DebFace: De-biasing Face Recognition

Sixue Gong Xiaoming Liu Anil K. Jain
Michigan State University, East Lansing MI 48824
{gongsixu, liuxm, jain}@msu.edu

(a) *BaseFace:* 0.45 *DebFace:* 0.32 (b) *BaseFace:* 0.49 *DebFace:* 0.31 (c) *BaseFace:* 0.33 *DebFace:* 0.25

(d) *BaseFace:* 0.17 *DebFace:* 0.36 (e) *BaseFace:* 0.19 *DebFace:* 0.39 (f) *BaseFace:* 0.20 *DebFace:* 0.34

Figure 1: Example image pairs of different demographic cohorts. The similarity scores below each row are obtained by a baseline method and our DebFace. First row lists imposter pairs (false accepted by BaseFace) and second row lists genuine pairs (false rejected by BaseFace).

Abstract

We address the problem of bias in automated face recognition algorithms, where errors are consistently lower on certain cohorts belonging to specific demographic groups. We present a novel de-biasing adversarial network 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.

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 on different demographic groups? In fact, it has been observed that many face recognition systems have lower performance for certain demographic groups than others [21, 27]. Such face recognition systems are said to be *biased* in terms of demographics.

At the time when face recognition systems are being deployed in real world for societal benefit, this type of bias [1] is not acceptable. Why does the bias problem exist in face recognition systems? First of all, 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 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 [4, 46]. 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

[1] This is different from the notion of machine learning bias to mean "any basis for choosing one generalization [hypothesis] over another, other than strict consistency with the observed training instances" [13].

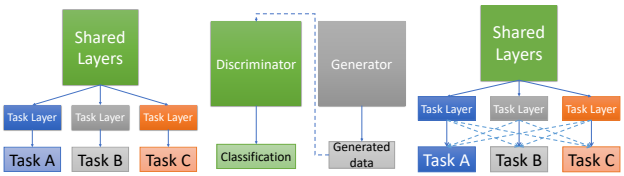
function will result in feature vectors of the specific demographics with lower discriminative ability. Klare *et al*. [27] shows the errors that are inherent to some demographics by studying non-trainable face recognition algorithms.

To address the bias issue, data re-sampling methods have been exploited to balance the data distribution by under-sampling the majority [14] or over-sampling the minority classes [7, 36]. Despite its simplicity, valuable information may be removed by under-sampling, and over-sampling may introduce noisy samples. Another common option for imbalanced data training is cost-sensitive learning that (i) assigns weights for different classes, (ii) samples based on their frequency [22] or the effective number of samples [5, 10]. To eschew the overfitting of Deep Neural Network (DNN) to minority classes, hinge loss is often used to train classifiers that increase margins among classification decision boundaries [19, 25]. The aforementioned methods have also been adopted for face recognition and attribute prediction on imbalanced datasets [23, 53]. 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 demographic bias on face recognition performance. In typical deep learning based face recognition frameworks, face feature encoders are trained on ample amounts of face data to generate a feature representation for each image. The large capacity of DNN enables the face representations to embed demographic details, including gender, race, and age [2, 15]. Thus, the biased demographic information is transmitted from the training dataset to the output representations. To tackle this issue, we assume that if 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. 2b). 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 bias (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 undesired 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 classifiers of each task to act as adversarial supervision for the other tasks (e.g., the dash arrows in Fig. 2c). These four classifiers help each other



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

Figure 2: Methods to learn different tasks simultaneously. Solid lines are typical feature flow in CNN, while dash lines are adversarial losses.

to achieve better feature disentanglement, resulting in unbiased feature representations for both the identity and demographic attributes. As shown in Fig. 2, our proposed framework is novel and in sharp contrast to prior works in 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 performance 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 obtained from the de-biasing network (DebFace-ID) shows lower bias on different demographic cohorts and also achieves SOTA face verification results on the cross-age face recognition and race-unbiased face recognition.

◇ The demographic estimations through DebFace are less biased across different.

◇ Combine ID with demographics to obtain robust features for face recognition on biased datasets.

