PASS: Protected Attribute Suppression System for Mitigating Bias in Face Recognition

Prithviraj Dhar*[1], Joshua Gleason*[2], Aniket Roy[1], Carlos D. Castillo[1], Rama Chellappa[1]
[1]Johns Hopkins University, [2]University of Maryland, College Park
{pdhar1,aroy28,carlosdc,rchella4}@jhu.edu, gleason@umd.edu

Abstract

Face recognition networks encode information about sensitive attributes while being trained for identity classification. Such encoding has two major issues: (a) it makes the face representations susceptible to privacy leakage (b) it appears to contribute to bias in face recognition. However, existing bias mitigation approaches generally require end-to-end training and are unable to achieve high verification accuracy. Therefore, we present a descriptor-based adversarial de-biasing approach called 'Protected Attribute Suppression System (PASS)'. PASS can be trained on top of descriptors obtained from any previously trained high-performing network to classify identities and simultaneously reduce encoding of sensitive attributes. This eliminates the need for end-to-end training. As a component of PASS, we present a novel discriminator training strategy that discourages a network from encoding protected attribute information. We show the efficacy of PASS to reduce gender and skintone information in descriptors from SOTA face recognition networks like Arcface. As a result, PASS descriptors outperform existing baselines in reducing gender and skintone bias on the IJB-C dataset, while maintaining a high verification accuracy.

1. Introduction

Over the past few years, the accuracy of face recognition networks has significantly improved [40, 41, 36, 15, 10, 17]. These improvements have led to the deployment of face recognition systems in a large number of applications. However, recent studies [16, 25, 43] have also shown that face recognition networks encode information about protected attributes such as race, gender, and age, while being trained for identity classification. Encoding of sensitive attributes raises concerns regarding privacy and bias.

Privacy concerns: Many large-scale face verification and identification systems employ a database that stores face descriptors of identities, as opposed to face images. Face descriptors refer to the features extracted from the penultimate layer of a previously trained face recognition network.

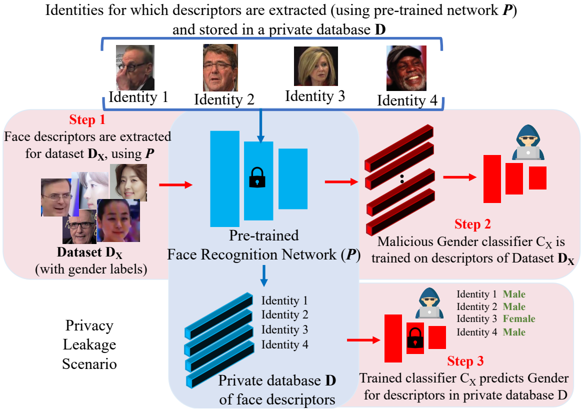

Identities for which descriptors are extracted (using pre-trained network ***P***)
and stored in a private database **D**
Identity 1
Identity 2
Identity 3
Identity 4
Step 1
Face descriptors are extracted for dataset $\mathbf{D_X}$, using ***P***
Dataset $\mathbf{D_X}$
(with gender labels)
Pre-trained
Face Recognition Network (***P***)
Step 2
Malicious Gender classifier C_X is trained on descriptors of Dataset $\mathbf{D_X}$
Privacy
Leakage
Scenario
Identity 1
Identity 2
Identity 3
Identity 4
Private database **D**
of face descriptors
Identity 1 **Male**
Identity 2 **Male**
Identity 3 **Female**
Identity 4 **Male**
Step 3
Trained classifier C_X predicts Gender for descriptors in private database D

Figure 1. Suppose a malicious agent X has gained access to a private database D (blue) which consists of a pre-trained network P and face descriptors of four identities. The agent can use P to extract descriptors (red) for a gender-labeled dataset D_X (Step 1). Using these descriptors, the agent can train a gender classifier C_X (Step 2). Using the trained C_X, the agent can predict the gender of the descriptors in D (Step 3) and thus cause privacy breach.

Storing descriptors, rather than images, allows for very fast gallery lookup and verification against known subjects. This also acts as an additional layer of security by not storing potentially sensitive information present in the original face images. However, since some sensitive information is still encoded in these descriptors (e.g. race, gender, age), a malicious agent with access to these descriptors can potentially extract this information and use it for nefarious purposes. An example scenario is presented in Figure 1.

Bias concerns: Encoding of protected attributes such as gender or race in face descriptors results in bias w.r.t. these attributes when used for face recognition. A recent study from NIST [22] found evidence that characteristics such as gender and ethnicity impact verification and matching performance of face descriptors. Similarly, it has been shown that most face-based gender classifiers perform significantly better on male faces with light skintone than female faces with dark skintone [12].

One method of addressing privacy and bias issues is by producing face descriptors that are independent of the pro-

*These authors have contributed equally to this work.

tected attribute(s). For instance, Debface [20] proposes an end-to-end method for producing face descriptors that are disentangled from protected attributes using an adversarial approach. Another common strategy for mitigating bias is to train face recognition systems using training datasets that are balanced in terms of sensitive attributes. However, building large datasets that are balanced in terms of the attributes we want to protect is difficult, expensive, and time-consuming. Moreover, once such a 'fair' dataset is constructed, we still need to perform the costly operation of training a large recognition network from scratch.

End-to-end training of a large-scale network requires access to a large dataset and computing power, and is time-consuming. Application of adversarial losses while training (as done in [20]), also slows down the training process. Several works [48, 11, 27] show that reducing the information of sensitive attributes while training a network results in a drop in overall performance. Even if a new network is trained to generate attribute-agnostic face descriptors, we need to replace the existing network (say, P in Fig 1), and re-compute the descriptors for all the identities by feeding in the respective face images.

In this work, we propose a solution that addresses the following four points: (i) reduces the opportunity for leakage of protected attributes in face descriptors. (ii) mitigates bias with respect to multiple attributes (gender and skintone). (iii) operates on existing descriptors and does not require expensive end-to-end training. (iv) does not require a balanced training dataset.

The proposed method trains a lightweight model that transforms face descriptors obtained from an existing face recognition model, and maps them to an attribute agnostic representation. We achieve this using a novel adversarial training procedure called **P**rotected **A**ttribute **S**uppression **S**ystem (PASS). Unlike other works that adversarially suppress protected attributes [20, 48] using end-to-end training, we operate on descriptor space. Once trained, PASS may be easily applied to other existing face descriptors. In summary, we make the following contributions in this paper:

1. We present PASS, an adversarial method that aims to reduce the information of sensitive attributes in face descriptors from any face recognition network, while maintaining high face verification performance. We show the efficacy of PASS to reduce gender and skintone information in face descriptors, and thus considerably reduce the associated biases. Moreover, PASS can be used on top of face descriptors obtained from *any* face recognition network. We show these results on two SOTA pre-trained networks: Arcface [15] and Crystalface [36].

