Jointly De-biasing Face Recognition and Demographic Attribute Estimation

Sixue Gong Xiaoming Liu Anil K. Jain
{gongsixu, liuxm, jain}@msu.edu

Michigan State University

Abstract. We address the problem of bias in automated face recognition and demographic attribute estimation algorithms, where errors are lower on certain cohorts belonging to specific demographic groups. We present a novel de-biasing adversarial network (DebFace) that learns to extract disentangled feature representations for both unbiased face recognition and demographics estimation. The proposed network consists of one identity classifier and three demographic classifiers (for gender, age, and race) that are trained to distinguish identity and demographic attributes, respectively. Adversarial learning is adopted to minimize correlation among feature factors so as to abate bias influence from other factors. We also design a new scheme to combine demographics with identity features to strengthen robustness of face representation in different demographic groups. The experimental results show that our approach is able to reduce bias in face recognition as well as demographics estimation while achieving state-of-the-art performance.

Keywords: Bias, Feature Disentanglement, Face Recognition, Fairness

1 Introduction

Automated face recognition has achieved remarkable success with the rapid developments of deep learning algorithms. Despite the improvement in the accuracy of face recognition, one topic is of significance. Does a face recognition system perform equally well in different demographic groups? In fact, it has been observed that many face recognition systems have lower performance in certain demographic groups than others [23, 29, 42]. Such face recognition systems are said to be *biased* in terms of demographics.

In a time when face recognition systems are being deployed in the real world for societal benefit, this type of bias [1] is not acceptable. Why does the bias problem exist in face recognition systems? First, state-of-the-art (SOTA) face recognition methods are based on deep learning which requires a large collection of face images for training. Inevitably the distribution of training data has a

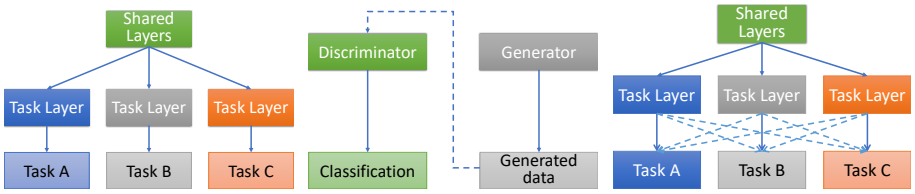
[1] This is different from the notion of machine learning bias, defined as "any basis for choosing one generalization [hypothesis] over another, other than strict consistency with the observed training instances" [15].

great impact on the performance of the resultant deep learning models. It is well understood that face datasets exhibit imbalanced demographic distributions where the number of faces in each cohort is unequal. Previous studies have shown that models trained with imbalanced datasets lead to biased discrimination [5, 49]. Secondly, the goal of deep face recognition is to map the input face image to a target feature vector with high discriminative power. The bias in the mapping function will result in feature vectors with lower discriminability for certain demographic groups. Moreover, Klare *et al.* [29] show the errors that are inherent to some demographics by studying non-trainable face recognition algorithms.

To address the bias issue, data re-sampling methods have been exploited to balance the data distribution by under-sampling the majority [16] or over-sampling the minority classes [8,39]. Despite its simplicity, valuable information may be removed by under-sampling, and over-sampling may introduce noisy samples. Naively training on a balanced dataset can still lead to bias [56]. Another common option for imbalanced data training is cost-sensitive learning that assigns weights to different classes based on (i) their frequency or (ii) the effective number of samples [6, 12]. To eschew the overfitting of Deep Neural Network (DNN) to minority classes, hinge loss is often used to increase margins among classification decision boundaries [21, 27]. The aforementioned methods have also been adopted for face recognition and attribute prediction on imbalanced datasets [24, 58]. However, such face recognition studies only concern bias in terms of *identity*, rather than our focus of *demographic bias*.

In this paper, we propose a framework to address the influence of bias on face recognition and demographic attribute estimation. In typical deep learning based face recognition frameworks, the large capacity of DNN enables the face representations to embed demographic details, including gender, race, and age [3, 17]. Thus, the biased demographic information is transmitted from the training dataset to the output representations. To tackle this issue, we assume that if the face representation does not carry discriminative information of demographic attributes, it would be unbiased in terms of demographics. Given this assumption, one common way to remove demographic information from face representations is to perform feature disentanglement via adversarial learning (Fig. 1b). That is, the classifier of demographic attributes can be used to encourage the identity representation to *not* carry demographic information. However, one issue of this common approach is that, the demographic classifier itself could be biased (*e.g.*, the race classifier could be biased on gender), and hence it will act differently while disentangling faces of different cohorts. This is clearly undesirable as it leads to demographic biased identity representation.

To resolve the chicken-and-egg problem, we propose to *jointly* learn unbiased representations for both the identity and demographic attributes. Specifically, starting from a multi-task learning framework that learns disentangled feature representations of gender, age, race, and identity, respectively, we request the classifier of each task to act as adversarial supervision for the other tasks (*e.g.*, the dash arrows in Fig. 1c). These four classifiers help each other to achieve better feature disentanglement, resulting in unbiased feature representations for



(a) Multi-task learning (b) Adversarial learning (c) DebFace

Fig. 1: Methods to learn different tasks simultaneously. Solid lines are typical feature flow in CNN, while dash lines are adversarial losses.

both the identity and demographic attributes. As shown in Fig. 1, our framework is in sharp contrast to either multi-task learning or adversarial learning.

Moreover, since the features are disentangled into the demographic and identity, our face representations also contribute to privacy-preserving applications. It is worth noticing that such identity representations contain little demographic information, which could undermine the recognition competence since demographic features are *part* of identity-related facial appearance. To retain the recognition accuracy on demographic biased face datasets, we propose another network that combines the demographic features with the demographic-free identity features to generate a new identity representation for face recognition.

The key contributions and findings of the paper are:

◇ A thorough analysis of deep learning based face recognition performance on three different demographics: (i) gender, (ii) age, and (iii) race.

◇ A de-biasing face recognition framework, called DebFace, that generates disentangled representations for both identity and demographics recognition while jointly removing discriminative information from other counterparts.

