A Comprehensive Study on Face Recognition Biases Beyond Demographics

Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, *Member, IEEE*, Naser Damer, *Member, IEEE*, Aythami Morales Moreno, Julian Fierrez, *Member, IEEE*, and Arjan Kuijper

***Abstract*—Face recognition (FR) systems have a growing effect on critical decision-making processes. Recent works have shown that FR solutions show strong performance differences based on the user's demographics. However, to enable a trustworthy FR technology, it is essential to know the influence of an extended range of facial attributes on FR beyond demographics. Therefore, in this work, we analyze FR bias over a wide range of attributes. We investigate the influence of 47 attributes on the verification performance of two popular FR models. The experiments were performed on the publicly available MAAD-Face attribute database with over 120M high-quality attribute annotations. To prevent misleading statements about biased performances, we introduced control group-based validity values to decide if unbalanced test data causes the performance differences. The results demonstrate that also many nondemographic attributes strongly affect recognition performance, such as accessories, hairstyles and colors, face shapes, or facial anomalies. The observations of this work show the strong need for further advances in making the FR system more robust, explainable, and fair. Moreover, our findings might help to a better understanding of how FR networks work, enhance the robustness of these networks, and develop more generalized bias-mitigating FR solutions.**

***Index Terms*—Bias estimation, bias, biometrics, face recognition (FR), fairness, performance differences, soft-biometrics.**

I. Introduction

LARGE-SCALE face recognition (FR) systems are spreading worldwide [13]. These systems have a growing effect on daily life [78] and are increasingly involved in critical decision-making processes, such as in forensics and law enforcement. However, recent works [4], [7], [24], [25], [49], [53], [63] showed that current FR solutions possess biases leading to discriminatory performance differences [59] based on the user's demographics [67], [74].

From a legal perspective, there are several regulations to prevent such discrimination, for instance, Article 7 of the Universal Declaration on Human Rights, Article 14 of the European Convention of Human Rights, or Article 71 of the general data protection regulation (GDPR) [77]. Driven by: 1) these legal efforts to guarantee fairness and 2) the findings that the performance of current FR solutions depends on the user's demographics, several approaches were proposed to mitigate demographics-bias in FR technologies. This was achieved through adversarial learning [27], [42], [47], [48], margin-based approaches [35], [79], data augmentation [40], [80], [81], metric-learning [73], or score normalization [72]. However, the research focus on demographic-bias does only tackle a minor proportion of all possible discriminatory effects. Knowing the influence of an extended set of facial attributes on the FR performance will enable the development of accurate and less discriminatory FR systems.

In this work, we aim at investigating the FR bias based on a wide range of attributes beyond demographics. These biases might affect the fairness [59] but also the security of FR systems [32]. Biases can be identified as learning weakness to be exploited by users with malicious intentions (i.e., vulnerability attacks [52]). To be precise, we analyze the differential outcome as defined by Howard *et al.* [34] of two popular FR models (FaceNet [58] and ArcFace [17]) with regard to 47 attributes. The experiments are conducted on the recently published and publicly available MAAD-Face[1] annotation database [69] based on VGGFace2 [8]. It consists of over 120M high-quality attribute annotations for 3.3M face images. For the experiments, several decision thresholds are taken into account to cover a wide range of applications. To prevent misinterpretations of the results origin from testing data with: 1) unbalanced label distributions or 2) attribute correlations, we: 1) introduce control groups to derive a validity value for the recognition performance in the presence of a specific attribute and 2) analyze the pairwise correlations of the attribute annotations. While 1) allowing us to quantify results that arise from unbalanced testing data and prevent falsified statements about the attribute-related bias and 2) emphasize if an attribute bias might originate from a different (correlated) attribute. Besides a detailed analysis, we present a visual

Manuscript received March 1, 2021; revised July 14, 2021; accepted August 26, 2021. Date of publication September 10, 2021; date of current version March 16, 2022. This work was supported in part 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, and in part by the Projects BIBECA under Grant RTI2018-101248-B-I00 MINECO/FEDER and PRIMA under Grant H2020-MSCA-ITN-2019-860315. *(Corresponding author: Philipp Terhörst.)*

Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, and Arjan Kuijper are with the Department of Smart Living & Biometric Technologies, Fraunhofer Institute for Computer Graphics Research IGD, 64283 Darmstadt, Germany, and also with the Interactive Graphics Systems Group, Technical University of Darmstadt, 64289 Darmstadt, Germany (e-mail: philipp.terhoerst@igd.fraunhofer.de).

Aythami Morales Moreno and Julian Fierrez are with the Biometrics and Data Pattern Analytics Lab—BiDA Lab, Universidad Autonoma de Madrid, 28049 Madrid, Spain.

Digital Object Identifier 10.1109/TTS.2021.3111823

[1]https://github.com/pterhoer/MAAD-Face

summary that states the performance difference between samples with and without a specific attribute over the validity of the results. This aims to present the results in a compact and simply understandable manner.

The results support the findings of previous works stating that FR systems have to deal with demographic biases [60]. We differentiate between explicit demographic attributes, such as gender, age, or ethnicity and nonexplicit demographic attributes, such as accessories, hairstyles and hair colors, face shapes, or facial anomalies. However, we have to consider that some of the nonexplicit demographic attributes might be affected by implicit demographic covariates. For example, the hairstyle is highly affected by gender or ethnicity. The results demonstrate that also many of the nondemographic attributes strongly affect recognition performance. Investigating two FR models that differ only in the loss function used during training, we showed the effect of the underlying training principles on recognition. While the triplet-loss-based FaceNet model showed attribute-related differential outcomes that are relatively constant on several decision thresholds, the angular margin-based ArcFace model showed differential outcomes that are often dependent on the used decision threshold. Many performance differences affected by attributes could be explained through the attribute's relation to the visibility of a face, the temporal variability, and the degree of abnormality. However, our experiment also reveals many unconventional results that future work has to address. Our findings strongly motivate further advances in making recognition systems more robust against covariates [28], [44], explainable [5], [50], and fair [51], [60]. We hope that these findings help to develop robust and bias-mitigating FR solutions and also help to move forward bias-aware and bias-mitigating technology in other AI application areas.

II. Related Work

Algorithmic bias is a phenomenon that occurs when an algorithm produces systematic errors that create unfair outcomes for an individual or groups of individuals [22]. In biometrics, this can be measured in terms of differences in the recognition performance based on the user's attributes. The exact set of attributes that can be included under the biometric bias term is still an open discussion issue as concluded in the EAB demographic fairness in the biometric systems workshop [55]. The phenomena of bias in face biometrics were found in several disciplines, such as presentation attack detection [19], [33], the estimation of facial characteristics [15], [70], and the assessment of face image quality [71]. In some previous works, factors that affect the recognition performance were also known as covariates [44], [45]. However, this phenomenon was addressed as bias or fairness in more recent works, especially when they refer to sensitive attributes (e.g., demographic or cultural characteristics). In general, one of the main reasons for bias might be the induction of nonequally distributed classes in training data [35], [40], [61] that leads to differences in the recognition performance and, thus, might have an unfair impact, e.g., on specific subgroups of the population. Howard *et al.* [34] introduced the terms differential performance and outcome for classifying biometric performance differentials that separately consider the effect of false-positive and false-negative outcomes. They show that the often-cited evidence regarding biometric equitability has focused primarily on false negatives.

Previous works on bias in FR [18] mainly focused on the influence of demographics. However, Terhörst *et al.* [68] demonstrated recently that more (nondemographic) characteristics are stored in face templates that might have an impact on the FR performance. In the following, we will shortly discuss related works on estimating and mitigating bias in face biometrics. For a more complete overview, we refer to [18].

