Exploring Racial Bias within Face Recognition via per-subject Adversarially-Enabled Data Augmentation

Seyma Yucer[1], Samet Akçay[1,3], Noura Al-Moubayed[1], Toby P. Breckon[1,2]
Department of {Computer Science[1], Engineering[2]}, Durham University, Durham, UK
COSMONiO[3], Durham, UK
{ seyma.yucer-tektas, samet.akcay, noura.al-moubayed, toby.breckon }@durham.ac.uk

Abstract

Whilst face recognition applications are becoming increasingly prevalent within our daily lives, leading approaches in the field still suffer from performance bias to the detriment of some racial profiles within society. In this study, we propose a novel adversarial derived data augmentation methodology that aims to enable dataset balance at a per-subject level via the use of image-to-image transformation for the transfer of sensitive racial characteristic facial features. Our aim is to automatically construct a synthesised dataset by transforming facial images across varying racial domains, while still preserving identity-related features, such that racially dependant features subsequently become irrelevant within the determination of subject identity. We construct our experiments on three significant face recognition variants: Softmax, CosFace and ArcFace loss over a common convolutional neural network backbone. In a side-by-side comparison, we show the positive impact our proposed technique can have on the recognition performance for (racial) minority groups within an originally imbalanced training dataset by reducing the per-race variance in performance.

1. Introduction

Numerous machine learning applications utilising facial attributes have proliferated in recent years as autonomous decision-making processes have become widely adopted by companies and governments [1]. A growing number of applications based on face analyses for surveillance [2], recruitment [3], and health-care [4] have increasingly become integrated into our daily lives.

However, the generalisation of such research and applications is problematic due to the prevalence of bias occurrences within face recognition [5]. The imbalance in specific demographic groups occurring with varying geographic locale globally, including race, age or gender, poses a challenge of transparent explanations and solutions for facial recognition applications. Hence, to cope with real-world diversity, it is crucial to have a profound understanding of this bias within every aspect [6].

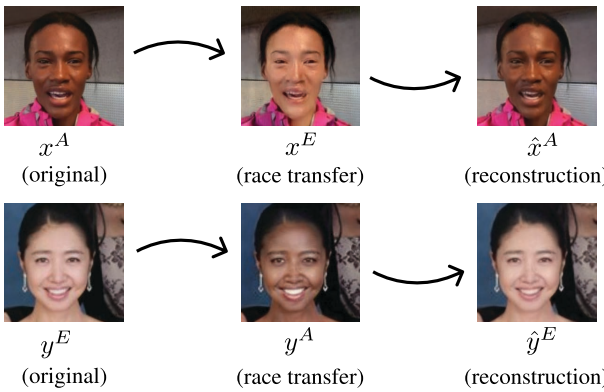

Figure 1. Racial transformation example using [7]. We transfer an African image x^A to Asian image y^E and obtain sythesised x^E in Asian domain and we reconstruct $\hat{x}^A$ from x^E image. Asian image y^E to African image x^A transformation follows the same procedure.

Bias in machine learning has been extensively studied for decades [8, 9]. These studies provide the fundamental understanding of the underlying reasons for face recognition bias which has also seen a surge of interest in recent years [10, 6]. Studies have addressed this problem in various perspectives such as data pre-processing [11, 12, 13], and adversarial training [14, 15, 16].

Meanwhile, recent advances in Generative Adversarial Networks (GAN), have led to realistic image generation [17] and even class generation [12]. Such advances in the field have a promising potential to overcome the bias in face recognition via realistic image generation as most of the face recognition datasets have a significantly imbalance distribution on either classes [18] or demographic groups [19].

In this study, we address the racial bias of face recog-

nition from an adversarial augmentation point of view. As most of the datasets [20, 21, 6] consist of four major racial groups, namely African, Asian, Caucasian and Indian, we seek group-fairness among these races, in terms of facial recognition performance, by utilising generative adversarial network (GAN) [22].

Previous work [14, 15, 16] has established adversarial techniques to minimise mutual information on identity features, which reveal sensitive attributes about race, gender and age of the subject. However, such approaches [14, 23], have failed to effectively address the trade-off between suppressing the use of such sensitive attributes and the loss of key identity-related features which pertain to the overall performance of the facial recognition approach. Our solution, instead, uses an adversarial image re-synthesise technique [7], to transform sensitive attributes across a set of synthetic images comprising the full range of races being considered within the facial recognition problem.

By doing so, we preserve the important identity-related features whilst making the racially dependent features of the face less prevalent due to the artificially synthesised distribution of these identity characteristics across the full range of race profiles for any given individual.

Figure 1 illustrates how we transform the identity characteristics, and hence features, any given individual across multiple racial profiles using a CycleGAN [7]. It proposes transformation across racial domains and reconstruction to produce an identical image from a transformed image during the cyclic adversarial training.

To show its robustness, we explore the performance of our approach using balanced and imbalanced training datasets. The main contributions of this paper are as follows:

- we propose an adversarial image-to-image transformation technique to mitigate racial bias based on the cyclic adversarial training approach of CycleGAN [7].
- we illustrate both quantitative and qualitative performance of our proposed facial data augmentation techniques over established benchmark datasets within the face recognition domain, establishing a statistical paradigm for the presentation of recognition results on a per-race basis.