◇ The identity representation from DebFace (DebFace-ID) shows lower bias on different demographic cohorts and also achieves SOTA face verification results on demographic-unbiased face recognition.

◇ The demographic attribute estimations via DebFace are less biased across other demographic cohorts.

◇ Combining ID with demographics results in more discriminative features for face recognition on biased datasets.

2 Related Work

Face Recognition on Imbalanced Training Data Previous efforts on face recognition aim to tackle class imbalance problem on training data. For example, in prior-DNN era, Zhang *et al.* [66] propose a cost-sensitive learning framework to reduce misclassification rate of face identification. To correct the skew of separating hyperplanes of SVM on imbalanced data, Liu *et al.* [33] propose Margin-Based Adaptive Fuzzy SVM that obtains a lower generalization error bound. In the DNN era, face recognition models are trained on large-scale face datasets with highly-imbalanced class distribution [63,65]. Range Loss [65] learns

a robust face representation that makes the most use of every training sample. To mitigate the impact of insufficient class samples, center-based feature transfer learning [63] and large margin feature augmentation [58] are proposed to augment features of minority identities and equalize class distribution. Despite their effectiveness, these studies ignore the influence of demographic imbalance on the dataset, which may lead to demographic bias. For instance, The FRVT 2019 [42] shows the demographic bias of over 100 face recognition algorithms. To uncover deep learning bias, Alexander *et al.* [4] develop an algorithm to mitigate the hidden biases within training data. Wang *et al.* [57] propose a domain adaptation network to reduce racial bias in face recognition. They recently extended their work using reinforcement learning to find optimal margins of additive angular margin based loss functions for different races [56]. To our knowledge, no studies have tackled the challenge of de-biasing demographic bias in DNN-based face recognition and demographic attribute estimation algorithms.

Adversarial Learning and Disentangled Representation Adversarial learning [44] has been well explored in many computer vision applications. For example, Generative Adversarial Networks (GANs) [18] employ adversarial learning to train a generator by competing with a discriminator that distinguishes real images from synthetic ones. Adversarial learning has also been applied to domain adaptation [36, 48, 52, 53]. A problem of current interest is to learn interpretable representations with semantic meaning [60]. Many studies have been learning factors of variations in the data by supervised learning [31,32,34,50,51], or semi-supervised/unsupervised learning [28,35,40,68], referred to as disentangled representation. For supervised disentangled feature learning, adversarial networks are utilized to extract features that only contain discriminative information of a target task. For face recognition, Liu *et al.* [34] propose a disentangled representation by training an adversarial autoencoder to extract features that can capture identity discrimination and its complementary knowledge. In contrast, our proposed DebFace differs from prior works in that each branch of a multi-task network acts as both a generator and discriminators of other branches (Fig. 1c).

3 Methodology

3.1 Problem Definition

The ultimate goal of unbiased face recognition is that, given a face recognition system, no statistically significant difference among the performance in different categories of face images. Despite the research on pose-invariant face recognition that aims for equal performance on all poses [51, 62], we believe that it is inappropriate to define variations like pose, illumination, or resolution, as the categories. These are instantaneous *image-related* variations with intrinsic bias. E.g., large-pose or low-resolution faces are inherently harder to be recognized than frontal-view high-resolution faces.

Rather, we would like to define *subject-related* properties such as demographic attributes as the categories. *A face recognition system is* ***biased*** *if it performs*

worse on certain demographic cohorts. For practical applications, it is important to consider what demographic biases may exist, and whether these are intrinsic biases across demographic cohorts or algorithmic biases derived from the algorithm itself. This motivates us to analyze the demographic influence on face recognition performance and strive to reduce algorithmic bias for face recognition systems. One may achieve this by training on a dataset containing uniform samples over the cohort space. However, the demographic distribution of a dataset is often imbalanced and underrepresents demographic minorities while overrepresenting majorities. Naively re-sampling a balanced training dataset may still induce bias since the diversity of latent variables is different across cohorts and the instances cannot be treated fairly during training. To mitigate demographic bias, we propose a face de-biasing framework that jointly reduces mutual bias over all demographics and identities while disentangling face representations into gender, age, race, and demographic-free identity in the mean time.

3.2 Algorithm Design

The proposed network takes advantage of the relationship between demographics and face identities. On one hand, demographic characteristics are highly correlated to face features. On the other hand, demographic attributes are heterogeneous in terms of data type and semantics [20]. A male person, for example, is not necessarily of a certain age or of a certain race. Accordingly, we present a framework that jointly generates demographic features and identity features from a single face image by considering both the aforementioned attribute correlation and attribute heterogeneity in a DNN.

While our main goal is to mitigate demographic bias from face representation, we observe that demographic estimations are biased as well (see Fig. 5). How can we remove the bias of face recognition when demographic estimations themselves are biased? Cook *et al.* [11] investigated this effect and found the performance of face recognition is affected by multiple demographic covariates. We propose a de-biasing network, DebFace, that disentangles the representation into gender (DebFace-G), age (DebFace-A), race (DebFace-R), and identity (DebFace-ID), to decrease bias of both face recognition and demographic estimations. Using adversarial learning, the proposed method is capable of jointly learning multiple discriminative representations while ensuring that each classifier cannot distinguish among classes through non-corresponding representations.

Though less biased, DebFace-ID loses demographic cues that are useful for identification. In particular, race and gender are two critical components that constitute face patterns. Hence, we desire to incorporate race and gender with DebFace-ID to obtain a more integrated face representation. We employ a lightweight fully-connected network to aggregate the representations into a face representation (DemoID) with the same dimensionality as DebFace-ID.