A. Estimating Bias in Face Recognition

In recent years, several works have been published that demonstrated the influence of demographics on commercial and open-sources FR algorithms. Studies [16], [46], [56], [66] analyzing the impact of age demonstrated a lower biometric performance on faces of children. Studies [2], [3], [61], [76] analyzing the effect of gender on FR showed that the recognition performance of females is weaker than the performance on male faces. Experiments without unbalanced data distributions and with an unbalanced toward female faces resulted in similar results [3]. In [2], experiments with a PCA-decomposition showed that female faces are intrinsically more similar than male ones. Research analyzing the impact of the user's ethnicity showed faces of ethnicities which were under-represented in the training process perform significantly weaker. The same was found for darker-skinned cohorts in general [41].

More recent studies [6], [9], [12], [26], [29], [34], [36], [39], [57], [65] focused on jointly investigating the effects of user demographics on FR. These studies showed that the effects lead to an exponential FR error increase when facing the same biased race, gender, and age factors [34]. Particular attention deserves the FR vendor test (FRVT) [29], a large-scale benchmark of commercial algorithms analyzing the FR performance with regards to demographics. They consistently elevated false positives for female subjects and subjects at the outer ends of the age spectrum. An overview of bias estimation in FR is shown in Table I.

B. Mitigating Bias in Face Recognition

The findings summarized in Section II-A motivated research toward mitigating demographic-bias in FR approaches. An early approach was presented by Zhang and Zhou [83] who formulated the face verification problem as a multiclass cost-sensitive learning task and demonstrated that this approach can reduce different kinds of faulty decisions of the system. In 2017, the range loss [82] was proposed to learn robust face representations that can deal with long-tailed training data. It is designed to reduce overall intrapersonal variations while enlarging interpersonal differences simultaneously. Recent works aimed at mitigating demographic-bias in FR through adversarial learning [27], [42], [47], [48], margin-based approaches [35], [79], data augmentation [40], [80], [81], metric-learning [73], or score normalization [72].

TABLE I
OVERVIEW OF RECENT WORKS ANALYZING BIAS IN FR. IDENTITIES AND IMAGES REFER TO THE USED TESTING DATA. IN CONTRAST TO PREVIOUS WORKS THAT ANALYZES SOME SPECIFIC DEMOGRAPHIC ATTRIBUTES, OUR WORK INVESTIGATES A LARGE RANGE OF DEMOGRAPHIC AND NONDEMOGRAPHIC ATTRIBUTES

Work	Identities	Images	Attributes (number classes)
Ricanek et al. [56]	0.7k	8.0k	Age (2)
Deb et al. [16]	0.9k	3.7k	Age (cont.)
Srinivas et al. [66]	1.7k	9.2k	Age (2)
Michalski et al. [46]	-	4.7M	Age (cont.)
Albiero et al. [2]	26.9k	151.6k	Gender (2)
Albiero et al. [3]	15.9k	101.3k	Gender (2)
Vera-Rodriguez et al. [76]	0.5k	169.4k	Gender (2)
Cavazos et al. [9]	0.4k	1.1k	Ethnicity (2)
Krishnapriya et al [41]	22.7k	3.3M	Gender (2), Ethnicity (2)
Serna et al. [60]	55k	1.4M	Gender (2), Ethnicity (4)
Acien et al. [1]	1.7k	13k	Gender (2), Ethnicity (3)
Hupont et al. [36]	0.6k	10.8k	Gender, Ethnicity (3)
Robinson et al. [57]	0.8k	2.0k	Gender (2), Ethnicity (4)
Srinivas et al. [65]	0.7k	8.0k	Age (cont.), Gender (2)
Klare et al. [39]	52.3k	102.9k	Age (3), Gender (2), Ethnicity (3)
Howard et al. [34]	1.1k	2.7k	Age (cont.), Gender (2), Ethnicity (2)
Grother et al. [29]	8.0M	18.0M	Age (5), Gender (2), Ethnicity (4)
Georgopoulos et al [26]	1.0k	41.0k	Age (5), Gender (2), Kinship (5)
Balakrishnan et al. [6]	1.3k	1.3k	Gender (2), Hair (cont.) Ethnicity (cont.)
Cook et al. [12]	1.1k	2.7k	Age (cont.), Gender (2), Ethnicity (4), Eyewear (2)
Lu et al. [44]	5.4k	162.5k	Demographics (3), Non-demographics (4)
This work	9.1k	3.3M	**Demographics (8)**, **Non-demographics (40)**

C. How This Work Contributes to the State of the Art

So far, the majority of research in estimating and mitigating bias in FR has focused on demographic factors, such as age, gender, and race. However, to achieve a generally accurate and fair FR model, it is necessary to know all potential origins of the differential outcome. Therefore, this work aims at closing this knowledge gap by analyzing the differential outcome on a much wider attribute range than previous works (see Table I). More precisely, this work investigates the influence of 47 attributes on the FR performance of two popular face embeddings. These 47 attributes represent a step forward in the literature in comparison with previous analyzes focused on no more than seven attributes [44].

III. EXPERIMENTS ON MEASURING DIFFERENTIAL OUTCOME

A. Database and Considered Attributes

To get reliable statements on the effect of different attributes on FR, we need a database that 1) provides a high number of face images with 2) many attribute annotations of 3) high quality. For the experiments, we choose the publicly available MAAD-Face[2] annotation database [69] based on the images of VGGFace2 [8] since this database fulfills our experimental requirements. MAAD-Face provides over 120M high-quality attribute annotations of 3.3M face images of over 9k individuals. It provides annotations for 47 distinct attributes of various kinds, such as demographics, skin types, hairstyles and hair colors, face geometry, annotations for the periocular, mouth, and nose area, as well as annotations for accessories. An exact list of the MAAD-Face annotation attributes can be seen in Table II with the number of images that are associated with (positive) and without (negative) the attributes. These attribute annotations proofed to have a higher quality than comparable face annotation databases [69].

B. Face Recognition Models

For the experiments, we use two popular FR models: 1) FaceNet [58] and 2) ArcFace [17]. To create a face embedding for a given face image, the image has to be aligned, scaled, and cropped. Then, the preprocessed image is passed to an FR model to extract the embeddings. For FaceNet, the preprocessing is done as described in [38]. To extract the embeddings, a pretrained model[3] was used. For ArcFace, the image preprocessing was done as described in [30] and a pretrained model[4] is used, which is provided by the authors of ArcFace. Both models use a ResNet-100 architecture and were trained on the MS1M database [31]. The identity verification is done by comparing two embeddings with the widely used cosine-similarity.

C. Evaluation Metrics

The face verification performance is reported in terms of 1) false nonmatch rates (FNMRs) at a fixed false match rate (FMR) and 2) equal error rates (EERs). The EER equals the FMR at the threshold, where FMR $=$ FNMR and is well known as a single-value indicator of the verification performance. The used error rates are specified for biometric verification evaluation in the international standard [37]. In the experiments, the face verification performance is reported on three operating points to cover a wide range of potential applications. This includes EER, as well as, the FNMR at 10^{-3} and 10^{-4} FMR as recommended by the best practice guidelines for automated border control of the European Boarder Guard Agency Frontex [23]. For each operating point and attribute, the verification performance is computed on all samples with positive and all samples with negative annotations. This will allow comparing the performance differences of face embeddings regarding binary attributes, such as bald versus nonbald faces.

