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

Prithviraj Dhar*1, Joshua Gleason*2, Aniket Roy1, Carlos D. Castillo1, Rama Chellappa1
1Johns Hopkins University, 2University of Maryland, College Park
{pdharl,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.

*These authors have contributed equally to this work.

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_\text{X} (Step 1). Using these descriptors, the agent can train a gender classifier C_\text{X} (Step 2). Using the trained C_\text{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-
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 Protected Attribute Suppression System (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.

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
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.

**Measuring 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:

**Gender Bias (F):**

$$\text{Gender Bias}^{(F)} = |\text{TPR}_{\text{f}}^{(F)} - \text{TPR}_{\text{d}}^{(F)}|$$

**Skintone Bias (F):**

$$\text{Skintone Bias}^{(F)} = |\text{TPR}_{\text{f}}^{(F)} - \text{TPR}_{\text{d}}^{(F)}|$$

where \(\text{TPR}_{\text{f}}^{(F)}\) and \(\text{TPR}_{\text{d}}^{(F)}\) 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.

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

Here, \(\text{TPR}^{(F)}_{\text{deb}}, \text{Bias}^{(F)}_{\text{deb}}\) refer to the overall TPR obtained by original descriptors and the corresponding bias (Gender/Skintone bias) at FPR of \(F\). \(\text{TPR}^{(F)}_{\text{st}}, \text{Bias}^{(F)}_{\text{st}}\) 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 \(\text{Bias}^{(F)}_{\text{st}} = \text{Bias}^{(F)}_{\text{deb}}\) and \(\text{TPR}^{(F)}_{\text{st}} = \text{TPR}^{(F)}_{\text{deb}}\)). 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 skin tone as ‘BPC\(_g\)’ and that for gender as ‘BPC\(_f\)’. 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 \(f_{\text{in}}\) using a pre-trained network \(P\).

$$f_{\text{in}} = P(I)$$

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 \(f_{\text{in}}\) from a pre-trained network \(P\), and generates a lower dimensional descriptor \(f_{\text{out}} \in \mathbb{R}^{256}\). \(M\) consists of a single linear layer with 256 units, followed by a PReLU [24] layer. The weights of \(M\) are denoted as \(\phi_M\).

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

3. **Ensemble of attribute classifiers \(E\):** An ensemble of
which acts as a discriminator and tries to predict the protected attribute. The objective of \( E \) is because the input to each model in \( E \) is \( f_{\text{out}} \) which is fed to the ensemble \( E \) of attribute-agnostic descriptor models \( E_k \). This is because the input to \( E \) are low-dimensional descriptors \( f_{\text{out}} \) and not images. 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 \( L_{\text{class}} \). In stage 2, \( E \) is initialized and trained to classify attribute using gradients of \( L_{\text{att}} \). In stage 3, \( M \) and \( C \) are trained using the gradients of \( L_{\text{deb}} \) to debias \( f_{\text{out}} \) 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 \( f_{\text{out}} \) using the gradients of \( L_{\text{att}} \). 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 \( E_1, E_2 \ldots E_K \) that take \( f_{\text{out}} \) 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 \( N_{\text{att}} \) units. Here, \( N_{\text{att}} \) denotes the number of classes in the attribute being considered. We collectively denote the weights of all the models in \( E \) as \( \phi_E \) and weights of \( k \)th model \( E_k \) as \( \phi_{E_k} \). 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 \( f_{\text{out}} \) and not images.

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

### 4.1.1 Bias reducing classification loss \( L_{\text{br}} \)

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

\[
f_{\text{out}} = M(f_{\text{in}}, \phi_M)
\]  

(5)

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

\[
L_{\text{class}}(\phi_M, \phi_C) = -y_{\text{id}} \cdot \log(C(f_{\text{out}}, \phi_C))
\]  

(6)

\( y_{\text{id}} \) is a one hot identity label and classifier \( C \) produces softmaxed outputs.