2. Related Work

Face Recognition on Imbalanced Training Data Previous efforts on face recognition aim to tackle the class imbalance problem on training data. For example, in prior-DNN era, Zhang *et al*. [59] 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*. [31] 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. Range Loss [58] 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 [56] and large margin feature augmentation [53] are proposed to augment features of minority identities and equalize class distribution. Huang *et al*. [23] propose cluster-based large margin local embedding that reduces local data imbalance. Despite their effectiveness, these studies ignore the influence of demographic imbalance issue on the face dataset, which may lead to demographic bias. For instance, both [21] and [27] show that face recognition algorithms consistently perform worse on certain demographic cohorts. To uncover deep learning bias, Alexander *et al*. [3] develop an algorithm to mitigate the hidden biases within training data. To our knowledge, no studies have tackled the challenge of debiasing DNN-based face recognition algorithms.

Adversarial Learning and Disentangled Representation Adversarial learning [41] has been well explored in many computer vision applications. For example, Generative Adversarial Networks (GANs) [16] 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 problems [48, 49, 33, 45]. A problem of current interest is to learn interpretable representations with semantic meaning. There have been many studies that learn factors of variations in the data by supervised learning [29, 30], or semi-supervised/unsupervised learning [26, 37, 32], referred 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*. [30] 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 prior works in that the each branch of a multi-task network act as both a generator and discriminators of other branches (Fig. 2c).

3. Methodology

3.1. Problem Definition

The concept of unbiased face recognition is that given a face recognition system, equal performances can be achieved in different categories of face images. Despite the research on pose-invariant face recognition that aims for equal performance on all poses, 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.

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. We aim to learn a face representation that carries equal discriminative information across demographic cohorts. 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 that under-represents demographic minorities while over-represents majorities. Naively re-sampling training data may still induce bias since the diversities of latent variables are 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 disentangles 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. Some demographic attributes, e.g., gender and race, are two of the factors that determine facial appearances and can provide identification-related information. On the other hand, demographic attributes are heterogeneous in terms of data type and semantics [18]. Individual attributes like race are fixed while age or gender may change individually over time. Meanwhile, the three demographic attributes are semantically independent. A male person, for example, is not necessary to be 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 goal is to diminish demographic bias from face representation, we observe that demographic estimations are biased as well (see Fig. 8). How can we remove the bias of face recognition when demographic estimations themselves are biased? To increase fairness of all demographic classifiers and decrease bias of both face recognition and demographic estimations, we propose a de-biasing network, DebFace, that disentangles the representation into gender, age, race, and identity (DebFace-ID), respectively.

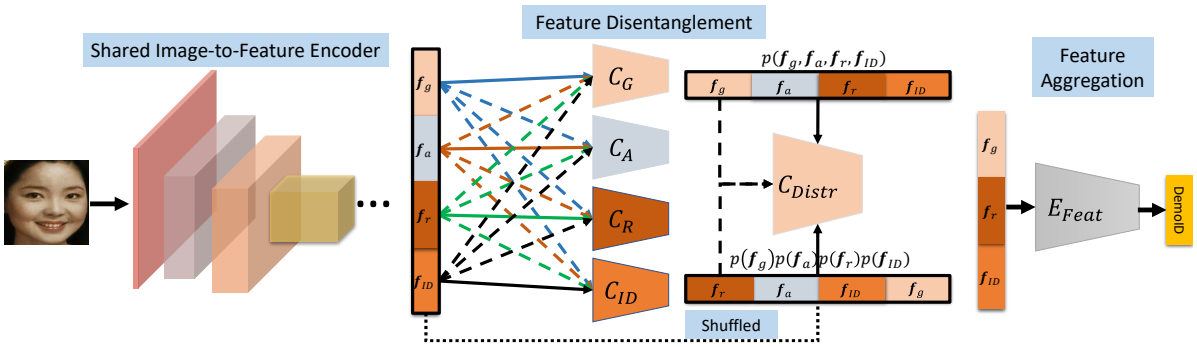


Figure 3: Overview of the proposed the De-biasing face network. The dashed arrows represent adversarial training.

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 is 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 light-weight fully-connected network that is trained to aggregate the representations into a face representation with the same dimensionality as DebFace-ID.