2. Our descriptor-based model cannot include CNN-based discriminators, which poses new challenges. We present a novel discriminator training strategy in PASS, to enforce the removal of sensitive information in the descriptors.

Method	Target task	Sensitive attribute
[49, 30]	Analogy completion	Gender
[47]	Object classification	Gender
[48]	Action classification	Identity, private attributes
[13]	Action recognition	Scene
[6]	Gender/Age prediction	Age/Gender
[19]	Preserve pose/illumination/expresssion	Identity
[27]	Smile, high-cheekbones	Gender, make-up
[7]	Face detection	Skintone
[35]	Face attractiveness	Gender
[46]	Face recognition	Race
[20]	Face recognition	Age,gender,race
PASS (Ours)	Face recognition	Gender, skintone

Table 1. Methods that adversarially remove sensitive attributes in general vision/NLP tasks (top) and face-related tasks (bottom)

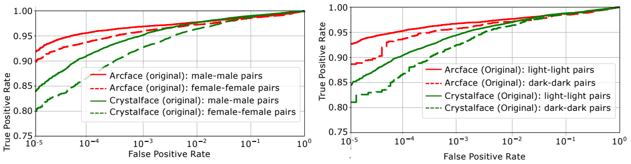
3. We extend PASS to reduce information of multiple attributes simultaneously, and show that such a framework (known as 'MultiPASS') also performs well in terms of reducing the leakage of sensitive attributes and bias in face descriptors, while maintaining reasonable face verification performance.

4. Since reducing the information of protected attributes in face descriptors also reduces their identity-classifying capability, we introduce a new metric called Bias Performance Coefficient (BPC), that measures the trade-off between bias reduction and drop in verification performance. We show that our PASS framework achieves better BPC values than existing baselines.

2. Related work

Bias in face recognition: Several empirical studies [22, 12, 18] have shown that many publicly available face recognition systems demonstrate bias towards attributes such as race and gender. [46, 45, 21] highlight the issue of racial bias in face recognition, and propose strategies to mitigate the same. In the context of gender bias [5, 29], most experiments show that the performance of face recognition on females is lower than that of males. Use of cosmetics by females [14, 26] and gendered hairstyles [3] has been assumed to play a major role in the resulting gender bias. However, [4] shows that cosmetics only play a minor role in the gender gap. [29] shows that face verification systems perform better on lighter skintones than darker skintones.[5, 47] show that the gender bias is not mitigated even if the training dataset is gender-balanced. [38, 46] presents an evaluation datasets that is balanced in terms of race and provide the verification protocols for the same.

Adversarial techniques to suppress attributes: A summary of works that adversarially remove sensitive attributes, while performing a target task is provided in Table 1. Most of these works do not operate on descriptor space. Also, in some of the these experiments, the attribute under consideration is ephemeral to the target task. For example, in [48], an action is not specific to an identity. In contrast, attributes



(a) (b)

Figure 2. (a) Gender-wise and (b) Skintone-wise verification plot for Arcface and Crystalface networks, on IJB-C dataset. We define bias as the difference between TPRs of males and females (or dark and light skintones) at a fixed FPR.

like gender and race may not be ephemeral to face recognition. A given identity can be generally tied to a single gender/skintone. Because of the dependence between identity and gender/skintone, disentangling them is harder.

Attribute privacy in face recognition: [32, 33] introduce techniques to synthesize perturbed face images using an adversarial approach so that gender classifiers are confounded, but the performance of a commercial face-matcher is preserved. [42, 11, 44] introduce techniques to suppress protected attributes like race, age and gender in face representations (as opposed to face images). However, the effect of such privacy preserving techniques on bias in face recognition is currently unclear.

3. Problem Statement

Our goal is to reduce gender and skintone information in face descriptors so that the ability of a classifier to predict gender and skintone from these descriptors is reduced. As an additional requirement, we constrain the gender and skintone-agnostic face descriptors to encode sufficient identity information, so that they can be used to perform face verification. We hypothesize that *reducing the ability to predict protected attributes (gender and skintones) in face descriptors will reduce gender/skintone bias in face verification tasks*. This hypothesis is built on the results of [20], which shows that adversarially removing sensitive information from face representations reduces bias. However, unlike [20], we approach the problem in descriptor space.

Mesuring bias: At this point, we quantitatively describe gender and skintone bias in the context of face verification. Most work on face verification [15, 36, 28, 17] report performance of a system by using an ROC (TPR vs FPR) curve, similar to Fig 2. Hence, we define gender and skintone bias, at a given false positive rate (FPR) as follows:

$$\text{Gender Bias}^{(F)} = |\text{TPR}_m^{(F)} - \text{TPR}_f^{(F)}| \tag{1}$$

$$\text{Skintone Bias}^{(F)} = |\text{TPR}_l^{(F)} - \text{TPR}_d^{(F)}| \tag{2}$$

where $(\text{TPR}_m^{(F)}, \text{TPR}_f^{(F)}, \text{TPR}_l^{(F)}, \text{TPR}_d^{(F)})$ denote the true positive rates for the verification of male-male, female-female, light-light and dark-dark pairs respectively at FPR F. In some works such as [20], bias is evaluated as the difference between area under ROC curves (AUC). While this can be viewed as an aggregate of our measure, such an aggregation fails to meaningfully capture the bias at realistic operating points as it marginalizes the performance at low FPR. In our experience, most real world verification systems tend to operate at very low FPR, i.e. less than 10^{-4}, which is not meaningfully captured with AUC. In this work, we focus on FPR values that we consider to be realistic operating conditions.

Measuring bias/performance trade-off: Several methods that reduce the information of sensitive attributes in images or representations demonstrate a slight drop in overall performance of the system [11, 27, 48, 39]. So, reducing gender/skintone information in descriptors for de-biasing may lead to a slight drop in face verification performance. Inspired by the metric in [11], we introduce a new metric called bias performance coefficient (BPC) to measure the trade-off between bias reduction and drop in verification performance.

$$\text{BPC}^{(F)} = \frac{\text{Bias}^{(F)} - \text{Bias}_{deb}^{(F)}}{\text{Bias}^{(F)}} - \frac{\text{TPR}^{(F)} - \text{TPR}_{deb}^{(F)}}{\text{TPR}^{(F)}} \tag{3}$$

Here, $(\text{TPR}^{(F)}, \text{Bias}^{(F)})$ refer to the *overall* TPR obtained by original descriptors and the corresponding bias (Gender/Skintone bias) at FPR of F. $(\text{TPR}_{deb}^{(F)}, \text{Bias}_{deb}^{(F)})$ denote their de-biased counterparts. We prefer an algorithm that obtains higher BPC since a higher BPC denotes high bias reduction and low drop in verification performance. The original face descriptors (without any de-biasing) would have a zero BPC (since $\text{Bias}^{(F)} = \text{Bias}_{deb}^{(F)}$ and $\text{TPR}^{(F)} = \text{TPR}_{deb}^{(F)}$). Note that a negative BPC denotes that the percentage drop in TPR is higher than the percentage reduction in bias. In our work, we denote the BPC for skintone as 'BPC_{st}' and that for gender as 'BPC_{g}'. In summary, we aim *to build systems that achieve high BPC values*.