The rest of this paper is structured as follows: in Section 2, we review the current solutions for face recognition bias in three different categories. We present a methodology for this study in Section 3 with our experimental setup and results in Section 4 and 5, respectively. An extended discussion on adversarial face recognition bias for both balanced and imbalanced datasets is presented within Section 5 with our final conclusions subsequently presented in Section 6.

2. Related Work

Bias and fairness in machine learning have been studied in the last decade, and significant research [24, 25] draws attention to bias for different fields like face recognition, action recognition or language processing.

As one of the most prominent fields of machine learning, face recognition has been extensively used across different areas [26, 27]. As the popularity of face recognition increases, we face more bias incidents [10]. Moreover, studies [6, 28, 29] point out the bias of current face recognition web services and state-of-the-art algorithms for demographic groups such as age, gender, and race. Although definitions of demographic attributes might be uncertain, it is still important to strive for group-fairness [30].

Studies of bias in face recognition which use contemporary deep learning approaches are categorised into three main groups: pre-processing (data preparation), in-processing (model training) and post-processing (output inference) techniques.

Pre-processing Methods. Previous studies [31, 19] revealed that the public face recognition datasets have more male and lighter skin tone subjects than respectively female and darker skin tone subjects. This is due to the images within these datasets are mostly from celebrities, including sports players, actors, politicians, collected from predominantly white male subjects. In other respects, the studies of [20, 31] released balanced datasets for four racial groups; they do not provide universal race coverage nor they are not openly and readily available for access.

To obtain fair datasets, studies [32, 12, 13] propose resampling methods by either dropping or augmenting samples in the datasets. Downsampling can be considered as a solution for avoiding bias despite the information loss it introduces. Augmentation techniques [12, 13] for image generation have improved significantly using adversarial learning. However, the limitations, as described in [33], are still a concern for mitigating bias. Feature transformation is another pre-processing approach [34] that improves the feature space of under-represented subjects by moving the distribution of the feature space closer to the regular, supposedly unbiased distribution.

In-processing Methods. In-processing methods are divided into three groups: (i) adversarial approaches [14, 23, 15, 16], (ii) domain adaptation methods [20] and (iii) cost-sensitive learning techniques [35, 36]. Adversarial methods focus on sensitive features on the image; with [14] proposing an adversarial feature learning approach rather than learning all the feature representations from the image. In this way, it minimises mutual information between bias features and characteristic features to decrease bias influence. The experiments of [14] are relatively simplistic compared to face recognition bias. Distinguishing demographic information within an image is a serious trade-off

of face recognition as demographic features (age, gender, race), and identity features overlap. Another approach in [23], addresses this problem by highlighting the difficulty of setting a demographic condition in realistic face generation. On the other hand, [15] debiases images by minimising correlation on disentangled features. Another study [16] reduces the dependence on sensitive attributes. Despite achieving state-of-the-art results on the test, there is still ample room for further understanding of bias.

A domain adaptation technique, [20], transfers the Caucasian domain to non-Caucasian domains during training but requires to have at least one source domain to transfer into others. Cost-sensitive solutions [35, 36] have been used for imbalanced learning and machine learning fairness in general. For face recognition, adaptive margin [37] or cluster large margin settings [18] are more frequently considerable since the aim is to have intra-class compactness and inter-class discrepancy for large scale datasets. Distinguishing the group features on hypersphere helps to avoid overfitting of under-represented groups. Adaptive margins [21] for each race improves the scatter of features of races.

Post-processing Methods. Post-processing studies are based on either detecting the bias or improving the fairness after training the model. For example, [38] proposes a Multiaccuracy-Boost algorithm for any machine learning algorithms to improve fairness. IBM [39] provides an extensive toolkit to detect bias and determine the current model fairness level. For broader explanations, [40, 29] give demographic bias level of current state-of-the-art face recognition algorithms.

Motivated by [7], our approach is based on adversarial image synthesise to mitigate bias. Unlike other adversarial studies [14, 16], we transform race information from one group to another for fair face recognition. We aim to augment sensitive attributes to make them irrelevant for face recognition solutions.

3. Proposed Method

We present our methodology in three parts: we first describe our problem definition in Section 3.1, explain image-to-image transfer method [7] for race transformation to mitigate face recognition bias in Section 3.2 and outline our comparator state-of-the-art face recognition algorithms [41, 37] in Section 3.3.

3.1. Problem Definition

In this section, we define our problem by introducing the general terms of machine learning bias. *Disparate impact*, as indirect discrimination, appears when there is a correlation between sensitive attributes (age, gender, race) and other attributes. It causes inequality on outcomes for different demographic groups, as observed on various machine learning applications, including face recognition web services [6].

Ideally, a machine learning algorithm should require that the conditional probability P of the output given input x does not depend on any *sensitive attributes* which is demographic features in our case. This *unawareness* can be formalized as $P(y \mid x) = P(y \mid x, s)$, where x is an input, y is the corresponding label and s is a sensitive attribute that does not alter the outcome. However, removing dependency is highly challenging for face recognition due to high *mutual information* between facial features and sensitive attributes, like race.

For a given face image dataset, $D = [x_1, x_2, x_3, \ldots, x_N]$ provides N number of face images. A feature embedding vector of an image, $z_i = [f_1, f_2, \ldots, f_d]$, where $z_i \in \mathbb{R}^d$, is commonly statistically dependent on sensitive attributes where it causes *indirect discrimination* for particular demographic groups which potentially form overlapping, subsets of D. Although the common approach for face recognition bias is to minimise this mutual information to remove the dependency on sensitive features; it is still an extremely difficult task using face features without sacrificing any prior information for face recognition as shown in [14, 23].