3.3. Network Architecture

Figure 3 gives an overview of the proposed de-biasing face recognition network. It consists of four components, namely, 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 $E_{Img}(\mathbf{x}^{(i)})$ with $4D$ dimensionality. 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 DebFace-ID $\mathbf{f}_{ID}^{(i)}$, 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 is the standard cross entropy loss function $\mathcal{L}_{C_{Demo}}(\mathbf{x}, y_{Demo}; E_{Img}, C_{Demo}) = -\sum_{k=1}^{K} \mathbb{I}(k = y_{Demo}) \log \frac{e^{C_{Demo}(\mathbf{f}_{Demo})_k}}{\sum_{j=1}^{K} e^{C_{Demo}(\mathbf{f}_{Demo})_j}}$, where $\mathbb{I}(x = y) = \begin{cases} 1 & \text{for } x = y \\ 0 & \text{for } x \neq y \end{cases}$ is an index function, $y_{Demo} = \{y_g, y_a, y_r\}$, $C_{Demo} = \{C_G, C_A, C_R\}$, and $\mathbf{f}_{Demo} = \{\mathbf{f}_g, \mathbf{f}_a, \mathbf{f}_r\}$. For the $n-$identity classification, we adopt AM-Softmax [50] as the objective function $\mathcal{L}_{C_{ID}}(\mathbf{x}, y_{id}; E_{Img}, C_{ID}) = -\sum_{k=1}^{n} \mathbb{I}(k = y_{id}) \cdot \log \frac{e^{s \cdot C_{ID}(\mathbf{f}_{ID})_k - m}}{e^{s \cdot C_{ID}(\mathbf{f}_{ID})_k - m} + \sum_{j=1, j \neq k}^{n} e^{s \cdot C_{ID}(\mathbf{f}_{ID})_j}}$, where s is the feature scale, and m is the angular margin.

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 representation 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 described in more details in Sec. 3.4.

Altogether, the proposed de-biasing face network endeavors to minimize the joint loss function:

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

where λ and ν are hyper-parameters determining how completely 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 $\mathcal{L}_{E_{Feat}} = \frac{1}{M}\sum_{i=1}^{M}[\|\mathbf{f}^{(i)}_{DemoID^a} - \mathbf{f}^{(i)}_{DemoID^p}\|_2^2 - \|\mathbf{f}^{(i)}_{DemoID^a} - \mathbf{f}^{(i)}_{DemoID^n}\|_2^2 + \alpha]_+$, where M is the number of hard triplets in a mini-batch, and $\{\mathbf{f}^{(i)}_{DemoID^a}, \mathbf{f}^{(i)}_{DemoID^p}, \mathbf{f}^{(i)}_{DemoID^n}\}$ is the i^{th} triplet consisting of an anchor, a positive, and a negative DemoID representation. $[x]_+ = \max(0, x)$, 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, which mitigates bias related to those attributes. Such a representation is invariant if a classifier trained on it cannot correctly classify the categories of the attribute 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. 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 $\mathcal{L}^G_{Adv}(\mathbf{x}, y_{Demo}, y_{id}; E_{Img}, C_G) = -\sum_{k=1}^{K_G}\frac{1}{K_G}\cdot(\log\frac{e^{C_G(\mathbf{f}_{Demo})_k}}{\sum_{j=1}^{K_G}e^{C_G(\mathbf{f}_{Demo})_j}} + \log\frac{e^{C_G(\mathbf{f}_{ID})_k}}{\sum_{j=1}^{K_G}e^{C_G(\mathbf{f}_{ID})_j}})$, 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 strives for gender-invariance by finding a representation that makes the gender classifier C_G perform poorly. To this end, we minimize the adversarial loss by only updating parameters in E_{Img}.

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

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 $\mathcal{L}_{Fact}(\mathbf{x}, y_{Distr}; E_{Img}, C_{Distr}) = -\sum_{i=1}^{2}\frac{1}{2}\log\frac{e^{C_{Distr}(\mathbf{f}_{Distr})_i}}{\sum_{j=1}^{2}e^{C_{Distr}(\mathbf{f}_{Distr})_j}}$. 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 then randomly shuffle the feature vectors of each attribute in a batch, and re-concatenate them into $4D$-dimensional vectors, 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 learn decomposed representations with minimum mutual information.

