Towards Explaining Demographic Bias through the Eyes of Face Recognition Models

Biying Fu[1,2], Naser Damer[1,2]
[1]Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2]Department of Computer Science, TU Darmstadt, Darmstadt, Germany
biying.fu@igd.fraunhofer.de

Abstract

Biases inherent in both data and algorithms make the fairness of widespread machine learning (ML)-based decision-making systems less than optimal. To improve the trustfulness of such ML decision systems, it is crucial to be aware of the inherent biases in these solutions and to make them more transparent to the public and developers. In this work, we aim at providing a set of explainability tool that analyse the difference in the face recognition models' behaviors when processing different demographic groups. We do that by leveraging higher-order statistical information based on activation maps to build explainability tools that link the FR models' behavior differences to certain facial regions. The experimental results on two datasets and two face recognition models pointed out certain areas of the face where the FR models react differently for certain demographic groups compared to reference groups. The outcome of these analyses interestingly aligns well with the results of studies that analyzed the anthropometric differences and the human judgment differences on the faces of different demographic groups. This is thus the first study that specifically tries to explain the biased behavior of FR models on different demographic groups and link it directly to the spatial facial features. The code is publicly available here[1].

1. Introduction

The performance and accuracy of automated Face Recognition (FR) systems have been boosted lately due to the advances made in deep-learning [31, 33, 32, 3] and large-scaled training face image datasets [15, 30, 42, 21]. Both algorithmic improvement and large-scaled datasets have contributed to the boom of FR systems being applied in diverse application areas. However, research revealed a bias problem in FR Systems. The face recognition vendor test (FRVT) in 2002 [23] and later in 2019 [13] showed that recognition accuracy differs between demographic groups. Among tested FR solutions, some algorithms perform well on Caucasians, while showing less superior performance on other demographics. This biased behavior of the FR solutions causes some problems in diverse applications. While it is less sensitive to make more failure verification on unlocking personal devices, it is more problematic to falsely identify a person as a criminal. Thus bias funds mistrust in the use of biometric recognition systems both by the scientific communities and the public. Therefore, understanding the biases and making them more transparent could instill trust, fairness, and security into the biometric systems. More importantly, it can help develop new solutions that are specifically designed to be fair.

To contribute toward explaining the demographic bias of FR Models, we propose a set of explainability tools. We show that the average activation mappings of different demographic groups are extremely similar and thus do not reflect the bias. Based on that, we take our explainability tool to a higher derivative of these maps by analyzing the difference (between demographic groups) in the variations in these activation maps. After motivating our analyses with Fairness analyses on two FR models and two datasets, we demonstrated our explainability pipelines on these models and datasets. Our analyses on gender differences and ethnic differences pointed out certain regions of the face where the FR models behave differently in comparison to a reference demographic group. The results are largely consistent across FR models and datasets. The results, very interestingly, were consistent with findings in previous studies on facial anthropometric differences and on the human judgment on gender from faces [47, 6]. This is thus the first attempt to explain the differences in the FR models' behavior on different demographic groups. To achieve that, this work provides explaination tools of the FR model behaviour towards a group of samples rather than single samples in the more conventional explainability tools.

[1]https://github.com/fbiying87/Demographic-Bias-Visualization.git

2. Related works

The biases inherently built into ML-based systems are a matter of concern, whether data-dependent or human-coded. These biases can be introduced either through data [46, 11], historical prejudices [22], or other proxies[29] which supposedly to be fair representations of the latent bias factors. For example, taking zip codes as a proxy could also include social or ethical bias.

Focusing on fairness/bias in face recognition, there are several works [10, 24, 34] pointing out that FR algorithms suffer from the "own-race bias" or the "other-race effect". Drozdowski et al. in [10] pointed out that this effect is visible in most FR algorithms developed in different countries. In [24] it is shown, that FR algorithms developed in Asia perform better in recognizing Asian individuals, while solutions developed in Europe perform well on Caucasians. Algorithmic bias in FR tends to over-perform on majority groups and thus making ethnicity a co-variate in the model. But also with balanced ethnicity data, the FR algorithms do not perform equally on all demographic groups, as shown in [18]. The performance of some groups is still inferior to other groups. This observation gives indications of the inherent and non-quantifiable characteristics of bias. These biases have been shown to extend to other demographic variations, end even non-demographic ones such as personal styling choices [37].

Methods to mitigate these demographic biases are proposed both from the algorithmic viewpoints or data perspective [35, 41, 46, 36]. Wang proposed in [41] a Meta-Learning approach to combat the algorithmic bias. The network tries to learn adaptive margins in the latent space for the model to be optimized and perform fairly across people of different skin tones. Later in [39], Wang et al. used reinforcement learning to optimize these adaptive margins. From the data perspective, there are works as in [46, 1]. In [46] a two-stream approach is used to learn discriminative face representation supervised by mining hard identities on long-tailed data. This iterative way of integrating hard samples from the tail data enables the network to learn through effective batch mining. Amini et al. [1] proposed an algorithm for mitigating bias during training by re-sampling the training data according to the automatically learned latent variables within the training stage. The idea is to select rarer data points more often. Other works target balancing the data by adding augmentation of adversarial data per-subject [44] or adding synthetic data to balance the ethnicity distribution [19].