4. Proposed Approach

4.1. PASS

The key idea in our proposed approach - PASS, is to train a model to classify identities while discouraging it from predicting a specific protected attribute. Firstly, for a given image I, we extract a face descriptor f_{in} using a pre-trained network P.

$$f_{in} = P(I) \tag{4}$$

We present the PASS architecture in Fig. 3. This architecture is inspired by the adversarial framework in [48]. PASS is composed of three components:

(1) **Generator model** M: A model that accepts face descriptor f_{in} from a pre-trained network P, and generates a lower dimensional descriptor $f_{out} \in \mathbb{R}^{256}$. M consists of a single linear layer with 256 units, followed by a PReLU [24] layer. The weights of M are denoted as ϕ_M.

(2) **Classifier** C: A classifier that takes in f_{out} and generates a prediction vector for identity classification. The weights of C are denoted as ϕ_C.

(3) **Ensemble of attribute classifiers** E: An ensemble of

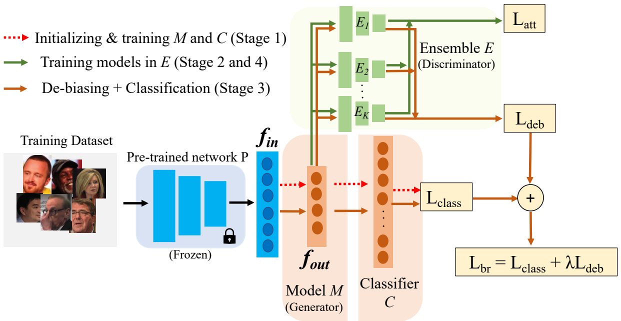


Figure 3. **PASS architecture**. Face descriptors f_{in} are extracted from a previously trained network P and are fed to a model M. M consists of a single linear layer with PReLU activation that outputs transformed face descriptor f_{out}. This is then fed to classifier C and ensemble E. The arrows indicate the dataflow at various training stages. In stage 1, M and C are initialized and trained to classify identity using the gradients of L_{class}. In stage 2, E is initialized and trained to classify attribute using gradients of L_{att}. In stage 3, M and C are trained using the gradients of L_{br} to debias f_{out} with respect to the target attribute, while simultaneously being able to classify identity. In stage 4, one member of ensemble E is trained to classify attribute from f_{out} using the gradients of L_{att}. Stages 3 and 4 are repeated in alternating fashion, where the ensemble member of E being trained in stage 4 changes at each iteration.

K attribute prediction models represented as $E_1, E_2 \dots E_K$ that take f_{out} as input. Each of these models is a two layer MLP with 128 and 64 hidden units respectively with SELU activations, followed by a sigmoid activated output layer with N_{att} units. Here, N_{att} denotes the number of classes in the attribute being considered. We collectively denote the weights of all the models in E as ϕ_E and weights of k^{th} model E_k as ϕ_{E_k}. Note that the attribute classifiers in E are simple MLP networks (and not CNNs as used in [48]). This is because the input to E are low-dimensional descriptors f_{out} and not images.

We now explain PASS as an adversarial approach. M can be viewed as a generator that should generate descriptors f_{out} that are agnostic to the attribute under consideration. f_{out} is fed to the ensemble E of attribute prediction models which acts as a discriminator and tries to predict the protected attribute. The objective of M is to generate descriptors f_{out} that can fool E in terms of attribute prediction, and can also be used to classify identities. Therefore, we impose two constraints on f_{out}: (i) a penalty for misidentification, and (ii) a penalty for attribute predictability from f_{out}. To this end, we propose a bias reducing classification loss L_{br} described in section 4.1.1.

4.1.1 Bias reducing classification loss L_{br}

After extracting the descriptor f_{in} from a pre-trained face recognition network, we pass it through M to obtain a lower dimensional descriptor f_{out}.

$$f_{out} = M(f_{in}, \phi_M) \tag{5}$$

First constraint: To make f_{out} proficient at classifying identities, we provide it to classifier C and use cross-entropy classification loss L_{class} to train both C and M.

$$L_{class}(\phi_M, \phi_C) = -\mathbf{y_{id}}.\log(C(f_{out}, \phi_C)) \tag{6}$$

$\mathbf{y_{id}}$ is a one hot identity label and classifier C produces softmaxed outputs.

Training discriminators: M generates f_{out} which is fed to ensemble E. Each of the attribute prediction models in E, denoted as E_k, is used for computing the cross entropy loss $L_{att}^{(E_k)}$ for attribute classification. L_{att} is computed as the sum of cross-entropy losses for each E_k.

$$L_{att}(\phi_M, \phi_E) = -\sum_{k=1}^{K}\sum_{i=1}^{N_{att}} y_{att,i}\log y_{att,i}^{(k)} \tag{7}$$

$y_{att,i}$ is the binary attribute label for the i^{th} attribute category associated with the input face descriptor, and $y_{att,i}^{(k)}$ represents the respective softmaxed outputs of E_k in the ensemble. N_{att} denotes the number of categories associated with the attribute under consideration.

Training generator (second constraint): After training E, M is trained to transform f_{in} into attribute-agnostic descriptor f_{out}. We then provide f_{out} to each model in E:

$$o_k = E_k(f_{out}, \phi_{E_k}) \text{ for } k = 1 \dots K \tag{8}$$

The outputs o_k are N_{att}-dimensional and represent the probability scores for different categories associated with the attribute. We refer to the i^{th} element of o_k as $o_{k,i}$.