4. Experiments

4.1. Datasets and Pre-processing

Datasets: We utilize 15 face datasets in this work, for learning the demographic estimation models, the baseline face recognition model, the de-biasing face model as well as for evaluating these models. To be specific, CACD [8], IMDB [40], UTKFace [60], AgeDB [35], AFAD [38], AAF [9], FG-NET [3], RFW [52], IMFDB-CVIT [42], Asian-DeepGlint [1], and PCSO [11] are the datasets for training and testing models of demographic estimations; and the datasets for learning and evaluating models of face verification are MS-Celeb-1M [17], LFW [24], IJB-A [28], and IJB-C [34].

Pre-Processing: All face images are detected by MTCNN [57]. Each face is cropped and resized to 112×112 pixels using a similarity transformation based on the detected five landmarks.

4.2. Implementation Details

We train the proposed de-biasing network on a cleaned version of MS-Celeb-1M [12], using the ArcFace architecture [12] with 50 layers for the encoder E_{Img}. Since there is no demographic labels in MS-Celeb-1M, we first train three demographic 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

[3] https://yanweifu.github.io/FG_NET_data

PCSO for race estimation training. All the demographic models use ResNet [20] with 34 layers for age, 18 layers for gender and race.

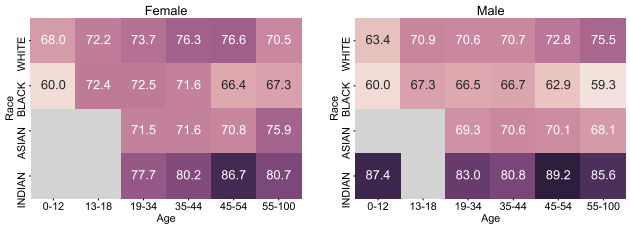
We predict the demographic labels of MS-Celeb-1M with the well-trained demographic models. Our DebFace is then trained on the re-labeled MS-Celeb-1M using SGD with a momentum of 0.9, a weight decay of 0.01, and a batch side of 256. The learning rate starts from 0.1 and drops to 0.0001 following the schedule at 8, 13, and 15 epochs. The model is trained for 30 epochs. The dimensionality of the embedding layer of E_{Img} is 4×512 so that each attribute representation (gender, age, race, ID) is a 512-dim vector. We keep the hyper-parameter setting of AM-Softmax as [12]: $s = 64$ 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 DebFace-ID are concatenated into a 3×512-dim vector, which is the input of E_{Feat}. The network is then trained to output a 512-dim feature representation for face recognition on biased datasets.

4.3. De-biasing Face Verification

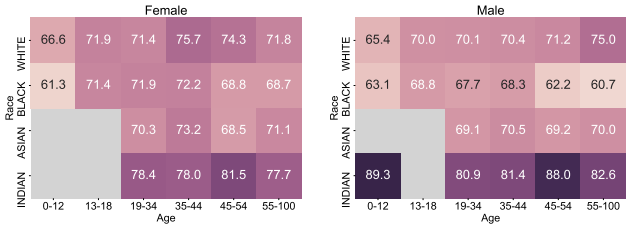
Baseline: We compare DebFace with a regular face representation model which has the same architecture as the shared feature encoder of DebFace. Referred as BaseFace, this baseline model is also trained on MS-Celeb-1M, with the representation dimension of 512.

To show the efficacy of DebFace on bias mitigation in face recognition, we evaluate the verification performance of both DebFace and BaseFace on faces from each demographic cohort separately. There are 48 total cohorts given the combination of demographic attributes including gender (male, female), race (Black, White, East Asian, Indian), and age group (0-12, 13-18, 19-34, 35-44, 45-54, 55-100). We combine IMDB, CACD, AgeDB, and CVIT as the testing set. Overlapping identities among these datasets are removed. Pre-defining a False Accept Rate (FAR) and comparing the corresponding True Accept Rate (TAR) may be biased due to the limited number of images in minority classes. Besides, the thresholds derived from FAR are susceptible to errors of the identity labels, especially to minorities. Therefore, we report the Area Under the Curve (AUC) - Receiver Operating Characteristics (ROC) that involves FAR from zero to one for each demographic group. We define the degree of bias, termed biasness, as the standard deviation of performance across cohorts.