Hence, we approach the problem from a completely different perspective by transferring sensitive attributes from one domain to another whilst simultaneously preserving prior information for recognition. On the other hand, we are aware that some features are more prevalent in some demographic groups than others. The sensitive information, in this case, may improve the prior information for the recognition task. Lighter skin allows the model to learn more detailed features given characteristics of modern cameras and common scene lumination conditions. A novel input mechanism which projects different sensitive information for one image to a model makes race modelling irrelevant. As a result, we ask a question; *What if we augment and transfer sensitive information rather than removing it*? To answer this question, we present a new pre-processing based method requires augmentation of sensitive attributes of an image.

Our new inputs consist of three generated images from different domains for each image. Given the race domains $\{A, E, C, I\}$ for $\{African, Asian, Caucasian, Indian\}$ respectively, we aim to transform an image x_i from one domain as an image x_j to another domain. For instance, we transform given x_i in A to another image from different domains such as E, C, I. If we use different images belonging to these domains to transform, we can define new generated input dataset as following list $x_i^+ = \left[x_i, x_i^E, x_i^C, x_i^I\right]$ where x_i is the original image and x_i^+ is a new input list including the original image.

Transferring sensitive information while keeping

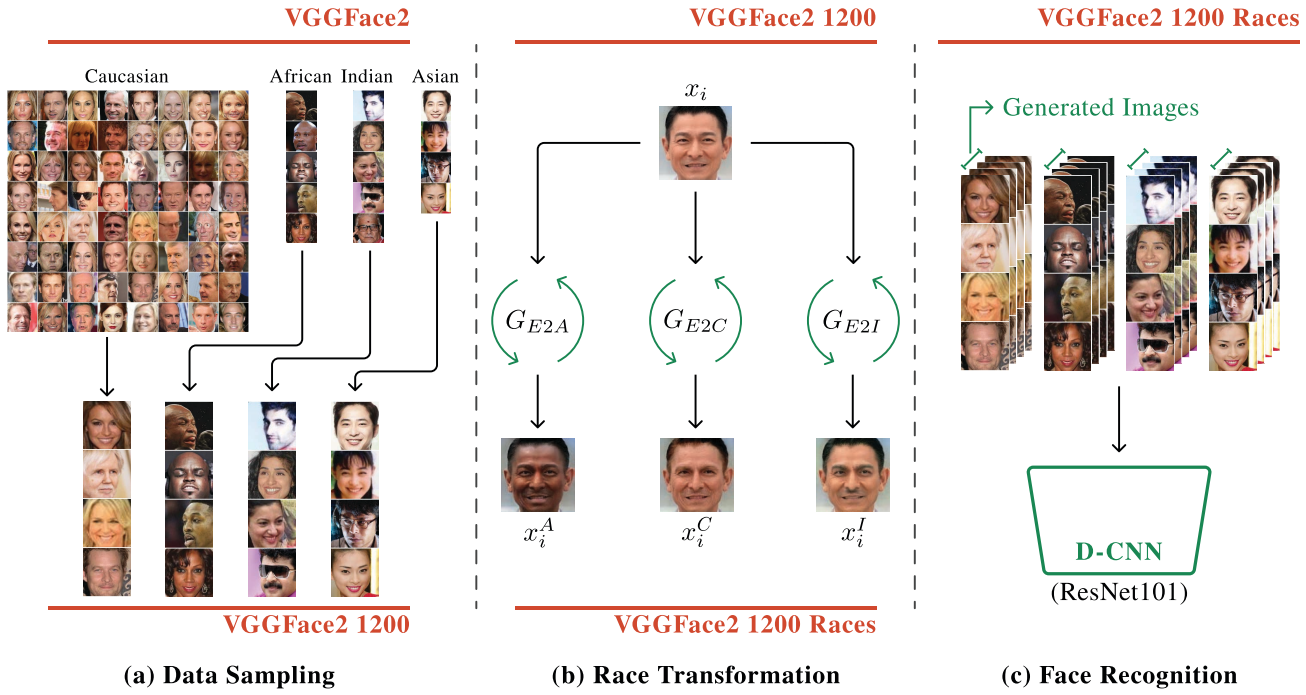

Figure 2. Overview of our solution in three phases: (a) describes imbalanced distribution of VGGFace2 [42] and downsampling it to VGGFace2 1200. (b) illustrates race domain transformation schema for a given image x_i (c) shows face recognition algorithms with Softmax [43], CosFace [41] and ArcFace [37] loss functions using VGGFace2 1200 Races.

prior information of the image is possible via adversarial methods, as they are capable of generating images from the training data distribution. To show that, we propose a solution of sensitive attribute transformation while keeping prior information for face recognition and present a new augmented dataset, $D^{+}_{image} = [x_i, x_i^A, x_i^C, x_i^I, \ldots x_i, x_i^E, x_i^C, x_i^I, \ldots, x_n, x_n^A, x_n^E, x_n^C,]$. In the next Section 3.2 we present our approach to the image synthesise process to obtain D^{+}_{image}.

3.2. Adversarial Image-to-Image Transfer

Our solution transforms these sensitive attributes using a cyclic adversarial domain transfer approach, CycleGAN [7]. We assume that learning a mapping function between two different race groups domain reduces the dependency on sensitive features.