Raising awareness of the bias issue both for the scientific community and the general public is already a start for building fairer and trustworthy solutions [26]. Cross-discipline collaboration of researchers and developers is a mjor requierment to enhance the ML fairness. A major effort in interpreting such phenomena in ML is the explainable artificial intelligence (XAI) program by the Defense Advanced Research Projects Agency [14], aiming at promoting innovation in AI in general, not only in privacy-related sectors but also in the fields of medical, finance and autonomous driving.

3. Methodology

Our explainability toolset is built upon the activation mappings (AM) of the FR solutions. These AMs are used to create heat maps for the input image, highlighting the important regions in terms of the network's output. However, from one side, these heat maps deviate largely between samples (due to variations in pose, expression, illumination, etc.), which makes making general conclusions on the effect of different demographic groups virtually pointless. On the other side, they are almost identical if averaged on large groups of samples, even if each sample represents a different demographic group, which limits their explainability utilization as will be shown later in this work. To avoid this, our explainability tools go beyond the base activation mappings into a higher derivative where one can notice statistical differences between groups of face samples. A similar concept has been lately used to derive reasoning for differences between face images of different qualities [12]. These explainability tools are presented in this section and demonstrated in Figure 1.

3.1. Activation CAM method

As a backbone of our explainability tool, we require a method to represent the special activation properties induced by a single sample in FR models. The AM visualization scheme used in this work is the Score-CAM proposed by Wang et al. in [38]. This method is designed to efficiently display visual explanations for CNNs. It re-weighted the final activation based on emphasizing the most relevant regions within each feature map according to the network's decision. These activation CAM methods surpass the inherent limitations in the gradient-based CAMs [28] and provide a more effective and faster way to calculate the salient map [25].

3.2. Our proposed explainability tools

Figure 1 illustrates the different processes that comprised our explainability tools below. For each input face image, the Score-CAM generates an output AM with respect to the two FR solutions. Each pixel value is denoted as $a_{i,j}$ with $\{i = 1 : 112, j = 1 : 112\}$. This saliency map is the up-sampled and re-weighted activation of the output feature layers according to the penultimate layer of the FR model. This penultimate layer is placed before the FC layer of the FR model. This layer is originally used to generate identity descriptors. For symmetry reasons, we also include the AM of the horizontally flipped image in our calculations.

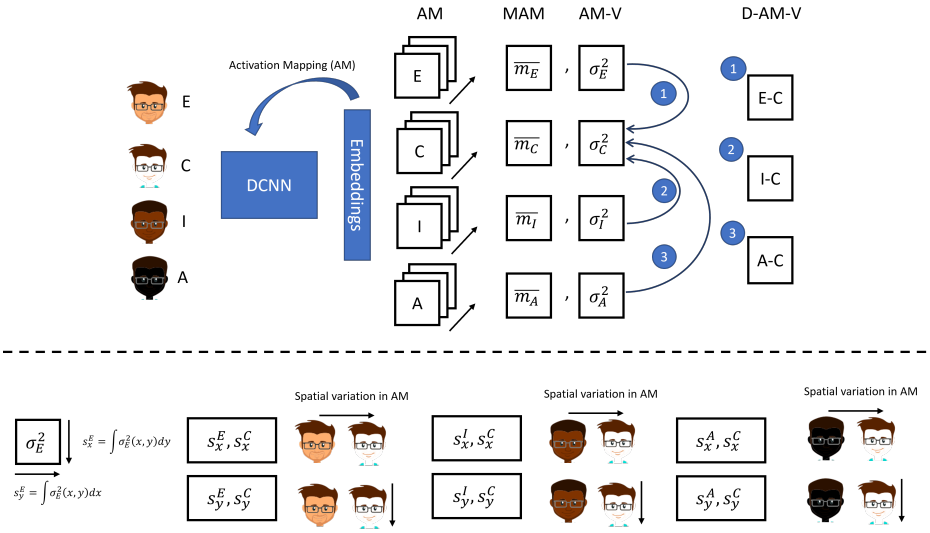

Figure 1. Pipeline illustrates the different processes in our proposed explainability tool. The considered demographic groups are E: East-Asian, C: Caucasian, I: Indian, and A: African. The same processes also apply to gender with males and females. It starts with the activation mapping (AM) from the embedding space for building group statistics using these AM maps. Then the mean activation mapping (MAM) and the activation mapping variation map (AM-V) are determined by building the group mean and group variation of the input AM maps per demographic group. Deviation from the demographic groups is always referred to Caucasian, as they proved (later in the paper) to lead to the highest FR performance. Differential activation mapping variation (D-AM-V) and the spatial variation map consider the spatial differentiation in activation variation between non-Caucasian and Caucasians. We apply a similar pipeline between the gender groups, female and male.

The exact FR models used in this work will be presented in Section 4.2.

We introduce the mean activation mappings (MAM) for each of the demographic groups. We denote them as $\text{MAM}_{dg/g}$ with $dg = \{E, C, I, A\}$ for demographic groups which include the Asian, Caucasian, Indian, and African and $g = \{m, f\}$ for gender and includes males and females. Each element in the MAM is denoted as $\overline{a_{i,j}}$ and it is derived from Eq. (1) using the activation value of each single AM $a_{i,j}$:

$$\overline{a_{i,j}} = \frac{1}{N}\sum_{k=1}^{N} a_{i,j}^{k}, \tag{1}$$

where N is the number of images within each demographic per database. The MAM will be shown later (in Section 5) to have no comprehensive pattern to study the FR behavior differences between demographic groups.