Figure 4 shows the face verification results of BaseFace and DebFace on each cohort. That is, for a particular face representation (e.g., DebFace), we report its AUC on each cohort within that demographic and put the number in the corresponding cell. For example, on the female heatmap,

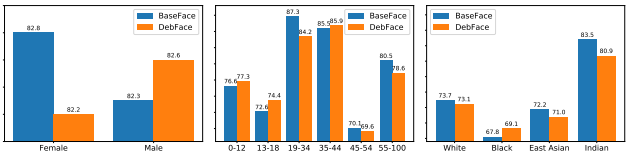


(a) BaseFace



(b) DebFace

Figure 4: Face Verification AUC in 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. The biasness of BaseFace and DebFace are 0.0726 and **0.0638**, respectively.

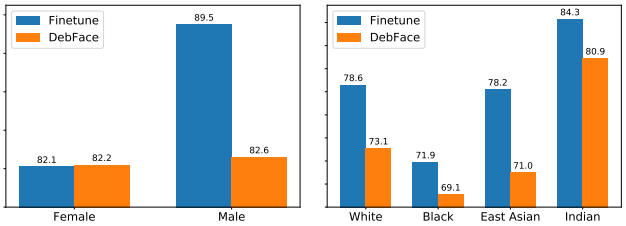


(a) Gender (b) Age (c) Race

Figure 5: The overall performance of face verification AUC on the gender, age, and race, respectively. The biasness of BaseFace and DebFace on gender is 0.0025 and **0.0020**; 0.0631 and **0.0555** on age; 0.0574 and **0.0449** on race.

the first cell represents the performance of BaseFace on faces of white female, aging from 0 to 12. From these heatmaps, we can observe that both DebFace and BaseFace present the bias issue in face verification, where the performance in some cohorts are significantly worse than others, especially the cohort of black children and elder people. Compared to BaseFace, DebFace suggests less bias and the difference of AUC on the cohorts is smaller, where the heatmap exhibits smoother edges. Note that the overall performance of DebFace declines compared to BaseFace. This is because part of the identity-related information like gender and race is disentangled from identity so that the discriminativeness of Debface-ID deteriorates.

Figure 5 shows the performance of face verification on 12 cohorts based on three demographic categories. Both DebFace and BaseFace present similar relative accuracies across cohorts. For example, both algorithms performs worse on the children cohort than the adults; and the perfor-



(a) Gender (b) Race

Figure 6: Face Verification AUC in each demographic cohort. The comparison is between the finetuned BaseFace and DebFace. The biasness of Finetune and DebFace on gender is 0.0037 and **0.0020**; **0.0439** and 0.0449 on race

mance on the Indian cohort is significantly higher than the other races. DebFace decreases the bias from demographics by gaining discriminative features of minorities in spite of the reduction in the performance of majorities.

To further demonstrate the intrinsic bias in different cohorts, we also finetune BaseFace using face images that only belong to a specific cohort. Since age is not informative in terms of identity, we only finetune BaseFace on six cohorts of gender and race separately. Figure 6 shows the performance of the finetuned models versus DebFace. Compared to BaseFace, the AUC increases on most of the cohorts by finetuning except female. However, there are still bias even after finetuning on each cohort. Our DebFace cannot do no better than finetuned models in terms of de-biasing the race influence. For gender groups, on the other hand, bias between male and female increases by finetuning, suggesting that the de-biasing factors in DebFace are capable of mitigating the gender bias in face verification.

4.4. De-biasing Demographic Estimation

Baseline: We further explore the bias of demographic estimation and compare DebFace with baseline estimation models. We train three demographic estimation models, namely, gender estimation (BaseGender), age estimation (BaseAge), and race estimation (BaseRace). For fairness, all three models have the same architecture and training dataset as the shared layers of DebFace. All the demographic estimations are mapped as classification problems, so classification accuracy is used as the performance metric.

We combine the four datasets mentioned in Sec. 4.3 with Asian-DeepGlint as the global testing set. Note that not all of the datasets include labels of all three demographics. Thus, we again employ the demographic models that were trained to label MS-Celeb-1M. For the dataset without certain demographic labels, we simply use the corresponding model to predict the labels.

