$

The transformations $T_{a\to r}^{\mathrm{FAR}}$ and $T_{a\to r}^{\mathrm{FRR}}$ are illustrated in Fig. 1.

Table 7: Evaluation metrics on FairFace at several FAR levels, for the pre-trained model ArcFace ResNet34. The ROC metric is expressed as a percentage (%). **Bold**=Best.

	$\text{FAR} = 10^{-4}$			$\text{FAR} = 10^{-3}$			$\text{FAR} = 10^{-2}$		
MODEL	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR
ARCFACE-R34	**29.48**	1.40	1.07	**22.21**	1.28	1.07	14.61	1.24	1.07
ARCFACE-R34 + CF	30.98	**1.07**	**1.05**	22.32	**1.08**	**1.05**	**14.21**	**1.03**	**1.05**

Table 8: Evaluation metrics on RFW at several FAR levels, for the pre-trained model ArcFace MobileFaceNet. The ROC metric is expressed as a percentage (%). **Bold**=Best.

	$\text{FAR} = 10^{-6}$			$\text{FAR} = 10^{-5}$			$\text{FAR} = 10^{-4}$		
MODEL	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR
ARCFACE-MBF	**36.65**	5.32	1.08	**24.32**	4.19	1.11	**13.33**	3.22	1.15
ARCFACE-MBF + CF	37.67	**4.18**	**1.07**	25.00	**3.66**	**1.09**	13.76	**2.86**	**1.10**

Table 9: Evaluation metrics on FairFace at several FAR levels, for the pre-trained model ArcFace MobileFaceNet. The ROC metric is expressed as a percentage (%). **Bold**=Best.

	$\text{FAR} = 10^{-4}$			$\text{FAR} = 10^{-3}$			$\text{FAR} = 10^{-2}$		
MODEL	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR
ARCFACE-MBF	**36.66**	1.77	1.06	**26.74**	1.31	**1.05**	**17.44**	1.12	1.06
ARCFACE-MBF + CF	48.27	**1.44**	**1.05**	28.24	**1.21**	**1.05**	17.49	**1.06**	**1.05**

Table 10: Evaluation metrics on RFW at several FAR levels, for the pre-trained model ArcFace ResNet100. The pre-trained model and its Fairness Module are trained on BUPT-Balancedface. The ROC metric is expressed as a percentage (%). **Bold**=Best.

	$\text{FAR} = 10^{-6}$			$\text{FAR} = 10^{-5}$			$\text{FAR} = 10^{-4}$		
MODEL	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR	ROC (%)	BFAR	BFRR
ARCFACE	**29.00**	4.68	1.29	**17.85**	4.13	1.36	**8.78**	3.39	1.48
ARCFACE + CF	29.21	**3.72**	**1.27**	17.94	**3.36**	**1.33**	8.85	**2.84**	**1.43**

Proposition B.1. *We have that:*

$$\sup_{t\in(-1,1)} |\overline{\text{FAR}}_{a\to r}(t) - \overline{\text{FAR}}_r(t)| \leq 1/|\overline{\mathcal{I}}_a|,$$

$$\sup_{t\in(-1,1)} |\overline{\text{FRR}}_{a\to r}(t) - \overline{\text{FRR}}_r(t)| \leq 1/|\overline{\mathcal{G}}_a|.$$

The proof is postponed to C.3. This result states that $T^{\text{FAR}}_{a\to r}$ (resp. $T^{\text{FRR}}_{a\to r}$) is an impostor (resp. genuine) pseudo-score transformation which aligns the curve $\overline{\text{FAR}}_a(t)$ (resp. $\overline{\text{FRR}}_a(t)$) on $\overline{\text{FAR}}_r(t)$ (resp. on $\overline{\text{FRR}}_r(t)$).