To take our explainability tool to a space where we can expect higher order differences between demographic groups, we measure the variability in the AM, resulting in the activation mapping variation map (AM-V) as higher order analysis, where each pixel is denoted as $\sigma_{i,j}^2$ and is determined in Eq. (2):

$$\sigma_{i,j}^2 = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(a_{i,j}^{k} - \overline{a_{i,j}})^2}, \tag{2}$$

, where N is the number of samples in each demographic per-database and $a_{i,j}$ is the element of the individual AM. AM-V thus aims at spatially showing the degree of variation in the activation of a group of samples.

MAM and AM-V look into the spatial areas where either a high activation or a relatively large variation in the activation of the FR occur, respectively. However, as we will see later, the differences between the MAM of images per demographic group do not reveal a lot of explainability information. Therefore, to uncover the spatially related difference between these demographic groups, we need to analyze the differences between the variations of activation in higher derivatives. We introduce the term Differential activation mapping variation (D-AM-V) as in Eq. (3)

$$\text{D-AM-V} = |\text{AM-V}_{dg,1} - \text{AM-V}_{dg,2}|, \tag{3}$$

where the term is calculated between two different demographic groups. Due to the symmetry constraint, this mapping is further mirrored and averaged to enhance the left-and-right symmetry of a face image. The D-AM-V thus graphically highlights the facial areas where two different demographic groups have large differences in their activation variations.

To further enable easier conclusions from the AM-V maps, we further integrate the AM-V map along both x- and y-direction to illustrate the spatial variations along the horizontal and vertical face axes by Eq. (4) and (5), namely the

spatial-variation-x (s_x^{dg}) and spatial-variation-y (s_y^{dg}). Using this measure of two demographic groups further provides locations with higher activation variation differences between them.

$$s_x^{dg} = \int_1^{112} \text{AM-V}(x, y)\, dy \tag{4}$$

$$s_y^{dg} = \int_1^{112} \text{AM-V}(x, y)\, dx \tag{5}$$

The details of the considered pairs of demographic groups will be discussed in more detail in the next section.

4. Experimental Setup

This section provides an overview of our experimental setup in terms of the ethnicity and gender-balanced face datasets, fairness evaluation metrics, considered FR models, and the investigated demographic differences.

4.1. Database

We performed our experiment on two publicly available face datasets especially designed for validating the demographic bias in FR algorithms. As opposed to other large-scaled face image datasets with heavily unbalanced and long-tailed distributions, these two datasets have a balanced number of subjects in each of the four ethnicity groups included.

Robinson et al. proposed the Balanced Faces in the Wild (BFW) in [27]. The data consists of four different ethnic groups (Asian, Black, Indian, and White). Each ethnicity is further split into two subgroups of balanced males and females. Each subgroup has 25 faces of 100 subjects and aggregated to 20K faces in total. This dataset is used both for the investigation of ethnicity bias and gender bias, where we combined all female and male subjects across all ethnicity groups. Five folds cross-validation is used in the BFW dataset, with in total more than 920K pairs of 240K genuine and 680K imposter comparisons. Based on these comparison pairs, we separate them further into Caucasian-Caucasian, Asian-Asian, African-African, and Indian-Indian pairs, as well as the female-female and male-male pairs.

The Racial Faces in-the-wild (RFW) in [40] also consists of four testing ethnicity groups, namely Caucasian, Asian, Indian, and African. Each subset contains around 10K images of 3K individuals. In RFW, the images are carefully balanced and cleaned. As no gender labels are provided for this dataset, we only use this dataset for the investigation of FR models on ethnicity differences. RFW dataset composes of 6000 pairs of equal genuine (3000) and imposter (3000) pairs for each ethnicity (Caucasian-Caucasian, Asian-Asian, African-African, and Indian-Indian pairs), which makes in a total of 24K pairs of genuine and imposter comparisons.

4.2. Face recognition models

Our experiments are performed on two FR models. Both FR models are trained with ArcFace loss [9] and publicly released by their creators [2]. Both backbones of the FR models are based on the ResNet architecture [16] of different scales. The larger backbone is the ResNet-100, which has deeper middle layers for the feature extraction compared to the smaller backbone with ResNet-50. This selection is motivated by the effect of the model scale on ML bias pointed out in [17]. We chose the ArcFace r100 model, as the solution shows high performance on face identification accuracy, across substantial changes in viewpoint, illumination, expression, and quality [9, 20]. We base our investigations on FR models of different scales to further show the behavioral patterns of demographic fairness across models of different scales.

To match the input expected by the FR models, each face image is first cropped and aligned (similarity transfer) using MTCNN [45, 9] into standardized face images of 112×112 pixels. For both ArcFace models, the output layer *net.layer4* is used for activation mapping. To mitigate the non-symmetry issue, we include both the horizontally flipped version of the input image and its original version for the activation mapping.

4.3. Fairness evaluation metrics

To measure the fairness of the FR models, as a motivation for our explainability efforts, we adopt the Fairness discrepancy ratio (FDR) proposed in [8]. The FDR takes both verification errors, namely the false match rate (FMR) and the false non-match rate (FNMR) into consideration for a given decision threshold. A biometric verification system is said to be fair only if, at a given decision threshold, statistical equality can be achieved for all pairs of demographic groups in terms of both FMR and FNMR. A higher FDR value indicates a fairer behavior between two demographic groups. The equation of FDR is given by Eq.

$$FDR(\tau) = 1 - (\alpha A(\tau) + (1 - \alpha)B(\tau)), \tag{6}$$

where the term $A(\tau)$ and $B(\tau)$ are the two premises in [8] considering both the FMR and FNMR measures. The equations are $A(\tau) = max(|\text{FMR}^{dg_i}(\tau) - \text{FMR}^{dg_j}(\tau)|) \leq \epsilon$ and $B(\tau) = max(|\text{FNMR}^{dg_i}(\tau) - \text{FNMR}^{dg_j}(\tau)|) \leq \epsilon$, where dg_i, dg_j are each from one demographic group. α is set to 0.5 in our experiments, giving both error types an equal contribution to FDR. ϵ is a relaxation constraint that puts a limit of when to consider a system "fair" [8], which we set in our analyses to zero.