If an optimal classifier operating on f_{out} were to always produce a posterior probability of $\frac{1}{N_{att}}$ for all categories in the attribute, then this implies that no attribute information is present in the descriptor. To this end, we define the adversarial loss $L_{adv}^{(E_k)}$ for the k^{th} model in E to be:

$$L_{adv}^{(E_k)}(\phi_M, \phi_{E_k}) = -\sum_{i=1}^{N_{att}} \frac{1}{N_{att}}\log(o_{k,i}) \tag{9}$$

Here, we use an ensemble of attribute prediction models, rather than a single model because, we want f_{out} to be constructed such that *no* model can predict the protected attribute. This approach was motivated by the work of [48] to solve 'the ∀ challenge'. After computing the adversarial loss for model M with respect to all the models in E, we select the one for which the loss is maximum. We term this loss as debiasing loss L_{deb}.

$$L_{deb}(\phi_M, \phi_E) = \max\{L_{adv}^{(E_k)}(\phi_M, \phi_{E_k})|_{k=1}^{K}\} \tag{10}$$

This loss function penalizes M with respect to the strongest attribute predictor which it was not able to fool. This approach was introduced in [48]. L_{deb} is then combined with L_{class} to compute a bias reducing classification loss L_{br}.

$$L_{br}(\phi_C, \phi_M, \phi_E) = L_{class}(\phi_C, \phi_M) + \lambda L_{deb}(\phi_M, \phi_E) \tag{11}$$

Here, λ is used to weight the de-biasing loss.

4.1.2 Stage-wise Training

We now discuss the various stages of training PASS.

Stage 1 - Initializing and training M and C: Using input

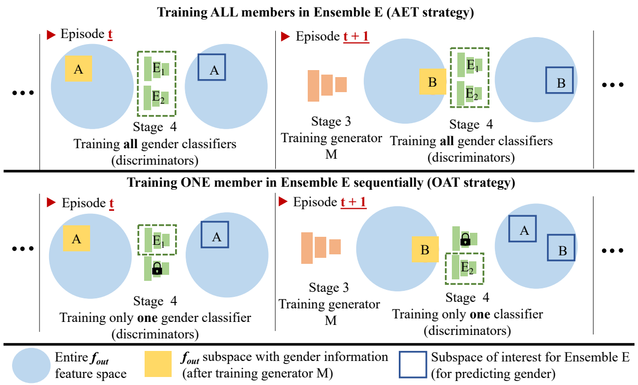

Training ALL members in Ensemble E (AET strategy)
Episode t
A
E1
E2
A
Stage 4
Training all gender classifiers
(discriminators)
Episode t + 1
Stage 3
Training generator
M
B
E1
E2
B
Stage 4
Training all gender classifiers
(discriminators)
Training ONE member in Ensemble E sequentially (OAT strategy)
Episode t
A
E1
A
Stage 4
Training only one gender classifier
(discriminators)
Episode t + 1
Stage 3
Training generator
M
B
E2
A
B
Stage 4
Training only one classifier
(discriminators)
Entire f_out
feature space
f_out subspace with gender information
(after training generator M)
Subspace of interest for Ensemble E
(for predicting gender)

Figure 4. Descriptor space for AET (top) versus OAT (bottom) strategies (example using 2 member ensemble). Using OAT, M is more restricted in how it may represent protected attribute information in descriptor space, encouraging it to instead remove information about the protected attribute all-together.

descriptors f_{in} from a pre-trained network, we train M and C from scratch for T_{fc} iterations using L_{class} (Eq. 6).

Stage 2 - Initializing and training E: Once M is trained to perform classification, we feed the outputs f_{out} of M to an ensemble E of K attribute prediction models. E is trained from scratch to classify attribute for T_{atrain} iterations using L_{att} (Eq. 7). ϕ_M, ϕ_C remain unchanged in this stage.

Stage 3 - Update model M **and classifier** C: Here, M is trained to generate descriptors f_{out} that are proficient in classifying identities and are relatively attribute-agnostic. f_{out} is fed to the ensemble E and the classifier C, the outputs of which result in L_{deb} (Eq. 10) and L_{class} (Eq. 6) respectively. We combine them to compute L_{br} (Eq. 11) for training M and C for T_{deb} iterations, while ϕ_E remains locked. While computing L_{br}, the gradient updates for L_{deb} are propagated to ϕ_M and those for L_{class} are propagated to ϕ_M and ϕ_C.

Stage 4 - Update ensemble E **(discriminator)**: In stage 4, members of E are trained to classify attribute using f_{out}. Therefore, we run stages 3 and 4 alternatively, for T_{ep} episodes, after which we re-initialize and re-train all the models in E (as done in stage 2). This re-initialization follows from [48], in order to prevent trivial overfitting between M and E. Here, one episode indicates an instance of running stages 3 and 4 consecutively. In stage 4, we choose one of the models in E, and train it for T_{plat} iterations or until it reaches an accuracy of A^* on the validation set. ϕ_M and ϕ_C remain locked in this stage. The detailed PASS algorithm is provided in the supplementary material.

4.1.3 One-At-a-time (OAT) vs All-Every-Time (AET)

We note that the method on which PASS is based [48], trains all the discriminators during stage 4 training. We call this 'All-Every-Time (AET)' strategy. However, in this section we present a conceptual argument describing how AET could produce descriptors that still contain sensitive information. The key ideas of this argument are visualized in Fig 4.

Consider the case where PASS consists of an ensemble E with two gender classifiers, and suppose that model M has distilled all gender information into a subspace, A, of descriptor space after stage 3 of episode t. Following the AET strategy, all classifiers in E are trained to classify gender, thus, encouraging them to focus on subspace A. In episode $t+1$, suppose M re-organizes the descriptor space to distill gender information into a new subspace B (orthogonal to A) in order to fool the classifiers in E. In stage 4 of episode $t+1$, all the gender classifiers will then be trained again to extract gender information, causing them to focus on subspace B and forget subspace A. Thus, in stage $t+2$, M could revert to its episode t state, once again distilling gender information back into subspace A without penalty.

To address this issue, we propose a novel discriminator training strategy that we call '**O**ne-**A**t-a-**T**ime (OAT)', where, during stage 4 we train one member in E, and freeze the rest. Using the same example from Fig 4 (bottom row), we describe how this encourages M to remove gender.

As before, suppose that after stage 3 of episode t, M has distilled all gender information into subpace A. However, unlike in the AET example, suppose only member E_1 of ensemble E is trained during stage 4. In stage 3 of episode $t+1$, suppose M again distills gender information into subspace B. During stage 4 of episode $t+1$, E_2 is trained, and the weights of E_1 are held constant. Thus, after 2 episodes the prediction of ensemble E depends on both subspace A and B (since E_1 is still dependent on subspace A). Our conclusion is that this strategy restricts M from reverting back to its episode t state after stage 3 of episode $t+2$, thus improving the chance that M removes gender information all-together.

For the PASS architecture with K classifiers in ensemble E, at episode i, we train the j^{th} classifier in the ensemble, where $j = i \bmod K$, and freeze the rest (thus sequentially choosing one discriminator). We conduct experiments to compare OAT and AET (in Section 5.5) and show that OAT leads to better attribute-removal as compared to AET.

4.2. MultiPASS

We also propose MultiPASS (Fig 5), by extending PASS to reduce the information of several sensitive attributes simultaneously. Here, we describe how to extend PASS to tackle two attributes.