B.3 Discussion on the weights $w^{(i,k)}_{\text{FAR}}$ and $w^{(i,k)}_{\text{FRR}}$

As detailed in Section 3.3, it is necessary to enforce $w^{(i,k)}_{\text{FAR}} \propto \mathbb{I}\{y_i \neq k\}\mathbb{I}\{a_{y_i} = a_k\}$ and $w^{(i,k)}_{\text{FRR}} \propto \mathbb{I}\{y_i = k\}\,\mathbb{I}\{a_{y_i} = a_k\}$.

Let $(\boldsymbol{x}_i, \boldsymbol{\mu}^{(0)}_k) \in \overline{\mathcal{G}}_{a_k}$. From the definition of $\overline{\text{FRR}}_{a_k}$, 90% of the pseudo-scores $\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}^{(0)}_k)$ satisfy $\overline{\text{FRR}}_{a_k}[\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}^{(0)}_k)] \in (10^{-1}, 10^0]$. Thus, 90% of the non-zero terms of the loss $\mathcal{L}_{\text{FRR}}$ enforce

the Fairness Module to reach FRR fairness at FRR levels in $(10^{-1}, 10^{0}]$. Similarly, 9% of the pseudo-scores $\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$ satisfy $\overline{\mathrm{FRR}}_{a_k}[\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})] \in (10^{-2}, 10^{-1}]$, enforcing FRR fairness at FRR levels in $(10^{-2}, 10^{-1}]$. As the FRR fairness should be reached at *all* FRR levels, the loss $\mathcal{L}_{\mathrm{FRR}}$ should give the same weight for the 9% than for the 90% of the pseudo-scores. Note that the 9% of the pseudo-scores are 10 times less present than the 90% in $\mathcal{L}_{\mathrm{FRR}}$ but that the quantity $\tilde{w}_{\mathrm{FRR}}^{(i,k)} := (\overline{\mathrm{FRR}}_{a_k}[\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})])^{-1}$ is nearly 10 times higher for the 9% than for the 90%. Thus, we choose to weight each $\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$ by this quantity $\tilde{w}_{\mathrm{FRR}}^{(i,k)}$, to ensure FRR fairness at all levels.

However, there is a drawback for this weighting. Consider the case of two groups $a, r \in \mathcal{A}$, r having many more images than a within the training set. This scenario is typical of imbalanced/realistic training sets. In addition, the reference group r is set to be the more populated subgroup in our experiment (see Section 4). This setting implies that $|\overline{\mathcal{G}}_r| \gg |\overline{\mathcal{G}}_a|$. For the sake of simplicity, let $|\overline{\mathcal{G}}_r| = 10^6$ and $|\overline{\mathcal{G}}_a| = 10^3$. Now, consider the pair $(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)}) \in \overline{\mathcal{G}}_a$ which has the lowest pseudo-score, *i.e.* which minimizes $\overline{\mathrm{FRR}}_a(t)$. From the definition of $\overline{\mathrm{FRR}}_a(t)$, this pair satisfies $\overline{\mathrm{FRR}}_a[\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})] = 1/|\overline{\mathcal{G}}_a|$. Similarly, let $(\boldsymbol{x}_j, \boldsymbol{\mu}_l^{(0)}) \in \overline{\mathcal{G}}_r$ be the pair satisfying $\overline{\mathrm{FRR}}_r[\overline{s}(\boldsymbol{x}_j, \boldsymbol{\mu}_l^{(0)})] = 1/|\overline{\mathcal{G}}_r|$. The previous weights for both pairs are $\tilde{w}_{\mathrm{FRR}}^{(i,k)} = 10^3$ and $\tilde{w}_{\mathrm{FRR}}^{(j,l)} = 10^6$. Note that these pairs maximize those weights among their group a or r. One can observe a favorable weighting for all pairs in $\overline{\mathcal{G}}_r$ compared to the pairs in $\overline{\mathcal{G}}_a$. This means that the loss $\mathcal{L}_{\mathrm{FRR}}$ would enforce all groups to align their performance with the reference subgroup, but some groups more than others *i.e.* those which have lots of images. To counteract this effect, we impose the maximum weighting $\tilde{w}_{\mathrm{FRR}}^{(i,k)}$ among a subgroup to be equal for all subgroups. All the previous considerations on the weights in $\mathcal{L}_{\mathrm{FRR}}$ can be applied similarly to the weights in $\mathcal{L}_{\mathrm{FAR}}$ and lead us to the final weights:

$$w_{\mathrm{FAR}}^{(i,k)} = \frac{\mathbb{I}\{y_i \neq k\}\mathbb{I}\{a_{y_i} = a_k\}}{|\overline{\mathcal{I}}_{a_k}| \, \overline{\mathrm{FAR}}_{a_k}[\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})]},$$
$$w_{\mathrm{FRR}}^{(i,k)} = \frac{\mathbb{I}\{y_i = k\}\mathbb{I}\{a_{y_i} = a_k\}}{|\overline{\mathcal{G}}_{a_k}| \, \overline{\mathrm{FRR}}_{a_k}[\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})]}.$$

B.4 Implementation Details for the training of ArcFace

For the training of ArcFace on BUPT, we follow their official implementation[4]. We train a ResNet100, with $d = 512$, 0.9 as the momentum, 5×10^{-4} as the weight decay, a batch size of 256, a learning rate equal to 10^{-1}, for 20 epochs, as listed within their paper Deng et al. (2019).

For the MobileFaceNet version of ArcFace (see A.3), the model is trained during 40 epochs with weight decay equal to 1×10^{-4}. The other parameters remain the same than for the ResNet100.

B.5 Implementation Details of the Centroid Fairness Loss

Using the pre-trained model f, we compute all its embeddings $f(\boldsymbol{x}_i)$ on the training set and store them. Those embeddings are the input of our Fairness Module and there is no need to recompute $f(\boldsymbol{x}_i)$ each time we need $g_\theta(\boldsymbol{x}_i)$. From those embeddings, we then compute the pre-trained centroids $\boldsymbol{\mu}_k^{(0)}$ and the pre-trained pseudo-scores $\overline{s}(\boldsymbol{x}_i, \boldsymbol{\mu}_k^{(0)})$. From those pseudo-scores, one gets the pre-trained pseudo-metrics $\overline{\mathrm{FAR}}_a(t)$ and $\overline{\mathrm{FRR}}_a(t)$ for all subgroups $a \in \mathcal{A}$. Then, using the definition of their generalized inverses in C.2, we are able to compute $\overline{\mathrm{FAR}}_r^{-1}$ and $\overline{\mathrm{FRR}}_r^{-1}$. All those steps allow us to define the target scores for our regression task. Those steps are achieved before the Centroid Fairness loss training.

On BUPT, we set the reference group as $r =$ Caucasian. The Fairness Module is trained with $\mathcal{L}_{\mathrm{CF}}$ for 20 epochs on the ArcFace embeddings of BUPT, with a batch size of 4096, a learning rate equal to 10^{-3}, using the Adam Kingma & Ba (2014) optimizer. Note that our loss function $\mathcal{L}_{\mathrm{CF}}$ does not have any hyperparameter.

[4] https://github.com/deepinsight/insightface/tree/master/recognition/arcface_torch

Table 11: Hyperparameters used to train PASS-s on BUPT.

PARAMETER	VALUE
λ	10
K	2
T_{fc}	10000
T_{deb}	1200
T_{atrain}	30000
T_{plat}	20000
A^*	0.95
α_1	10^{-2}
α_2	10^{-3}
α_3	10^{-4}
T_{ep}	40
N_{ep}	50

In addition to those parameters, we use a certain image sampling. The probability of sampling an image/embedding $\boldsymbol{x}_i$ (so that it appears within the next batch in the loss $\mathcal{L}_{\mathrm{CF}}$) is inversely proportional to the number of images sharing the attribute a_{y_i} of $\boldsymbol{x}_i$ within the training set. Then, once an image is sampled for the current batch, all the pseudo-scores $\overline{s}_\theta(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$ are computed (for all K centroids) and one is able to compute the loss.