[2] https://github.com/deepinsight/insightface

(a) FDR on RFW over 4 ethnicity (b) FDR on BFW over 4 ethnicity (c) FDR on BFW gender

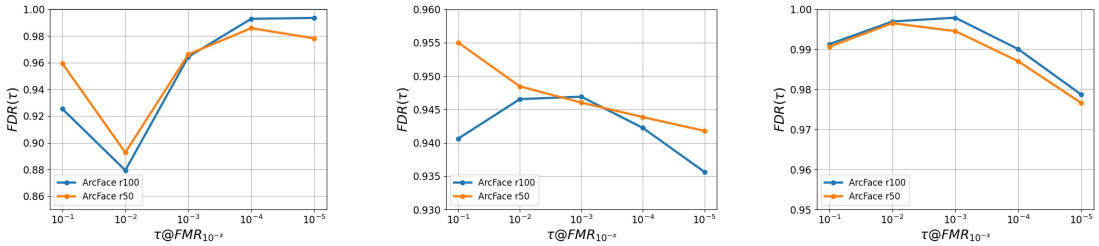

Figure 2. Fairness Discrepancy Rate (FDR) of two FR models over multiple decision thresholds τ to compare the considered FR models over ethnicity and gender groups.

	RFW						BFW					
	ArcFace r100			ArcFace r50			ArcFace r100			ArcFace r50		
$\tau@FMR_{10^x}$	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}
Demographic	FMR			FMR			FMR			FMR		
Caucasian	1.5E-1	1.0E-2	1.6E-3	1.9E-1	1.5E-2	0.7E-3	0.7E-2	0.4E-3	0.1E-4	0.9E-2	0.6E-3	0.4E-4
Asian	3.8E-1	8.0E-2	12E-3	4.1E-1	7.4E-2	9E-3	2.6E-2	3.6E-3	3.9E-4	1.9E-2	2.6E-3	3.1E-4
African	4.2E-1	9.0E-2	17E-3	5.1E-1	10E-2	13E-3	1.7E-2	1.8E-3	2.2E-4	1.9E-2	2.2E-3	2.1E-4
Indian	3.9E-1	8.1E-2	11E-3	4.6E-1	9.0E-2	8E-3	2.2E-2	2.1E-3	1.8E-4	2.3E-2	2.4E-3	2.4E-4
	FNMR			FNMR			FNMR			FNMR		
Caucasian	0.7E-3	0.005	0.008	0.003	0.013	0.043	0.040	0.057	0.090	0.044	0.073	0.112
Asian	1.3E-3	0.004	0.012	0.004	0.022	0.063	0.127	0.160	0.205	0.137	0.178	0.224
African	0.7E-3	0.002	0.005	0.002	0.011	0.037	0.085	0.109	0.136	0.090	0.121	0.155
Indian	0.7E-3	0.003	0.006	0.003	0.014	0.044	0.082	0.105	0.131	0.087	0.119	0.151
FDR	0.879	0.964	0.992	0.892	0.966	0.985	0.946	0.946	0.942	0.948	0.946	0.943
FDR AUC	0.949			0.953			0.943			0.947		

Table 1. FNMR(τ), FMR(τ), and FDR(τ) are given per demographic group, where the operational points are defined as τ at FMR_x. It is to note that τ is set using the entire test dataset as a global threshold.

To better compare the two FR models, we plot the FDR as a function of an operational threshold τ. The global threshold τ is determined on the entire test dataset from all demographics following [8]. For the FDR curve, we vary τ, the global threshold that results in an FMR of 10^{-1} to 10^{-5} in 5 steps, the τ will be noted by the FMR threshold is calculated. We also provide the area under the curve of FDR (within the same range) as another measure to assess the fairness of a certain FR model.

4.4. Investigation

The experiments are designed to address the two main demographic variations of our study (1) ethnic differences, and (2) gender differences.

Before introducing the findings of the research scope, we first motivate the need by looking at the demographic fairness issue in both considered FR models in terms of verification performance and FDR metric. Then, we applied our proposed explainability tools to the BFW and RFW datasets using both FR models of different scales to back-propagate the network's decision via activation mapping on the input data. Both FR models have the same DCNN-based backbone ResNet-100 and RestNet-50. We build our analysis on these sets of activation mappings. As we only have gender labels for the BFW dataset, the experiments on gender bias addressing the second aspect are only conducted on the BFW dataset. Similarly, we applied our chain of tools to the gender-balanced dataset to draw implications on the network's commonality on the gender aspect.

As both FR solutions are trained on face datasets with Caucasian males as the majority class and as will be demonstrated, perform the best on Caucasians and males, the experiments conducted in this study compare the demographic groups against the Caucasian group when considering demographics as a reference, and the female group against the male group as a reference when considering gender.

5. Results

In this section, we discuss the results of the two explainability aspects (ethnic and gender differences) using our presented tools with respect to both considered FR models.