D. Control Groups

During the experiments, the number of testing samples with positive and negative labels might be significantly different. To prevent misleading conclusions from such unbalanced annotation distributions, we introduce positive and negative control groups for each attribute. For each attribute, six positive and six negative control groups are created by randomly selecting samples from the database. To construct a positive (negative) control, sample a randomly selected without replacement until

[2]https://github.com/pterhoer/MAAD-Face

[3]https://github.com/davidsandberg/facenet

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

TABLE II
DISTRIBUTION OF THE MAAD-FACE DATA. FOR EACH ATTRIBUTE THE NUMBER OF POSITIVE AND NEGATIVE LABELS ARE SHOWN. SINCE THE LABEL DISTRIBUTION FOR AN ATTRIBUTE IS OFTEN UNBALANCED, WE INTRODUCE THE CONCEPT OF CONTROL GROUPS IN SECTION III-D TO PREVENT FAULTY DECISIONS ARISING FROM THE DISTRIBUTION OF THE TESTING DATA

Attribute	Negative	Positive	Attribute	Negative	Positive	Attribute	Negative	Positive
Male	1,349,127	1,958,913	Brown Hair	1,196,846	817,910	Bushy Eyebrows	2,119,007	1,071,154
Young	989,321	1,250,114	Gray Hair	2,839,278	316,839	Arched Eyebrows	1,814,707	762,116
Middle Aged	2,395,142	354,968	No Beard	466,498	2,108,546	Mouth Closed	485,079	139,989
Senior	3,013,551	260,687	Mustache	2,629,842	16,629	Smiling	1,034,713	625,844
Asian	3,048,755	115,021	5 o Clock Shadow	1,834,468	434,288	Big Lips	1,532,764	939,155
White	642,933	2,136,057	Goatee	2,655,062	9,229	Big Nose	1,202,627	503,066
Black	2,973,783	157,109	Oval Face	793,888	466,869	Pointy Nose	1,044,887	1,816,441
Rosy Cheeks	2,321,058	33,990	Square Face	1,585,311	1,709,811	Heavy Makeup	2,314,175	982,666
Shiny Skin	1,110,002	581,133	Round Face	2,287,232	5,905	Wearing Hat	2,946,013	256,130
Bald	3,004,817	207,554	Double Chin	2,326,091	605,454	Wearing Earrings	1,961,832	992,962
Wavy Hair	2,193,351	856,616	High Cheekbones	1,328,748	857,224	Wearing Necktie	2,162,852	350,886
Receding Hairline	1,948,374	513,859	Chubby	2,410,459	406,896	Wearing Lipstick	2,138,389	1,126,676
Bangs	2,701,346	355,048	Obstructed Forehead	2,316,315	195,722	No Eyewear	199,386	2,597,310
Sideburns	2,198,368	1,097,130	Fully Visible Forehead	845,847	1,668,763	Eyeglasses	2,854,252	339,032
Black Hair	2,067,750	514,619	Brown Eyes	401,359	1,303,174	Attractive	2,301,934	884,429
Blond Hair	2,574,286	347,723	Bags Under Eyes	1,367,622	917,779	Total	87,929,447	35,969,435

the positive (negative) control group consists of the same number of samples as the positive (negative) groups of the real data. For instance, if the real data consist of 100 k samples labeled with attribute *a* (positive samples) and 500 k samples labeled without attribute *a* (negative samples), the positive and negative control groups consist of 100 k and 500 k samples that were randomly selected from the data.

Comparing the verification performance of the positive and negative control groups allows us to state the validity of the (real) attribute-based verification performance. If the performances of the negative and positive control groups are very similar, the (real) attribute recognition performance is treated as valid. In this case, the unbalanced testing data distribution shows no effect on the performance. If the relative performance of the control groups differs strongly, the recognition performance might be significantly affected from the unbalanced distribution of the positively and negatively annotated samples. In this case, the (real) attribute recognition performance might be affected as well. Consequently, statements about the influence of this attribute on the recognition performance are of low validity. In the experiments, the validity *val* of an attribute *a*

$$\mathrm{val}(a) = 1 - \frac{\mathrm{err}^{(+)}_{\mathrm{control}}(a)}{\mathrm{err}^{(-)}_{\mathrm{control}}(a)} \tag{1}$$

is defined over the relative performance differences between the control groups. The terms $\mathrm{err}^{(+)}_{\mathrm{control}}(a)$ and $\mathrm{err}^{(-)}_{\mathrm{control}}(a)$ represent the recognition errors of the positive $(+)$ and the negative $(-)$ control groups of attribute *a*. For the experiments, we consider attributes with a validity of < 0.9 as *not valid*. However, we will also present the performance differences with the corresponding validity values so that the operators are able to choose a more suitable validity threshold for their applications.

The idea of control groups is similar to stratified sampling. In statistics, stratified sampling refers to sampling from a population that can be partitioned into distinct subgroups. For sampling, the ratio of the subgroup's size to the total data population is computed and the samples are taken from each subgroup. In contrast to this, the control group-based concept selects the samples from the total population and, thus, allows us to measure the impact of unbalanced testing data on the recognition performance.

E. Investigations

To analyze the influence of different attributes on the recognition performance of two FR models, the investigations are divided into several parts.

1) To emphasize if an attribute bias might originate from correlated attribute annotations, we analyze the correlation between the attribute annotations.
2) For each attribute, the recognition performance of its positively labeled and negatively labeled attribute groups are compared to investigate the influence of this attribute on the recognition performance. The results are discussed in the context of the corresponding validity values to avoid misinterpretations occurring from unbalanced testing data.
3) A visual summary is provided to relate the impact of the attributes on the FR systems to the validity of the results. This aims at providing a compact and easily understandable overview of the findings of this work.
4) We provide possible explanations causing the differential outcome and discuss these differences between both FR systems.
5) Finally, we use the observations to derive future research directions for FR systems.

IV. RESULTS

A. Investigating the Correlation of Facial Attributes

To understand the quality of the used labels and potential biases in the attribute space, Fig. 1 shows a selection of specific attribute-label correlations. The attributes are chosen to show the 15 most positive and negative pairwise correlations. It can be seen that *Wearing Lipstick*, *Wearing Earrings*, *Heavy Makeup*, *Young*, and *Attractive* correlates highly positively with *Arched Eyebrows*, *Wavy Hair*, and *Rosy Cheeks*.

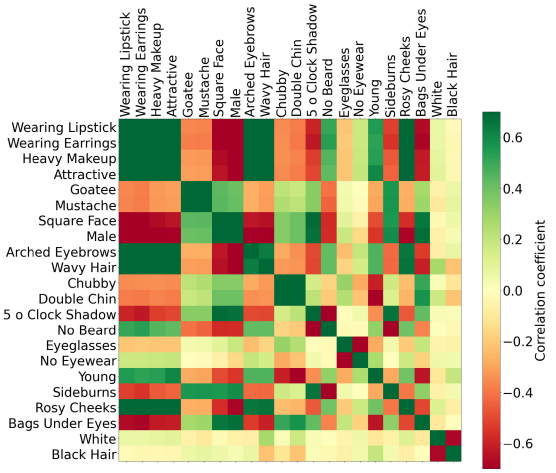


Fig. 1. Compressed annotation correlations of the used MAAD-Face database. The attributes are chosen such that the 15 most positive and negative pairwise are visible. Green indicates positive correlations, while red indicates a negative correlation. The correlation is based on the Pearson coefficient. When interpreting the results from Section IV-B, highly correlated attributes should be considered to prevent misinterpretations.

