

# Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning

Ignacio Serna<sup>a,\*</sup>, Aythami Morales<sup>a</sup>, Julian Fierrez<sup>a</sup>, Nick Obradovich<sup>b</sup>

<sup>a</sup> School of Engineering, Universidad Autonoma de Madrid, Spain

<sup>b</sup> Center for Humans & Machines, Max Planck Institute for Human Development, Berlin, Germany

## ARTICLE INFO

### Article history:

Received 8 October 2020

Received in revised form 27 January 2022

Accepted 8 February 2022

Available online 14 February 2022

### Keywords:

Machine behavior

Bias

Fairness

Discrimination

Machine learning

Learning representations

Face

Biometrics

## ABSTRACT

We propose a discrimination-aware learning method to improve both the accuracy and fairness of biased face recognition algorithms. The most popular face recognition benchmarks assume a distribution of subjects without paying much attention to their demographic attributes. In this work, we perform a comprehensive discrimination-aware experimentation of deep learning-based face recognition. We also propose a notational framework for algorithmic discrimination with application to face biometrics. The experiments include three popular face recognition models and three public databases composed of 64,000 identities from different demographic groups characterized by sex and ethnicity. We experimentally show that learning processes based on the most used face databases have led to popular pre-trained deep face models that present evidence of strong algorithmic discrimination. Finally, we propose a discrimination-aware learning method, Sensitive Loss, based on the popular triplet loss function and a sensitive triplet generator. Our approach works as an add-on to pre-trained networks and is used to improve their performance in terms of average accuracy and fairness. The method shows results comparable to state-of-the-art de-biasing networks and represents a step forward to prevent discriminatory automatic systems.

© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

Artificial Intelligence (AI) is developed to meet human needs that can be represented in the form of objectives. To this end, the most popular machine learning algorithms are designed to minimize a loss function that defines the cost of wrong solutions over a pool of samples. This is a simple but very successful scheme that has enhanced the performance of AI in many fields, such as Computer Vision, Speech Technologies, and Natural Language Processing. But this optimization of specific computable objectives may not lead to the behavior one may expect or desire from AI. International agencies, academia, and industry are alerting policymakers and the public to the unforeseen effects and behaviors of AI agents, not initially considered during the design phases [1]. In this context, aspects such as trustworthiness and fairness should be included as learning objectives and not taken for granted. (See Fig. 1.)

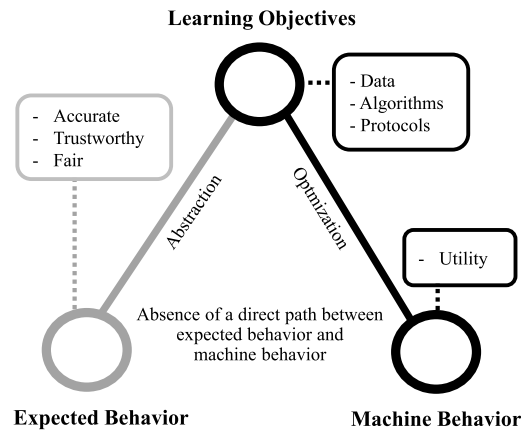
Machine vision in general and face recognition algorithms, in particular, are good examples of recent advances in AI [3–6]. The performance of automatic face recognition has been boosted during the last decade, achieving very competitive

\* Corresponding author.

E-mail addresses: [ignacio.serna@uam.es](mailto:ignacio.serna@uam.es) (I. Serna), [aythami.morales@uam.es](mailto:aythami.morales@uam.es) (A. Morales), [julian.fierrez@uam.es](mailto:julian.fierrez@uam.es) (J. Fierrez), [obradovich@mpib-berlin.mpg.de](mailto:obradovich@mpib-berlin.mpg.de) (N. Obradovich).

<https://doi.org/10.1016/j.artint.2022.103682>

0004-3702/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



**Fig. 1.** The objective of the learning process is an abstraction of the expected behavior of an AI. There is usually no direct path between the machine expected behavior and the machine behavior, which is normally evaluated in terms of its utility. Learning objectives are usually determined by factors such as task, data, algorithms, and experimental protocols, losing sight of key aspects of the expected behavior such as fairness. Figure inspired by the standard model proposed in [2].

accuracies in the most challenging scenarios [7]. These improvements have been made possible due to advances in machine learning (e.g., deep learning), powerful computation (e.g., GPUs), and larger databases (e.g., on a scale of millions of images). However, recognition accuracy is not the only aspect to be considered when designing biometric systems. There is currently a growing need to study AI behavior in order to better understand its impact on our society [1]. Face recognition systems are especially sensitive due to the personal information present in face images (e.g., identity, sex, ethnicity, and age). The number of published works pointing out the potential discriminatory effects in the results of face detection and recognition algorithms is large [8–17].

In this environment, only a limited number of works analyze how biases affect the learning process of algorithms dealing with personal information [18,19]. There is a lack of understanding regarding how demographic information affects popular and widely used pre-trained AI models beyond their performance.

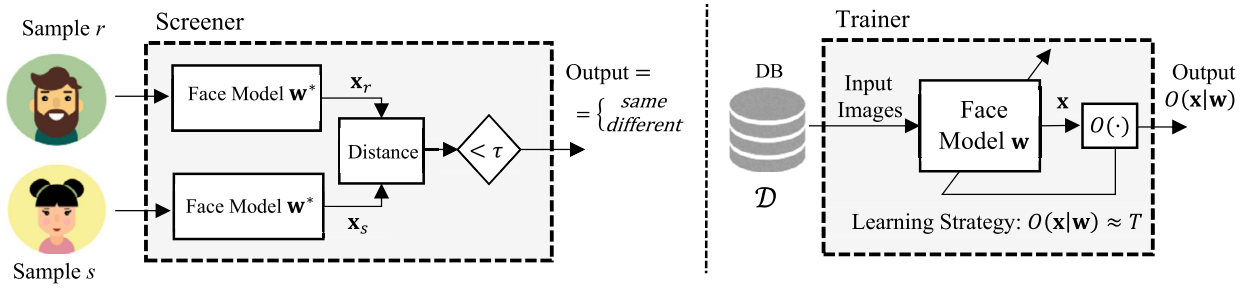
On the other hand, the right to non-discrimination is deeply rooted in the normative framework that underlies various national and international regulations, and can be found, for example, in Article 7 of the Universal Declaration of Human Rights and Article 14 of the European Convention on Human Rights, among others. As evidence of these concerns, the European Parliament passed the General Data Protection Regulation (GDPR)<sup>1</sup> in April 2018, a set of laws aimed at regulating the collection, storage, and use of personal information. According to paragraph 71 of GDPR, controllers of sensitive data processing, have to “implement appropriate technical and organizational measures” that “prevent, inter alia, discriminatory effects”.

The aim of this work is to analyze face recognition models using a discrimination-aware perspective and to demonstrate that learning processes involving such a discrimination-aware perspective can be used to train more accurate and fairer algorithms. The main contributions of this work are:

- A comprehensive analysis of the causes and effects of biased learning processes, including: (i) discrimination-aware performance analysis based on three public datasets, with 64K identities equally distributed across demographic groups; (ii) study of deep representations and the role of sensitive attributes such as sex and ethnicity; (iii) complete analysis of demographic diversity present in some of the most popular face databases, and analysis of new databases available to train models based on diversity.
- Based on our analysis of the causes and effects of biased learning algorithms, we propose an efficient discrimination-aware learning method to mitigate bias in deep face recognition models: Sensitive Loss. The method is based on the inclusion of demographic information in the popular triplet loss representation learning. Sensitive Loss incorporates fairness as a learning objective in the training process of the algorithm. The method works as an add-on that is applied over pre-trained representations to improve their performance and fairness without requiring a complete re-training. We evaluated the method in three public databases, showing an improvement in both overall accuracy and fairness. Our results show how to incorporate discrimination-aware learning rules to significantly reduce bias in deep learning models.

Preliminary work in this research line was presented in [20]. Key improvements here over [20] include: (i) in-depth analysis of the state-of-the-art, including an extensive survey of face recognition databases; (ii) inclusion of two new datasets in

<sup>1</sup> EU 2016/679 (General Data Protection Regulation). Available online at: <https://gdpr-info.eu/>.



**Fig. 2.** Face recognition block diagrams. The screener is an algorithm that given two face images decides if they belong to the same person. The trainer is an algorithm that generates the best data representation for the screener.

the experiments involving 40,000 new identities and more than 1M images; and (iii) a novel discrimination-aware learning method called Sensitive Loss.

The rest of the paper is structured as follows: Section 2 summarizes the related work. Section 3 presents our general formulation of algorithmic discrimination. Section 4 presents the proposed discrimination-aware learning method. Section 5 describes the evaluation procedure. Section 6 presents the experimental results. Finally, Section 7 summarizes the main conclusions.

## 2. Related work

### 2.1. Face recognition: methods

A face recognition algorithm, like other machine learning systems, can be divided into two different algorithms: screener and trainer. Both algorithms are used for different purposes. [21].

The screener takes the characteristics of an individual and returns a prediction of that individual's outcome, while the trainer produces the screener itself. In our case, the screener (see Fig. 2) is an algorithm that, given two face images, generates an output associated with the probability that they belong to the same person. This probability is obtained by comparing the two learned representations from a face model defined by the parameters  $\mathbf{w}$ . These parameters are previously trained from a given dataset  $\mathcal{D}$  (see Fig. 2). If properly trained, the output of the trainer would be a model with parameters  $\mathbf{w}^*$ , capable of representing the input data (e.g., face images) in a highly discriminant feature space  $\mathbf{x}$ .