As shown in Fig. 8, all demographic estimations present significant bias. For gender estimation, both algorithms perform worse on the White and Black cohorts than the East Asian and Indian cohorts. In addition, the performance on young children is significantly worse than adults. In general, the race estimation models perform better on the male cohort than female. Compared to gender, race estimation shows higher bias in terms of age cohorts. Both the baseline method and DebFace perform worse on cohorts with age between 13 to 44 than other age groups. Similar to race, age estimation still achieves better performance on the male cohort than female. Moreover, the white cohort shows dominant advantages over other races in age estimation. In spite of the existing bias in demographic estimations, the proposed DebFace is still able to diminish the bias derived from algorithms. Compared to Fig. 8a, 8b, 8c, cells in Fig. 8d, 8e, 8f present more uniform colors.

Table 1: footnotesizePerformance on LFW and IJB-A, with verification accuracy on LFW and TAR@0.1% FAR on IJB-A.

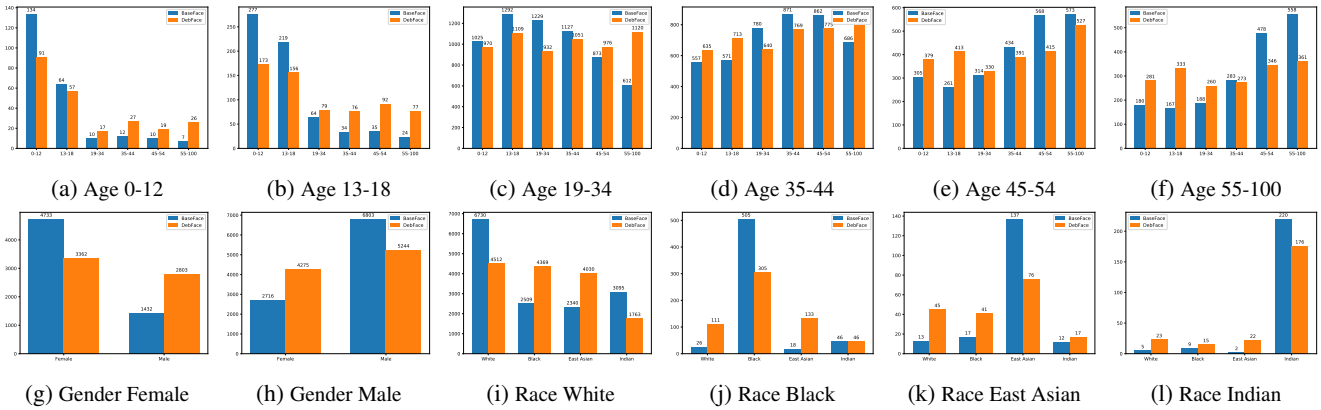
Method	LFW (%)	Method	IJB-A (%)
DeepFace+ [44]	97.35	DR-GAN [47]	53.9 ± 4.3
CosFace [51]	99.73	Yin *et al.* [55]	73.9 ± 4.2
L2-Face [39]	99.78	Cao *et al.* [6]	90.4 ± 1.4
ArcFace [12]	99.83	Multicolumn [54]	92.0 ± 1.3
PFE [43]	99.82	PFE [43]	95.3 ± 0.9
BaseFace	99.38	*BaseFace*	90.2 ± 1.1
DebFace	98.97	*DebFace*	87.6 ± 0.9
DemoID	99.50	*DemoID*	92.2 ± 0.8

4.5. Face Verification on Public Protocols

We compare the face verification performance of the proposed method with SOTA methods, on three public benchmarks: LFW, IJB-A, and IJB-C. All three datasets exhibit imbalanced data distribution in terms of demographics.

Ablations: We report the performance of three different settings, using 1) BaseFace, the same baseline in Sec. 4.3, 2) the ID representation output by DebFace, and 3) the fused representation DemoID.