In contrast, these attributes correlates negatively with *Square Face*, *Male*, and *Bags Under Eyes*. These correlations have to be considered when comparing the differential outcome for the different attributes. However, the correlation matrix also approves the quality of some labels that semantically exclude each other. For instance, *5 o Clock Shadow* negatively correlates with *No Beard* and *Eyeglasses* negatively correlates with *No Eyewear*.

B. Impact of Facial Attributes on Recognition

The main contribution of this work is an analysis of the effect of 47 distinct attributes on two popular FR models. This aims at investigating model biases. For each attribute, the face verification performance is calculated on positively labeled samples, as well as on negatively labeled samples. This is done on three operating points as explained in Section III-C. The relative performance between the positive and negative groups allows us to investigate potential biases of the FR model toward the analyzed attribute. To determine if differential outcome results from unbalanced data distributions, we introduced control groups as explained in Section III-D.

In Tables III–VI the performance of the positive and negative class is shown for each attribute. The performance of the annotated data is referred to as Real while the performance of the control groups is referred to as Control. The relative performance (Rel. Perf.) shows the relative performance difference between the positive and negative attribute classes. If the relative performances between the control classes are below 10% (val ≥ 0.9), the result is considered as *valid* (green highlighting). Otherwise, the result is considered as *not valid* indicated by a gray highlighting. Positive values for the relative performance of an attribute represent a positive effect of the attribute on the FR performance. Negative values indicate a negative influence of the attribute on the recognition performance. In the following, we present the results of our study on bias on FaceNet and ArcFace embeddings.

1) Biases in FaceNet Embeddings: The results of our attribute-related study on differential outcome of the FaceNet model are shown in Tables III and IV.

Previous works focused on differential outcomes affected by the user's demographics. The results on FaceNet confirm the observations of these works. Demographics strongly affect recognition performance. One of the strongest impacts on FaceNet is observed for ethnicities. For the investigated FaceNet model, *Asian* and *Black* faces lead to significantly lower the recognition rates than *White* faces. Also, *Young* ones perform significantly weaker than *Middle-aged* faces. Concerning gender, we observe that the *Male* face performs better than the *Female* ones. These findings are intensively discussed in previous works [2], [3]. However, the experimental results show that there are many more aspects that strongly affect recognition performance.

One factor leading to differential outcomes is the user's hair. While *Bald* faces and *Receding Hairlines* lead to improved recognition performance, *Wavy Hair* styles or *Bangs* are observed to degrade the performance. This can be explained by the visibility of the face. In general, *Wavy Hair* and *Bangs* are more likely to cover parts of the face while *Bald* faces or faces with *Receding Hairlines* do not occlude part of the faces.

A contradictory observation can be made for facial hair. Faces with *No-Beard* perform worse than faces with a beard, such as a *5 o Clock Shadow*. A reason for this can be that people might keep their beards over a long period of time and, thus, the training and testing data might be biased such that the recognition networks consider the beards for recognition.

Also, the color of the hair has an impact on the FaceNet embeddings. While *Blond Hair* shows a strongly degraded FR performance, *Gray Hair* leads to the strongest performances.

The results indicate that the shape of a face only had a minor impact on the FR performance. For *Oval Faces*, no significant differences to nonoval faces could be observed. Although, a positive effect on recognition performance is shown for *Square Faces*, in Section IV-A a strong correlation between *Square Face* and *Male* was shown. This might explain the behavior.

Faces with *High Cheekbones*, *Double Chins*, and *Chubby* faces also perform better for FaceNet features than the inverted counterparts. Probably because these properties provide additional information that can be used for recognition. In contrast to this, an *Obstructed Forehead* strongly degrades recognition performance while a *Fully Visible Forehead* provides additional (uncovered) information that supports the recognition process.

Anomalous properties in the periocular area, such as *Bags Under Eyes*, *Bushy Eyebrows*, or *Arched Eyebrows*, lead to better recognition rates compared to face images without these attributes. The same goes for *Big Nose* and *Pointy Nose*.

The reason that *Smiling* and a *Mouth Closed* lead to stronger recognition performances than other expressions might be explainable through the used face databases that mainly contain faces with these expressions. Damer *et al.* [14] showed the opposite effect by demonstrating that crazy faces result in low comparison scores. However, they considered extreme expressions with the aim of avoiding identification.

Interestingly, accessories have a strong influence on recognition performance of FaceNet. *Wearing Hat*, *Wearing Earrings*,