5.1. Demographic fairness

Demographic fairness requires the automatic FR algorithms to perform equally on all different demographic groups for any τ. However, recent studies [13, 37] show that depending on the underlying FR model, the system does not perform equally well for all ethnic groups.

Here, to build a basis for our explainability analyses, we analyse the fairness of the considered FR models. We first discuss the fairness in demographics in terms of verification performance comparison for our datasets. Table 1 shows that both FR models produce different performances for different ethnicity groups. For the BFW benchmark, both FR models performed the best (in terms of FMR and FNMR)

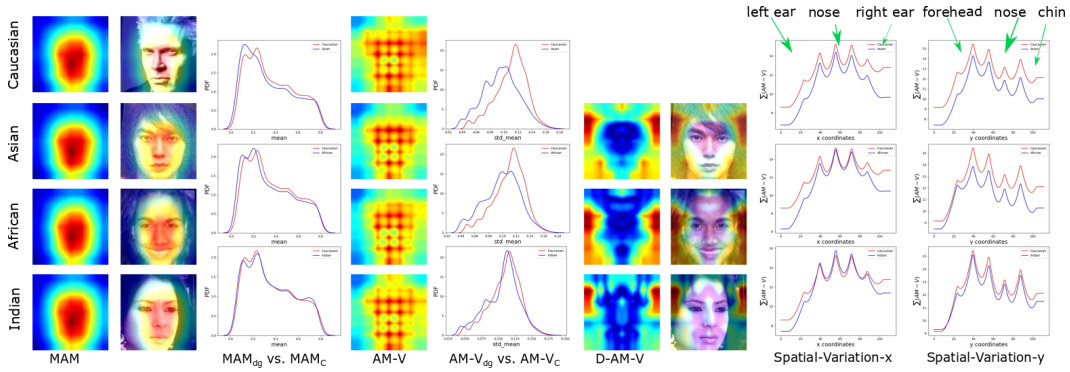

Figure 3. The Explainability tool is applied to all four ethnic groups in the BFW dataset, using ArcFace r100 as the FR model. The 1st column shows the MAM of each ethnic group, while the 2nd column shows the MAM map superimposed on a sample image of each group. The 3rd column shows the distribution of the values of the MAM map of two groups. The 4th column is the AM-V. The 5th column visualizes the distribution of AM-V values in two demographic groups. The D-AM-V maps in column 6 and overlaid with sample images in column 7, showing the main areas that trigger different behavior in the FR models. The last two columns, 8 and 9, summarize the AM-V by summing its values on the horizontal and vertical face axes for two different demographic groups. Higher values indicate higher local variation in activation for this ethnic group.

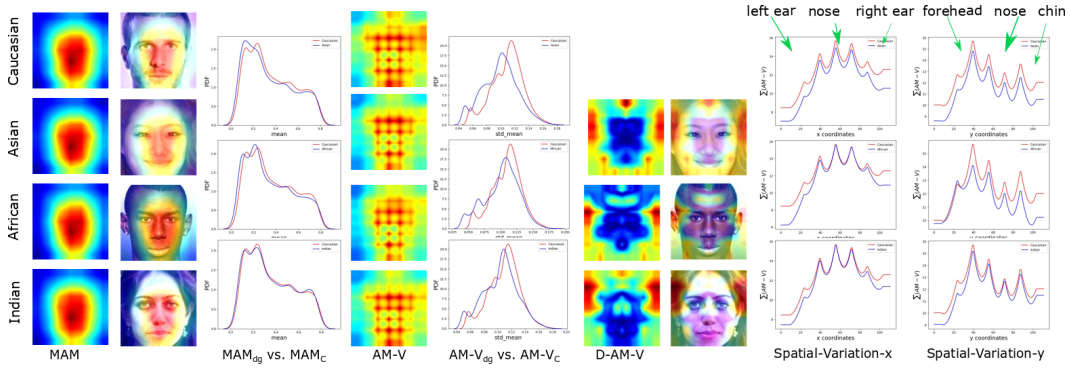

Figure 4. The Explainability tool is applied to all four ethnic groups in the RFW dataset, using ArcFace r100 as the FR model. The different columns follow the explanation in Figure 3 caption.

on faces from the Caucasian demographic group. Other ethnicity groups in BFW do perform worse than the Caucasian group by scoring higher FNMR and FMR values at most global operational thresholds. Similar conclusion can be made from the RFW benchmarks, especially from the FMR values, with the FNMR is slightly less consistent due to the lower number of the genuine pairs (lower statistical significance) in the RFW benchmark in comparison to BFW (See Section 4.1). In general, ArcFace r100 outperforms the smaller model ArcFace r50, as expected, in most experimental settings on both the BFW and RFW benchmarks.

Fairness, or rather the lack of it, is observed for both FR models on both datasets by looking at Figure 2 (a) and (b), as well as the FDR and FDR AUC values in Table 1. The FDR score varies widely over a wide range of τ in the RFW dataset. For BFW dataset, the FDR values are generally slightly higher for the smaller ArcFace r50 model, as seen in Figure 2 (b). However, the FDR AUC specifies the smaller model ArcFace r50 as having slightly higher fairness towards ethnicity groups compared to ArcFace r100 in both datasets.

In Table 2, we see that males perform slightly better compared to females in terms of FNMR and FMR across multiple thresholds, indicating some form of inherent gender bias in both FR models. Now, looking at Figure 2(c), the FDR curve shows the slightly higher gender fairness of the larger ArcFace r100 model.