For example, given an African face image $x_i \in A$, and a Caucasian image $x_j \in C$, we assume that the two different data distributions from these image race groups $x_i \sim p_{data}(x_i)$ and $x_j \sim p_{data}(x_j)$ can be transferable between each other. To map these two distributions between domain A and C, we introduce two mapping functions F and G, respectively from African to Caucasian domains and from Caucasian to African domains using CycleGAN [7]. Within a GAN framework, these two directional transformations need two discriminators D_A and D_C, to distinguish between x_i and $F(x_j)$, x_j and $G(x_i)$, respectively. Moreover, as an additional control on adversarial training, a cycle-consistency loss is introduced to ensure that the mapping function can transfer an individual input x_i to the desired output x_j.

$$L_{GAN}(G, D_C, A, C) = \mathbb{E}_{x_j} \sim p_d(x_j)\,[logD_C(x_j)] + \mathbb{E}_{x_i} \sim p_d(x_i)\,[log(1 - D_C(G(x_i))] \quad (1)$$

For the first part of race transformation, an adversarial loss is used as defined in Equation 1 where A and C are the African and Caucasian group domains, respectively. While the generator G synthesise images using source domain A to associate to target domain C, discriminator D_C distinguishes between the real image and x_j from the synthesised image, $G(x_i)$. The same process is applied with generator F and discriminator D_A to transform domains from C to A.

The key premise of CycleGAN [7] is a controlled mechanism of adversarial training which allows us to synthesise more accurate images from the desired images in the domain. To achieve this, cycle consistency loss is introduced as defined in Equation 2 , where $F(G(x_i))$ is reconstructed x_i from synthesised $G(x_i)$ new image. In this case, generators F and G are able to reconstruct the original images. The $L1$ norm in this loss measures the difference between

the original image and reconstructed image as follows:

$$L_{cyc}(G,F) = \mathbb{E}_{x_i} \sim p_d(x_i)\left[\| F(G(x_i)) - x_i \|_1\right] + \mathbb{E}_{x_j} \sim p_d(x_j)\left[\| G(F(x_j)) - x_j \|_1\right] \quad (2)$$

The overall loss function, as defined in Equation 3, consists of two adversarial loss within the cycle-consistency loss where λ is a term to control the relative importance of the cycle-consistency loss.

$$L(G,F,D_A,D_C) = L_{GAN}(G,D_C,A,C) + L_{GAN}(F,D_A,C,A) + \lambda L_{cyc}(G,F)] \quad (3)$$

Subsequently, overall adversarial training of this objective function aims to solve the following equation:

$$G^*, F^* = \underset{G,F}{\operatorname{argmin}} \max_{D_A,D_C} L(G,F,D_A,D_C). \quad (4)$$

In the intermediate step $G(x_i)$ and $F(x_j)$, the generator encodes features of inputs x_i and x_j and then $F(x_j)$ and $G(x_i)$ decodes back to obtain original images again. With reference to this set of transform Equations 1-4, we can transform both, domain A into domain C and C into A similarly for other domain pairings.

3.3. Face Recognition

Recent state-of-the-art face recognition algorithms [43, 41, 37, 42] achieve outstanding results for both face verification and identification tasks on public datasets. However, they are not as reliable for real-world racial diversity as their performance is lower for under-represented racial groups [20].

In Section 3.2, we presented our proposed approach to address racial bias within face recognition using an adversarial image-to-image transformation technique. To assess this proposed approach, we first present current face recognition loss functions namely Softmax, CosFace, ArcFace that underpin leading state-of-the-art face recognition algorithms [43, 41, 37], then we utilise each of these three methods in conjunction with our cyclic adversarial domain transfer approach.

The Softmax [43], CosFace [41] and ArcFace [37] methods are based on loss functions that operate on the outputs of the last fully connected layer of the selected backbone Deep Convolutional Neural Network [44] (DCNN). In essence, after feeding an image forward through a DCNN, we obtain the feature space representation of the image. These loss functions enforce different representations of features to predict if they belong to a given subject. First, Softmax loss is formulated as follows,

$$\mathcal{L}_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T z_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T z_i + b_j}} \quad (5)$$

where z_i is the feature representation of the image $x_i \in \mathbb{R}^d$ in the dataset D belonging to y_i-th subject class. The number of samples is N labelled with n classes. W_j is the j-th column of the weights and b_j is the j-th column of the bias term in the last fully-connected layer. Weights and bias term dimensions are $W \in \mathbb{R}^{dxn}$ and $b_j \in \mathbb{R}^n$, respectively.

Softmax loss [43] is one of the most widely used objective function to learn optimal feature representations from images. It discriminates deep representations from different classes by maximizing the posterior probability of the ground-truth class. Once large-scale datasets have high similarity on intra-class samples and diversity on inter-class samples, Softmax loss entangles features [45]. To address this problem, CosFace [41] proposes to use both norm and angle of the feature representation to contribute to the posterior probability such that:

$$\mathcal{L}_2 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\|z\|(\cos(\theta_{y_i,i})-m)}}{e^{\|z\|(\cos(\theta_{y_i,i})-m)} + \sum_{j\neq y_i}^{n} e^{\|z\|(\cos(\theta_{j,i}))}} \quad (6)$$

where $\cos(\theta_j, i) = {W^T}_j z_i$ and z_i, y_i,n N, W_i denote i-th feature representation with all other definitions as per previously defined. For CosFace loss, the bias term is removed, and the weights W and embeddings z are normalized using the L_2 normalization. To cope with incorrect classified samples, a cosine margin m is applied to the classification boundary.