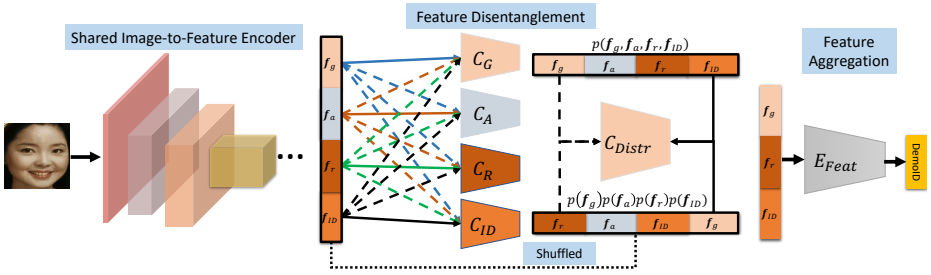


Fig. 2: Overview of the proposed De-biasing face (DebFace) network. DebFace is composed of three major blocks, *i.e.*, a shared feature encoding block, a feature disentangling block, and a feature aggregation block. The solid arrows represent the forward inference, and the dashed arrows stand for adversarial training. During inference, either DebFace-ID (*i.e.*, $\mathbf{f}_{ID}$) or DemoID can be used for face matching given the desired trade-off between biasness and accuracy.

3.3 Network Architecture

Figure 2 gives an overview of the proposed DebFace network. It consists of four components: the shared image-to-feature encoder E_{Img}, the four attribute classifiers (including gender C_G, age C_A, race C_R, and identity C_{ID}), the distribution classifier C_{Distr}, and the feature aggregation network E_{Feat}. We assume access to N labeled training samples $\{(\mathbf{x}^{(i)}, y_g^{(i)}, y_a^{(i)}, y_r^{(i)}, y_{id}^{(i)})\}_{i=1}^{N}$. Our approach takes an image $\mathbf{x}^{(i)}$ as the input of E_{Img}. The encoder projects $\mathbf{x}^{(i)}$ to its feature representation $E_{Img}(\mathbf{x}^{(i)})$. The feature representation is then decoupled into four D-dimensional feature vectors, gender $\mathbf{f}_g^{(i)}$, age $\mathbf{f}_a^{(i)}$, race $\mathbf{f}_r^{(i)}$, and identity $\mathbf{f}_{ID}^{(i)}$, respectively. Next, each attribute classifier operates the corresponding feature vector to correctly classify the target attribute by optimizing parameters of both E_{Img} and the respective classifier C_*. For a demographic attribute with K categories, the learning objective $\mathcal{L}_{C_{Demo}}(\mathbf{x}, y_{Demo}; E_{Img}, C_{Demo})$ is the standard cross entropy loss function. For the $n-$identity classification, we adopt AM-Softmax [54] as the objective function $\mathcal{L}_{C_{ID}}(\mathbf{x}, y_{id}; E_{Img}, C_{ID})$. To de-bias all of the feature representations, adversarial loss $\mathcal{L}_{Adv}(\mathbf{x}, y_{Demo}, y_{id}; E_{Img}, C_{Demo}, C_{ID})$ is applied to the above four classifiers such that each of them will not be able to predict correct labels when operating irrelevant feature vectors. Specifically, given a classifier, the remaining three attribute feature vectors are imposed on it and attempt to mislead the classifier by only optimizing the parameters of E_{Img}. To further improve the disentanglement, we also reduce the mutual information among the attribute features by introducing a distribution classifier C_{Distr}. C_{Distr} is trained to identify whether an input representation is sampled from the joint distribution $p(\mathbf{f}_g, \mathbf{f}_a, \mathbf{f}_r, \mathbf{f}_{ID})$ or the multiplication of margin distributions $p(\mathbf{f}_g)p(\mathbf{f}_a)p(\mathbf{f}_r)p(\mathbf{f}_{ID})$ via a binary cross entropy loss $\mathcal{L}_{C_{Distr}}(\mathbf{x}, y_{Distr}; E_{Img}, C_{Distr})$, where y_{Distr} is the distribution label. Similar to

adversarial loss, a factorization objective function $\mathcal{L}_{Fact}(\mathbf{x}, y_{Distr}; E_{Img}, C_{Distr})$ is utilized to restrain the C_{Distr} from distinguishing the real distribution and thus minimizes the mutual information of the four attribute representations. Both adversarial loss and factorization loss are detailed in Sec. 3.4. Altogether, DebFace endeavors to minimize the joint loss:

$$\begin{aligned}\mathcal{L}(\mathbf{x}, y_{Demo}, y_{id}, y_{Distr}; E_{Img}, C_{Demo}, C_{ID}, C_{Distr}) = \\ \mathcal{L}_{C_{Demo}}(\mathbf{x}, y_{Demo}; E_{Img}, C_{Demo}) \\ + \mathcal{L}_{C_{ID}}(\mathbf{x}, y_{id}; E_{Img}, C_{ID}) \\ + \mathcal{L}_{C_{Distr}}(\mathbf{x}, y_{Distr}; E_{Img}, C_{Distr}) \\ + \lambda \mathcal{L}_{Adv}(\mathbf{x}, y_{Demo}, y_{id}; E_{Img}, C_{Demo}, C_{ID}) \\ + \nu \mathcal{L}_{Fact}(\mathbf{x}, y_{Distr}; E_{Img}, C_{Distr}),\end{aligned} \tag{1}$$

where λ and ν are hyper-parameters determining how much the representation is decomposed and decorrelated in each training iteration.

The discriminative demographic features in DebFace-ID are weakened by removing demographic information. Fortunately, our de-biasing network preserves all pertinent demographic features in a disentangled way. Basically, we train another multilayer perceptron (MLP) E_{Feat} to aggregate DebFace-ID and the demographic embeddings into a unified face representation DemoID. Since age generally does not pertain to a person's identity, we only consider gender and race as the identity-informative attributes. The aggregated embedding, $\mathbf{f}_{DemoID} = E_{feat}(\mathbf{f}_{ID}, \mathbf{f}_g, \mathbf{f}_r)$, is supervised by an identity-based triplet loss:

$$\mathcal{L}_{E_{Feat}} = \frac{1}{M}\sum_{i=1}^{M}[\|\mathbf{f}^{(i)}_{DemoID^a} - \mathbf{f}^{(i)}_{DemoID^p}\|_2^2 - \|\mathbf{f}^{(i)}_{DemoID^a} - \mathbf{f}^{(i)}_{DemoID^n}\|_2^2 + \alpha]_+, \tag{2}$$

where $\{\mathbf{f}^{(i)}_{DemoID^a}, \mathbf{f}^{(i)}_{DemoID^p}, \mathbf{f}^{(i)}_{DemoID^n}\}$ is the i^{th} triplet consisting of an anchor, a positive, and a negative DemoID representation, M is the number of hard triplets in a mini-batch. $[x]_+ = \max(0, x)$, and α is the margin.