In summary, both FR models have consistent performance trends and less-than-perfect fairness in both the gender and ethnicity demographic groups. Thus, explaining the differences in the FR model's reaction to these groups, and the consistency of this explainability, is highly relevant to understanding their behaviour.

5.2. Explainability of ethnic differences in FR

In Figures 3, 4, 5 and 6, the MAM of different ethnicity groups are visually very similar. This goes as well to the histogram of the MAM values comparison between the

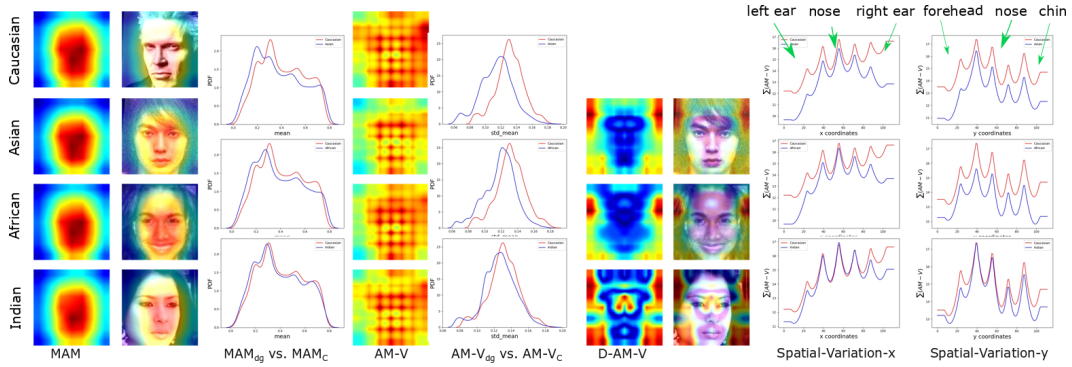

Figure 5. The Explainability tool is applied to all four ethnic groups in the BFW dataset, using ArcFace r50 as the FR model. The different columns follow the explanation in Figure 3 caption.

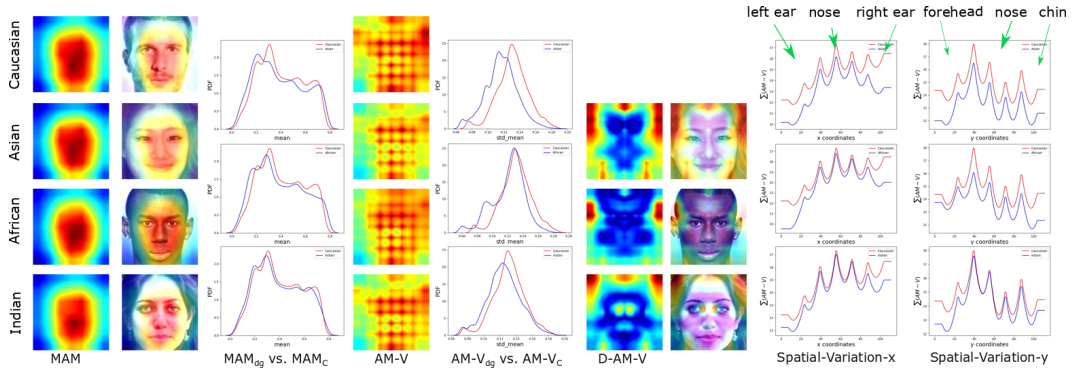

Figure 6. The Explainability tool is applied to all four ethnic groups in the RFW dataset, using ArcFace r50 as the FR model. The different columns follow the explanation in Figure 3 caption.

	ArcFace r100			ArcFace r50		
$\tau @ FMR_1 0^x$	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}
Gender	FMR			FMR		
Males	1.1E-2	1.1E-3	1.2E-4	1.1E-2	1.1E-3	1.1E-4
Females	1.6E-2	1.8E-3	1.7E-4	1.6E-2	1.8E-3	1.8E-4
	FNMR			FNMR		
Males	0.085	0.106	0.131	0.089	0.118	0.147
Females	0.083	0.110	0.151	0.091	0.128	0.173
FDR	0.997	0.998	0.990	0.996	0.994	0.987
FDR AUC	0.992			0.990		

Table 2. FNMR(τ), FMR(τ), and FDR(τ) are given per gender as subgroup, where the operational points are defined as τ at FMR_x. It is to note that τ is set using the entire test dataset, due to missing development set.

ethnic groups. This inability to see differences in the FR model's reaction to different groups, which is expected due to the demonstrated lack of fairness, is the main motivation behind our explainability tools. Rather than searching for differences in the model activation maps, we look for differences in the variation of these activations, thus the AM-V and its derivative, the D-AM-V.

The D-AM-V reveals better the spatially related difference between these demographic groups. Local areas with a higher difference in the activation variation indicate a higher difference between the way FR models see different ethnic groups. As indicated by the pipeline in Figure 1, we build our investigations always to the Caucasian group as reference. Figure 3, 4, 5 and 6 show the same shape of the D-AM-V maps for all demographics sets (E-C, A-C, and I-C) over two different datasets within the same FR model. D-AM-V maps demonstrate for Indian strong differences on the nose and outer eye corners, while for Africans the focus lies around the mouth, chin, and forehead, and for Asian on the cheeks and forehead area. Mapping the D-AM-V along the x- and y- direction show a higher difference in the activation variation between Caucasians (red) than non-Caucasians (blue) ethnic groups. Looking at the s_y maps (in E-C, A-C, and I-C) for the forehead, nose, and chin areas, the same findings as before can be obtained while observing large gaps in the cheek areas for the Asian group, between the mouth-chin area in Africans, and between the nose region in Indians. In general, Asians show higher D-AM-V values, which is probably related to them scoring some of the worse verification performances across ethnicity groups on both FR models (see Table 1).