An alternative loss function, ArcFace [37] differs from CosFace [41] based on its distinct margin. ArcFace has more accurate geodesic distance due to it has constant linear angular margin, m penalty throughout the interval while CosFace has a nonlinear angular margin. It also normalizes the weights and embeddings and fixes the bias term to zero. In Equation 7, the ArcFace loss function is formulized as follows:

$$\mathcal{L}_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\|z\|(\cos(\theta_{y_i,i}+m))}}{e^{\|z\|(\cos(\theta_{y_i,i}+m))} + \sum_{j\neq y_i}^{n} e^{\|z\|(\cos(\theta_{j,i}))}} \quad (7)$$

where all definitions are as per Equation 6. Overall the key Softmax, CosFace [41] and ArcFace [37] differences lie in their use of deep feature representation, weight vectors and approach to their margin penalty. Within the scope of this study, we only use these methods as experimental vehicles to illustrate our per-subject data augmentation methodology to address face recognition race bias within such state-of-the-art face recognition algorithms.

An overview of our approach is shown in Figure 2. Figure 2 (a) describes the real-world dataset imbalanced distribution for different racial groups. As an initial experimental exercise, we, downsample this imbalanced distribution to

understand the relationship between bias and data. In Figure 2 (b), we explain the image transformation process for one exemplar Asian subject. We introduce our x_i to three different CycleGAN and obtain three different synthesised images x_i^A, x_i^C, x_i^I. Subsequently, the training dataset has changed, and then we use our newly augmented dataset for face recognition using algorithms with Softmax, CosFace [41], ArcFace [37] in Figure 2 (c).

4. Experimental Setup

This section provides overview of our experimental evaluation in terms of the face recognition datasets used, the race classification used for racial annotation and the implementation details of our proposed approach.

4.1. Datasets

To validate our approach, we utilise BUPT-Transferface [20], VGGFace2 [42] and RFW [20].

BUPT-Transferface [20] provides 50K African, Asian and Indian face images and over 460K Caucasian face images. We use BUPT-Transferface dataset for two different purposes: (i) race transfer, (ii) race classification.

VGGFace2 [42] contains 3.3M+ images for over 9K subjects (8631 subjects training examples, 500 testing examples). We train the face recognition methods which we introduced in Section 3.3 on VGGFace2.

VGGFace2 1200 is a subsampled version of VGGFace2 which is racially balanced and contains 300 subjects per-race. We evaluate our approach on both VGGFace2 1200 and VGGFace2.

Racial Faces in-the-Wild (RFW) [20] is a face verification test set which provides 6K pairs of images for each race. We compare the verification accuracy of our proposed approach on different races using the same protocol in [46].

4.2. Race Annotations

We obtain racial annotation labels for VGGFace2 [42] dataset using fine-grained classification to solely support our development of a technique to mitigate bias.

The work of [47] proposes attention-guided data augmentation to improve the spatial representation of discriminative image parts using its cropping and dropping mechanism. We adopt this solution for a race classification problem where discriminative image parts are facial attributes of eyes, nose, mouth, and forehead. Via this approach [47], we obtain racial annotations of VGGFace2 [42] and we manually check the least certain subjects according to the majority of image labels for each subject and additionally exclude some subjects who are not in the four-race set $\{Caucasian, African, Asian, Indian\}$. After this semi-automatic process, the subject distribution for training and testing sets is shown in Figure 3 whereby the inherent racial and gender imbalance is clearly illustrated.

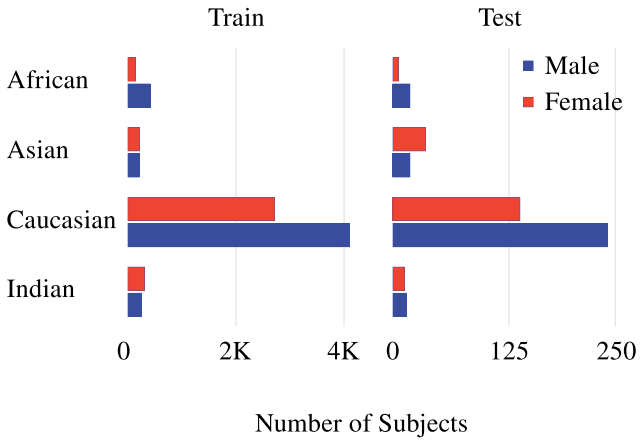

Figure 3. VGGFace2 dataset gender and race distribution for train and test.

4.3. Race Transfer

Our proposed image-to-image transformation approach creates a new dataset D^+_{image}, to transfer race attributes from one race group to another. To achieve that, we define separate mappings for each pair of the four different race groups. The set of 12 mappings are: {African $\rightarrow$ Asian, African $\rightarrow$ Caucasian, African $\rightarrow$ Indian, Asian $\rightarrow$ African, Asian $\rightarrow$ Caucasian, Asian $\rightarrow$ Indian, Caucasian $\rightarrow$ African, Caucasian $\rightarrow$ Asian, Caucasian $\rightarrow$ Indian, Indian $\rightarrow$ African, Indian $\rightarrow$ Asian, Indian $\rightarrow$ Caucasian}. As our CycleGAN based approach provides two-way transformations between source and target domains, we train six models to find these two directional mappings following the approach outlined in Section 3.2.