TABLE III
FACENET—PART 1/2. FR PERFORMANCE BASED ON SEVERAL ATTRIBUTES

Category	Attribute	Class	EER		FNMR@FMR=10^{-3}		FNMR@FMR=10^{-4}	
			Real	Control	Real	Control	Real	Control
Demographics	Male	Positive	6.64%	6.49%	33.28%	32.51%	53.64%	52.44%
		Negative	7.87%	6.46%	42.47%	32.40%	62.55%	52.32%
		Rel. Perf.	15.56%	-0.42%	21.63%	-0.35%	14.24%	-0.21%
	Young	Positive	6.91%	6.46%	39.39%	32.39%	60.37%	52.30%
		Negative	5.73%	6.47%	28.93%	32.37%	48.97%	52.18%
		Rel. Perf.	-20.58%	0.12%	-36.19%	-0.08%	-23.27%	-0.22%
	Middle_Aged	Positive	5.41%	6.33%	28.77%	31.70%	48.75%	51.25%
		Negative	6.96%	6.48%	36.77%	32.52%	57.70%	52.45%
		Rel. Perf.	22.29%	2.33%	21.74%	2.52%	15.52%	2.28%
	Senior	Positive	6.01%	6.23%	30.26%	31.19%	50.52%	50.52%
		Negative	6.69%	6.49%	34.19%	32.54%	54.58%	52.53%
		Rel. Perf.	10.16%	3.93%	11.50%	4.16%	7.44%	3.82%
	Asian	Positive	11.16%	5.91%	69.46%	29.52%	88.48%	48.20%
		Negative	6.33%	6.49%	31.91%	32.55%	51.27%	52.54%
		Rel. Perf.	-76.30%	8.88%	-117.66%	9.33%	-72.58%	8.28%
	White	Positive	5.97%	6.48%	31.28%	32.51%	50.15%	52.50%
		Negative	7.51%	6.44%	46.82%	32.16%	69.44%	51.94%
		Rel. Perf.	20.54%	-0.56%	33.18%	-1.11%	27.79%	-1.07%
	Black	Positive	8.85%	6.02%	52.50%	30.20%	73.61%	49.32%
		Negative	6.61%	6.49%	33.47%	32.54%	53.34%	52.52%
		Rel. Perf.	-33.98%	7.14%	-56.89%	7.19%	-37.99%	6.09%
Skin	Rosy_Cheeks	Positive	1.29%	5.46%	3.76%	26.05%	9.46%	42.72%
		Negative	7.36%	6.48%	37.03%	32.51%	57.65%	52.47%
		Rel. Perf.	82.42%	15.80%	89.86%	19.87%	83.59%	18.59%
	Shiny_Skin	Positive	6.08%	6.41%	36.43%	32.05%	57.46%	51.83%
		Negative	7.90%	6.47%	41.33%	32.37%	62.43%	52.29%
		Rel. Perf.	23.06%	0.85%	11.86%	0.99%	7.97%	0.88%
Hair	Bald	Positive	5.10%	6.13%	30.52%	30.69%	52.37%	49.93%
		Negative	6.70%	6.49%	34.13%	32.54%	54.47%	52.52%
		Rel. Perf.	23.89%	5.55%	10.59%	5.70%	3.85%	4.94%
	Wavy_Hair	Positive	7.55%	6.46%	40.97%	32.29%	60.69%	52.10%
		Negative	6.82%	6.48%	34.23%	32.50%	54.84%	52.48%
		Rel. Perf.	-10.68%	0.34%	-19.69%	0.65%	-10.65%	0.73%
	Receding_Hairline	Positive	4.93%	6.43%	26.02%	32.06%	44.95%	51.75%
		Negative	7.35%	6.47%	39.92%	32.46%	61.12%	52.43%
		Rel. Perf.	32.90%	0.74%	34.82%	1.23%	26.46%	1.29%
	Bangs	Positive	6.82%	6.34%	45.53%	31.63%	69.28%	51.14%
		Negative	6.43%	6.49%	32.02%	32.54%	51.86%	52.52%
		Rel. Perf.	-5.99%	2.25%	-42.17%	2.79%	-33.59%	2.62%
	Sideburns	Positive	6.68%	6.46%	34.33%	32.34%	54.08%	52.23%
		Negative	6.75%	6.49%	35.47%	32.52%	55.96%	52.46%
		Rel. Perf.	1.04%	0.43%	3.24%	0.53%	3.35%	0.44%
	Black_Hair	Positive	7.13%	6.42%	42.35%	32.06%	65.73%	51.73%
		Negative	6.20%	6.48%	32.46%	32.49%	52.06%	52.47%
		Rel. Perf.	-15.04%	1.02%	-30.47%	1.34%	-26.26%	1.40%
	Blond_Hair	Positive	9.63%	6.34%	52.00%	31.66%	71.71%	51.18%
		Negative	6.45%	6.48%	32.63%	32.52%	52.96%	52.48%
		Rel. Perf.	-49.35%	2.17%	-59.37%	2.66%	-35.41%	2.48%
	Brown_Hair	Positive	7.40%	6.45%	39.73%	32.26%	59.12%	52.06%
		Negative	6.19%	6.47%	35.13%	32.41%	57.09%	52.30%
		Rel. Perf.	-19.52%	0.26%	-13.08%	0.49%	-3.55%	0.48%
	Gray_Hair	Positive	5.32%	6.29%	26.00%	31.50%	44.11%	50.99%
		Negative	6.72%	6.49%	34.60%	32.54%	55.25%	52.52%
		Rel. Perf.	20.83%	3.05%	24.83%	3.20%	20.17%	2.90%
Beard	No_Beard	Positive	7.20%	6.48%	37.97%	32.49%	58.83%	52.44%
		Negative	6.13%	6.40%	31.07%	31.94%	51.01%	51.60%
		Rel. Perf.	-17.53%	-1.38%	-22.20%	-1.74%	-15.33%	-1.62%
	Mustache	Positive	6.45%	4.93%	50.77%	22.55%	73.71%	36.74%
		Negative	6.90%	6.48%	35.54%	32.52%	56.12%	52.49%
		Rel. Perf.	6.41%	23.94%	-42.88%	30.68%	-31.34%	30.01%
	5_o_Clock_Shadow	Positive	6.16%	6.38%	30.98%	31.87%	50.01%	51.53%
		Negative	7.49%	6.48%	39.91%	32.46%	60.86%	52.39%
		Rel. Perf.	17.78%	1.54%	22.37%	1.83%	17.83%	1.64%
	Goatee	Positive	2.59%	4.69%	18.78%	20.11%	38.17%	32.98%
		Negative	6.92%	6.49%	35.49%	32.54%	56.11%	52.55%
		Rel. Perf.	62.59%	27.63%	47.09%	38.19%	31.97%	37.24%

or *Eyeglasses* degrade the FR performance significantly and might be explained by the fact that these accessories cover parts of the face.

2) Biases in ArcFace Embeddings: The results of our attribute-related study on the differential outcome of the ArcFace model are shown in Tables V and VI.