B.6 Implementation Details for PASS-s

For the training of PASS-s on the ArcFace embeddings of BUPT, we follow their official implementation[5] and use the hyperparameters which they provide within their paper Dhar et al. (2021), as listed in Table 11.

C Technical Details

C.1 A Note on the Generalized Inverse of the FAR Quantity

In 2.1, we introduced the FAR metric, defined as:

$$\mathrm{FAR}(t) = \frac{1}{|\mathcal{I}|} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{I}} \mathbb{I}\{s_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) > t\},$$

and the ROC curve as $\mathrm{ROC}\colon \alpha \in (0,1) \mapsto \mathrm{FRR}\big[\ \mathrm{FAR}^{-1}(\alpha)\ \big]$.

The generalized inverse of any cumulative distribution function (cdf) $\kappa(t)$ on $\mathbb{R}$ is defined as

$$\kappa^{-1}(\alpha) = \inf\{t \in \mathbb{R} :\ \kappa(t) \geq \alpha\}, \quad \text{for } \alpha \in (0,1). \tag{12}$$

If κ is a cdf, the generalized inverse κ^{-1} is its quantile function.

Note that the quantity $\mathrm{FAR}(t)$ is not a cdf, so that its generalized inverse is not well defined. However, the opposite of $\mathrm{FAR}(t)$, the True Rejection Rate (TRR), is a proper cdf:

$$\mathrm{TRR}(t) = 1 - \mathrm{FAR}(t) = \frac{1}{|\mathcal{I}|} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{I}} \mathbb{I}\{s_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq t\}.$$

As such, the generalized inverse $\mathrm{TRR}^{-1}(\alpha)$ is well defined for the TRR quantity with Eq. (12).

This allows to define the generalized inverse for FAR. Indeed, for any $\alpha \in (0,1)$ satisfying $\mathrm{FAR}(t) = \alpha$, one would get the following TRR

$$\mathrm{TRR}(t) = 1 - \alpha.$$

The threshold t of interest is found using the generalized inverse of TRR (as in Hsieh & Turnbull (1996)):

$$\mathrm{FAR}^{-1}(\alpha) \colon = \mathrm{TRR}^{-1}(1-\alpha) = (1 - \mathrm{FAR})^{-1}(1-\alpha).$$

[5] https://github.com/Prithviraj7/PASS

C.2 A Note on the Generalized Inverses of the $\overline{\mathrm{FAR}}_a$ and $\overline{\mathrm{FRR}}_a$ Quantities

The generalized inverse of either $\overline{\mathrm{FAR}}_a(t)$ or $\overline{\mathrm{FRR}}_a(t)$ is defined similarly as in C.1.

Note that $\overline{\mathrm{FRR}}_a(t)$ is a proper cdf, thus its generalized inverse $\overline{\mathrm{FRR}}_a^{-1}$ is defined with Eq. (12). For $\overline{\mathrm{FAR}}_a(t)$, the same idea than in C.1 is used. Its generalized inverse is defined using the cdf $1 - \overline{\mathrm{FAR}}_a(t)$.

C.3 Proof of Proposition B.1

We start by proving the following result:

$$\sup_{t\in(-1,1)} |\overline{\mathrm{FRR}}_{a\to r}(t) - \overline{\mathrm{FRR}}_r(t)| \leq 1/|\overline{\mathcal{G}}_a|.$$