The most popular architecture used to model face attributes is the Convolutional Neural Network (CNN) [22]. The pre-trained models are used as an embedding extractor where  $\mathbf{x}$  is a  $l_2$ -normalised learned representation of a face image. The similarity between two face descriptors  $\mathbf{x}_r$  and  $\mathbf{x}_s$  is calculated as the Euclidean distance  $\|\mathbf{x}_r - \mathbf{x}_s\|$ . Two faces are assigned the same identity if their distance is smaller than a threshold  $\tau$ . The recognition accuracy is obtained by comparing distances between positive matches (i.e.,  $\mathbf{x}_r$  and  $\mathbf{x}_s$  belong to the same person) and negative matches (i.e.,  $\mathbf{x}_r$  and  $\mathbf{x}_s$  belong to different persons).

### 2.2. Bias in face databases

Following the trainer-screener division, bias is rooted in the trainer. The trainer is a common algorithm that usually varies in the loss function [23], the optimization algorithm, and in the data it uses for training. Bias is traditionally associated with the unequal representation of classes in a dataset. The history of automatic face recognition has been linked to the history of the databases used for algorithm training during the last two decades. The number of publicly available databases is high, and they allow the training of models using millions of face images.

Table 1 summarizes the demographic statistics of some of the most frequently cited face databases. In order to obtain demographic statistics, sex and ethnicity classification algorithms were trained based on a ResNet-50 model [24] and 12K identities from the DiveFace database (equally distributed between the six demographic groups). The models were evaluated in 20K labeled images of Celeb-A with a performance greater than 97%.

Each of these databases is characterized by its own biases (e.g., image quality, pose, backgrounds, and aging). In this work, we highlight the unequal representation of demographic information in very popular face recognition databases. As can be seen, the differences between ethnic groups are serious. Even though the people of Asia constitute more than 35% of the world's population, they account for only 9% of the content of these popular face recognition databases.

Biased databases imply a double penalty for underrepresented classes. On the one hand, models are trained according to non-representative diversity. On the other hand, accuracy is measured in privileged classes and overestimates the real performance in a diverse society.

Recently, diverse and discrimination-aware databases have been proposed in [13,15,18,25–27]. These databases are valuable resources for exploring how diversity can be used to improve face biometrics. However, some of these databases do not

**Table 1**

Demographic statistics of state-of-the-art face databases (ordered by number of images). In order to obtain demographic statistics, sex and ethnicity classification algorithms were trained based on a ResNet-50 model [24] and 12K identities of DiveFace database (equally distributed between the six demographic groups). Models were evaluated in 20K labeled images of Celeb-A with performance over 97%. The table includes the averaged demographic statistics for the most popular face databases in the literature.

Dataset [ref]	# images	# identities	# avg. images per identity	Caucasian		African/Indian		Asian	
				Male	Female	Male	Female	Male	Female
FRVT2018 [28]	27M	12M	2	48.4%	16.5%	19.9%	7.4%	1.2%	0.4%
MSCeleb1M [29]	8.5M	100K	85	52.4%	19.2%	12.1%	3.9%	7.7%	4.5%
MegaFace [30]	4.7M	660K	7	40.0%	30.3%	6.2%	4.7%	10.6%	8.1%
VGGFace2 [31]	3.3M	9K	370	45.9%	30.2%	10.5%	6.3%	3.4%	3.6%
VGGFace [32]	2.6M	2.6K	1K	43.7%	38.6%	5.8%	6.9%	2.1%	2.9%
YouTube [33]	621K	1.6K	390	56.9%	20.3%	7.7%	4.0%	7.9%	3.0%
CASIA [34]	500K	10.5K	48	48.8%	33.2%	7.2%	5.7%	2.6%	2.6%
CelebA [35]	203K	10.2K	20	33.9%	41.5%	6.4%	8.2%	4.4%	5.5%
PubFig [36]	58K	200	294	49.5%	35.5%	6.5%	5.5%	2.0%	1.0%
IJB-C [37]	21K	3.5K	6	40.3%	30.2%	11.8%	6.0%	5.4%	6.2%
UTKface [38]	24K	-	-	26.2%	20.0%	21.5%	16.3%	7.1%	8.9%
LFW [39]	13K	5.7K	2	58.9%	18.7%	9.6%	3.3%	7.2%	2.2%
BioSecure [40]	2.7K	667	4	50.1%	36%	3.1%	2.1%	4.3%	4.5%
Average				46%	29%	10%	6%	5%	4%
Databases for discrimination-aware learning									
BUPT-B [18]	1.3M	28K	46	33.33%		33.33%		33.33%	
DiveFace [41]	125K	24K	5	16.7%	16.7%	16.7%	16.7%	16.7%	16.7%
FairFace [27]	100K	-	-	25.0%	20.0%	14.4%	13.9%	13.6%	13.1%
RFW [25]	40K	12K	3	33.33%		33.33%		33.33%	
DemogPairs [15]	10.8K	600	18	16.7%	16.7%	16.7%	16.7%	16.7%	16.7%

include identities [13,26,27], and face images cannot be matched to other images. Therefore, these databases do not allow for adequate training and testing of face recognition algorithms.

Note that the groups into which the databases have been divided are heterogeneous, and they include people of different ethnicities. We are aware of the limitations of grouping all human ethnic origins into only three categories. According to studies, there are more than 5,000 ethnic groups in the world. Our experiments are similar to those reported in the literature, and include only three groups in order to maximize differences between classes. Automatic classification algorithms based on these reduced categories show performances of up to 98% accuracy [10].

### 2.3. Bias in face recognition

Facial recognition systems can suffer from a variety of biases, ranging from those arising from unconstrained environmental variables such as illumination, pose, expression, and face resolution, from systematic errors such as image quality, and from demographic factors like age, sex, and race [9].

An FBI-coauthored study [12] tested three commercial algorithms of supplier companies to various public organizations in the US. In all three algorithms, African Americans were less likely to be successfully identified —i.e., more likely to be falsely rejected— than other demographic groups. A similar decline surfaced for females compared to males and younger subjects compared to older subjects.

More recently, the latest NIST evaluation of commercial face recognition technology, the Face Recognition Vendor Test (FRVT) Ongoing, shows that at sensitivity thresholds that resulted in white men being falsely matched once in 1K, out of a list of 167 algorithms, all but two were more than twice as likely to misidentify black women, some reaching 40 times more [28]. The number of academic studies analyzing the fairness of face recognition algorithms has grown in recent years [16].

There are other published studies analyzing face recognition performance over demographic groups, but [12] and [28] are probably the most systematic, comprehensive, thorough, and up-to-date.

### 2.4. De-biasing face recognition

Originally, statistical sampling methods have been used to tackle unbalanced dataset bias in face recognition [42,43]. Recently, other attempts to eliminate bias in facial recognition are emerging.

Some semi-supervised approaches try to use unlabeled faces to reduce racial bias, but their performance falls short of supervised ones [25,44]. Others aim to remove potential sources of variation from the learned representations. This is the case presented in [14]: a method to learn a primary classification task (which can be sex recognition) whilst unlearning other spurious variables that represent undesirable sources of bias (such as age, ancestral origin, or pose). On the other hand, Das et al. proposed a Multi-Task CNN that also managed to improve performance across subgroups of sex, race, and

age [45]. These two methods were intended for classification tasks, not face recognition. Finally, Morales et al. developed an extension of the triplet loss function to remove sensitive information in feature embeddings for face recognition [41].

In [18], researchers proposed a race-balanced reinforcement learning network to find appropriate margins losses for different demographic groups. Their model significantly reduced the performance difference between demographic groups. However, to generate an adaptive margin policy to train the convolutional network, their approach requires two ancillary networks: an offline sampling network and a deep Q-learning network. Similar to [18], [46] proposed a new race-adaptive margin loss function based on a multi-task face recognition network with an auxiliary task of race classification.

Gong et al. presented in [19] an adversarial network that learns a representation of disentangled features of sex, age, race, and face recognition, minimizing their correlation to de-bias both face recognition and demographic attribute estimation. These same authors have then devised a group-adaptive classifier. Presented in [47], the new classifier focuses its attention on extracting the features that best discriminate from each demographic group. To do that, they use adaptive convolution kernels and attention mechanisms. On top of that, they introduced a new objective function to reduce the variation of the average intra-class distance between demographic groups.

The aforementioned methods [18,19,46,47] were applied to train de-biasing deep architectures for face recognition from scratch. They consist of complex architectures that may require significant work to exploit or combine with other pre-existing networks or knowledge, an important limitation that we alleviate with our proposed Sensitive Loss approach.

### 3. Problem statement

There is no formal, agreed-upon definition of algorithmic discrimination in the scientific literature. Most approaches rely on the inequality of outcome of an automatic system for a given sensitive attribute (e.g., sex or ethnicity). Discrimination is defined by the Cambridge Dictionary as treating a person or particular group of people differently, especially in a worse way than the way in which you treat other people, because of their skin color, sex, sexuality, etc.

For the purpose of studying discrimination in face recognition systems and machine learning at large, we now present our re-formulation of *Algorithmic Discrimination* based on the above dictionary definition. Even though ideas similar to those included in our formulation can be found elsewhere [48,49], we didn't find this kind of formulation in related works. We hope that the formalization of these concepts can be beneficial in fostering further research and discussion on this topic.