These observations are rather consistent to a large degree on both FR models and both databases (see Figures 3, 4, 5 and 6). However, comparing the AM-V distributions across non-Caucasians and Caucasians for BFW in Figure 3 and Figure 5, one sees a stronger variation in ArcFace r50 than

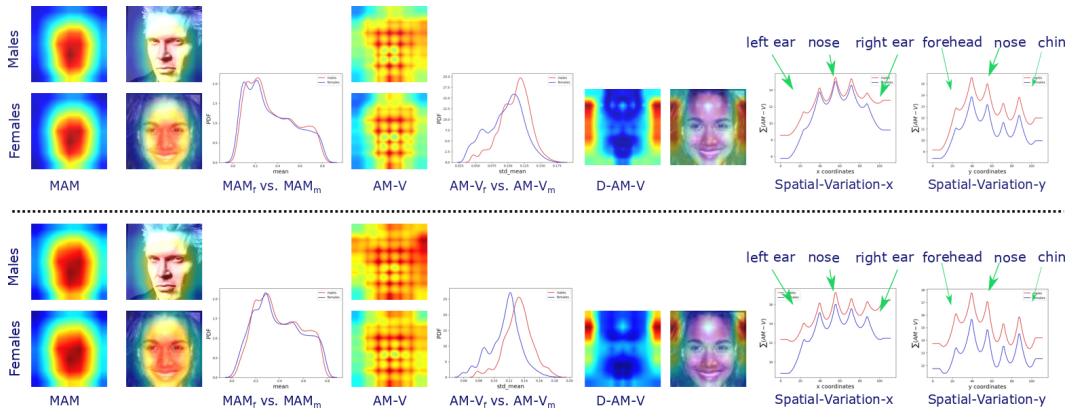

Figure 7. The Explainability tool is applied to both gender demographic groups in the BFW dataset, using ArcFace r100 (upper) and ArcFace r50 (below) as the FR model. The different columns follow the explanation in Figure 3 caption.

in ArcFace r100, suggesting that a smaller backbone causes stronger variations in activation, which can be related to its general lower verification performance. The same trend is also observed for RFW across ArcFace r100 and ArcFace r50 in Figure 4 and Figure 6.

In summary, the observations made on the most important areas of difference in the activation variation between the different ethnic groups and the Caucasian group are to a large degree consistent for both FR models and datasets. Different demographics show characteristic patterns in terms of the D-AM-V map in comparison to Caucasians highlighting specific areas of interest across demographic groups. In a study on the facial anthropometric differences among ethnicity [47], the authors pointed out that the chin arc and sub-nasal arc (mouth region) were on the top in terms of average change between African Americans and Caucasians, which is consistent with our observations through the eyes of FR models and confirms the validity of our explainability tools. Unfortunately, the study did not include information about the Indian and Asian groups.

5.3. Explainability of gender differences in FR

Figure 7 shows the group characteristics of the AM maps for the gender aspect. Only considering the shape of the MAM, no clear deviation is visible between both genders. However, when we focus on higher-order analysis, such as the D-AM-V, we observe a differentiation in the forehead and chin regions. This is rather consistent for both FR models. Very interestingly, in a study that analyse the facial areas that mostly affect the human judgment on the gender of the face [6], the chin and brow (forehead) areas came on top. Which, to some degree points out the sanity of our proposed explainability tools. Additional study on the facial anthropometric differences among genders [47] pointed out that the chin arc and frontal arc (forehead) were on the top in terms of average change between females and males' facial measurements.

One limitation of our study is the number of experimental variations that can fit in such a work. We chose to build variations in the FR model architecture and the dataset to avoid "bias" outcomes in these regards. However, interesting questions regarding the explainability tool's outcome across FR training losses [9, 3, 4], network architectures [16, 2, 43, 5], FR training datasets [15, 7], the set of analysed demographics (or even non-demographic) variations [37], the combined analyses of demographic groups (e.g. African females in comparison to Indian males), and the pairing of the compared demographic groups (we chose the top performer as a reference here), are yet to be explored.

6. Conclusion

In this work, we aimed at explaining the difference in the perspective of FR models between different demographic groups. Towards that, we presented a set of explainability tools visualizing the ethnic and gender differences for the underlying FR models. In general, both considered FR models show ethnic bias in both datasets in terms of unequal verification performance in different demographic groups. Our tools and analyzing the results on two datasets and two FR models pointed out certain regions that might cause the FR model's behavior differences between certain ethnic groups and the Caucasian ethnicity on one hand, and between males and females on the other hand. Interestingly, the outcome is, to a large degree, consistent with the available clues from facial anthropometric differences studies and studies on the human judgment of gender from faces.

Acknowledgements: This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

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

[2] F. Boutros, N. Damer, M. Fang, F. Kirchbuchner, and A. Kuijper. Mixfacenets: Extremely efficient face recognition networks. In *International IEEE Joint Conference on Biometrics, IJCB 2021, Shenzhen, China, August 4-7, 2021*, pages 1–8. IEEE, 2021.