TABLE IV
FACENET—PART 2/2. FR PERFORMANCE BASED ON SEVERAL ATTRIBUTES

Category	Attribute	Class	EER		FNMR@FMR=10^{-3}		FNMR@FMR=10^{-4}	
			Real	Control	Real	Control	Real	Control
Face Geometry	Oval_Face	Positive	8.14%	6.40%	45.16%	31.97%	64.96%	51.64%
		Negative	8.26%	6.46%	45.11%	32.30%	67.44%	52.08%
		Rel. Perf.	1.45%	1.01%	-0.11%	1.00%	3.68%	0.84%
	Square_Face	Positive	6.32%	6.48%	31.37%	32.49%	51.25%	52.44%
		Negative	7.81%	6.47%	41.51%	32.43%	61.90%	52.38%
		Rel. Perf.	19.13%	-0.12%	24.42%	-0.16%	17.20%	-0.12%
	Round_Face	Positive	16.53%	4.52%	88.11%	19.03%	93.33%	31.40%
		Negative	5.31%	6.49%	27.06%	32.52%	45.05%	52.46%
		Rel. Perf.	-211.14%	30.27%	-225.65%	41.49%	-107.17%	40.14%
	Double_Chin	Positive	5.45%	6.44%	26.28%	32.15%	44.43%	51.85%
		Negative	7.09%	6.48%	38.20%	32.51%	59.43%	52.46%
		Rel. Perf.	23.05%	0.71%	31.20%	1.10%	25.24%	1.18%
	High_Cheekbones	Positive	5.99%	6.46%	33.69%	32.27%	53.73%	52.10%
		Negative	8.10%	6.47%	41.66%	32.41%	62.29%	52.32%
		Rel. Perf.	26.11%	0.20%	19.13%	0.43%	13.73%	0.43%
	Chubby	Positive	5.11%	6.38%	26.98%	31.81%	47.76%	51.48%
		Negative	6.85%	6.49%	36.65%	32.54%	57.76%	52.49%
		Rel. Perf.	25.35%	1.62%	26.38%	2.23%	17.31%	1.92%
	Obstructed_Forehead	Positive	8.85%	6.11%	60.01%	30.67%	80.51%	49.92%
		Negative	6.02%	6.49%	31.14%	32.52%	50.70%	52.50%
		Rel. Perf.	-46.87%	5.75%	-92.69%	5.69%	-58.79%	4.91%
	Fully_Visible_Forehead	Positive	5.47%	6.48%	28.25%	32.46%	47.36%	52.35%
		Negative	7.82%	6.45%	44.34%	32.28%	66.70%	52.09%
		Rel. Perf.	30.01%	-0.43%	36.29%	-0.55%	28.99%	-0.49%
Periocular	Brown_Eyes	Positive	7.54%	6.48%	42.04%	32.44%	63.89%	52.36%
		Negative	6.12%	6.36%	33.59%	31.83%	52.03%	51.50%
		Rel. Perf.	-23.28%	-1.81%	-25.15%	-1.94%	-22.78%	-1.67%
	Bags_Under_Eyes	Positive	5.90%	6.45%	31.51%	32.31%	52.50%	52.16%
		Negative	8.03%	6.47%	42.47%	32.42%	62.85%	52.31%
		Rel. Perf.	26.47%	0.36%	25.79%	0.37%	16.48%	0.29%
	Bushy_Eyebrows	Positive	5.66%	6.47%	29.86%	32.36%	49.67%	52.29%
		Negative	7.26%	6.48%	37.79%	32.51%	58.28%	52.45%
		Rel. Perf.	22.03%	0.23%	21.00%	0.44%	14.77%	0.31%
	Arched_Eyebrows	Positive	5.99%	6.46%	33.71%	32.28%	52.99%	52.06%
		Negative	7.59%	6.48%	38.64%	32.47%	59.96%	52.40%
		Rel. Perf.	21.10%	0.37%	12.75%	0.58%	11.62%	0.64%
Mouth	Mouth_Closed	Positive	5.25%	5.99%	27.84%	29.97%	46.77%	48.87%
		Negative	7.05%	6.41%	46.08%	32.00%	68.38%	51.71%
		Rel. Perf.	25.49%	6.53%	39.60%	6.34%	31.60%	5.50%
	Smiling	Positive	6.08%	6.44%	34.06%	32.17%	53.51%	51.91%
		Negative	8.67%	6.46%	47.88%	32.36%	70.12%	52.23%
		Rel. Perf.	29.86%	0.28%	28.87%	0.58%	23.68%	0.61%
	Big_Lips	Positive	6.79%	6.45%	39.95%	32.33%	61.39%	52.19%
		Negative	6.97%	6.47%	34.09%	32.44%	53.99%	52.36%
		Rel. Perf.	2.58%	0.32%	-17.20%	0.31%	-13.72%	0.31%
Nose	Big_Nose	Positive	6.28%	6.42%	36.68%	32.04%	59.22%	51.82%
		Negative	8.40%	6.48%	46.15%	32.43%	67.05%	52.32%
		Rel. Perf.	25.23%	0.90%	20.52%	1.20%	11.67%	0.94%
	Pointy_Nose	Positive	6.04%	6.48%	32.67%	32.48%	51.66%	52.44%
		Negative	7.80%	6.46%	43.90%	32.32%	65.97%	52.22%
		Rel. Perf.	22.56%	-0.33%	25.57%	-0.49%	21.69%	-0.42%
Accessories	Heavy_Makeup	Positive	6.25%	6.46%	35.96%	32.31%	55.91%	52.17%
		Negative	7.08%	6.49%	34.76%	32.52%	54.97%	52.48%
		Rel. Perf.	11.70%	0.46%	-3.44%	0.62%	-1.71%	0.59%
	Wearing_Hat	Positive	9.01%	6.24%	55.58%	31.23%	77.17%	50.65%
		Negative	6.05%	6.49%	30.40%	32.54%	49.86%	52.55%
		Rel. Perf.	-48.74%	3.78%	-82.84%	4.03%	-54.77%	3.60%
	Wearing_Earrings	Positive	7.54%	6.46%	41.92%	32.35%	61.83%	52.25%
		Negative	6.78%	6.48%	33.84%	32.49%	54.34%	52.45%
		Rel. Perf.	-11.15%	0.25%	-23.89%	0.43%	-13.79%	0.37%
	Wearing_Necktie	Positive	3.99%	6.36%	19.72%	31.65%	37.81%	51.23%
		Negative	7.53%	6.48%	41.03%	32.52%	62.50%	52.47%
		Rel. Perf.	47.05%	1.88%	51.93%	2.65%	39.51%	2.37%
	Wearing_Lipstick	Positive	6.74%	6.46%	38.36%	32.37%	58.49%	52.29%
		Negative	7.01%	6.49%	34.54%	32.51%	54.78%	52.49%
		Rel. Perf.	3.91%	0.39%	-11.05%	0.44%	-6.78%	0.39%
	No_Eyewear	Positive	5.77%	6.48%	29.39%	32.53%	48.75%	52.51%
		Negative	6.64%	6.11%	37.21%	30.64%	63.01%	49.90%
		Rel. Perf.	13.10%	-6.09%	21.03%	-6.16%	22.63%	-5.24%
	Eyeglasses	Positive	7.79%	6.33%	43.15%	31.57%	65.99%	51.15%
		Negative	5.70%	6.49%	29.16%	32.54%	48.78%	52.52%
		Rel. Perf.	-36.65%	2.51%	-47.99%	3.00%	-35.27%	2.61%
Other	Attractive	Positive	6.27%	6.45%	36.28%	32.31%	56.11%	52.10%
		Negative	7.05%	6.49%	34.77%	32.51%	54.96%	52.50%
		Rel. Perf.	11.16%	0.51%	-4.35%	0.61%	-2.09%	0.77%

Similar to FaceNet, the results on ArcFace confirm the observed demographic performance differences shown by previous works. For the investigated ArcFace model, *Young* faces perform weaker than *Middle-aged* or *Senior* faces. Interestingly, the intensively discussed gender bias is strongly dependent on the used decision threshold. Especially, for