Let's begin with notation and preliminary definitions. Assume  $\mathbf{x}_s^i$  is a learned representation of individual  $i$  (out of  $I$  different individuals) corresponding to an input image  $\mathbf{I}_s^i$  ( $s = 1, \dots, S$  samples per individual). That representation  $\mathbf{x}$  is assumed to be useful for task  $T$ , e.g., face authentication or emotion recognition. That representation  $\mathbf{x}$  is generated from the input image  $\mathbf{I}$  using an artificial intelligence approach with parameters  $\mathbf{w}$ . We also assume that there is a goodness criterion  $G$  in that task that maximizes some real-valued performance function  $f$  over a given dataset  $\mathcal{D}$  (collection of multiple images), in the form:

$$G(\mathcal{D}) = \max_{\mathbf{w}} f(\mathcal{D}, \mathbf{w}) \quad (1)$$

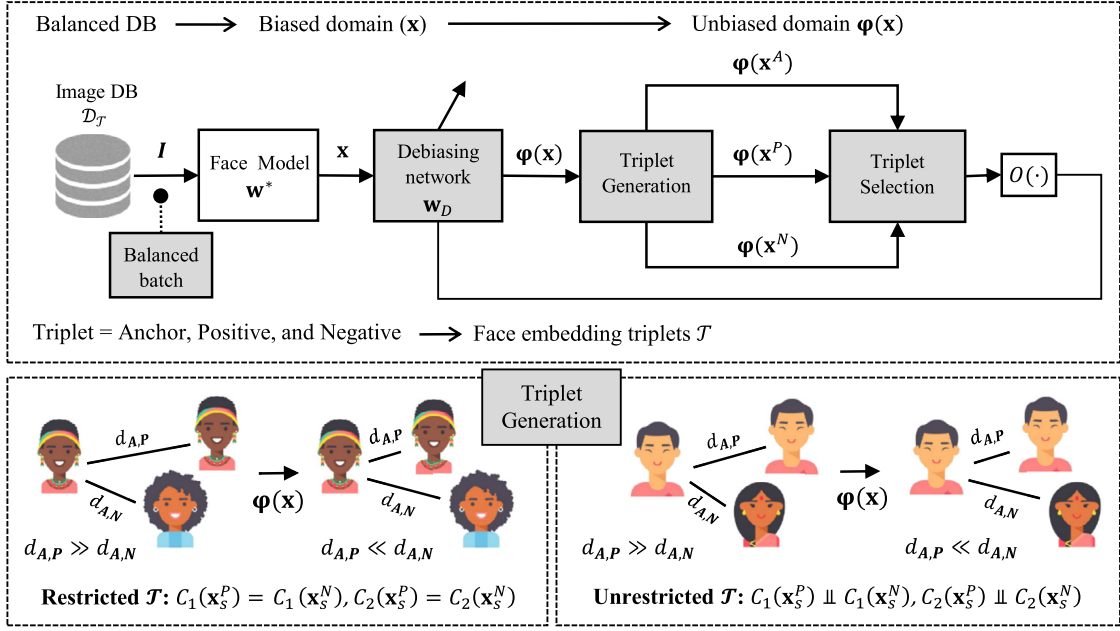
The most popular form of the previous expression minimizes a loss function  $\mathcal{L}$  over a set of training images  $\mathcal{D}$  in the form:

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \sum_{\mathbf{I}_s^i \in \mathcal{D}} \mathcal{L}(O(\mathbf{I}_s^i | \mathbf{w}), T_s^i) \quad (2)$$

where  $O$  is the output of the learning algorithm that we seek to bring closer to the target function (or groundtruth)  $T$  defined by the task at hand. On the other hand, an individual  $i$  can be classified according to a demographic criterion  $d$ , e.g., *Sex*, whose classes are  $C_{Sex} = \{Male, Female\}$ . This criterion  $d$  can be a source of discrimination. The particular class  $k$  for a given demographic criterion  $d$  and a given sample is noted as  $C_d(\mathbf{x}_s^i)$ , e.g.,  $C_{Sex}(\mathbf{x}_s^i) = Male$ . We say that all classes by criterion  $d$  are well represented in dataset  $\mathcal{D}$ , when the number of samples for each class is significant.  $\mathcal{D}_d^k \in \mathcal{D}$  represents all the samples corresponding to class  $k$  of demographic criterion  $d$ .

Finally, in our experimental framework, Algorithmic Discrimination is defined as: a significantly larger difference when performing task  $T$  (e.g., face verification) between the goodness  $G$  upon considering the full set of data  $\mathcal{D}$  (including multiple samples from multiple individuals) and the goodness  $G(\mathcal{D}_d^k)$  in the subset of data corresponding to class  $k$  (e.g., *Female*) of the demographic criterion  $d$ .

The representation  $\mathbf{x}$  and the model parameters  $\mathbf{w}$  will typically be real-valued vectors, but they can be any set of features combining real and discrete values. Note that the above formulation can easily be extended to the case of a variable number of samples  $S_i$  for different subjects, which is the common case; or to classes  $K$  that are not disjoint. Note also that the previous formulation is based on average performances over groups of individuals. In many artificial intelligence tasks it is common to have different performance between specific individuals due to various reasons, e.g., specific users who were not sensed properly [50], even in the case of algorithms that, on average, may have similar performance for the different classes that are the source of discrimination. Therefore, in our formulation and definition of Algorithmic Discrimination we opted to use average performances in demographic groups.



**Fig. 3.** (Top) Block diagram of the domain-adaptive learning process that allows us to generate an unbiased representation  $\varphi(\mathbf{x})$  from a biased representation  $\mathbf{x}$ . A Balanced Dataset  $\mathcal{D}_T$  is preferable as input for Sensitive Loss to select the triplets  $\mathcal{T}$ . This  $\mathcal{D}_T$  can be different or a subset of the (generally Unbalanced) Dataset  $\mathcal{D}$  used to train the biased model  $\mathbf{w}^*$  that appears in Eq. (1). (Bottom) Discrimination-aware generation of triplets given an under-represented (unfavored) demographic group: the representation  $\varphi(\mathbf{x})$  increases the distance  $d$  between Anchor and Negative samples while reducing the distance between Anchor and Positive, thus seeking to improve the performance of the unfavored group.  $\perp$  stands for independence.

Other related works, for example [51,52], are beginning to investigate discrimination effects in AI with user-specific methods, but still lack a mathematical framework with clear definitions of User-specific Algorithmic Discrimination (U-AD), in comparison to our defined Group-based Algorithmic Discrimination (G-AD). We will study and augment our framework with an analysis of U-AD in future work.

#### 4. Proposed discrimination-aware learning approach: sensitive loss

Models trained and evaluated over privileged demographic groups may fail to generalize when the model is evaluated over groups other than the privileged one. This is a behavior caused by the wrong assumption of homogeneity in the facial characteristics of the world population. In this work we propose to reduce the bias in face recognition models by incorporating a discrimination-aware learning process.

The methods proposed in this work to reduce bias are based on two strategies:

- Use of balanced and heterogeneous data to train and evaluate the models. The literature shows that training with a balanced dataset does not guarantee bias-free results [12,18,25] but can partially reduce such bias.
- A modified loss function (Sensitive Loss) that incorporates demographic information to guide the learning process into a more inclusive feature space. The development of new cost functions capable of incorporating discrimination-aware elements into the training process is another way to reduce bias. Our approach is based on the popular triplet loss function and it can be applied to pre-trained models without needing full re-training of the network.

##### 4.1. Discrimination-aware learning into triplet loss

Triplet loss was proposed as a distance metric in the context of nearest neighbor classification [53] and adapted to improve the performance of face descriptors in verification algorithms [54,32]. In this work, we propose to incorporate demographic data to generate discrimination-aware triplets to train a new representation that mitigates biased learning.

Assume that an image is represented by an embedding descriptor  $\mathbf{x}_s^i$  obtained by a pre-trained model (see Section 3 for notation). That image corresponds to the demographic group  $C_d(\mathbf{x}_s^i)$ . A triplet is composed of three different images of two different people: Anchor (A) and Positive (P) are different images of the same person, and Negative (N) is an image of a different person. The Anchor and Positive share the same demographic labels,  $C_d(\mathbf{x}_s^A) = C_d(\mathbf{x}_s^P)$  but these labels may differ for the Negative sample  $C_d(\mathbf{x}_s^N)$ . The transformation  $\varphi(\mathbf{x})$  represented by parameters  $\mathbf{w}_D$  (D for De-biasing) is trained to minimize the loss function:



$$\min_{\mathbf{w}_D} \sum_{\mathbf{x}_s \in \mathcal{T}} (||\boldsymbol{\varphi}(\mathbf{x}_s^A) - \boldsymbol{\varphi}(\mathbf{x}_s^N)||^2 - ||\boldsymbol{\varphi}(\mathbf{x}_s^A) - \boldsymbol{\varphi}(\mathbf{x}_s^P)||^2 + \Delta) \quad (3)$$

where  $|| \cdot ||$  is the Euclidean Distance,  $\Delta$  is a margin between genuine and impostor distances, and  $\mathcal{T}$  is a set of triplets generated by an online sensitive triplet generator that guides the learning process (see details in Section 4.2). The effects of biased training include a representation that fails to model properly the distance between faces of different people ( $||\mathbf{x}_s^A - \mathbf{x}_s^N||$ ) belonging to the same minority demographic groups (e.g.,  $C_d(\mathbf{x}_s^A) = C_d(\mathbf{x}_s^N) = \text{Asian Female}$ ). The proposed triplet loss function considers both genuine and impostor comparisons and also allows for the introduction of demographic-aware information. In order to guide the learning process in that discrimination-aware spirit, demographic groups with the worst performing triplets are prioritized in the online sensitive triplet generator (e.g., for *Asian Females*). Fig. 3 shows the block diagram of the learning algorithm.