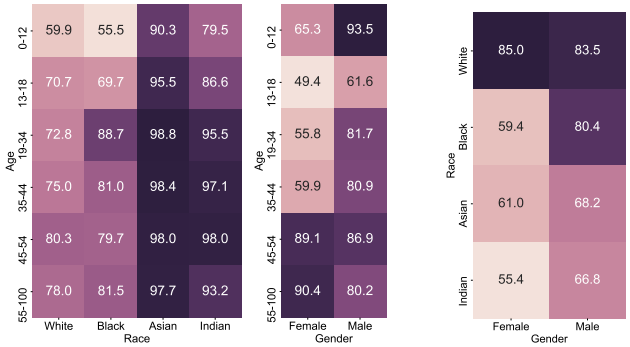
As shown in Tabs. 1, 2, the ID representation of DebFace is less discriminative than BaseFace, or DemoID, since race and gender are essential components of identity-related face features. Thus, the performance improves by simply concatenating race and gender features with DebFace-ID. On the other hand, re-introducing race and gender features to the face representation through the aggregation model may inevitably lead to demographic bias. In the sense of de-biasing, it is preferable to concatenate race and gender directly with the de-biased ID. 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-



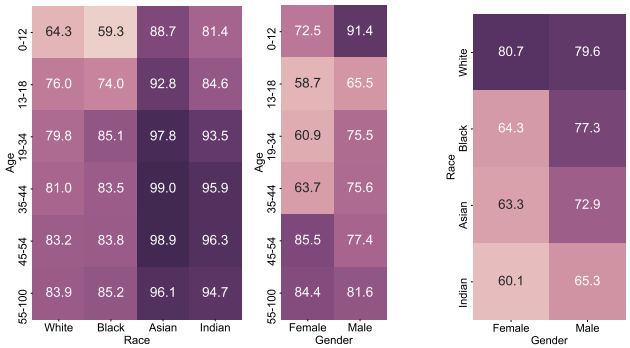
(a) Age 0-12 (b) Age 13-18 (c) Age 19-34 (d) Age 35-44 (e) Age 45-54 (f) Age 55-100

(g) Gender Female (h) Gender Male (i) Race White (j) Race Black (k) Race East Asian (l) Race Indian

Figure 7: Feature distribution of the representation output by BaseFace and DebFace.



(a) BaseGender (b) BaseRace (c) BaseAge



(d) DebFace-Gender (e) DebFace-Race (f) DebFace-Age

Figure 8: Classification accuracy of the demographic estimations on faces of different cohorts, for the baseline models and DebFace. The biasness of the baseline model and DebFace is 0.1238 and **0.1022** on gender; 0.1458 and **0.1000** on race; 0.1083 and **0.0761** on age.

biasing. Fortunately our algorithm design offers the flexibility in handling this trade-off.

4.6. Qualitative Analysis of Disentanglement

To demonstrate the feature disentanglement by DebFace, we plot the distribution of the nearest neighbors of the face images in the feature space. For example, Fig. 7g illustrates the gender distribution of the nearest neighbors of all the female faces in the dataset. In the feature space of DebFace, there are $3,362$ points that are nearest to the females faces belong to the female cohorts, and $2,803$ points belong to the male cohorts. As shown in Fig. 7, the DebFace representation presents more uniform distribution compared to BaseFace, indicating that faces within different demographic groups are converged together and the demographic information is disentangled from the face representation.

Table 2: Verification performance on IJB-C.

Method	TAR @ FAR (%)		
	0.001%	0.01%	0.1%
Yin *et al.* [55]	-	-	69.3
Cao *et al.* [6]	74.7	84.0	91.0
Multicolumn [54]	77.1	86.2	92.7
PFE [43]	89.6	93.3	95.5
BaseFace	80.2	88.0	92.9
DebFace	82.0	88.1	89.5
DemoID	83.2	89.4	92.9

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 not only DebFace can reduce bias in face recognition but in demographic estimation as well. Our future work will explore an aggregation scheme to combine race, gender, and identity without introducing algorithmic and dataset bias.

References

[1] http://trillionpairs.deepglint.com/overview. 5

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016. 2

[3] Alexander Amini, Ava Soleimany, Wilko Schwarting, Sangeeta Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. AAAI/ACM Conference on AI, Ethics, and Society, 2019. 3

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016. 1

[5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv preprint arXiv:1906.07413*, 2019. 2

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 2018. 7, 8

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of Artificial Intelligence research*, 16:321–357, 2002. 2

[8] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *ECCV*, 2014. 5

[9] Jingchun Cheng, Yali Li, Jilong Wang, Le Yu, and Shengjin Wang. Exploiting effective facial patches for robust gender recognition. *Tsinghua Science and Technology*, 24(3):333–345, 2019. 5

[10] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 2

[11] Debayan Deb, Lacey Best-Rowden, and Anil K Jain. Face recognition performance under aging. In *CVPRW*, 2017. 5