By construction, the jumps of the increasing stepwise function $\overline{\mathrm{FRR}}_{a\to r}(t)$ occur at points included in

$$\{\overline{\mathrm{FRR}}_r^{-1}(l/|\overline{\mathcal{G}}_r|) :\ l = 1,\ \ldots,\ |\overline{\mathcal{G}}_r|\},$$

like $\overline{\mathrm{FRR}}_r(t)$. Hence, we have:

$$\sup_{t\in(-1,\,+1)} \left|\overline{\mathrm{FRR}}_{a\to r}(t) - \overline{\mathrm{FRR}}_r(t)\right| = \max_{1\leq l\leq|\overline{\mathcal{G}}_r|} \left|\overline{\mathrm{FRR}}_{a\to r}(\overline{\mathrm{FRR}}_r^{-1}(l/|\overline{\mathcal{G}}_r|)) - l/|\overline{\mathcal{G}}_r|\right|.$$

Therefore, for all $l \in \{1,\ \ldots,\ |\overline{\mathcal{G}}_r|\}$, we have

$$\begin{aligned}\overline{\mathrm{FRR}}_{a\to r}(\overline{\mathrm{FRR}}_r^{-1}(l/|\overline{\mathcal{G}}_r|)) &= (1/|\overline{\mathcal{G}}_a|) \times \max\Big\{k \in \{0,\ \ldots,\ |\overline{\mathcal{G}}_a|\} :\ k/|\overline{\mathcal{G}}_a| \leq l/|\overline{\mathcal{G}}_r|\Big\} \\ &= \lfloor l|\overline{\mathcal{G}}_a|/|\overline{\mathcal{G}}_r|\rfloor \ /\ |\overline{\mathcal{G}}_a|.\end{aligned}$$

We obtain that

$$\left|\overline{\mathrm{FRR}}_{a\to r}(\overline{\mathrm{FRR}}_r^{-1}(l/|\overline{\mathcal{G}}_r|)) - l/|\overline{\mathcal{G}}_r|\right| < 1/|\overline{\mathcal{G}}_a|.$$

The second result of Proposition B.1 (on the FAR quantity) is proved with the same arguments.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contribution of the paper consists of a description of a novel methodological framework to improve fairness properties of similarity-based systems and empirical evidence of its performance based on experimental work, exactly as claimed in the abstract/introduction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the analysis are discussed in the experimental section, insofar as it where the performance of the method promoted is empirically evaluated and details about the implementation are described. As explained, the promising results obtained is an encouragement to apply the general principles proposed to other datasets/tasks.We also wrote a paragraph about the limitations of our method at the end of Section 3. The computational efficiency and scaling with dataset size are discussed. The method is empirically validated with 4 pre-trained models, 2 training sets and 2 test sets.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes] .

   Justification: The methodology proposed is mainly empirical. However, the statistical framework is precisely described and the theoretical arguments explaining the rationale behind the method are rigorously established (see Proposition B.1 in the Supplementary).

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The very simple architecture of the Fairness Module is presented, as well as all hyperarameters used to train it. Complete details about the experiments are provided in the Supplementary Material (Appendix B.4, B.5, B.6). The trained models will be made public, as well as the code, upon acceptance.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Links to the public datasets used in this paper are given, permitting to reproduce fully the experiments we have carried out. The code has also been provided.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The very simple architecture of the Fairness Module is presented, as well as all hyperarameters used to train it. Complete details about the experiments are provided in the supplementary material (Appendix B.4, B.5, B.6). The trained models will be made public.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The significance of the results is assessed thanks to several evalution data sets, whose characteristics are described in the paper and in the Supplementary Material.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computing resources are discussed in Section 4 and in the Supplementary Material (Appendix A.4).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have read the NeurIPS code of ethics and abide by it.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We wrote an impact statement just before the references. This paper does not promote the use of any technology, it is a methodological attempt to improve the fairness properties of similarity-based scoring techniques without impairing predictive performance, accompanied by some empirical evidence (with the limitations that this implies, which are discussed in the article).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: The paper poses no such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: As described in the paper and Supplementary Material, the code, data and models used in the paper (all public) have been properly credited and their license and terms of use explicitly mentioned and properly respected.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: the code has been communicated through a .zip archive, with the relevant documentation.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.