3.4 Adversarial Training and Disentanglement

As discussed in Sec. 3.3, the adversarial loss aims to minimize the task-independent information semantically, while the factorization loss strives to dwindle the interfering information statistically. We employ both losses to disentangle the representation extracted by E_{Img}. We introduce the adversarial loss as a means to learn a representation that is invariant in terms of certain attributes, where a classifier trained on it cannot correctly classify those attributes using that representation. We take one of the attributes, *e.g.*, gender, as an example to illustrate the adversarial objective. First of all, for a demographic representation $\mathbf{f}_{Demo}$, we learn a gender classifier on $\mathbf{f}_{Demo}$ by optimizing the classification loss $\mathcal{L}_{C_G}(\mathbf{x}, y_{Demo}; E_{Img}, C_G)$. Secondly, for the same gender classifier, we intend to maximize the chaos of the predicted distribution [26]. It is well known that a uniform distribution has the highest entropy and presents the most randomness.

Hence, we train the classifier to predict the probability distribution as close as possible to a uniform distribution over the category space by minimizing the cross entropy:

$$\mathcal{L}^{G}_{Adv}(\mathbf{x}, y_{Demo}, y_{id}; E_{Img}, C_G) = -\sum_{k=1}^{K_G} \frac{1}{K_G} \cdot (\log \frac{e^{C_G(\mathbf{f}_{Demo})_k}}{\sum_{j=1}^{K_G} e^{C_G(\mathbf{f}_{Demo})_j}} + \log \frac{e^{C_G(\mathbf{f}_{ID})_k}}{\sum_{j=1}^{K_G} e^{C_G(\mathbf{f}_{ID})_j}}), \quad (3)$$

where K_G is the number of categories in gender [2], and the ground-truth label is no longer an one-hot vector, but a K_G-dimensional vector with all elements being $\frac{1}{K_G}$. The above loss function corresponds to the dash lines in Fig. 2. It strives for gender-invariance by finding a representation that makes the gender classifier C_G perform poorly. We minimize the adversarial loss by only updating parameters in E_{Img}.

We further decorrelate the representations by reducing the mutual information across attributes. By definition, the mutual information is the relative entropy (KL divergence) between the joint distribution and the product distribution. To increase uncorrelation, we add a distribution classifier C_{Distr} that is trained to simply perform a binary classification using $\mathcal{L}_{C_{Distr}}(\mathbf{x}, y_{Distr}; E_{Img}, C_{Distr})$ on samples $\mathbf{f}_{Distr}$ from both the joint distribution and dot product distribution. Similar to adversarial learning, we factorize the representations by tricking the classifier via the same samples so that the predictions are close to random guesses,

$$\mathcal{L}_{Fact}(\mathbf{x}, y_{Distr}; E_{Img}, C_{Distr}) = -\sum_{i=1}^{2} \frac{1}{2} \log \frac{e^{C_{Distr}(\mathbf{f}_{Distr})_i}}{\sum_{j=1}^{2} e^{C_{Distr}(\mathbf{f}_{Distr})_j}}. \quad (4)$$

In each mini-batch, we consider $E_{Img}(\mathbf{x})$ as samples of the joint distribution $p(\mathbf{f}_g, \mathbf{f}_a, \mathbf{f}_r, \mathbf{f}_{ID})$. We randomly shuffle feature vectors of each attribute, and reconcatenate them into $4D$-dimension, which are approximated as samples of the product distribution $p(\mathbf{f}_g)p(\mathbf{f}_a)p(\mathbf{f}_r)p(\mathbf{f}_{ID})$. During factorization, we only update E_{Img} to minimize mutual information between decomposed features.

4 Experiments

4.1 Datasets and Pre-processing

We utilize 15 total face datasets in this work, for learning the demographic estimation models, the baseline face recognition model, DebFace model as well as their evaluation. To be specific, CACD [9], IMDB [43], UTKFace [67], AgeDB [38], AFAD [41], AAF [10], FG-NET [1], RFW [57], IMFDB-CVIT [45], Asian-DeepGlint [2], and PCSO [13] are the datasets for training and testing demographic estimation models; and MS-Celeb-1M [19], LFW [25], IJB-A [30], and IJB-C [37] are for learning and evaluating face verification models. All faces are detected by MTCNN [64]. Each face image is cropped and resized to 112×112 pixels using a similarity transformation based on the detected landmarks.

[2] In our case, $K_G = 2$, *i.e.*, male and female.

4.2 Implementation Details

DebFace is trained on a cleaned version of MS-Celeb-1M [14], using the ArcFace architecture [14] with 50 layers for the encoder E_{Img}. Since there are no demographic labels in MS-Celeb-1M, we first train three demographic attribute estimation models for gender, age, and race, respectively. For age estimation, the model is trained on the combination of CACD, IMDB, UTKFace, AgeDB, AFAD, and AAF datasets. The gender estimation model is trained on the same datasets except CACD which contains no gender labels. We combine AFAD, RFW, IMFDB-CVIT, and PCSO for race estimation training. All three models use ResNet [22] with 34 layers for age, 18 layers for gender and race.