For training, we generate 25K image pairs using the BUPT-Transfer [20] dataset. All face images are aligned and have a size of 256×256. To avoid gender domain differences, we only match images of the same gender as pairs. Using these six CycleGAN models, we synthesise new images and denote extended dataset as VGGFace2 1200 Races [42] which contains the original VGGFace2 1200 images and synthesised race images. Each image has three different transformed images that belong to other race domains in addition to the original. As a result, we partially absorb the downsampling effect on VGGFace2 1200. Subsequently, we synthesise all non-Caucasians images on original VGGFace2 and call the new dataset VGGFace2 8631 Races, D^+_{image}. We do not transform Caucasian images to other racial domains; they are already dominant in the original dataset.

4.4. Face Recognition

We train a common DCNN, ResNet [44] on proposed augmented datasets; VGGFace 2 1200, VGGFace 2 8631. We utilise ResNet100 explored by [37] with

$BatchNorm - Dropout - FC - BatchNorm$ structure to get the final 512-D feature space representation after the last convolutional layer. We use same architecture for Softmax [43], CosFace [41] and ArcFace [37] loss functions.

5. Results

To evaluate the performance of the proposed approach, we use LFW face verification protocol [46], which measures whether two images belong to the same subject or not.

We assess synthesised image quality by feeding them through a race classifier introduced in Section 4.2. We show examples of the correctly classified images and the misclassified images in Figure 4 (top and bottom parts are separated). Each column of Figure 4 show an image transformation example where the original image is represented with green and red borders, and synthesised images are laid in the corresponding racial domain label in the y-axis. As can be seen in the bottom part of Figure 4, image transformation is prone to fail on poor illumination and pose variations.

Loss	Training Dataset	LFW	RFW					
			African	Asian	Caucasian	Indian	AVG	STDV
Softmax	VGGFace2 1200	96.13	69.10	73.70	79.25	76.78	74.71	4.37
Softmax	VGGFace2 1200 Races	**96.27**	**70.65**	**75.68**	**80.27**	**78.28**	**76.22**	**4.16**
CosFace	VGGFace2 1200	98.16	82.78	82.68	87.53	85.41	84.60	2.33
CosFace	VGGFace2 1200 Races	**98.65**	**83.22**	83.23	**87.95**	**85.77**	**85.04**	**2.28**
Arcface	VGGFace2 1200	98.16	80.91	81.78	**86.86**	83.70	83.31	2.64
Arcface	VGGFace2 1200 Races	**98.63**	**81.28**	82.83	85.95	**84.72**	**83.69**	**2.06**

Table 1. Verification performance (%) of Softmax, CosFace, and ArcFace with ResNet-101 [44] on LFW [46] and RFW [20] when trained on VGGFace2 1200 and proposed VGGFace2 1200 Races datasets.

For face recognition, we first test our performance on **balanced datasets** VGGFace2 1200 and VGGFace2 1200 Races. We compare our results on RFW [20] using three different loss functions; Softmax, CosFace [41] and ArcFace [37] as shown in Table 1. Proposed facial image augmentation approach improves performance in all three methods by 0.38-1.51 %. As non-Caucasian results are improved, the standard deviation among groups is decreased. We also share LFW results in Table 1 to show the improvement of our solution on the imbalanced dataset. Second, we use the **imbalanced dataset** with the ArcFace as shown in Table 2. While LFW verification performance remains the same, RFW African and Asian performances are improved, and the standard deviation declines from 2.91 to 2.45.

Training Dataset	LFW	RFW					
		African	Asian	Caucasion	Indian	Average	STDV
VGGFace2	**99.51**	89.45	87.61	**94.71**	**91.21**	**90.75**	2.91
VGGFace2 8631 Races	**99.51**	**90.10**	**87.73**	93.72	90.50	90.51	**2.45**

Table 2. Verification performance (%) of ArcFace using ResNet-101 [44] trained on VGGFace2 [42] and VGGFace2 8631 Races with syntesised images of non-Caucasian subjects on VGGFace2, tested on LFW [46] and RFW [20].

Method	African	Asian	Caucasian	Indian	AVG	STDV
Softmax	67.95	73.5	77.77	75.78	73.75	4.24
CosFace	77.15	78	82.8	80.42	79.59	2.55
ArcFace	74.75	77.63	83.18	80.97	79.13	3.71

Table 3. RFW dataset verification performance using the LFW protocol [46] for state-of-the-art algorithms trained on per-subject, per-race and per-gender balanced data.

5.1. Ablation Study

Q: This study provides experiments on both balanced and imbalanced training datasets. Why do you not use only the imbalanced datasets? Does balancing datasets help to decrease bias?
A: Imbalanced data may seem to be the main reason for face recognition bias. However, when we train algorithms on completely equally distributed data, the results still appear to exhibit performance bias. To show this, we downsample VGGFace2 and obtain 1000 subjects with 100 images on each subject. We also keep the race and gender groups balanced. As shown in Table 3, there is still about eight per cent gap between African and Caucasian on average. Another study experiments on a large and nearly balanced dataset and again differs on Caucasians and non-Caucasians [21]. Subsequently, we focus on a novel per-subject racial data balancing approach to understanding its impact on the face recognition bias.