---

**Algorithm 1:** Sensitive Triplet Generation and Selection.

---

```

Input: Training data batch  $\mathcal{B} = \{\boldsymbol{\varphi}(\mathbf{x}_s^i)\}_{i=1,\dots,B; s=1,\dots,S_i}$ 
Output: Resulting Sensitive Loss  $\mathcal{L}$ 
 $\mathcal{L}: \emptyset$ 
if Restricted then
  for  $k$  in all demographic classes do
    /*  $\mathcal{R}$  : all the possible triplets within the demographic group */
     $\mathcal{R} : \{(\boldsymbol{\varphi}(\mathbf{x}_n^i), \boldsymbol{\varphi}(\mathbf{x}_m^i), \boldsymbol{\varphi}(\mathbf{x}_l^j)) \text{ such that } C_{\text{demographic}}(\mathbf{x}_n^i) = C_{\text{demographic}}(\mathbf{x}_m^i) = C_{\text{demographic}}(\mathbf{x}_l^j) = k, \forall i, j \in B \forall n, m \in S_i \forall l \in S_j \text{ where } i \neq j \text{ and } n \neq m\}$ 
     $\mathcal{L} \leftarrow \text{add}(\text{Triplet Selection}(\mathcal{R}))$ 
  end
else
  /*  $\mathcal{U}$  : all the possible triplets in the batch */
   $\mathcal{U} : \{(\boldsymbol{\varphi}(\mathbf{x}_n^i), \boldsymbol{\varphi}(\mathbf{x}_m^i), \boldsymbol{\varphi}(\mathbf{x}_l^j)) \mid \forall i, j \in B \forall n, m \in S_i \forall l \in S_j \text{ where } i \neq j \text{ and } n \neq m\}$ 
   $\mathcal{L} \leftarrow \text{add}(\text{Triplet Selection}(\mathcal{U}))$ 
end
function Triplet Selection( $\mathcal{T}$ ):
   $\mathcal{L}: \emptyset$ 
  foreach triplet  $\mathbf{t}_s \in \mathcal{T}$  where  $\mathbf{t}_s = (\boldsymbol{\varphi}(\mathbf{x}_s^A), \boldsymbol{\varphi}(\mathbf{x}_s^P), \boldsymbol{\varphi}(\mathbf{x}_s^N))$  do
    if  $0 < ||\boldsymbol{\varphi}(\mathbf{x}_s^A) - \boldsymbol{\varphi}(\mathbf{x}_s^N)||^2 - ||\boldsymbol{\varphi}(\mathbf{x}_s^A) - \boldsymbol{\varphi}(\mathbf{x}_s^P)||^2 < \Delta$  then
       $\mathcal{L} \leftarrow \text{add} (||\boldsymbol{\varphi}(\mathbf{x}_s^A) - \boldsymbol{\varphi}(\mathbf{x}_s^P)||^2 - ||\boldsymbol{\varphi}(\mathbf{x}_s^A) - \boldsymbol{\varphi}(\mathbf{x}_s^N)||^2 + \Delta)$ 
    end
  end
  return  $\mathcal{L}$ 
end

```

---

#### 4.2. Sensitive loss: sensitive triplets

Inspired by the semi-hard selection proposed in [32,54], we propose an online selection of triplets that automatically prioritizes, at each learning step, the triplets of those demographic groups where the algorithm performs less well (see Fig. 3). Our approach results in a supervised learning framework in which the loss function is minimized assuming a heterogeneous population (i.e., divided into demographic groups). On the one hand, triplets within the same demographic group improve the ability to discriminate between samples with similar anthropometric characteristics (e.g., reducing the false acceptance rate in *Asian Females*). On the other hand, heterogeneous triplets (i.e., triplets involving different demographic groups) improve the generalization capacity of the model (i.e., the overall accuracy).

Each batch is generated with images of different identities evenly distributed among the different demographic groups and with the same images per identity. In our experiments, we used 300 identities (50 per group) and 3 images per identity in each batch. For efficiency, we form the triplets after passing the images through the network when computing the loss function.

In triplet formation, we can distinguish between generation and selection of triplets (see Algorithm 1):

- **Triplet Generation:** This is where all possible triplets are formed and joined within a training batch. We evaluated two types of triplet generation (see Fig. 3):
  - **Unrestricted (U):** The generator allows triplets with mixed demographic groups (i.e.,  $C_d(\mathbf{x}_s^A) = C_d(\mathbf{x}_s^N)$  or  $C_d(\mathbf{x}_s^A) \neq C_d(\mathbf{x}_s^N)$ ). Thus, with 300 identities and 3 images per identity, around 135K triplets are generated (from which the semi-hard ones will be selected).
  - **Restricted (R):** The generator does not allow triplets with mixed demographic groups (i.e.,  $C_d(\mathbf{x}_s^P) = C_d(\mathbf{x}_s^N)$ ). Thus, with 300 identities from 6 groups and 3 images per identity, more than 22K triplets are generated (from which the semi-hard ones will be selected).

**Table 2**

Computational load of inference for the three face models and also for our method, Sensitive Loss (SL), applied to each of them. The computational load has been measured in number of Giga-FLOPs (GFLOPs): floating point operations.

	ResNet-50	SL-ResNet-50	VGG-Face	SL-VGG-Face	ArcFace	SL-ArcFace
Nº of GFLOPs	3.8	0.0042	15.5	0.0168	24.2	0.0002

- **Triplet Selection:** Among all the triplets generated within a batch, the triplet selection chooses those for which:  $0 < ||\varphi(\mathbf{x}_s^A) - \varphi(\mathbf{x}_s^N)||^2 - ||\varphi(\mathbf{x}_s^A) - \varphi(\mathbf{x}_s^P)||^2 < \Delta$ . These are semi-hard triplets that are critical for adequate convergence and avoiding bad local minima [54]. If a demographic group is not well modeled by the network (both in terms of genuine or impostor comparisons), more triplets from this group are likely to be included. This selection is purely guided by performance over each demographic group and could change for each batch depending on model deficiencies.

We chose triplet loss as the basis for Sensitive Loss because it allows us to incorporate demographic-aware learning in a natural way. The process is data-driven and does not require a large number of images per identity (e.g., while softmax requires a large number of samples per identity, we only use 3 images per identity). Another advantage is that it is not necessary to train the entire network, and triplet loss can be applied as a domain adaptation technique. In our case, we trained the model to move from a biased domain  $\mathbf{x}$  to an unbiased domain  $\varphi(\mathbf{x})$ .

Our results demonstrate that biased representations  $\mathbf{x}$  that exhibit clear performance differences contain the information necessary to reduce such differences. In other words, bias can be at least partially corrected from representations obtained from pre-trained networks, and new models trained from scratch are not necessary. Similar strategies might be applied to other loss functions.

## 5. Evaluation procedure

### 5.1. Databases for discrimination-aware learning

*DiveFace* [41] contains annotations equally distributed among six classes related to sex and ethnicity. There are 24K identities (4K per class) and 3 images per identity, for a total number of images equal to 72K. Users are grouped according to their sex (male or female) and three categories related to ethnic physical characteristics: *Caucasian*: people with ancestral origins in Europe, North-America, and Latin-America (with European origin). *African/Indian*: people with ancestral origins in Sub-Saharan Africa, India, Bangladesh, Bhutan, and others. *Asian*: people with ancestral origin in Japan, China, Korea, and other countries in that region.

*Racial Faces in the Wild (RFW)* [25] is divided into four demographic classes: Caucasian, Indian, Asian, and African. Each class has about 10K images of 3K individuals. There are no major differences in pose, age, and sex distribution between Caucasian, Asian, and Indian groups. The African group has a smaller age difference than the others, and while females account for approximately 35% in the other groups, they account for less than 10% in the African group.

*BUPT-Balancedface (BUPT-B)* [18] contains 1.3M images from 28K celebrities obtained from MS-Celeb-1M [29]. Divided into 4 demographic groups, it is roughly balanced by race, with 7K subjects per race: Caucasian, Indian, Asian, and African; with 326K, 325K, 275K, and 324K images, respectively. No sex data is available for this dataset.

### 5.2. Face recognition models

*VGG-Face* [32]: Model with 138M parameters based on the VGG-Very-Deep-16 CNN traditional architecture. We used a pre-trained model<sup>2</sup> trained with the VGGFace2 dataset according to the details provided in [31]. The VGG models were developed by the Visual Geometry Group (VGG) at the University of Oxford for face recognition and demonstrated on benchmark computer vision datasets [32].