**Training discriminators:** \( M \) generates \( f_{\text{out}} \) 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 \( L^{(E_k)}_{\text{att}} \) for attribute classification. \( L_{\text{att}} \) is computed as the sum of cross-entropy losses for each \( E_k \).

\[
L_{\text{att}}(\phi_M, \phi_E) = -\sum_{k=1}^{K} \sum_{i=1}^{N_{\text{att}}} y_{\text{att},i}^{(k)} \log y_{\text{att},i}^{(k)}
\]  

(7)

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

**Training generator (second constraint):** After training \( E \), \( M \) is trained to transform \( f_{\text{in}} \) into attribute-agnostic descriptor \( f_{\text{out}} \). We then provide \( f_{\text{out}} \) to each model in \( E_k \): \( o_k = E_k(f_{\text{out}}, \phi_{E_k}) \) for \( k = 1 \ldots K \) (8)

The outputs \( o_k \) are \( N_{\text{att}} \)-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 \( f_{\text{out}} \) were to always produce a posterior probability of \( \frac{1}{N_{\text{att}}} \) 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 \( L^{(E_k)}_{\text{adv}} \) for the \( k \)th model in \( E \) to be:

\[
L^{(E_k)}_{\text{adv}}(\phi_M, \phi_{E_k}) = -\sum_{i=1}^{N_{\text{att}}} \frac{1}{N_{\text{att}}} \log(o_{k,i})
\]  

(9)

Here, we use an ensemble of attribute prediction models, rather than a single model because, we want \( f_{\text{out}} \) to be constructed such that no model can predict the protected attribute. This approach was motivated by the work of [48] to solve the ‘\( \forall \) 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 \( L_{\text{deb}} \).

\[
L_{\text{deb}}(\phi_M, \phi_E) = \max \{ L^{(E_k)}_{\text{adv}}(\phi_M, \phi_{E_k}) \}_{k=1}^{K}
\]  

(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]. \( L_{\text{deb}} \) is then combined with \( L_{\text{class}} \) to compute a bias reducing classification loss \( L_{\text{br}} \).

\[
L_{\text{br}}(\phi_C, \phi_M, \phi_E) = L_{\text{class}}(\phi_C, \phi_M) + \lambda L_{\text{deb}}(\phi_M, \phi_E)
\]  

Here, \( \lambda \) 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
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 ‘One-At-a-Time (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 subspace \( A \). 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 \( j = i \mod K \), 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 \( E^{(a)} \).
We provide the detailed MultiPASS algorithm in the supplementary material.


5.1. Pre-trained networks and evaluation dataset

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


**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 Arcface 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 \( f_{in} \) to a pre-trained face recognition network on a training dataset that consists of appropriate labels. \( f_{in} \) 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.

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
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.

### 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

<table>
<thead>
<tr>
<th>Method</th>
<th>TPR$_m$</th>
<th>TPR$_f$</th>
<th>Bias$_m$</th>
<th>Bias$_f$</th>
<th>TPR$_m$</th>
<th>TPR$_f$</th>
<th>Bias$_m$</th>
<th>Bias$_f$</th>
<th>TPR$_m$</th>
<th>TPR$_f$</th>
<th>Bias$_m$</th>
<th>Bias$_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arcface[15]</td>
<td>72.43</td>
<td>72.43</td>
<td>0.017</td>
<td>0.017</td>
<td>0.974</td>
<td>0.974</td>
<td>0.006</td>
<td>0.006</td>
<td>0.976</td>
<td>0.976</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>W/o hair[3]</td>
<td>71.36</td>
<td>71.36</td>
<td>0.012</td>
<td>0.012</td>
<td>0.975</td>
<td>0.975</td>
<td>0.008</td>
<td>0.008</td>
<td>0.976</td>
<td>0.976</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>IVE[42]</td>
<td>70.54</td>
<td>70.54</td>
<td>0.015</td>
<td>0.015</td>
<td>0.976</td>
<td>0.976</td>
<td>0.005</td>
<td>0.005</td>
<td>0.978</td>
<td>0.978</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>PASS-g (ours)</td>
<td>79.22</td>
<td>79.22</td>
<td>0.006</td>
<td>0.006</td>
<td>0.951</td>
<td>0.951</td>
<td>0.000</td>
<td>0.000</td>
<td>0.974</td>
<td>0.974</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>MultiPASS (ours)</td>
<td>81.21</td>
<td>81.21</td>
<td>0.005</td>
<td>0.005</td>
<td>0.976</td>
<td>0.976</td>
<td>0.002</td>
<td>0.002</td>
<td>0.974</td>
<td>0.974</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>