We consider two attributes : Attribute a, with $N_{att}^{(a)}$ categories and attribute b, with $N_{att}^{(b)}$ categories. In contrast to PASS, we include two ensembles of discriminators in MultiPASS: one for attribute a, denoted as $E^{(a)}$ and one for attribute b, denoted as $E^{(b)}$. Let $E^{(a)}$ and $E^{(b)}$ consist of K_a and K_b adversary classifiers respectively. The stage 1 training for model M in MultiPASS is same as that in PASS. In stage 2, we train both $E^{(a)}$ and $E^{(b)}$. In stage 3, we compute the outputs $o_k^{(a)}$ from all the classifiers in $E^{(a)}$

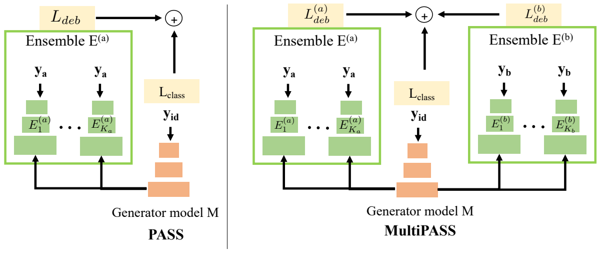


Figure 5. We build MultiPASS by extending PASS to tackle two attributes simultaneously.

by extending Eq 8.

$$o_k^{(a)} = E_k(f_{out}, \phi_{E_k^{(a)}}) \text{ for } k = 1 \dots K_a \tag{12}$$

Using $o_k^{(a)}$ and extending Eq 9 and 10, we compute the adversarial loss $L_{adv}^{(E_k^{(a)})}$ and debiasing loss $L_{deb}^{(a)}$ with respect to $E^{(a)}$ as follows:

$$L_{adv}^{(E_k^{(a)})} = -\sum_{i=1}^{N_{att}^{(a)}} \frac{1}{N_{att}^{(a)}} \log(o_{k,i}^{(a)}) \tag{13}$$

$$L_{deb}^{(a)} = \max\{L_{adv}^{(E_k^{(a)})}|_{k=1}^{K_a}\} \tag{14}$$

We compute the adversarial loss $L_{deb}^{(b)}$ with respect to $E^{(b)}$ in a similar way. Using weights λ_a for $L_{deb}^{(a)}$ and λ_b for $L_{deb}^{(b)}$, we compute the bias reducing classification as follows:

$$L_{br} = L_{class} + \lambda_a L_{deb}^{(a)} + \lambda_b L_{deb}^{(b)} \tag{15}$$

We provide the detailed MultiPASS algorithm in the supplementary material.

5. Experiments

5.1. Pre-trained networks and evaluation dataset

We evaluate the face descriptors obtained from the penultimate layer of following two pre-trained networks:

Arcface[1] : Resnet-101 trained on MS1MV2[2] with Additive Angular margin (Arcface) loss [15].

Crystalface : Resnet-101 trained on a mixture of UMDFaces[9], UMDFaces-Videos[8] and MS1M [23], with crystal loss [36].

The aforementioned Arcface [15] network achieves state-of-the-art performance in face verification and identification. Hence, we construct the baselines and our PASS framework on top of the Arcface descriptors, and provide detailed analysis for the same (in Sec. 5.4). To evaluate the generalizability of PASS and baselines, we also perform similar experiments with Crystalface [36] descriptors (in Sec. 5.4.4).

For evaluation, we use aligned faces from IJB-C, and follow the 1:1 face verification protocol defined in [31]. The alignment is done using [37]. This dataset provides gender (male/female) and skintone labels. There are six classes for the skintone attribute which we reorganize into three groups, (i) *Light* ('light pink' ∪ 'light yellow'), (ii) *Medium* ('medium pink' ∪ 'medium yellow'), (iii) *Dark* ('medium dark' ∪ 'dark brown'). For evaluating gender bias, we compute the verification performance of face descriptors for male-male and female-female pairs separately (out of all the pairs defined in the IJB-C protocol [31]). To compute skintone bias, we compute the verification performance of face descriptors for dark-dark and light-light pairs.

Using Arcface and Crystalface, we extract 512 dimensional descriptors for the aligned faces in the IJB-C dataset which are then used for gender-wise and skintone-wise verification, the plots for which are provided in Fig. 2.

5.2. PASS for gender and skintone

In Section 4, we present PASS as a general approach to de-bias face descriptors with respect to any attribute. Here, we show the effectiveness of PASS by using it to reduce information about gender and skintone (separately). We term the PASS framework trained to reduce gender information from descriptors as PASS-g, and its skintone counterpart as PASS-s. Additionally, we build another variant of PASS (called 'MultiPASS') to reduce the predictability of gender and skintone simultaneously. To train PASS-g, PASS-s and MultiPASS, we first need to extract f_{in} from a pre-trained face recognition network on a training dataset that consists of appropriate labels. f_{in} is extracted using the Arface network, described in Section 5.1.

PASS-g : For training PASS-g, we extract f_{in} for a combination of UMDFaces[9], UMDFaces-Videos[8] and MS1M[23]. There are 39,712 male and 18,308 female identities in the dataset. Face alignment and gender labels are obtained using [37]. For PASS-g, $N_{att} = 2$ (male/female).

PASS-s : To the best of our knowledge, we currently do not have a large dataset with skintone labels. So, we train PASS-s using f_{in} extracted for a dataset with race labels instead, since there is some correlation between race and skintone [34]. We use the BUPT-BalancedFace [45] for training PASS-s (aligned using [37]). The dataset consists of 1.3 million images for 28k identities. Each identity is associated with one of the four races : African, Asian, Indian and Caucasian. So, for PASS-s, $N_{att} = 4$.

MultiPASS: We design MultiPASS by combining the adversarial ensembles in PASS-s and PASS-g. MultiPASS is trained using the descriptors for BUPT-BalancedFace dataset, which consists of race labels. The gender labels for this dataset are predicted using [37].

After training PASS/MultiPASS, we feed the 512-dimensional descriptor f_{in} for test (IJB-C) images to the trained model M which generates 256-dimensional f_{out}. f_{out} is then used for face verification. Additional information on the hyperparameters required for training PASS is provided in the supplementary material, where we also analyze the effect of important hyperparameters on bias mitigation and verification performance. The code for implementing PASS will be made publicly available upon publication.

5.3. Baseline methods

5.3.1 Incremental Variable Elimination (IVE)

IVE [42] is an attribute suppression algorithm that excludes variables in the face representation that affect attribute classification. We build a two variants of IVE: IVE(g) and IVE(s). IVE(g) is trained to reduce gender information using Arcface descriptors descriptors from MS1M and gender labels predicted using [37]. Similarly, IVE(s) is trained to reduce skintone information using Arcface descriptors and labels from BUPT-BalancedFace [45]. Additional training details are provided in the supplementary material.