TABLE V
ARCFACE—PART 1/2. FR PERFORMANCE BASED ON SEVERAL ATTRIBUTES

Category	Attribute	Class	EER		FNMR@FMR=10^{-3}		FNMR@FMR=10^{-4}	
			Real	Control	Real	Control	Real	Control
Demographics	Male	Positive	3.98%	3.98%	7.07%	7.22%	9.71%	10.17%
		Negative	3.82%	3.96%	7.99%	7.20%	12.33%	10.13%
		Rel. Perf.	-4.35%	-0.38%	11.54%	-0.38%	21.24%	-0.37%
	Young	Positive	3.74%	3.97%	7.30%	7.20%	11.08%	10.14%
		Negative	3.70%	3.95%	6.32%	7.17%	8.52%	10.11%
		Rel. Perf.	-0.86%	-0.46%	-15.42%	-0.46%	-30.08%	-0.28%
	Middle_Aged	Positive	3.01%	3.81%	5.05%	6.93%	6.93%	9.80%
		Negative	4.07%	3.98%	7.79%	7.22%	11.36%	10.17%
		Rel. Perf.	26.14%	4.05%	35.20%	4.04%	39.04%	3.56%
	Senior	Positive	2.95%	3.62%	4.52%	6.58%	6.15%	9.38%
		Negative	4.02%	3.98%	7.47%	7.24%	10.62%	10.18%
		Rel. Perf.	26.60%	9.02%	39.44%	9.09%	42.13%	7.87%
	Asian	Positive	7.99%	3.29%	16.68%	6.01%	22.59%	8.69%
		Negative	3.73%	3.98%	6.75%	7.23%	9.61%	10.18%
		Rel. Perf.	-114.49%	17.22%	-147.13%	16.84%	-134.94%	14.60%
	White	Positive	3.27%	3.98%	5.84%	7.23%	8.55%	10.18%
		Negative	5.80%	3.91%	11.69%	7.10%	16.03%	10.01%
		Rel. Perf.	43.50%	-1.66%	50.08%	-1.87%	46.66%	-1.74%
	Black	Positive	5.72%	3.40%	10.90%	6.21%	15.02%	8.95%
		Negative	3.85%	3.98%	7.06%	7.23%	10.11%	10.18%
		Rel. Perf.	-48.63%	14.53%	-54.43%	14.16%	-48.64%	12.08%
Skin	Rosy_Cheeks	Positive	0.98%	2.91%	1.17%	5.12%	1.31%	7.47%
		Negative	4.39%	3.98%	8.33%	7.23%	11.77%	10.16%
		Rel. Perf.	77.61%	26.88%	85.99%	29.13%	88.86%	26.51%
	Shiny_Skin	Positive	3.50%	3.93%	6.33%	7.13%	9.27%	10.04%
		Negative	4.17%	3.96%	8.13%	7.18%	11.89%	10.11%
		Rel. Perf.	16.14%	0.61%	22.13%	0.72%	22.04%	0.73%
Hair	Bald	Positive	2.79%	3.50%	4.48%	6.38%	6.07%	9.14%
		Negative	4.01%	3.98%	7.43%	7.23%	10.62%	10.18%
		Rel. Perf.	30.40%	12.13%	39.77%	11.78%	42.83%	10.21%
	Wavy_Hair	Positive	3.03%	3.95%	6.34%	7.17%	10.28%	10.09%
		Negative	4.35%	3.97%	7.92%	7.23%	10.82%	10.17%
		Rel. Perf.	30.46%	0.73%	19.95%	0.84%	4.96%	0.80%
	Receding_Hairline	Positive	3.03%	3.92%	4.68%	7.12%	6.13%	10.04%
		Negative	4.10%	3.97%	8.20%	7.22%	12.25%	10.15%
		Rel. Perf.	26.21%	1.18%	42.90%	1.28%	49.98%	1.17%
	Bangs	Positive	4.03%	3.80%	8.79%	6.91%	13.94%	9.78%
		Negative	3.83%	3.98%	6.77%	7.23%	9.42%	10.17%
		Rel. Perf.	-5.11%	4.43%	-29.80%	4.44%	-47.96%	3.89%
	Sideburns	Positive	3.72%	3.97%	6.51%	7.21%	9.10%	10.13%
		Negative	3.98%	3.97%	7.62%	7.22%	11.10%	10.16%
		Rel. Perf.	6.58%	0.08%	14.56%	0.12%	18.07%	0.30%
	Black_Hair	Positive	5.12%	3.92%	9.85%	7.11%	13.47%	10.01%
		Negative	3.48%	3.97%	6.36%	7.21%	9.28%	10.15%
		Rel. Perf.	-47.25%	1.28%	-54.86%	1.46%	-45.17%	1.38%
	Blond_Hair	Positive	3.09%	3.81%	7.38%	6.92%	12.43%	9.76%
		Negative	4.09%	3.98%	7.34%	7.23%	10.16%	10.18%
		Rel. Perf.	24.53%	4.22%	-0.57%	4.25%	-22.38%	4.07%
	Brown_Hair	Positive	3.24%	3.96%	6.46%	7.18%	10.26%	10.10%
		Negative	4.12%	3.97%	7.59%	7.20%	10.59%	10.14%
		Rel. Perf.	21.36%	0.35%	14.93%	0.26%	3.11%	0.36%
	Gray_Hair	Positive	2.68%	3.76%	4.01%	6.82%	5.40%	9.67%
		Negative	4.07%	3.98%	7.57%	7.23%	10.77%	10.17%
		Rel. Perf.	34.09%	5.58%	47.01%	5.70%	49.87%	4.97%
Beard	No_Beard	Positive	4.13%	3.98%	8.10%	7.23%	11.93%	10.18%
		Negative	3.31%	3.89%	5.61%	7.06%	7.90%	9.95%
		Rel. Perf.	-25.05%	-2.23%	-44.32%	-2.49%	-50.91%	-2.28%
	Mustache	Positive	4.89%	2.63%	9.62%	4.61%	13.54%	6.68%
		Negative	4.06%	3.98%	7.62%	7.23%	10.97%	10.17%
		Rel. Perf.	-20.46%	33.85%	-26.25%	36.24%	-23.41%	34.31%
	5_o_Clock_Shadow	Positive	2.96%	3.90%	4.94%	7.06%	7.08%	9.96%
		Negative	4.24%	3.96%	8.55%	7.20%	12.68%	10.14%
		Rel. Perf.	30.18%	1.74%	42.23%	1.97%	44.16%	1.75%
	Goatee	Positive	1.18%	2.46%	1.68%	4.00%	2.67%	5.89%
		Negative	4.08%	3.98%	7.67%	7.23%	11.03%	10.18%
		Rel. Perf.	71.16%	38.16%	78.13%	44.74%	75.83%	42.12%

lower FMRs, the differential outcome between *Male* and *Female* increases. Concerning the ethnic-bias on the ArcFace model, we are not able to confirm the observations from previous works. For *White* faces, the performance is significantly higher than for nonwhite faces. For *Asian* and *Black* faces, a strong degradation in recognition performance can be