### 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

<table>
<thead>
<tr>
<th>Method</th>
<th>TPR$_l$</th>
<th>TPR$_d$</th>
<th>Bias$_l$</th>
<th>Bias$_d$</th>
<th>TPR$_l$</th>
<th>TPR$_d$</th>
<th>Bias$_l$</th>
<th>Bias$_d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arcface[36]</td>
<td>86.73</td>
<td>86.73</td>
<td>0.833</td>
<td>0.833</td>
<td>0.951</td>
<td>0.951</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>W/o hair[3]</td>
<td>86.04</td>
<td>86.04</td>
<td>0.589</td>
<td>0.589</td>
<td>0.901</td>
<td>0.901</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>IVE[42]</td>
<td>88.33</td>
<td>88.33</td>
<td>0.833</td>
<td>0.833</td>
<td>0.991</td>
<td>0.991</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>PASS-g</td>
<td>80.54</td>
<td>80.54</td>
<td>0.786</td>
<td>0.786</td>
<td>0.901</td>
<td>0.901</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>MultiPASS</td>
<td>79.21</td>
<td>79.21</td>
<td>0.708</td>
<td>0.708</td>
<td>0.910</td>
<td>0.910</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

### 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

<table>
<thead>
<tr>
<th>Method</th>
<th>TPR$_m$</th>
<th>TPR$_f$</th>
<th>Bias$_m$</th>
<th>Bias$_f$</th>
<th>TPR$_m$</th>
<th>TPR$_f$</th>
<th>Bias$_m$</th>
<th>Bias$_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crystalface[36]</td>
<td>89.30</td>
<td>89.30</td>
<td>0.910</td>
<td>0.910</td>
<td>0.974</td>
<td>0.974</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>IVE[42]</td>
<td>88.26</td>
<td>88.26</td>
<td>0.910</td>
<td>0.910</td>
<td>0.974</td>
<td>0.974</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>PASS-g</td>
<td>83.84</td>
<td>83.84</td>
<td>0.914</td>
<td>0.914</td>
<td>0.919</td>
<td>0.919</td>
<td>0.125</td>
<td>0.125</td>
</tr>
<tr>
<td>MultiPASS</td>
<td>79.44</td>
<td>79.44</td>
<td>0.890</td>
<td>0.890</td>
<td>0.968</td>
<td>0.968</td>
<td>0.994</td>
<td>0.994</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method/PRR</th>
<th>TPR$_l$</th>
<th>TPR$_d$</th>
<th>Bias$_l$</th>
<th>Bias$_d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arcface[37][50][51][52]</td>
<td>92.5</td>
<td>92.5</td>
<td>95.3</td>
<td>95.3</td>
</tr>
<tr>
<td>Demo-ID[20]</td>
<td>83.2</td>
<td>83.2</td>
<td>94.9</td>
<td>94.9</td>
</tr>
<tr>
<td>Debrief-ID[20]</td>
<td>82.0</td>
<td>82.0</td>
<td>89.5</td>
<td>89.5</td>
</tr>
<tr>
<td>GAC[21]</td>
<td>83.5</td>
<td>83.5</td>
<td>92.7</td>
<td>92.7</td>
</tr>
<tr>
<td>PASS-g w/ AF</td>
<td>88.1</td>
<td>88.1</td>
<td>93.4</td>
<td>93.4</td>
</tr>
<tr>
<td>MultiPASS w/ AF</td>
<td>88.1</td>
<td>88.1</td>
<td>93.4</td>
<td>93.4</td>
</tr>
</tbody>
</table>

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

### 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 pre-computed 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 ensure 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

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