5.3.2 Obscuring hair - similar to [3]

It is shown in [3] that obscuring hair in facial images during evaluation helps to reduce gender bias by improving the similarity scores of genuine female-female pairs. We construct a similar pipeline for gender-bias mitigation. We compute the face border keypoints using [37] for the images in the evaluation dataset (IJB-C) and obscure all hair regions using these keypoints. Finally, we extract Arcface descriptors for these hair-obscured images. More details for [3] are provided in the supplementary material.

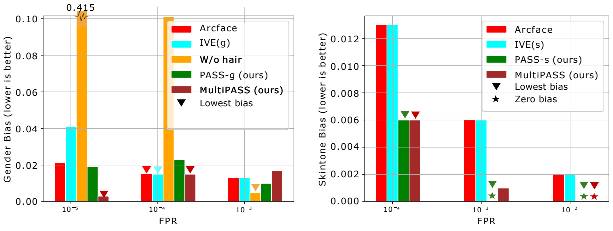
5.4. Results

5.4.1 Evaluating leakage of gender and skintone

To evaluate gender-leakage, we train an MLP classifier on Arcface descriptors and its de-biased counterparts (PASS variants/IVE). These descriptors are extracted for a training set with 60k images (30k males and females), sampled from IJB-C. The MLP classifier is a two hidden layer MLP with 128 and 64 hidden units respectively with SELU activations, followed by a sigmoid activated output layer. Subsequently, we test the MLP on descriptors extracted for 20k non-training images (10k males and females) in IJB-C. Finally, we compute the gender classification accuracy of the MLP. Using the same experimental setup with respect to skintone, we also train an MLP (with the same architecture) to predict skintone (dark/medium/light). In Tables 2 and 3, we find that for both gender and skintone, *the classification accuracy is lowest when the face descriptors are produced using MultiPASS*. We also find that classifiers trained on PASS-g and PASS-s descriptors obtain the second lowest classification accuracy. This indicates that PASS variants are capable of reducing gender and skintone information in face descriptors.

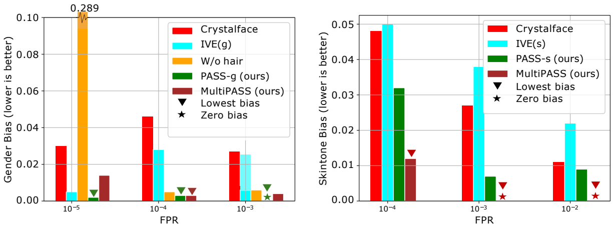
5.4.2 Evaluating bias

We provide the gender-wise and skintone-wise verification TPRs and the corresponding bias on IJB-C for all the methods in Tables 2 and 3 respectively. From Fig 6, we infer that Arcface descriptors transformed using PASS/MultiPASS obtain lowest gender/skintone bias at most FPRs. Moreover, from Tables 2 and 3 , we also infer that *PASS/MultiPASS-based frameworks obtain higher BPCs* (Eq 3) *than the baselines at most FPRs*. This shows that PASS variants are effective in reducing bias while maintaining high verification performance. We provide the gender-wise, skintone-wise ROC plots (similar to the ROC curves in Fig 2), along with overall verification plots in the supplementary material.



(a) (b)

Figure 6. (a) Gender and (b) skintone bias in Arcface descriptors and their de-biased counterparts on IJB-C.



(a) (b)

Figure 7. (a) Gender and (b) skintone bias in Crystalface descriptors and their de-biased counterparts on IJB-C.

5.4.3 End-to-end vs PASS

One subtlety when operating in an end-to-end fashion is that, in order to establish a baseline, one is generally required to retrain an entire face recognition system from scratch. Training such systems to achieve SOTA performance is technically challenging. Other works often report results using a weaker baseline system. For example, GAC [21] uses a ResNet50 version of Arcface that achieves lower overall performance in IJB-C, than the original ArcFace, as shown in Table 6. Alternatively, PASS operates on pre-trained models, allowing us to start with an existing SOTA model, and maintaining nearly SOTA performance.

5.4.4 PASS with Crystalface

To evaluate the generalizability of PASS and other baselines, we perform all of the aforementioned experiments on the Crystalface descriptors (mentioned in Sec. 5.1). We present the corresponding results of gender/skintone leakage in IJB-C in Tables 4 and 5. We find that *PASS and MultiPASS-transformed descriptors have the least gender/skintone predictability*. Similarly, *Crystalface descriptors transformed with PASS/MultiPASS obtain the lowest bias* (Fig. 7) and *highest BPC values on IJB-C* (as shown in Tables 4 and 5) *at **all** FPRs, for both gender and skintone*. The hyperparameter information and detailed results for all the methods are provided in the supplementary material.

5.5. OAT vs AET results

We train PASS-g systems with OAT and AET strategy on top of Arcface and Crystalface descriptors. We ensure that both OAT and AET approaches have the same number

FPR		10^{-5}					10^{-4}					10^{-3}				
Method	Acc-g (↓)	TPR_m	TPR_f	TPR	Bias(↓)	BPC_g(↑)	TPR_m	TPR_f	TPR	Bias(↓)	BPC_g(↑)	TPR_m	TPR_f	TPR	Bias(↓)	BPC_g(↑)
Arcface[15]	82.06	0.921	0.900	0.929	0.021	0.000	0.962	0.947	0.953	**0.015**	**0.000**	0.969	0.956	0.974	0.013	0.000
W/o hair[3]	80.77	0.418	0.833	0.616	0.415	-19.099	0.788	0.889	0.864	0.101	-5.827	0.933	0.928	0.925	**0.005**	**0.565**
IVE(g[42])	80.20	0.922	0.881	0.925	0.041	-0.957	0.962	0.947	0.950	**0.015**	-0.003	0.969	0.956	0.966	0.013	-0.008
PASS-g (ours)	73.65	0.900	0.881	0.919	0.019	0.084	0.948	0.925	0.946	0.023	-0.541	0.957	0.947	0.962	0.010	0.218
MultiPASS (ours)	**68.43**	0.871	0.874	0.881	**0.003**	**0.805**	0.934	0.919	0.934	**0.015**	-0.019	0.953	0.936	0.950	0.017	-0.332

Table 2. *Gender* bias analysis and accuracy ('Acc-g') of gender classifier for *Arcface* descriptors, and their transformed counterparts on IJB-C. TPR: overall True Positive rate, TPR_m: male-male TPR, TPR_f: female-female TPR. **Bold**=Best, Underlined=Second best