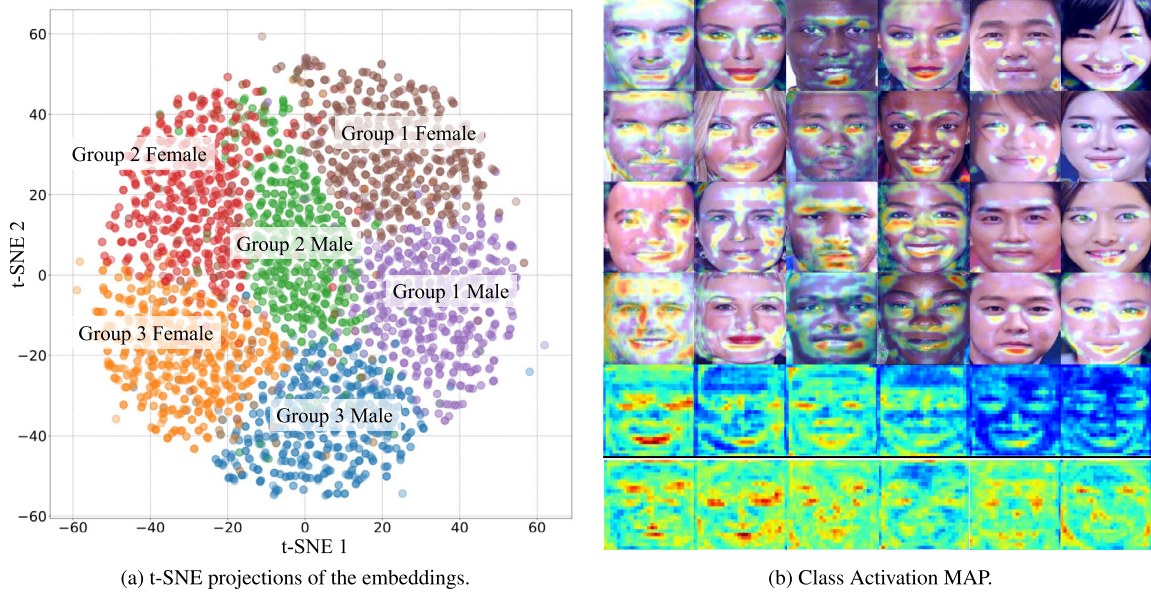
*ResNet-50* [24]: ResNet-50 is a CNN model with 25M parameters initially proposed for general-purpose image recognition tasks [24]. It combines convolutional neural networks with residual connections to allow information to skip layers and improve gradient flow. These models have been tested in competitive evaluations and public benchmarks [32,31]. The model<sup>2</sup> we used was trained with the VGGFace2 dataset.

*ArcFace* [55]: With a ResNet architecture and 64M parameters, ArcFace obtains state-of-the-art results on multiple datasets (e.g., 99.80% accuracy on LFW [39]). We used a publicly available<sup>3</sup> pre-trained ArcFace model trained on MS-Celeb-1M [29].

<sup>2</sup> Available on <https://github.com/rcmalli/keras-vggface>.

<sup>3</sup> <https://github.com/deepinsight/insightface>.





**Fig. 4.** (a) Projections of the ResNet-50 embeddings into the 2D space generated with t-SNE. (b) Examples of the six classes available in the DiveFace database (different columns). Rows 5 and 6 show the averaged Class Activation MAP (first filter of the third convolutional block of ResNet-50) with and without our method obtained from 20 random face images from each of the classes. Rows 1–4 show Class Activation MAPs for each of the face images. Maximum and minimum activations are represented by red and blue colors respectively. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

### 5.3. Implementation details

The proposed de-biasing method, Sensitive Loss, does not require retraining the entire pre-trained model (see Fig. 3). The sensitive triplets are used to train a dense layer with the following characteristics: number of units equal to the size of the pre-trained representation  $\mathbf{x}$  (4,096, 2,048 and 512 units for VGG-Face, ResNet-50 and ArcFace respectively), dropout (of 0.5 for VGG-Face and Resnet-50 and 0.05 for ArcFace), linear activation, random initialization, and  $L_2$  normalization. This layer, relatively easy to train (10 epochs and Adam optimizer), will be used to generate the new representation  $\phi(\mathbf{x})$ .

Table 2 shows the FLOPs (floating point operations) for the three models evaluated and for their Sensitive Loss layer. Sensitive Loss depends on the size of the pre-trained representation  $\mathbf{x}$ , so for each model it is different. Nevertheless, the computational load of inference added to the face models is three orders of magnitude less in all cases.

Experiments are conducted with  $k$ -fold cross-validation across users and three images per identity (therefore 3 genuine and  $3 \times (\text{users} - 1)$  impostor combinations per identity), with five folds. Thus, the three databases are divided into a training set (80%) and a test set (20%) in every fold. This results in a total of 192K genuine comparisons (DiveFace = 72K, RFW = 36K, and BUPT = 84K) and 98M impostor comparisons (DiveFace = 28.7M, RFW 10.8M, and BUPT = 58.7M).

Note that ArcFace is only evaluated on DiveFace, since that model was trained with MS1M [29], which overlaps with RFW and BUPT-B datasets (both databases obtained from it). Before applying the face recognition models, we cropped the face images using the algorithms proposed in [56,57].

## 6. Experiments

### 6.1. Demographic bias in learned representations

We applied a popular data visualization algorithm to better understand the importance of ethnic features in the embedding space generated by deep models. t-SNE is an algorithm to visualize high-dimensional data. This algorithm minimizes the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data.

Fig. 4a shows the projection of each face into a 2D space generated from ResNet-50 embeddings and the t-SNE algorithm. This t-SNE projection is unsupervised and its inputs are just the face embeddings without any labels. After running t-SNE, we have colored each projected point according to its ethnic attribute. As we can see, the consequent face representation results in three clusters highly correlated with the ethnicity attributes. Note that ResNet-50 has been trained for face recognition, not ethnicity detection. However, information on sex and ethnicity is highly embedded in the feature space, and the unsupervised t-SNE algorithm reveals the presence of this information.

**Table 3**

Face verification Performance (Equal Error Rate EER in %) on the face datasets described in Section 5.1 for matchers VGG-Face, ResNet-50 and ArcFace without and with our de-biasing Sensitive Loss module (U = Unrestricted Triplet Generation; R = Restricted Triplet Generation). Also shown: Average EER across demographic groups and Standard deviation (lower means fairer).

Model	DiveFace							
	Caucasian		Indian/African		Asian		Avg	Std
	Male	Female	Male	Female	Male	Female		
VGG-Face	1.62	1.76	2.06	2.33	2.53	3.15	<b>2.24</b>	<b>0.51</b>
VGG-Face-U	1.84	1.98	1.63	1.77	1.38	1.44	1.67 (↓25%)	0.21 (↓58%)
VGG-Face-R	1.80	1.97	1.65	1.77	1.42	1.42	1.67 (↓25%)	0.20 (↓61%)
ResNet-50	0.63	0.73	0.88	1.41	0.99	1.26	<b>0.98</b>	<b>0.28</b>
ResNet-50-U	0.84	0.90	0.74	1.21	0.58	0.60	0.81 (↓17%)	0.21 (↓24%)
ResNet-50-R	0.90	0.93	0.78	1.22	0.61	0.62	0.84 (↓14%)	0.21 (↓25%)
ArcFace	0.79	0.85	1.11	1.98	1.34	1.27	<b>1.22</b>	<b>0.39</b>
ArcFace-U	0.71	0.67	1.08	1.79	1.24	1.17	1.11 (↓9%)	0.37 (↓5%)
ArcFace-R	0.69	0.65	0.96	1.88	1.22	1.19	1.10 (↓10%)	0.41 (↑5%)

Model	RFW						Std
	Caucasian	Indian	African	Asian	Avg		
VGG-Face	8.22	10.38	17.24	13.67	<b>12.38</b>		<b>3.41</b>
VGG-Face-U	7.34	7.78	13.09	9.47	9.42 (↓24%)		2.27 (↓34%)
VGG-Face-R	7.26	7.75	12.79	9.05	9.21 (↓26%)		2.17 (↓36%)
ResNet-50	3.62	4.72	5.75	5.96	<b>5.01</b>		<b>0.93</b>
ResNet-50-U	3.02	3.29	3.99	3.83	3.53 (↓30%)		0.40 (↓58%)
ResNet-50-R	3.02	3.22	4.06	3.92	3.56 (↓29%)		0.44 (↓53%)

Model	BUPT-Balancedface						Std
	Caucasian	Indian	African	Asian	Avg		
VGG-Face	7.18	7.44	9.78	12.56	<b>9.24</b>		<b>2.17</b>
VGG-Face-U	6.49	4.97	7.73	8.29	6.87 (↓26%)		1.28 (↓41%)
VGG-Face-R	6.48	5.03	7.68	8.20	6.85 (↓26%)		1.22 (↓44%)
ResNet-50	3.24	2.65	3.80	5.56	<b>3.82</b>		<b>1.09</b>
ResNet-50-U	2.62	1.69	2.72	3.19	2.56 (↓33%)		0.54 (↓50%)
ResNet-50-R	2.62	1.72	2.77	3.12	2.56 (↓32%)		0.52 (↓52%)

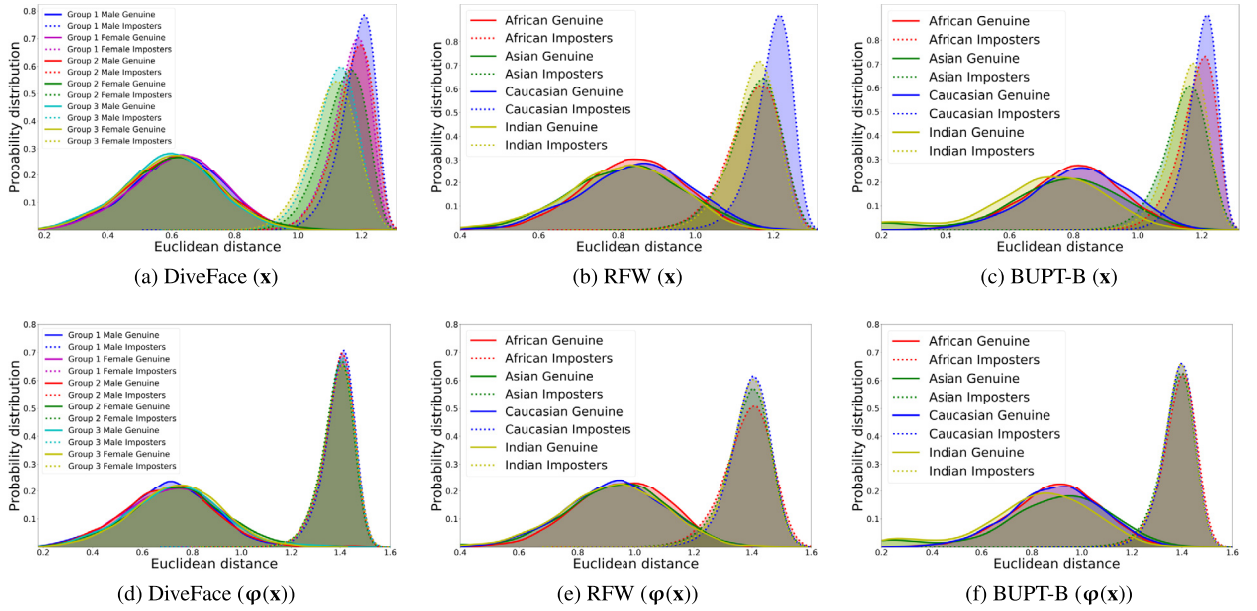
We have also used the t-SNE algorithm on the embeddings of our method, and the result is the same as in Fig. 4a. This means that ethnic information is still present. We do not propose to remove racial information. This information is key to recognition, and it is not within our goals to remove it. In fact, removing it decreases performance, as shown in [19].