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 5, 6, 7

[13] Thomas G Dietterich and Eun Bae Kong. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, 1995. 1

[14] Chris Drummond, Robert C Holte, 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. 2

[15] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *the 22nd ACM SIGSAC*, 2015. 2

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 3

[17] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*. Springer, 2016. 5

[18] H. Han, K Jain A, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Trans. Pattern Analysis Machine Intelligence*, PP(99):1–1, 2017. 3

[19] Munawar Hayat, Salman Khan, Waqas Zamir, Jianbing Shen, and Ling Shao. Max-margin class imbalanced learning with gaussian affinity. *arXiv preprint arXiv:1901.07711*, 2019. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[21] J Howard, Y Sirotin, and A Vemury. 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. 1, 3

[22] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, 2016. 2

[23] Chen Huang, Yining Li, Change Loy Chen, and Xiaoou Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2019. 2, 3

[24] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008. 5

[25] Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. In *CVPR*, 2019. 2

[26] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018. 3

[27] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. Face recognition performance: Role of demographic information. *IEEE Trans. Information Forensics and Security*, 7(6):1789–1801, 2012. 1, 2, 3

[28] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR*, 2015. 5

[29] Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Wassell. Multi-task adversarial network for disentangled feature learning. In *CVPR*, 2018. 3

[30] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Exploring disentangled feature representation beyond face identification. In *CVPR*, 2018. 3

[31] Yi-Hung Liu and Yen-Ting Chen. Face recognition using total margin-based adaptive fuzzy support vector machines. *IEEE Transactions on Neural Networks*, 18(1):178–192, 2007. 2

[32] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018. 3

[33] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NIPS*, 2018. 3

[34] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 ICB*, 2018. 5

[35] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *CVPRW*, 2017. 5

[36] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Generative adversarial minority oversampling. *arXiv preprint arXiv:1903.09730*, 2019. 2

[37] Siddharth Narayanaswamy, T Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning disentangled representations with semi-supervised deep generative models. In *NIPS*, 2017. 3

[38] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, 2016. 5

[39] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017. 7

[40] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 2018. 5

[41] Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992. 3

[42] Parisa Beham Jyothi Gudavalli Menaka Kandasamy Radhesyam Vaddi Vidyagouri Hemadri J C Karure Raja Raju Rajan Vijay Kumar Shankar Setty, Moula Husain and C V Jawahar. Indian Movie Face Database: A Benchmark for Face Recognition Under Wide Variations. In *NCVPRIPG*, 2013. 5

[43] Yichun Shi, Anil K Jain, and Nathan D Kalka. Probabilistic face embeddings. *arXiv preprint arXiv:1904.09658*, 2019. 7, 8

[44] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 7

[45] Chaofan Tao, Fengmao Lv, Lixin Duan, and Min Wu. Minimax entropy network: Learning category-invariant features for domain adaptation. *arXiv preprint arXiv:1904.09601*, 2019. 3

[46] Antonio Torralba, Alexei A Efros, et al. Unbiased look at dataset bias. In *CVPR*, 2011. 1

[47] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017. 7

[48] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *CVPR*, 2015. 3

[49] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 3

[50] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 4

[51] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, 2018. 7

[52] Mei Wang, Weihong Deng, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang. Racial faces in-the-wild: Reducing racial bias by deep unsupervised domain adaptation. *arXiv preprint arXiv:1812.00194*, 2018. 5

[53] Pingyu Wang, Fei Su, Zhicheng Zhao, Yandong Guo, Yanyun Zhao, and Bojin Zhuang. Deep class-skewed learning for face recognition. *Neurocomputing*, 2019. 2, 3

[54] Weidi Xie and Andrew Zisserman. Multicolumn networks for face recognition. *arXiv preprint arXiv:1807.09192*, 2018. 7, 8

[55] Xi Yin and Xiaoming Liu. Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Trans. Image Processing*, 27(2):964–975, 2017. 7, 8

[56] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *CVPR*, 2019. 3

[57] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 5

[58] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *CVPR*, 2017. 3

[59] Yin Zhang and Zhi-Hua Zhou. Cost-sensitive face recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(10):1758–1769, 2009. 2

[60] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*. IEEE, 2017. 5