FPR		10^{-4}					10^{-3}					10^{-2}				
Method	Acc-st (↓)	TPR_l	TPR_d	TPR	Bias(↓)	BPC_st(↑)	TPR_l	TPR_d	TPR	Bias(↓)	BPC_st(↑)	TPR_l	TPR_d	TPR	Bias(↓)	BPC_st(↑)
Arcface [15]	87.15	0.951	0.938	0.953	0.013	0.000	0.974	0.968	0.974	0.006	0.000	0.976	0.974	0.976	0.002	0.000
IVE(s)[42]	88.23	0.951	0.938	0.953	0.013	0.000	0.973	0.967	0.974	0.006	0.000	0.976	0.974	0.976	0.002	0.000
PASS-s (ours)	83.86	0.925	0.919	0.934	**0.006**	**0.519**	0.949	0.949	0.950	**0.000**	**0.975**	0.974	0.974	0.973	**0.000**	**0.997**
MultiPASS (ours)	**79.22**	0.925	0.919	0.934	**0.006**	**0.519**	0.950	0.949	0.950	0.001	0.809	0.974	0.974	0.973	**0.000**	**0.997**

Table 3. *Skintone* bias analysis and accuracy ('Acc-st') of skintone classifier for *Arcface* descriptors, and their transformed counterparts on IJB-C. TPR: overall True Positive rate, TPR_l: light-light TPR, TPR_d: dark-dark TPR. **Bold**=Best, Underlined=Second best

FPR		10^{-5}		10^{-4}		10^{-3}	
Method	Acc-g(↓)	TPR	BPC_g (↑)	TPR	BPC_g(↑)	TPR	BPC_g(↑)
Crystalface[36]	86.73	0.833	0.000	0.910	0.000	0.951	0.000
W/o hair[3]	86.04	0.589	-8.926	0.780	0.823	0.899	0.731
IVE(g)[42]	86.10	0.833	0.833	0.910	0.391	0.951	0.071
PASS-g	80.54	0.761	**0.847**	0.839	**0.857**	0.910	**0.956**
MultiPASS	**76.31**	0.708	0.383	0.809	0.823	0.881	0.784

Table 4. *Gender* bias analysis and accuracy ('Acc-g') of gender classifier of *Crystalface* descriptors, and their transformed counterparts on IJB-C. **Bold**=Best, Underlined=Second best

FPR		10^{-4}		10^{-3}		10^{-2}	
Method	Acc-st (↓)	TPR	BPC_st(↑)	TPR	BPC_st(↑)	TPR	BPC_st(↑)
Crystalface[36]	89.30	0.910	0.000	0.951	0.000	0.974	0.000
IVE(s)[42]	88.26	0.910	-0.041	0.951	-0.407	0.974	-1.000
PASS-s	83.84	0.844	0.261	0.914	0.702	0.919	0.125
MultiPASS	**79.44**	0.809	**0.639**	0.881	**0.927**	0.968	**0.994**

Table 5. *Skintone* bias analysis and accuracy ('Acc-st') of skintone classifier for *Crystalface* descriptors, and their transformed counterparts in IJB-C. **Bold**=Best, Underlined=Second best

Method/FPR	10^{-5}	10^{-4}	10^{-3}	Training method	Training attributes
Arcface [15](SOTA)	92.9	95.3	97.4	-	-
Demo-ID$^+$ [20]	83.2	89.4	92.9	End-to-End	Age
Debface-ID$^+$ [20]	82.0	88.1	89.5	End-to-End	Age,gender,race
GAC$^+$ [21]	83.5	89.2	93.7	End-to-End	Race
PASS-s w/ AF	88.1	93.4	95.0	Descriptor-based	Race
PASS-g w/ AF	91.9	94.6	96.2	Descriptor-based	Gender
MultiPASS w/ AF	88.1	93.4	95.0	Descriptor-based	Race, gender

Table 6. IJB-C verification performance (TPR% @ given FPR). AF refers to *Arcface*.$^+$ = Numbers copied from original paper.

of classifiers ($K = 3$ for Arcface, and $K = 4$ for Crystalface) in ensemble E. We conduct the same gender-leakage experiment as done in Sec 5.4.1, and report the gender classification accuracy of the trained MLP in Table 7. For both Arcface and Crystalface, *MLP classifiers trained on descriptors from 'PASS-g (OAT)' obtain lower accuracy than their AET counterparts*. Moreover, in Table 7, we find that the *gender bias demonstrated by 'PASS-g (OAT)' is lower than that of PASS-g (AET) at most FPRs*. In fact, from Table 7, it is clear that AET frameworks hardly reduce gender bias.

FPR		10^{-5}			10^{-4}			10^{-3}		
Method	Acc-g	TPR_m	TPR_f	Bias	TPR_m	TPR_f	Bias	TPR_m	TPR_f	Bias
Arcface	82.06	0.921	0.900	0.021	0.962	0.947	**0.015**	0.969	0.956	0.013
AET	81.84	0.922	0.900	0.022	0.962	0.947	**0.015**	0.969	0.956	0.013
OAT	**73.65**	0.900	0.881	**0.019**	0.948	0.925	0.023	0.957	0.947	**0.010**
Crystlfce	86.73	0.836	0.806	0.030	0.913	0.867	0.046	0.952	0.924	0.028
AET	86.42	0.834	0.806	0.028	0.912	0.867	0.045	0.952	0.924	0.028
OAT	**80.54**	0.751	0.749	**0.002**	0.831	0.828	**0.003**	0.909	0.909	**0.000**

Table 7. Comparison of AET vs OAT strategies for gender bias reduction on Arcface (top) and Crystalface (bottom). Acc-g refers to gender classification accuracy (lower is better).

Therefore, we conclude that our novel discriminator training strategy - OAT is an important component of PASS, and effectively removes sensitive attributes in descriptors.

6. Conclusion

We present an adversarial approach called PASS that can reduce the information of any protected attribute in face descriptors, while making them proficient in identity classification. Our approach allows the user to re-use the precomputed descriptors for de-biasing them, without the need for expensive end-to-end training. In PASS, we also propose a novel discriminator training strategy called OAT to enforce removal of sensitive attributes and show that OAT is an important component of PASS. PASS can also be extended (as MultiPASS) to reduce the information of multiple attributes simultaneously.

Acknowledgement

The authors would like to thank Dr. P. Jonathon Phillips (NIST) and Dr. Rajeev Ranjan (Amazon) for their helpful suggestions. This research is based upon work supported by a MURI from the Army Research Office under the Grant No. W911NF-17-1-0304. This is part of the collaboration between US DOD, UK MOD and UK Engineering and Physical Research Council (EPSRC) under the Multidisciplinary University Research Initiative. Carlos D. Castillo was supported by funding provided by National Eye Institute Grant R01EY029692-03.