On a different front, CNNs are composed of a large number of stacked filters. These filters are trained to extract the richest information for a pre-defined task (e.g., face recognition). Since face recognition models are trained to identify individuals, it is reasonable to think that the response of the models may vary slightly from one person to another. In order to visualize the response of the model to different faces, we consider the specific Class Activation MAP (CAM) proposed in [58], named Grad-CAM. This visualization technique uses the gradients of a target flowing into the selected convolutional layer to produce a coarse localization map. The resulting heatmap highlights the activated regions in the image for the selected target (e.g., an individual identity in our case).

Fig. 4b represents the heatmaps obtained in the first filter of the third convolutional block of ResNet-50 for faces from the six demographic groups included in DiveFace. Each column corresponds to a demographic group. The first rows contain face images with their overlaid heatmap. The last two rows represent the heatmaps obtained in the same ResNet-50 filter without and with our method after averaging results from 120 different individuals. For a better visualization the 120 images chosen are all frontal. We only averaged a small group of individuals because if we did it with the whole dataset nothing would be seen, since the images vary widely in pose and morphology.

The activation maps show clear differences between ethnic groups, with the highest activation for Caucasians and the lowest for Asians. These differences suggest that features extracted by the model are, at least partially, affected by the ethnic attributes. However, with our method (last row), the activations are more homogeneous across demographic groups. This homogeneous activation suggests a better representation across the different ethnic groups. Recent work has shown that there is a correlation between high activations and performance in CNN's architectures [59,60]. The activation maps obtained with the VGG-Face and ArcFace models are similar to those of ResNet-50.

These two experiments illustrate the presence and importance of ethnic attributes in the feature space generated by face deep models.



**Fig. 5.** ResNet-50 face verification distance score distributions for all DiveFace, RFW and BUPT-B demographic groups using the original representation  $\mathbf{x}$  (top) and the proposed representation  $\boldsymbol{\varphi}(\mathbf{x})$  (bottom). Note how the proposed Sensitive Loss representation  $\boldsymbol{\varphi}(\mathbf{x})$  reduces the gap between impostor distributions (dotted lines) across demographic groups. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

## 6.2. Performance of sensitive loss

Table 3 shows the performance (Equal Error Rate EER in %) for each demographic group, as well as the average EER on the DiveFace, RFW, and BUPT test sets for the baseline models (VGG-Face, ResNet-50, and ArcFace), and the Sensitive Loss methods described in Section 4.1 (Unrestricted and Restricted). In order to measure fairness, Table 3 includes the standard deviation of the EER across demographic groups (Std). These measures were proposed in [18,19] to analyze the performance of de-biasing algorithms.

If we focus our attention on the results obtained by the Baseline systems (denoted as VGG-Face, ResNet-50, and ArcFace), we see different performances in each database. This is caused by the particular characteristics of each database (e.g., age or pose distributions of the samples).

The results reported in Table 3 clearly show that sex and ethnicity significantly affect the performance of biased models. These effects are particularly high for ethnicity, with a very large degradation in performance for the class less represented in the training data. In DiveFace, the relative increment of the Equal Error Rate (EER) is 94%, 124%, and 150% for VGG-Face, ResNet-50, and ArcFace, respectively, with regard to the best class (*Caucasian Male*). For RFW and BUPT-Balancedface, the differences between demographic groups are similar but not so large, because the distinction between the demographic groups is only of ethnic origin and not of sex.

These differences are important as they mark the percentage of faces successfully matched and faces incorrectly matched for a certain threshold. Let us consider the performance function  $f$  described in Eq. (1) as the accuracy of the face recognition model, and recall that  $G(\mathcal{D}_d^k) = f(\mathcal{D}_d^k, \mathbf{w}^*)$  is the goodness for an algorithm  $\mathbf{w}^*$  trained on the entire set of data  $\mathcal{D}$  (see Eq. (2)) considering to compute the goodness of all samples corresponding to class  $k$  of the demographic criterion  $d$ . The findings indicate significant differences in the goodness  $G(\mathcal{D}_d^k)$  for different classes  $k$  of demographic criterion  $d$ , especially between classes  $k = \text{Caucasian}$  and *Asian*, with a difference of 2% in DiveFace, 10% in RFW, and 6% in BUPT-Balancedface.

Concerning the triplet generation method (Unrestricted or Restricted, see Section 4.2), both methods show competitive performances with similar improvements over the baseline approaches. The higher number of triplets generated by the Unrestricted method (about 6 times more) does not show clear improvements compared to the Restricted method. We can see that the biggest improvements are achieved for VGG-Face, and that ArcFace barely improves (in fact, ArcFace-R worsens its standard deviation by 5%). This is due to the fact that ArcFace's performance is already highly optimized, so the room for improvement in VGG-Face and ResNet-50 is much greater. The size of the embedding obtained from VGG-Face is eight times larger than that of ArcFace.

One would think that ResNet-50 is not part of the state of the art. Yet, in Table 3, you can see that ResNet-50 has a better average accuracy (lower average EER) than ArcFace in the DiveFace database. Normally, performance evaluations are done on unbalanced datasets (see Table 1), so they don't show a full picture of their performance. For example, a model that does not perform that well in the *Asian Female* demographic group, if evaluated on a test set that barely contains samples from this group, will see little or no effect on its overall performance and will appear to be a good model.

**Table 4**

EER (in %) comparison with SOTA de-biasing approaches tested on RFW dataset for face verification, and trained with BUPT-Balancedface database. The results are directly copied from their work. In parentheses we show the relative improvement with respect to the baseline approach used in each work. NA = Not Available. \*Our Sensitive Loss method is applied over the feature embedding space of a ResNet-50 model pre-trained with VGGFace2 [31].

Method [ref]	Model	Caucasian	Indian	African	Asian	Avg	Std
RL-RBN [18]	ResNet-34	3.73	5.32	5.00	5.18	4.81 (10%)	0.63 (34%)
DebFace [19]	ResNet-50	4.05	5.22	6.33	5.67	5.32 (NA)	0.83 (NA)
GAC [47]	ResNet-50	3.80	5.02	5.23	5.13	4.79 (11%)	0.58 (39%)
RamFace [46]	ResNet-50	2.60	3.42	3.75	4.50	3.57 (NA)	0.68 (NA)
<b>Sensitive Loss [Ours]</b>	ResNet-50*	2.77	3.05	4.18	3.50	3.37 (33%)	0.54 (42%)

The relatively low performance in some groups seems to be originated by a limited ability to capture the best discriminant features for underrepresented samples in the training databases. ResNet-50 seems to learn better discriminant features as it performs better than VGG-Face. Additionally, ResNet-50 shows a smaller difference between demographic groups. The results suggest that features capable of reaching high accuracy for a specific demographic group may be less competitive in others.

Let's now analyze the causes behind this degradation. Fig. 5 represents the probability distributions of genuine and impostor distance scores for all demographic groups. A comparison between genuine and impostor distributions reveals large differences for impostors. The genuine distribution (intra-class variability) between groups is similar, but the impostor distribution (inter-class variability) is significantly different.

Fig. 5 shows the score distributions obtained for the ResNet-50 model without and with our Sensitive Loss de-biasing method (with Unrestricted sensitive triplet generation). Table 3 showed the performance for a specific decision threshold (at the EER) for face verification. Now Fig. 5 provides all the information (the indicators commonly used, such as EER, FMR, FNMR, etc., are obtained by setting a decision threshold in these distributions). The improvements in Accuracy and Fairness caused by our Sensitive Loss discrimination-aware representation  $\phi(\mathbf{x})$  come mainly from a better alignment of impostor score distributions across demographic groups. To a large extent, the proposed Sensitive Loss learning method was able to correct the biased behavior of the baseline model.

The results obtained by Sensitive Loss outperform the baseline approaches by:

- i) Improving fairness (Std). The standard deviation of performance across all demographic groups is lower. Fairness improvements in terms of EER Std vary by model and database, ranging from 5% to 61% relative improvements, with an average improvement of 44%.
- ii) Reducing the Average EER in each of the three databases (see Table 3). The results show that discrimination-aware learning not only helps to generate fairer, but also more accurate representations. Our Sensitive Loss discrimination-aware learning yields better representations for specific demographic groups and collectively for all groups.