Q: How does the training of CycleGAN affect overall accuracy?
A: We assess the quality of our synthesised images by testing them using a race classifier (Section 4.2). We would expect the race classifier to recognise them as the correct transformed racial label. Our overall accuracy is 49% across all transformations, but when we increase this accuracy using more pairs, and longer training, this results in an overall reduction in face recognition performance. The trade-off is complex because after transforming the main racial attributes of the face such as skin colour, eye structure and hair colour, CycleGAN proceeds to translate all facial features including those which implicitly encode unique subject identity. Other notable negatives are variations in pose and illumination on the synthesised images which could alternatively be addressed via [17] in future work.

6. Conclusion

Although the usage of face recognition applications is increasing every day, state-of-the-art-methods are still suffering from racial bias in terms of performance. To address this issue, in this study, we explore racial bias in face recognition and present a novel adversarial derived data augmentation methodology. Transferring racial attributes of a human

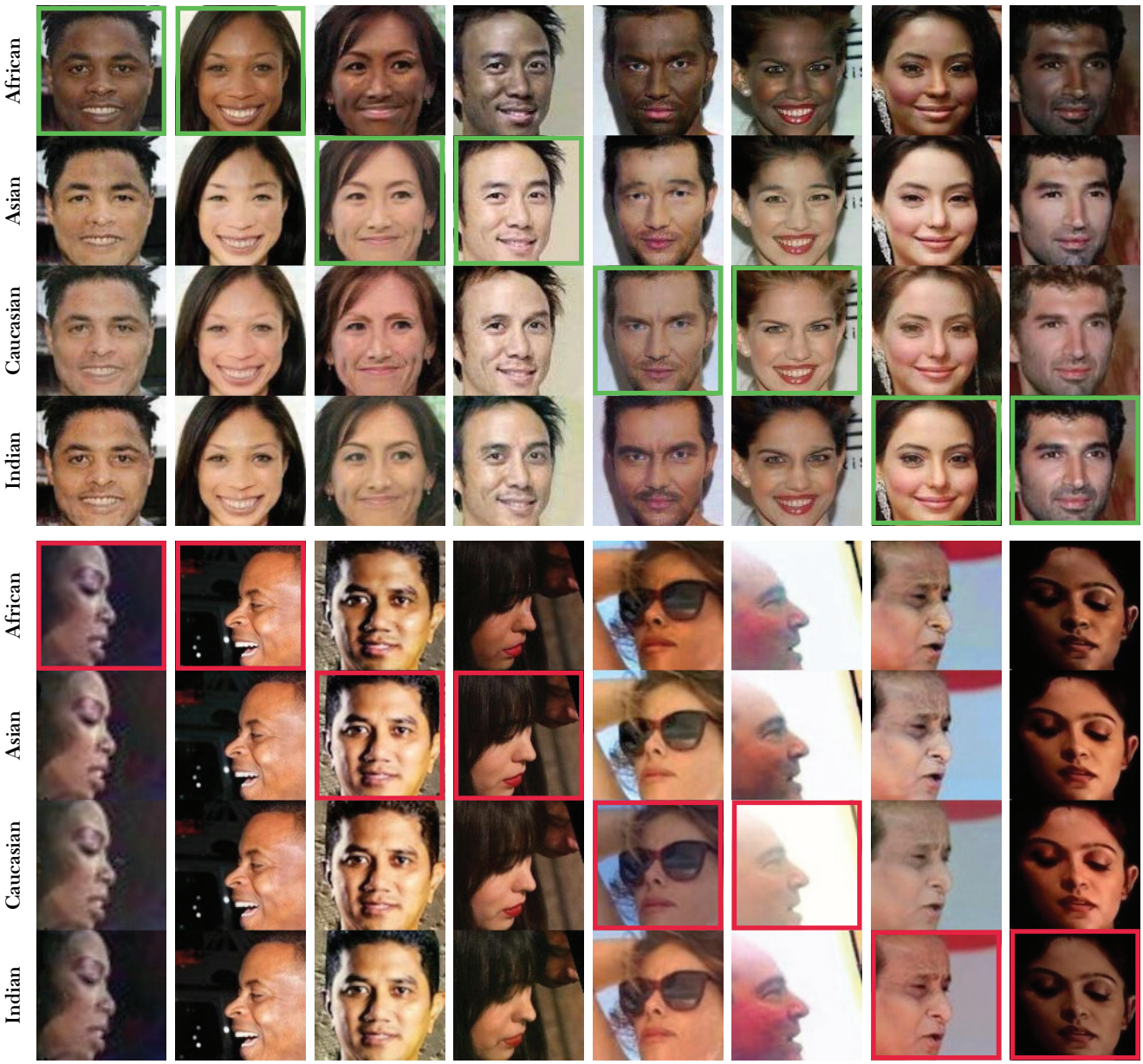

Figure 4. A selection of successful (top) and failure (bottom) examples of the CycleGAN racial domain transformation of VGGFace2 dataset. Each column contains an original and sythesised face images of the same subject where the green (top) and red (bottom) borders indicate the original image and the corresponding race labels are laid out on the y-axis.

face whilst preserving identity features in the face recognition datasets makes face recognition algorithms more robust and less race-dependant. We demonstrate that our proposed technique improves face recognition accuracy on minority groups by 1% using imbalanced and balanced training datasets. On our manually balanced dataset, we also compare three significant face recognition variants: Softmax [43], CosFace [41] and ArcFace [37] loss functions with a common convolutional neural network backbone ResNet-101 [44]. Although illumination, pose, and light challenge the quality of the image transformation; our technique not only improves the overall face recognition accuracy but also suppresses inter-group performance variation.