References

[1] Arcface pretrained resnet-101 model. https://www.dropbox.com/s/tj96fsm6t6rq8ye/model-r100-arcface-ms1m-refine-v2.zip?dl=0, 2018. 6

[2] Dataset. https://github.com/deepinsight/insightface/wiki/Dataset-Zoo, 2018. 6

[3] V Albiero and KW Bowyer. Is face recognition sexist? no, gendered hairstyles and biology are. *BMVC*, 2020. 2, 7, 8

[4] V Albiero, Krishnapriya KS, K Vangara, K Zhang, MC King, and KW Bowyer. Analysis of gender inequality in face recognition accuracy. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops*, pages 81–89, 2020. 2

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

[6] M Alvi, A Zisserman, and C Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2

[7] A Amini, AP Soleimany, W Schwarting, SN Bhatia, and D Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019. 2

[8] A Bansal, CD Castillo, R Ranjan, and R Chellappa. The do's and don'ts for CNN-based face verification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2545–2554, 2017. 6

[9] A Bansal, A Nanduri, C D Castillo, R Ranjan, and R Chellappa. Umdfaces: An annotated face dataset for training deep networks. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 464–473. IEEE, 2017. 6

[10] A Bansal, R Ranjan, C D Castillo, and R Chellappa. Deep features for recognizing disguised faces in the wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 10–106. IEEE, 2018. 1

[11] B Bortolato, M Ivanovska, P Rot, J Križaj, Philipp Terhörst, Naser Damer, Peter Peer, and Vitomir Štruc. Learning privacy-enhancing face representations through feature disentanglement. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 45–52. IEEE Computer Society, 2020. 2, 3

[12] J Buolamwini and T Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018. 1, 2

[13] J Choi, C Gao, JCE Messou, and JB Huang. Why Can't I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition. In *Advances in Neural Information Processing Systems*, pages 851–863, 2019. 2

[14] CM Cook, JJ Howard, YB Sirotin, and JL Tipton. Fixed and varying effects of demographic factors on the performance of eleven commercial facial recognition systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 40(1), 2019. 2

[15] J Deng, J Guo, X Niannan, and S Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1, 2, 3, 6, 8

[16] P Dhar, A Bansal, CD Castillo, J Gleason, PJ Phillips, and R Chellappa. How are attributes expressed in face dcnns? In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 85–92. IEEE, 2020. 1

[17] P Dhar, C Castillo, and R Chellappa. On measuring the iconicity of a face. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2137–2145. IEEE, 2019. 1, 3

[18] P Drozdowski, C Rathgeb, A Dantcheva, N Damer, and C Busch. Demographic bias in biometrics: A survey on an emerging challenge. *arXiv preprint arXiv:2003.02488*, 2020. 2

[19] O Gafni, L Wolf, and Y Taigman. Live face de-identification in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9378–9387, 2019. 2

[20] S Gong, X Liu, and AK Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *European Conference on Computer Vision*, pages 330–347. Springer, 2020. 2, 3, 8

[21] S Gong, X Liu, and AK Jain. Mitigating face recognition bias via group adaptive classifier. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Nashville, TN, June 2021. 2, 7, 8

[22] P. Grother et al. Face recognition vendor test (FRVT) part 3: Demographic effects. *NIST*, 2019. 1, 2

[23] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016. 6

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 3

[25] MQ Hill, CJ Parde, CD Castillo, YI Colon, R Ranjan, JC Chen, V Blanz, and AJ O'Toole. Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence*, 1(11):522–529, 2019. 1

[26] BF Klare, MJ Burge, JC Klontz, RWV Bruegge, and AK Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012. 2

[27] A Li, J Guo, H Yang, and Y Chen. Deepobfuscator: Adversarial training framework for privacy-preserving image classification. *arXiv preprint arXiv:1909.04126*, 2019. 2, 3

[28] W Liu, Y Wen, Z Yu, M Li, B Raj, and L Song. Sphereface: Deep hypersphere embedding for face recognition. *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[29] B Lu, JC Chen, CD Castillo, and R Chellappa. An experimental evaluation of covariates effects on unconstrained face verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):42–55, 2019. 2

[30] D Madras, E Creager, T Pitassi, and R Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018. 2

[31] B Maze, J Adams, J A Duncan, N Kalka, T Miller, C Otto, A K Jain, W T Niggel, J Anderson, J Cheney, et al. IARPA janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018. 6

[32] V Mirjalili, S Raschka, and A Ross. Gender privacy: An ensemble of semi adversarial networks for confounding arbitrary gender classifiers. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10. IEEE, 2018. 3

[33] V Mirjalili and A Ross. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In *2017 IEEE International joint conference on biometrics (IJCB)*, pages 564–573. IEEE, 2017. 3

[34] KJE Norwood. *Color matters: Skin tone bias and the myth of a postracial America.* Routledge/Taylor & Francis Group, 2014. 6

[35] N Quadrianto, V Sharmanska, and O Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8227–8236, 2019. 2

[36] R Ranjan, A Bansal, J Zheng, H Xu, J Gleason, B Lu, A Nanduri, J-C Chen, C D Castillo, and R Chellappa. A fast and accurate system for face detection, identification, and verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(2):82–96, 2019. 1, 2, 3, 6, 8

[37] R Ranjan, S Sankaranarayanan, C D Castillo, and R Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 17–24. IEEE, 2017. 6, 7

[38] JP Robinson, G Livitz, Y Henon, C Qin, Y Fu, and S Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–1, 2020. 2

[39] P Sattigeri, SC Hoffman, V Chenthamarakshan, and KR Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3:1–3:9, 2019. 3

[40] F Schroff, D Kalenichenko, and J Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 1

[41] Y Taigman, M Yang, M Ranzato, and L Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. 1

[42] P Terhörst, N Damer, F Kirchbuchner, and A Kuijper. Suppressing gender and age in face templates using incremental variable elimination. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019. 3, 7, 8

[43] P Terhörst, D Fährmann, N Damer, F Kirchbuchner, and A Kuijper. Beyond identity: What information is stored in biometric face templates? *arXiv preprint arXiv:2009.09918*, 2020. 1

[44] P. Terhörst et. al. Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations. *Applied Intelligence*, 49, 2019. 3

[45] M Wang and W Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9322–9331, 2020. 2, 6, 7

[46] M Wang, W Deng, J Hu, X Tao, and Y Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 692–702, 2019. 2

[47] T Wang, J Zhao, M Yatskar, KW Chang, and V Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5310–5319, 2019. 2

[48] Z Wu, Z Wang, Z Wang, and H Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 606–624, 2018. 2, 3, 4, 5

[49] BH Zhang, B Lemoine, and M Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. 2