The discrimination-aware learning method proposed in this work, Sensitive Loss, is a step forward to prevent discriminatory effects in the usage of automatic face recognition systems. The representation  $\phi(\mathbf{x})$  reduces the discriminatory effects of the original representation  $\mathbf{x}$ , since the differences between goodness criteria  $G(\mathcal{D}_d^k)$  across demographic groups are reduced. However, differences still exist and should be considered in the deployment of these technologies.

#### 6.2.1. Comparison with the state of the art

Table 4 shows the comparison of our approach with four recent state-of-the-art de-biasing techniques [18,19,46,47]. These four methods consist of full networks trained specifically to avoid bias, whereas what we propose here with Sensitive Loss is not an entire network, but rather an add-on method to reduce the biased outcome of a given network.

The results of this comparison should be interpreted with care, because the arrangements are different. Still, the comparison gives us a rough idea of the range of bias mitigation in the three methods.

The four approaches we compared ourselves with have trained their networks with only one database: BUPT-Balancedface. We have instead taken a network already trained with VGGFace2 and added a layer that we have trained with BUPT-Balancedface. Our network may have an advantage because one part has been trained with VGGFace2 and the other with BUPT-Balancedface, and therefore the average performance is better. However, we are not looking to improve performance, but to reduce discrimination, and with our experiments we want to demonstrate that complex models are not always needed.

DebFace, RL-RBN, GAC, or RamFace cannot be compared to our ArcFace-based method because the RFW database is included in MS1M (ArcFace training data set). In fact, both the EER and Std of ArcFace in RFW are up to 10 times lower than those achieved with these. That is why we used the ResNet-50 network for the comparison.

From Table 4, it can be seen that in terms of fairness (measured as performance differences among demographic groups), our approach is at least comparable to that of dedicated networks trained from scratch to produce unbiased models. With such similar behavior from a fairness perspective, our proposed Sensitive Loss is superior to the compared methods in terms of simplicity and applicability, as it can be directly applied to already trained networks without the need for complete retraining.



## 7. Summary and conclusions

We have presented a comprehensive analysis of face recognition models based on deep learning according to a new discrimination-aware perspective. We started by presenting a new general re-formulation of Algorithmic Discrimination with application to face recognition. Next, we showed the high bias introduced by training deep models with the most popular face databases employed in the literature. Our analysis of some of the most popular face databases in the literature revealed a large gap between the number of samples from different ethnic groups. Classes are unevenly represented in the most popular face databases. New databases and benchmarks are needed to train more diverse and heterogeneous algorithms. Evaluation over representative populations from different demographic groups is important to prevent discriminatory effects.

We also looked at the interior of the tested models, revealing different activation patterns of the networks for different demographic groups. This corroborates the biased nature of these popular pre-trained face models. Popular deep models trained on databases biased towards certain classes (according to criterion  $d$ ), result in feature spaces with strong differentiation between those classes. This differentiation affects the obtained representation  $\mathbf{x}$  and enables its use for tasks other than those it was trained for.

We then evaluated three popular pre-trained face models (VGG-Face, ResNet-50, and ArcFace), according to the proposed formulation. The experiments were carried out on three public databases (DiveFace, RFW, and BUPT-B) comprising 64,000 identities and 1.5M images. The results showed that the two tested face models are highly biased across demographic groups. In particular, we observed large performance differences in face recognition across sex and ethnic groups. These performance gaps reached up to 200% of relative error degradation between the best class and the worst. This means that false positives are 200% more likely for some demographic groups than for others when using the popular face models evaluated in this work.

After the bias analysis, we proposed a novel discrimination-aware training method, Sensitive Loss, based on a triplet loss function and online selection of sensitive triplets. Unlike existing related de-biasing methods, Sensitive Loss works as an add-on to pre-trained networks, facilitating its application to problems (like face recognition) where hard-worked models with excellent performance exist, but where little attention to fairness aspects was paid at their inception. Experiments with Sensitive Loss demonstrate how simple discrimination-aware rules can guide the learning process towards fairer and more accurate representations. The results of the proposed Sensitive Loss representation outperform the baseline models for the three evaluated databases, both in terms of average accuracy and fairness metrics. These results encourage the training of more diverse models and the development of methods capable of dealing with the inherent differences in demographic groups.

The framework analyzed in this work is focused on the analysis of Group-based Algorithmic Discrimination (G-AD). Future work will investigate how to incorporate User-specific [61] Algorithmic Discrimination (U-AD) into the proposed framework [51]. Additionally, the analysis of other covariates such as facial expression [62,63] or age will be included in the study. Discrimination by age is an important concern in applications such as automatic recruitment tools [64]. Other future directions include the study of new methods to detect bias in the training process with low input information [65] or the application of privacy-preserving techniques [66,67]. We hope this line of research in algorithmic discrimination enable a future of more transparent and explainable AI [68].

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

The authors would like to thank Manuel Cebrian and Iyad Rahwan for their constructive feedback and inspiring talks. This work has been supported by projects: TRESPASS-ETN (MSCA-ITN-2019-860813), PRIMA (MSCA-ITN-2019-860315), BIBECA (RTI2018-101248-B-I00 MINECO/FEDER), and BBforTAI (PID2021-127641OB-I00 MICINN/FEDER). I. Serna is supported by a research fellowship from the Universidad Autónoma de Madrid (FPI-UAM-2020). A. Morales is supported by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with Autonomous University of Madrid in the line Encouragement of the Research of Young Researchers, in the context of the V PRICIT (Regional Programme of Research and Technological Innovation).

## References

- [1] I. Rahwan, M. Cebrian, N. Obradovich, et al., Machine behaviour, *Nature* 568 (7753) (2019) 477–486.
- [2] S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson, 2016.
- [3] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J. Chen, V.M. Patel, C.D. Castillo, R. Chellappa, Deep learning for understanding faces: machines May be just as good, or better, than humans, *IEEE Signal Process. Mag.* 35 (1) (2018) 66–83.
- [4] X. Akhtar, A. Hadid, M. Nixon, M. Tistarelli, J. Dugelay, S. Marcel, Biometrics: in search of identity and security (Q & A), *IEEE Multimed.* 25 (3) (2018) 22–35.
- [5] B. Bhanu, A. Kumar, *Deep Learning for Biometrics*, *Advances in Computer Vision and Pattern Recognition (ACVPR)*, Springer, 2017.