References

[1] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *Conference on Graphics, Patterns and Images*. IEEE, 2018. 1

[2] Saman Bashbaghi, Eric Granger, Robert Sabourin, and Mostafa Parchami. Deep learning architectures for face

recognition in video surveillance. In *Deep Learning in Object Detection and Recognition*. Springer, 2019. 1

[3] Léo Hemamou, Ghazi Felhi, Vincent Vandenbussche, Jean-Claude Martin, and Chloé Clavel. Hirenet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews. In *AAAI Conference on Artificial Intelligence*, 2019. 1

[4] Md Azher Uddin, Joolekha Bibi Joolee, and Young-Koo Lee. Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm. *IEEE Transactions on Affective Computing*, 2020. 1

[5] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Conference on AI, Ethics, and Society*, 2019. 1

[6] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, 2018. 1, 2, 3

[7] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017. 1, 2, 3, 4

[8] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *International Conference on Knowledge Discovery and Data Mining*, 2008. 1

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Conference on Neural Information Processing Systems*, 2017. 1

[10] Raul Vicente Garcia, Lukasz Wandzik, Louisa Grabner, and Joerg Krueger. The harms of demographic bias in deep face recognition research. In *International Conference on Biometrics*. IEEE, 2019. 1, 2

[11] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Conference on Fairness, Accountability, and Transparency*, 2020. 1

[12] Adamu Ali-Gombe and Eyad Elyan. Mfc-gan: class-imbalanced dataset classification using multiple fake class generative adversarial network. *Neurocomputing*, 2019. 1, 2

[13] Payel Sadhukhan. Learning minority class prior to minority oversampling. In *International Joint Conference on Neural Networks*. IEEE, 2019. 1, 2

[14] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 3

[15] Sixue Gong, Xiaoming Liu, and Anil K Jain. Debface: De-biasing face recognition. *arXiv preprint arXiv:1911.08080*, 2019. 1, 2, 3

[16] Xiaoqian Wang and Heng Huang. Approaching machine learning fairness through adversarial network. *arXiv preprint arXiv:1909.03013*, 2019. 1, 2, 3

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 7

[18] Chen Huang, Yining Li, Change Loy Chen, and Xiaoou Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 3

[19] Isabelle Hupont and Carles Fernández. Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019. 1, 2

[20] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *IEEE International Conference on Computer Vision*, 2019. 2, 3, 5, 6, 7

[21] Mei Wang and Weihong Deng. Mitigate bias in face recognition using skewness-aware reinforcement learning. *arXiv preprint arXiv:1911.10692*, 2019. 2, 3, 7

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 2

[23] Daniel McDuff, Shuang Ma, Yale Song, and Ashish Kapoor. Characterizing bias in classifiers using generative models. In *Advances in Neural Information Processing Systems*, 2019. 2, 3

[24] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019. 2

[25] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019. 2

[26] Ya Wang, Tianlong Bao, Chunhui Ding, and Ming Zhu. Face recognition in real-world surveillance videos with deep learning method. In *International Conference on Image, Vision and Computing*. IEEE, 2017. 2

[27] Ian D Stephen, Vivian Hiew, Vinet Coetzee, Bernard P Tiddeman, and David I Perrett. Facial shape analysis identifies valid cues to aspects of physiological health in caucasian, asian, and african populations. *Frontiers in Psychology*, 2017. 2

[28] Shruti Nagpal, Maneet Singh, Richa Singh, Mayank Vatsa, and Nalini Ratha. Deep learning for face recognition: Pride or prejudiced? *arXiv preprint arXiv:1904.01219*, 2019. 2

[29] Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *arXiv preprint arXiv:1912.07398*, 2019. 2, 3

[30] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, 2013. 2

[31] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019. 2

[32] Faisal Kamiran and Toon Calders. Classification with no discrimination by preferential sampling. In *Machine Learning Conference Belgium and The Netherlands*, 2010. 2

[33] Niharika Jain, Alberto Olmo, Sailik Sengupta, Lydia Manikonda, and Subbarao Kambhampati. Imperfect imaganation: Implications of gans exacerbating biases on facial data augmentation and snapchat selfie lenses. *arXiv preprint arXiv:2001.09528*, 2020. 2

[34] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[35] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8), 2017. 2, 3

[36] Bahram K Baloch, Sateesh Kumar, Sanjay Haresh, Abeerah Rehman, and Tahir Syed. Focused anchors loss: cost-sensitive learning of discriminative features for imbalanced classification. In *Asian Conference on Machine Learning*, 2019. 2, 3

[37] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2019. 3, 4, 5, 6, 7, 8

[38] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Conference on AI, Ethics, and Society*, 2019. 3

[39] Rachel KE Bellamy, Kuntal Dey, Hind, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018. 3

[40] Nisha Srinivas, Karl Ricanek, Dana Michalski, David S Bolme, and Michael King. Face recognition algorithm bias: Performance differences on images of children and adults. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 3

[41] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2018. 3, 4, 5, 6, 7, 8

[42] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 2018. 4, 5, 6, 7

[43] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 4, 5, 7, 8

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 5, 6, 7, 8

[45] Lanqing He, Zhongdao Wang, Yali Li, and Shengjin Wang. Softmax dissection: Towards understanding intra-and inter-clas objective for embedding learning. *arXiv preprint arXiv:1908.01281*, 2019. 5

[46] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. 2008. 6, 7

[47] Tao Hu and Honggang Qi. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *arXiv preprint arXiv:1901.09891*, 2019. 6