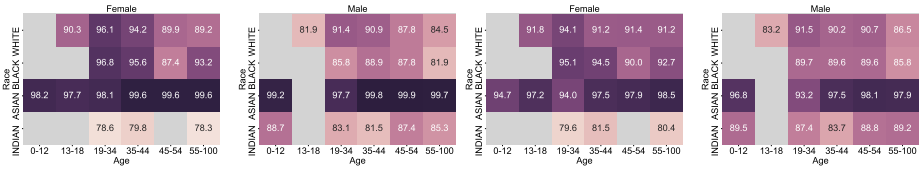
We predict the demographic labels of MS-Celeb-1M with the well-trained demographic models. Our DebFace is then trained on the re-labeled MS-Celeb-1M using SGD with a momentum of 0.9, a weight decay of 0.01, and a batch size of 256. The learning rate starts from 0.1 and drops to 0.0001 following the schedule at 8, 13, and 15 epochs. The dimensionality of the embedding layer of E_{Img} is 4×512, *i.e.*, each attribute representation (gender, age, race, ID) is a 512-*dim* vector. We keep the hyper-parameter setting of AM-Softmax as [14]: $s = 64$ and $m = 0.5$. The feature aggregation network E_{Feat} comprises of two linear residual units with P-ReLU and BatchNorm in between. E_{Feat} is trained on MS-Celeb-1M by SGD with a learning rate of 0.01. The triplet loss margin α is 1.0. The disentangled features of gender, race, and identity are concatenated into a 3×512-*dim* vector, which inputs to E_{Feat}. The network is then trained to output a 512-*dim* representation for face recognition on biased datasets. Our source code is available at `https://github.com/gongsixue/DebFace.git`.

4.3 De-biasing Face Verification

Baseline: We compare DebFace-ID with a regular face representation model which has the same architecture as the shared feature encoder of DebFace. Referred to as BaseFace, this baseline model is also trained on MS-Celeb-1M, with the representation dimension of 512.

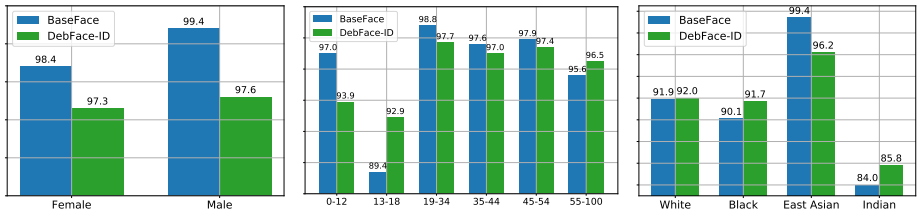
To show the efficacy of DebFace-ID on bias mitigation, we evaluate the verification performance of DebFace-ID and BaseFace on faces from each demographic cohort. There are 48 total cohorts given the combination of demographic attributes including 2 gender (male, female), 4 race [3] (Black, White, East Asian, Indian), and 6 age group ($0-12$, $13-18$, $19-34$, $35-44$, $45-54$, $55-100$). We combine CACD, AgeDB, CVIT, and a subset of Asian-DeepGlint as the testing set. Overlapping identities among these datasets are removed. IMDB is excluded from the testing set due to its massive number of wrong ID labels. For the dataset without certain demographic labels, we simply use the corresponding models to predict the labels. We report the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC). We define the degree of bias, termed *biasness*, as the standard deviation of performance across cohorts.

[3] To clarify, we consider two race groups, Black and White; and two ethnicity groups, East Asian and Indian. The word race denotes both race and ethnicity in this paper.



(a) BaseFace (b) DebFace-ID

Fig. 3: Face Verification AUC (%) on each demographic cohort. The cohorts are chosen based on the three attributes, *i.e.*, gender, age, and race. To fit the results into a 2D plot, we show the performance of male and female separately. Due to the limited number of face images in some cohorts, their results are gray cells.



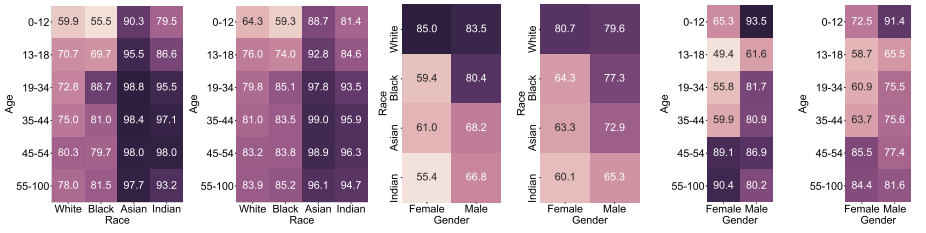
(a) Gender (b) Age (c) Race

Fig. 4: The overall performance of face verification AUC (%) on gender, age, and race.

Figure 3 shows the face verification results of BaseFace and DebFace-ID on each cohort. That is, for a particular face representation (*e.g.*, DebFace-ID), we report its AUC on each cohort by putting the number in the corresponding cell. From these heatmaps, we observe that both DebFace-ID and BaseFace present bias in face verification, where the performance on some cohorts are significantly worse, especially the cohorts of Indian female and elderly people. Compared to BaseFace, DebFace-ID suggests less bias and the difference of AUC is smaller, where the heatmap exhibits smoother edges. Figure 4 shows the performance of face verification on 12 demographic cohorts. Both DebFace-ID and BaseFace present similar relative accuracies across cohorts. For example, both algorithms perform worse on the younger age cohorts than on adults; and the performance on the Indian is significantly lower than on the other races. DebFace-ID decreases the bias by gaining discriminative face features for cohorts with less images in spite of the reduction in the performance on cohorts with more samples.

4.4 De-biasing Demographic Attribute Estimation

Baseline: We further explore the bias of demographic attribute estimation and compare demographic attribute classifiers of DebFace with baseline estimation models. We train three demographic estimation models, namely, gender estimation (BaseGender), age estimation (BaseAge), and race estimation (BaseRace),



(a) BaseGender (b) DebFace-G (c) BaseAge (d) DebFace-A (e) BaseRace (f) DebFace-R

Fig. 5: Classification accuracy (%) of demographic attribute estimations on faces of different cohorts, by DebFace and the baselines. For simplicity, we use DebFace-G, DebFace-A, and DebFace-R to represent the gender, age, and race classifier of DebFace.

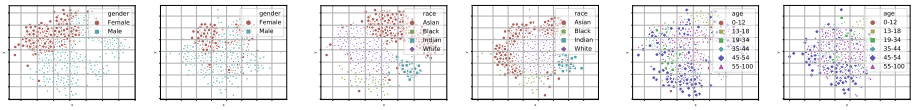
Table 1: Biasness of Face Recognition and Demographic Attribute Estimation.