[3] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, Louisiana, USA, June 19-24, 2022*. Computer Vision Foundation / IEEE, 2022.

[4] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper. Self-restrained triplet loss for accurate masked face recognition. *Pattern Recognit.*, 124:108473, 2022.

[5] F. Boutros, P. Siebke, M. Klemt, N. Damer, F. Kirchbuchner, and A. Kuijper. Pocketnet: Extreme lightweight face recognition network using neural architecture search and multi-step knowledge distillation. *IEEE Access*, 10:46823–46833, 2022.

[6] V. Bruce, A. M. Burton, E. Hanna, P. Healey, O. Mason, A. Coombes, R. Fright, and A. Linney. Sex discrimination: How do we tell the difference between male and female faces? *Perception*, 22(2):131–152, 1993. PMID: 8474840.

[7] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pages 67–74. IEEE Computer Society, 2018.

[8] T. de Freitas Pereira and S. Marcel. Fairness in biometrics: A figure of merit to assess biometric verification systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(1):19–29, 2022.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE, 2019.

[10] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020.

[11] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper. Demographic bias in presentation attack detection of iris recognition systems. In *28th European Signal Processing Conference, EUSIPCO 2020, Amsterdam, Netherlands, January 18-21, 2021*, pages 835–839. IEEE, 2020.

[12] B. Fu and N. Damer. Explainability of the implications of supervised and unsupervised face image quality estimations through activation map variation analyses in face recognition models. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV - Workshops, Waikoloa, HI, USA, January 4-8, 2022*, pages 349–358. IEEE, 2022.

[13] P. Grother, M. Ngan, and K. Hanaoka. *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology, 2019.

[14] D. Gunning and D. Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.

[15] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world. *Electronic imaging*, 2016(11):1–6, 2016.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[17] S. Hooker, N. Moorosi, G. Clark, S. Bengio, and E. Denton. Characterising bias in compressed models. *CoRR*, abs/2010.03058, 2020.

[18] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

[19] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[20] S. Mallick, G. Jeckeln, C. J. Parde, C. D. Castillo, and A. J. O'Toole. The influence of the other-race effect on susceptibility to face morphing attacks. *arXiv preprint arXiv:2204.12591*, 2022.

[21] D. Miller, I. Kemelmacher-Shlizerman, and S. M. Seitz. Megaface: A million faces for recognition at scale. *CoRR*, abs/1505.02108, 2015.

[22] R. Oda. Biased face recognition in the prisoner's dilemma game. *Evolution and human behavior*, 18(5):309–315, 1997.

[23] P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002. In *2003 IEEE International SOI Conference. Proceedings (Cat. No. 03CH37443)*, page 44. IEEE, 2003.

[24] P. J. Phillips, F. Jiang, A. Narvekar, J. H. Ayyad, and A. J. O'Toole. An other-race effect for face recognition algorithms. *ACM Trans. Appl. Percept.*, 8(2):14:1–14:11, 2011.

[25] H. G. Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 983–991, 2020.

[26] C. Rathgeb, P. Drozdowski, N. Damer, D. C. Frings, and C. Busch. Demographic fairness in biometric systems: What do the experts say? *CoRR*, abs/2105.14844, 2021.

[27] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, pages 0–1, 2020.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations

from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[29] I. Serna, A. Peña, A. Morales, and J. Fierrez. Insidebias: Measuring bias in deep networks and application to face gender biometrics. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3720–3727, 2021.

[30] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database. In *5th IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2002), with CD-ROM, 20-21 May 2002, Washington, D.C., USA*, pages 53–58. IEEE Computer Society, 2002.

[31] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1988–1996, 2014.

[32] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10, 000 classes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1891–1898. IEEE Computer Society, 2014.

[33] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1701–1708. IEEE Computer Society, 2014.

[34] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. In *2020 IEEE International Joint Conference on Biometrics, IJCB 2020, Houston, TX, USA, September 28 - October 1, 2020*, pages 1–11. IEEE, 2020.

[35] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognit. Lett.*, 140:332–338, 2020.

[36] P. Terhörst, M. L. Tran, N. Damer, F. Kirchbuchner, and A. Kuijper. Comparison-level mitigation of ethnic bias in face recognition. In *8th International Workshop on Biometrics and Forensics, IWBF 2020, Porto, Portugal, April 29-30, 2020*, pages 1–6. IEEE, 2020.

[37] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. M. Moreno, J. Fierrez, and A. Kuijper. A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, 3(1):16–30, 2022.

[38] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020.

[39] M. Wang and W. Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9322–9331, 2020.

[40] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 692–702, 2019.

[41] M. Wang, Y. Zhang, and W. Deng. Meta balanced network for fair face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[42] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 529–534. IEEE Computer Society, 2011.

[43] M. Yan, M. Zhao, Z. Xu, Q. Zhang, G. Wang, and Z. Su. Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 2647–2654. IEEE, 2019.

[44] S. Yucer, S. Akçay, N. Al-Moubayed, and T. P. Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020.

[45] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503, 2016.

[46] Y. Zhong, W. Deng, M. Wang, J. Hu, J. Peng, X. Tao, and Y. Huang. Unequal-training for deep face recognition with long-tailed noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7812–7821, 2019.

[47] Z. Zhuang, D. Landsittel, S. Benson, R. Roberge, and R. Shaffer. Facial Anthropometric Differences among Gender, Ethnicity, and Age Groups. *The Annals of Occupational Hygiene*, 54(4):391–402, 03 2010.