- [6] L. Shao, P. Hubert, T. Hospedales, Special issue on machine vision with deep learning, *Int. J. Comput. Vis.* 128 (2020) 771–772.
- [7] P.J. Grother, M.L. Ngan, K.K. Hanaoka, Ongoing Face Recognition Vendor Test (FRVT) Part 2: Identification, NIST Internal Report, National Institute of Standards and Technology, 2018.
- [8] C.M. Cook, J.J. Howard, Y.B. Sirotin, J.L. Tipton, A.R. Vemury, Demographic effects in facial recognition and their dependence on image acquisition: an evaluation of eleven commercial systems, *IEEE Trans. Biometr. Behav. Ident. Sci.* 1 (1) (2019) 32–41.
- [9] B. Lu, J.-C. Chen, C.D. Castillo, R. Chellappa, An experimental evaluation of covariates effects on unconstrained face verification, *IEEE Trans. Biometr. Behav. Ident. Sci.* 1 (1) (2019) 42–55.
- [10] A. Acien, A. Morales, R. Vera-Rodriguez, I. Bartolome, J. Fierrez, Measuring the gender and ethnicity bias in deep models for face recognition, in: *Iberoamerican Congress on Pattern Recognition*, Springer, Madrid, Spain, 2018, pp. 584–593.
- [11] K. Krishnapriya, V. Albiero, K. Vangara, M. King, K. Bowyer, Issues related to face recognition accuracy varying based on race and skin tone, *IEEE Trans. Technol. Soc.* 1 (2020) 8–20.
- [12] B.F. Klare, M.J. Burge, J.C. Klontz, R.W.V. Bruegge, A.K. Jain, Face recognition performance: role of demographic information, *IEEE Trans. Inf. Forensics Secur.* 7 (6) (2012) 1789–1801.
- [13] J. Buolamwini, T. Gebru, Gender shades: intersectional accuracy disparities in commercial gender classification, in: S.A. Friedler, C. Wilson (Eds.), *Conference on Fairness, Accountability and Transparency*, in: *Proceedings of Machine*, vol. 81, Learning Research, New York, NY, USA, 2018, pp. 77–91.
- [14] M. Alvi, A. Zisserman, C. Nellåker, Turning a blind eye: explicit removal of biases and variation from deep neural network embeddings, in: *European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 556–572.
- [15] I. Hupont, C. Fernandez, DemogPairs: quantifying the impact of demographic imbalance in deep face recognition, in: *International Conference on Automatic Face & Gesture Recognition (FG)*, Lille, France, 2019, pp. 1–7.
- [16] P. Drozdzowski, C. Rathgeb, A. Dantcheva, N. Damer, C. Busch, Demographic bias in biometrics: a survey on an emerging challenge, *IEEE Trans. Technol. Soc.* 1 (2) (2020) 89–103.
- [17] P. Terhöst, J.N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. Morales, J. Fierrez, A. Kuijper, A comprehensive study on face recognition biases beyond demographics, *IEEE Trans. Technol. Soc.* (2022), <https://doi.org/10.1109/TTS.2021.3111823>, in press.
- [18] M. Wang, W. Deng, Mitigate bias in face recognition using skewness-aware reinforcement learning, in: *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Seattle, Washington, USA, 2020, pp. 9319–9328.
- [19] S. Gong, X. Liu, A. Jain, Jointly de-biasing face recognition and demographic attribute estimation, in: *European Conference on Computer Vision, Virtual*, 2020, pp. 330–347.
- [20] I. Serna, A. Morales, J. Fierrez, N. Cebrian, M. Obradovich, I. Rahwan, Algorithmic discrimination: formulation and exploration in deep learning-based face biometrics, in: *AAAI Workshop on Artificial Intelligence Safety (SafeAI)*, New York, NY, USA, 2020.
- [21] J. Kleinberg, J. Ludwig, S. Mullainathan, C.R. Sunstein, Discrimination in the age of algorithms, *J. Legal Anal.* 10 (2019) 113–174.
- [22] R. Ranjan, S. Sankaranarayanan, et al., Deep learning for understanding faces: machines May be just as good, or better, than humans, *IEEE Signal Process. Mag.* 35 (1) (2018) 66–83.
- [23] A. Morales, J. Fierrez, A. Acien, R. Tolosana, I. Serna, SetMargin loss applied to deep keystroke biometrics with circle packing interpretation, *Pattern Recognit.* 122 (2022) 108283, <https://doi.org/10.1016/j.patcog.2021.108283>.
- [24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Las Vegas, NV, USA, 2016, pp. 770–778.
- [25] M. Wang, W. Deng, J. Hu, X. Tao, Y. Huang, Racial faces in the wild: reducing racial bias by information maximization adaptation network, in: *International Conference on Computer Vision (ICCV)*, IEEE, Seoul, Korea, 2019, pp. 692–702.
- [26] M. Merler, N. Ratha, R.S. Feris, J.R. Smith, Diversity in faces, *arXiv:1901.10436*, 2019, pp. 1–29.
- [27] K. Karkkainen, J. Joo, FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation, in: *Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Virtual, 2021, pp. 1548–1558.
- [28] P.J. Grother, M.L. Ngan, K.K. Hanaoka, Ongoing Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects, NIST Internal Report, National Institute of Standards and Technology, 2019.
- [29] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, Ms-celeb-1m: a dataset and benchmark for large-scale face recognition, in: *European Conference on Computer Vision (ECCV)*, Springer, Amsterdam, the Netherlands, 2016, pp. 87–102.
- [30] I. Kemelmacher-Shlizerman, S.M. Seitz, D. Miller, E. Brossard, The megaface benchmark: 1 million faces for recognition at scale, in: *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Las Vegas, Nevada, USA, 2016, pp. 4873–4882.
- [31] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, Vggface2: a dataset for recognising faces across pose and age, in: *International Conference on Automatic Face & Gesture Recognition (FG)*, IEEE, Lille, France, 2018, pp. 67–74.
- [32] O.M. Parkhi, A. Vedaldi, A. Zisserman, et al., Deep face recognition, in: *British Machine Vision Conference (BMVC)*, Swansea, UK, 2015, pp. 41.1–41.12.
- [33] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, in: *Computer Vision and Pattern Recognition (CVPR)*, IEEE, Colorado Springs, CO, USA, 2011, pp. 529–534.
- [34] D. Yi, Z. Lei, S. Liao, S.Z. Li, Learning face representation from scratch, *arXiv:1411.7923*, 2014, 1–9.
- [35] S. Yang, P. Luo, C.-C. Loy, X. Tang, From facial parts responses to face detection: a deep learning approach, in: *International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 3676–3684.
- [36] N. Kumar, A. Berg, P.N. Belhumeur, S. Nayar, Describable visual attributes for face verification and image search, *IEEE Trans. Pattern Anal. Mach. Intell.* 33 (10) (2011) 1962–1977.
- [37] 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: *International Conference on Biometrics (ICB)*, IEEE, Gold Coast, Australia, 2018, pp. 158–165.
- [38] Z. Zhang, Y. Song, H. Qi, Age progression/regression by conditional adversarial autoencoder, in: *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Honolulu, Hawaii, USA, 2017, pp. 5810–5818.
- [39] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: a Database for Studying Face Recognition in Unconstrained Environments, *Tech. Rep.* 07-49, University of Massachusetts, Amherst, October 2007.
- [40] J. Ortega-García, J. Fierrez, et al., The multisenario multienvironment biosecure multimodal database (BMDb), *IEEE Trans. Pattern Anal. Mach. Intell.* 32 (6) (2009) 1097–1111.
- [41] A. Morales, J. Fierrez, R. Vera-Rodriguez, R. Tolosana, SensitiveNets: learning agnostic representations with application to face recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* 43 (6) (2021) 2158–2164, <https://doi.org/10.1109/TPAMI.2020.3015420>.
- [42] S. Khan, M. Hayat, S.W. Zamir, J. Shen, L. Shao, Striking the right balance with uncertainty, in: *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Long Beach, California, USA, 2019, pp. 103–112.
- [43] C. Huang, Y. Li, C.C. Loy, X. Tang, Deep imbalanced learning for face recognition and attribute prediction, *IEEE Trans. Pattern Anal. Mach. Intell.* 42 (11) (2019) 2781–2794.
- [44] H. Qin, Asymmetric rejection loss for fairer face recognition, *arXiv:2002.03276*.
- [45] A. Das, A. Dantcheva, F. Bremond, Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach, in: *European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 573–585.



- [46] Z. Yang, X. Zhu, C. Jiang, W. Liu, L. Shen, RamFace: race adaptive margin based face recognition for racial bias mitigation, in: International Joint Conference on Biometrics (IJCB), IEEE, 2021.
- [47] S. Gong, X. Liu, A. Jain, Mitigating face recognition bias via group adaptive classifier, in: Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, 2021.
- [48] T. Calders, S. Verwer, Three naive Bayes approaches for discrimination-free classification, *Data Min. Knowl. Discov.* 21 (2) (2010) 277–292.
- [49] I.D. Raji, J. Buolamwini, Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products, in: Conference on AI Ethics and Society (AIES), AAAI/ACM, New York, NY, USA, 2019, pp. 429–435.
- [50] F. Alonso-Fernandez, J. Fierrez, J. Ortega-Garcia, Quality measures in biometric systems, *IEEE Secur. Priv.* 10 (6) (2011) 52–62.
- [51] M. Bakker, H.R. Valdes, D.P. Tu, K. Gummadi, K. Varshney, A. Weller, A. Pentland Fair, Enough: improving fairness in budget-constrained decision making using confidence thresholds, in: AAAI Workshop on Artificial Intelligence Safety (SafeAI), New York, NY, USA, 2020, pp. 41–53.
- [52] Y. Zhang, R. Bellamy, K. Varshney, Joint optimization of AI fairness and utility: a human-centered approach, in: Conference on AI, Ethics and Society (AIES), AAAI/ACM, New York, NY, USA, 2020, pp. 400–406.
- [53] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, in: Advances in Neural Information Processing Systems (NIPS), MIT Press, 2006, pp. 1473–1480.
- [54] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering, in: Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 815–823.
- [55] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: additive angular margin loss for deep face recognition, in: Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, California, USA, 2019, pp. 4690–4699.
- [56] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, *IEEE Signal Process. Lett.* 23 (10) (2016) 1499–1503.
- [57] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, RetinaFace: single-shot multi-level face localisation in the wild, in: Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, Washington, USA, 2020, pp. 5202–5211.
- [58] R.R. Selvaraju, M. Cogswell, et al., Grad-CAM: visual explanations from deep networks via gradient-based localization, in: International Conference on Computer Vision (CVPR), IEEE, Honolulu, Hawaii, USA, 2017, pp. 618–626.
- [59] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, A. Torralba, Understanding the role of individual units in a deep neural network, *Proc. Natl. Acad. Sci.* (2020) 1–8.
- [60] I. Serna, A. Peña, A. Morales, J. Fierrez, InsideBias: measuring bias in deep networks and application to face gender biometrics, in: IAPR Intl. Conf. on Pattern Recognition (ICPR), IEEE, 2021, pp. 3720–3727.
- [61] J. Fierrez, A. Morales, R. Vera-Rodriguez, D. Camacho, Multiple classifiers in biometrics. Part 2: Trends and challenges, *Inf. Fusion* 44 (2018) 103–112.
- [62] A. Pena, J. Fierrez, A. Lapedriza, A. Morales, Learning emotional-blinded face representations, in: IAPR Intl. Conf. on Pattern Recognition (ICPR), 2021.
- [63] A. Pena, I. Serna, A. Morales, J. Fierrez, A. Lapedriza, Facial expressions as a vulnerability in face recognition, in: IEEE Intl. Conf. on Image Processing (ICIP), 2021, pp. 2988–2992.
- [64] A. Pena, I. Serna, A. Morales, J. Fierrez, Bias in multimodal AI: testbed for fair automatic recruitment, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRw), 2020.
- [65] I. Serna, D. DeAlcala, A. Morales, J. Fierrez, J. Ortega-Garcia, IFBiD: inference-free bias detection, in: AAAI Workshop on Artificial Intelligence Safety (SafeAI), 2022.
- [66] V. Mirjalili, S. Raschka, A. Ross, FlowSAN: privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers, *IEEE Access* 7 (2019) 99735–99745.
- [67] M. Ghafourian, J. Fierrez, R. Vera-Rodriguez, I. Serna, A. Morales, OTB-morph: one-time biometrics via morphing applied to face templates, in: IEEE/CVF Winter Conf. on Applications of Computer Vision Workshops (WACVw), 2022.
- [68] A. Ortega, J. Fierrez, A. Morales, Z. Wang, M. Cruz, C.L. Alonso, T. Ribeiro, Symbolic AI for XAI: evaluating LFIT inductive programming for explaining biases in machine learning, *Computers* 10 (11) (2021) 154, <https://doi.org/10.3390/computers10110154>.