Method	Face Verification				Demographic Estimation		
	All	Gender	Age	Race	Gender	Age	Race
Baseline	6.83	0.50	3.13	5.49	12.38	10.83	14.58
DebFace	**5.07**	**0.15**	**1.83**	**3.70**	**10.22**	**7.61**	**10.00**

on the same training set as DebFace. For fairness, all three models have the same architecture as the shared layers of DebFace.

We combine the four datasets mentioned in Sec. 4.3 with IMDB as the global testing set. As all demographic estimations are treated as classification problems, the classification accuracy is used as the performance metric. As shown in Fig. 5, all demographic attribute estimations present significant bias. For gender estimation, both algorithms perform worse on the White and Black cohorts than on East Asian and Indian. In addition, the performance on young children is significantly worse than on adults. In general, the race estimation models perform better on the male cohort than on female. Compared to gender, race estimation shows higher bias in terms of age. Both baseline methods and DebFace perform worse on cohorts in age between 13 to 44 than in other age groups.

Similar to race, age estimation still achieves better performance on male than on female. Moreover, the white cohort shows dominant advantages over other races in age estimation. In spite of the existing bias in demographic attribute estimations, the proposed DebFace is still able to mitigate bias derived from algorithms. Compared to Fig. 5a, 5e, 5c, cells in Fig. 5b, 5f, 5d present more uniform colors. We summarize the biasness of DebFace and baseline models for both face recognition and demographic attribute estimations in Tab. 1. In general, we observe DebFace substantially reduces biasness for both tasks. For the task with larger biasness, the reduction of biasness is larger.



(a) BaseFace (b) DebFace-ID (c) BaseFace (d) DebFace-ID (e) BaseFace (f) DebFace-ID

Fig. 6: The distribution of face identity representations of BaseFace and DebFace. Both collections of feature vectors are extracted from images of the same dataset. Different colors and shapes represent different demographic attributes. Zoom in for details.

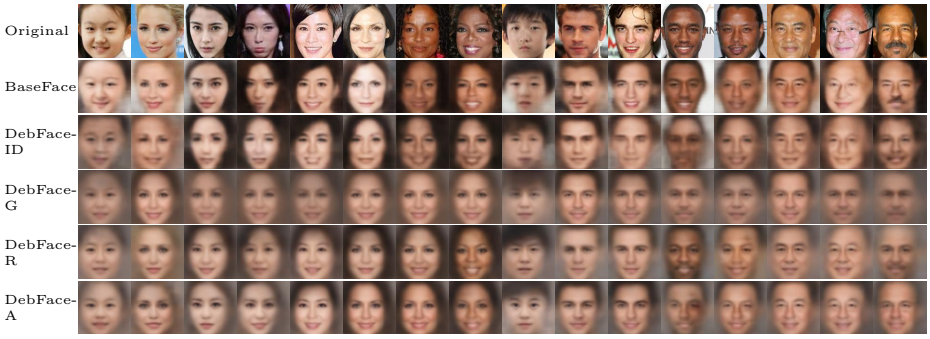


Fig. 7: Reconstructed Images using Face and Demographic Representations. The first row is the original face images. From the second row to the bottom, the face images are reconstructed from 2) BaseFace; 3) DebFace-ID; 4) DebFace-G; 5) DebFace-R; 6) DebFace-A. Zoom in for details.

4.5 Analysis of Disentanglement

We notice that DebFace still suffers unequal performance in different demographic groups. It is because there are other latent variables besides the demographics, such as image quality or capture conditions that could lead to biased performance. Such variables are difficult to control in pre-collected large face datasets. In the framework of DebFace, it is also related to the degree of feature disentanglement. A fully disentangling is supposed to completely remove the factors of bias from demographic information. To illustrate the feature disentanglement of DebFace, we show the demographic discriminative ability of face representations by using these features to estimate gender, age, and race. Specifically, we first extract identity features of images from the testing set in Sec. 4.1 and split them into training and testing sets. Given demographic labels, the face features are fed into a two-layer fully-connected network, learning to classify one of the demographic attributes. Tab. 2 reports the demographic classification accuracy on the testing set. For all three demographic estimations, DebFace-ID presents much lower accuracies than BaseFace, indicating the decline of demographic information in DebFace-ID. We also plot the distribution of identity representations in the feature space of BaseFace and DebFace-ID. From the testing set in Sec. 4.3, we randomly select 50 subjects in each demo-

Table 2: Demographic Classification Accuracy (%) by face features.

Method	Gender	Race	Age
BaseFace	95.27	89.82	78.14
DebFace-ID	73.36	61.79	49.91

Table 3: Face Verification Accuracy (%) on RFW dataset.

Method	White	Black	Asian	Indian	Biasness
[56]	96.27	94.68	94.82	95.00	0.93
DebFace-ID	95.95	93.67	94.33	94.78	0.83

Table 4: Verification Performance on LFW, IJB-A, and IJB-C.

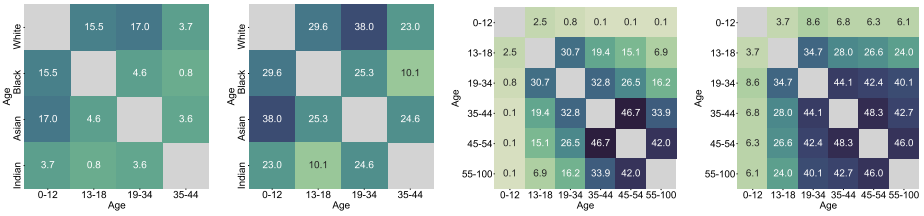
Method	LFW (%)	Method	IJB-A (%)	IJB-C @ FAR (%)		
			0.1% FAR	0.001%	0.01%	0.1%
DeepFace+ [47]	97.35	Yin *et al.* [61]	73.9 ± 4.2	-	-	69.3
CosFace [55]	99.73	Cao *et al.* [7]	90.4 ± 1.4	74.7	84.0	91.0
ArcFace [14]	99.83	Multicolumn [59]	92.0 ± 1.3	77.1	86.2	92.7
PFE [46]	99.82	PFE [46]	95.3 ± 0.9	89.6	93.3	95.5
BaseFace	99.38	*BaseFace*	90.2 ± 1.1	80.2	88.0	92.9
DebFace-ID	98.97	*DebFace-ID*	87.6 ± 0.9	82.0	88.1	89.5
DemoID	99.50	*DemoID*	92.2 ± 0.8	83.2	89.4	92.9

graphic group and one image of each subject. BaseFace and DebFace-ID are extracted from the selected image set and are then projected from 512-*dim* to 2-*dim* by T-SNE. Fig. 6 shows their T-SNE feature distributions. We observe that BaseFace presents clear demographic clusters, while the demographic clusters of DebFace-ID, as a result of disentanglement, mostly overlap with each other.

To visualize the disentangled feature representations of DebFace, we train a decoder that reconstructs face images from the representations. Four face decoders are trained separately for each disentangled component, *i.e.*, gender, age, race, and ID. In addition, we train another decoder to reconstruct faces from BaseFace for comparison. As shown in Fig. 7, both BaseFace and DebFace-ID maintain the identify features of the original faces, while DebFace-ID presents less demographic characteristics. No race or age, but gender features can be observed on faces reconstructed from DebFace-G. Meanwhile, we can still recognize race and age attributes on faces generated from DebFace-R and DebFace-A.

4.6 Face Verification on Public Testing Datasets

We report the performance of three different settings, using 1) BaseFace, the same baseline in Sec. 4.3, 2) DebFace-ID, and 3) the fused representation DemoID. Table 4 reports face verification results on on three public benchmarks: LFW, IJB-A, and IJB-C. On LFW, DemoID outperforms BaseFace while maintaining similar accuracy compared to SOTA algorithms. On IJB-A/C, DemoID outperforms all prior works except PFE [46]. Although DebFace-ID shows lower discrimination, TAR at lower FAR on IJB-C is higher than that of BaseFace. To evaluate DebFace on a racially balanced testing dataset RFW [57] and compare with the work [56], we train a DebFace model on BUPT-Balancedface [56] dataset. The new model is trained to reduce racial bias by disentangling ID and



(a) BaseFace: Race (b) DebFace-ID: Race (c) BaseFace: Age (d) DebFace-ID: Age

Fig. 8: The percentage of false accepted cross race or age pairs at 1% FAR.

race. Tab. 3 reports the verification results on RFW. While DebFace-ID gives a slightly lower face verification accuracy, it improves the biasness over [56].

We observe that DebFace-ID is less discriminative than BaseFace, or DemoID, since demographics are essential components of face features. To understand the deterioration of DebFace, we analyse the effect of demographic heterogeneity on face verification by showing the tendency for one demographic group to experience a false accept error relative to another group. For any two demographic cohorts, we check the number of falsely accepted pairs that are from different groups at 1% FAR. Fig. 8 shows the percentage of such falsely accepted demographic-heterogeneous pairs. Compared to BaseFace, DebFace exhibits more cross-demographic pairs that are falsely accepted, resulting in the performance decline on demographically biased datasets. Due to the demographic information reduction, DebFace-ID is more susceptible to errors between demographic groups. In the sense of de-biasing, it is preferable to decouple demographic information from identity features. However, if we prefer to maintain the overall performance across all demographics, we can still aggregate all the relevant information. It is an application-dependent trade-off between accuracy and de-biasing. DebFace balances the accuracy vs. bias trade-off by generating both debiased identity and debiased demographic representations, which may be aggregated into DemoID if bias is less of a concern.

5 Conclusion

We present a de-biasing face recognition network (DebFace) to mitigate demographic bias in face recognition. DebFace adversarially learns the disentangled representation for gender, race, and age estimation, and face recognition simultaneously. We empirically demonstrate that DebFace can not only reduce bias in face recognition but in demographic attribute estimation as well. Our future work will explore an aggregation scheme to combine race, gender, and identity without introducing algorithmic and dataset bias.

Acknowledgement: This work is supported by U.S. Department of Commerce (#60NANB19D154), National Institute of Standards and Technology. The authors thank reviewers, area chairs, Dr. John J. Howard, and Dr. Yevgeniy Sirotin for offering constructive comments.

References

1. https://yanweifu.github.io/FG_NET_data
2. http://trillionpairs.deepglint.com/overview
3. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. pp. 308–318 (2016)
4. Amini, A., Soleimany, A., Schwarting, W., Bhatia, S., Rus, D.: Uncovering and mitigating algorithmic bias through learned latent structure. AAAI/ACM Conference on AI, Ethics, and Society (2019)
5. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Advances in neural information processing systems. pp. 4349–4357 (2016)
6. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. arXiv preprint arXiv:1906.07413 (2019)
7. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: IEEE International Conference on Automatic Face & Gesture Recognition. IEEE (2018)
8. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of Artificial Intelligence research **16**, 321–357 (2002)
9. Chen, B.C., Chen, C.S., Hsu, W.H.: Cross-age reference coding for age-invariant face recognition and retrieval. In: ECCV (2014)
10. Cheng, J., Li, Y., Wang, J., Yu, L., Wang, S.: Exploiting effective facial patches for robust gender recognition. Tsinghua Science and Technology **24**(3), 333–345 (2019)
11. Cook, C.M., Howard, J.J., Sirotin, Y.B., Tipton, J.L., Vemury, A.R.: Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. IEEE Transactions on Biometrics, Behavior, and Identity Science (2019)
12. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: CVPR (2019)
13. Deb, D., Best-Rowden, L., Jain, A.K.: Face recognition performance under aging. In: CVPRW (2017)
14. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)
15. Dietterich, T.G., Kong, E.B.: Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Tech. rep. (1995)
16. Drummond, C., Holte, R.C., et al.: C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In: Workshop on Learning from Imbalanced Datasets II. Citeseer (2003)
17. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: the 22nd ACM SIGSAC (2015)
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
19. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: ECCV. Springer (2016)
20. Han, H., A, K.J., Shan, S., Chen, X.: Heterogeneous face attribute estimation: A deep multi-task learning approach. IEEE Trans. Pattern Analysis Machine Intelligence **PP**(99), 1–1 (2017)

21. Hayat, M., Khan, S., Zamir, W., Shen, J., Shao, L.: Max-margin class imbalanced learning with gaussian affinity. arXiv preprint arXiv:1901.07711 (2019)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
23. Howard, J., Sirotin, Y., Vemury, A.: The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In: IEEE BTAS (2019)
24. Huang, C., Li, Y., Chen, C.L., Tang, X.: Deep imbalanced learning for face recognition and attribute prediction. IEEE Trans. Pattern Analysis and Machine Intelligence (2019)
25. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments (2008)
26. Jourabloo, A., Yin, X., Liu, X.: Attribute preserved face de-identification. In: ICB (2015)
27. Khan, S., Hayat, M., Zamir, S.W., Shen, J., Shao, L.: Striking the right balance with uncertainty. In: CVPR (2019)
28. Kim, H., Mnih, A.: Disentangling by factorising. arXiv preprint arXiv:1802.05983 (2018)
29. Klare, B.F., Burge, M.J., Klontz, J.C., Bruegge, R.W.V., Jain, A.K.: Face recognition performance: Role of demographic information. IEEE Trans. Information Forensics and Security **7**(6), 1789–1801 (2012)
30. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: CVPR (2015)
31. Liu, F., Zeng, D., Zhao, Q., Liu, X.: Disentangling features in 3D face shapes for joint face reconstruction and recognition. In: CVPR (2018)
32. Liu, Y., Wang, Z., Jin, H., Wassell, I.: Multi-task adversarial network for disentangled feature learning. In: CVPR (2018)
33. Liu, Y.H., Chen, Y.T.: Face recognition using total margin-based adaptive fuzzy support vector machines. IEEE Transactions on Neural Networks **18**(1), 178–192 (2007)
34. Liu, Y., Wei, F., Shao, J., Sheng, L., Yan, J., Wang, X.: Exploring disentangled feature representation beyond face identification. In: CVPR (2018)
35. Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. arXiv preprint arXiv:1811.12359 (2018)
36. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: NIPS (2018)
37. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 ICB (2018)
38. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: CVPRW (2017)
39. Mullick, S.S., Datta, S., Das, S.: Generative adversarial minority oversampling. arXiv preprint arXiv:1903.09730 (2019)
40. Narayanaswamy, S., Paige, T.B., Van de Meent, J.W., Desmaison, A., Goodman, N., Kohli, P., Wood, F., Torr, P.: Learning disentangled representations with semi-supervised deep generative models. In: NIPS (2017)
41. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output cnn for age estimation. In: CVPR (2016)

42. Patrick Grother, M.N., Hanaoka, K.: Face recognition vendor test (frvt) part 3: Demographic effects. In: Technical Report, National Institute of Standards and Technology (2019)
43. Rothe, R., Timofte, R., Van Gool, L.: Deep expectation of real and apparent age from a single image without facial landmarks. IJCV (2018)
44. Schmidhuber, J.: Learning factorial codes by predictability minimization. Neural Computation **4**(6), 863–879 (1992)
45. Shankar Setty, Moula Husain, P.B.J.G.M.K.R.V.V.H.J.C.K.R.R.R.V.K., Jawahar, C.V.: Indian Movie Face Database: A Benchmark for Face Recognition Under Wide Variations. In: NCVPRIPG (2013)
46. Shi, Y., Jain, A.K., Kalka, N.D.: Probabilistic face embeddings. arXiv preprint arXiv:1904.09658 (2019)
47. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: CVPR (2014)
48. Tao, C., Lv, F., Duan, L., Wu, M.: Minimax entropy network: Learning category-invariant features for domain adaptation. arXiv preprint arXiv:1904.09601 (2019)
49. Torralba, A., Efros, A.A., et al.: Unbiased look at dataset bias. In: CVPR (2011)
50. Tran, L., Yin, X., Liu, X.: Disentangled representation learning GAN for pose-invariant face recognition. In: CVPR (2017)
51. Tran, L., Yin, X., Liu, X.: Representation learning by rotating your faces. IEEE Trans. on Pattern Analysis and Machine Intelligence **41**(12), 3007–3021 (2019)
52. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: CVPR (2015)
53. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR (2017)
54. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Processing Letters **25**(7), 926–930 (2018)
55. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: CVPR (2018)
56. Wang, M., Deng, W.: Mitigating bias in face recognition using skewness-aware reinforcement learning. In: CVPR (2020)
57. Wang, M., Deng, W., Hu, J., Tao, X., Huang, Y.: Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In: ICCV (2019)
58. Wang, P., Su, F., Zhao, Z., Guo, Y., Zhao, Y., Zhuang, B.: Deep class-skewed learning for face recognition. Neurocomputing (2019)
59. Xie, W., Zisserman, A.: Multicolumn networks for face recognition. arXiv preprint arXiv:1807.09192 (2018)
60. Yin, B., Tran, L., Li, H., Shen, X., Liu, X.: Towards interpretable face recognition. In: ICCV (2019)
61. Yin, X., Liu, X.: Multi-task convolutional neural network for pose-invariant face recognition. IEEE Trans. Image Processing **27**(2), 964–975 (2017)
62. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Towards large-pose face frontalization in the wild. In: ICCV (2017)
63. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Feature transfer learning for face recognition with under-represented data. In: CVPR (2019)
64. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
65. Zhang, X., Fang, Z., Wen, Y., Li, Z., Qiao, Y.: Range loss for deep face recognition with long-tailed training data. In: CVPR (2017)

66. Zhang, Y., Zhou, Z.H.: Cost-sensitive face recognition. IEEE Trans. Pattern Analysis and Machine Intelligence **32**(10), 1758–1769 (2009)
67. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: CVPR. IEEE (2017)
68. Zhang, Z., Tran, L., Yin, X., Atoum, Y., Wan, J., Wang, N., Liu, X.: Gait recognition via disentangled representation learning. In: CVPR (2019)