TABLE VI
ARCFACE—PART 2/2. FR PERFORMANCE BASED ON SEVERAL ATTRIBUTES

Category	Attribute	Class	EER		FNMR@FMR=10^{-3}		FNMR@FMR=10^{-4}	
			Real	Control	Real	Control	Real	Control
Face Geometry	Oval_Face	Positive	2.73%	3.90%	5.69%	7.07%	9.65%	9.97%
		Negative	5.40%	3.96%	11.10%	7.19%	15.61%	10.12%
		Rel. Perf.	49.55%	1.59%	48.72%	1.67%	38.22%	1.39%
	Square_Face	Positive	3.73%	3.97%	6.37%	7.22%	8.68%	10.16%
		Negative	4.13%	3.97%	8.61%	7.21%	13.02%	10.14%
		Rel. Perf.	9.65%	0.03%	25.96%	-0.10%	33.37%	-0.15%
	Round_Face	Positive	7.04%	2.30%	22.68%	3.89%	35.87%	5.53%
		Negative	3.17%	3.98%	5.30%	7.22%	7.43%	10.16%
		Rel. Perf.	-122.46%	42.18%	-328.09%	46.22%	-383.05%	45.63%
	Double_Chin	Positive	3.34%	3.93%	5.32%	7.12%	7.00%	10.04%
		Negative	4.08%	3.98%	7.84%	7.23%	11.50%	10.17%
		Rel. Perf.	18.22%	1.23%	32.24%	1.43%	39.15%	1.29%
	High_Cheekbones	Positive	3.34%	3.95%	5.96%	7.17%	8.63%	10.10%
		Negative	4.28%	3.97%	8.60%	7.20%	12.70%	10.13%
		Rel. Perf.	21.87%	0.48%	30.76%	0.42%	32.08%	0.34%
	Chubby	Positive	3.70%	3.87%	6.11%	7.01%	7.86%	9.90%
		Negative	3.90%	3.97%	7.37%	7.22%	10.79%	10.17%
		Rel. Perf.	5.18%	2.62%	17.14%	2.85%	27.15%	2.58%
	Obstructed_Forehead	Positive	5.48%	3.51%	13.03%	6.39%	20.40%	9.17%
		Negative	3.52%	3.97%	6.10%	7.21%	8.56%	10.15%
		Rel. Perf.	-55.61%	11.62%	-113.74%	11.37%	-138.28%	9.66%
	Fully_Visible_Forehead	Positive	3.30%	3.97%	5.47%	7.21%	7.49%	10.15%
		Negative	4.64%	3.95%	9.98%	7.16%	15.06%	10.06%
		Rel. Perf.	28.85%	-0.46%	45.15%	-0.70%	50.30%	-0.86%
Periocular	Brown_Eyes	Positive	4.69%	3.97%	9.13%	7.21%	12.88%	10.14%
		Negative	2.63%	3.85%	5.36%	6.98%	8.73%	9.85%
		Rel. Perf.	-78.48%	-2.96%	-70.17%	-3.28%	-47.51%	-2.92%
	Bags_Under_Eyes	Positive	3.78%	3.96%	6.37%	7.19%	8.48%	10.11%
		Negative	3.87%	3.96%	8.17%	7.19%	12.63%	10.12%
		Rel. Perf.	2.20%	0.13%	22.05%	0.01%	32.84%	0.10%
	Bushy_Eyebrows	Positive	3.51%	3.96%	6.05%	7.19%	8.28%	10.11%
		Negative	4.00%	3.97%	7.81%	7.22%	11.58%	10.17%
		Rel. Perf.	12.35%	0.20%	22.54%	0.54%	28.46%	0.60%
	Arched_Eyebrows	Positive	3.21%	3.94%	6.08%	7.15%	9.29%	10.05%
		Negative	4.42%	3.97%	8.38%	7.23%	11.79%	10.17%
		Rel. Perf.	27.52%	0.81%	27.46%	1.05%	21.20%	1.14%
Mouth	Mouth_Closed	Positive	3.06%	3.37%	5.40%	6.13%	7.70%	8.85%
		Negative	3.88%	3.90%	7.79%	7.08%	11.93%	9.99%
		Rel. Perf.	21.21%	13.69%	30.62%	13.38%	35.48%	11.39%
	Smiling	Positive	3.35%	3.94%	5.93%	7.14%	8.48%	10.05%
		Negative	4.62%	3.96%	9.65%	7.19%	14.57%	10.12%
		Rel. Perf.	27.32%	0.56%	38.59%	0.71%	41.81%	0.65%
	Big_Lips	Positive	4.15%	3.94%	8.12%	7.17%	11.91%	10.10%
		Negative	3.88%	3.97%	7.00%	7.22%	9.84%	10.16%
		Rel. Perf.	-6.93%	0.78%	-16.08%	0.75%	-21.01%	0.63%
Nose	Big_Nose	Positive	4.39%	3.90%	7.89%	7.07%	10.48%	9.95%
		Negative	3.90%	3.95%	8.62%	7.18%	13.58%	10.10%
		Rel. Perf.	-12.67%	1.49%	8.52%	1.51%	22.80%	1.46%
	Pointy_Nose	Positive	3.15%	3.97%	5.84%	7.22%	8.86%	10.16%
		Negative	5.28%	3.96%	10.46%	7.18%	14.62%	10.11%
		Rel. Perf.	40.44%	-0.43%	44.19%	-0.63%	39.44%	-0.49%
Accessories	Heavy_Makeup	Positive	3.08%	3.96%	5.79%	7.20%	9.00%	10.13%
		Negative	4.32%	3.97%	8.02%	7.22%	11.14%	10.16%
		Rel. Perf.	28.75%	0.18%	27.75%	0.24%	19.27%	0.32%
	Wearing_Hat	Positive	5.51%	3.66%	12.28%	6.62%	18.45%	9.44%
		Negative	3.71%	3.98%	6.53%	7.23%	9.14%	10.18%
		Rel. Perf.	-48.79%	8.09%	-88.01%	8.45%	-101.94%	7.29%
	Wearing_Earrings	Positive	3.25%	3.95%	6.64%	7.17%	10.59%	10.10%
		Negative	4.08%	3.98%	7.33%	7.23%	10.10%	10.17%
		Rel. Perf.	20.23%	0.83%	9.44%	0.80%	-4.92%	0.78%
	Wearing_Necktie	Positive	2.72%	3.82%	3.84%	6.92%	4.72%	9.79%
		Negative	4.25%	3.97%	8.52%	7.22%	12.68%	10.16%
		Rel. Perf.	35.95%	4.00%	54.94%	4.24%	62.77%	3.73%
	Wearing_Lipstick	Positive	3.28%	3.96%	6.38%	7.19%	9.93%	10.11%
		Negative	4.27%	3.98%	7.85%	7.23%	10.83%	10.18%
		Rel. Perf.	23.21%	0.40%	18.74%	0.57%	8.25%	0.65%
	No_Eyewear	Positive	3.64%	3.98%	6.39%	7.23%	8.92%	10.18%
		Negative	3.86%	3.50%	6.62%	6.35%	8.99%	9.12%
		Rel. Perf.	5.75%	-13.57%	3.42%	-13.84%	0.82%	-11.60%
	Eyeglasses	Positive	4.60%	3.79%	9.13%	6.88%	13.03%	9.75%
		Negative	3.68%	3.98%	6.45%	7.23%	8.99%	10.18%
		Rel. Perf.	-25.08%	4.88%	-41.53%	4.89%	-44.86%	4.24%
Other	Attractive	Positive	2.95%	3.96%	5.49%	7.19%	8.60%	10.10%
		Negative	4.30%	3.97%	8.02%	7.22%	11.14%	10.17%
		Rel. Perf.	31.44%	0.20%	31.57%	0.47%	22.82%	0.65%

observed. However, we have to consider these results as *not valid*, since we can observe strong performance differences in the control groups. This indicates that these results are strongly influenced by the unbalanced testing data.

Similar to FaceNet, the user's hair shows to have a significant impact on the FR performance. While, *Receding Hairlines*, *Wavy Hair*, and *Sideburns* support the recognition process, faces with *Bangs* show a strong degradation. Again,

the performance differences on ArcFace show to be threshold-dependent. For *Wavy Hair*, the positive effect on FR vanishes for lower FMRs, and for *Bangs*, the negative effect increases drastically for lower FMRs.

Also, the color of the user's hair has an impact on recognition performance. *Gray Hair* performs significantly above average, while *Black Hair* performs significantly below average. *Blond Hair* and *Brown Hair* lead to differential outcome depending on the decision threshold. For high FMRs, *Blond Hair* improves recognition performance, while for lower FMRs, recognition performance changes to below average. For faces with *Brown Hair*, the positive effect on recognition vanishes for lower FMRs.

The effect of wearing a beard on the performance of ArcFace is similar to FaceNet. Having *No Beard* decreases recognition performance and having a beard, such as a *5 o Clock Shadow*, enhances the recognition. These effects are clearer for lower FMRs.

In contrast to FaceNet, the face shape affects recognition performance of ArcFace. Both, *Oval Faces* and *Square Faces* have a positive effect on recognition performance, which is dependent on the utilized decision threshold. *Round Faces* show a strongly degraded recognition. However, a large fraction of these performance differences can be explained by the unbalanced data distribution and, thus, we have to neglect the results for *Round Faces*.

Similar to FaceNet, *High Cheekbones*, *Double Chin*, *Chubby*, and a *Fully Visible Forehead* lead to improved FR performances. While a *Fully Visible Forehead* refers to no partial occlusions of the face that might negatively infer, the other attributes provide anomalous characteristics that might help for recognition.

Surprisingly, faces with *Brown Eyes* perform drastically weaker than faces with nonbrown eyes. For *Bags Under Eyes*, *Bushy Eyebrows*, and *Arched Eyebrows*, an improved FR performance can be observed. These attributes can be treated as anomalies and, thus, can support the recognition process. The same goes for *Big Nose* and *Pointy Nose*.

Similar to FaceNet, accessories have a strong impact on the differential outcome of ArcFace. While having *Heavy Makeup*, such as *Wearing Lipstick*, improves the recognition, faces with *Eyeglasses* or *Wearing Hat* lead to strong degradations in the FR performance. A reason for this might be that people using *Heavy Makeup* frequently. Consequently, a person in the training data might either have no or only *Heavy Makeup* images. On the other side, people tend to change their *Eyeglasses* or (Wearing) *Hats* more frequently. Moreover, these attributes might lead to partial occlusions of the face leading to less identity-information available and, thus, to a degraded FR performance.

As stated in Section III-A, the experiments were performed on the MAAD-Face annotations dataset since it is, to the best of our knowledge, the only publicly available dataset that meets the requirements for this analysis. More precisely, it 1) provides a high number of face images with 2) many attribute annotations of 3) high quality. Even though most databases include nondemographic attributes, these are seldom available in the form of annotations and, thus, prevent research in the direction of nondemographic biases in FR.

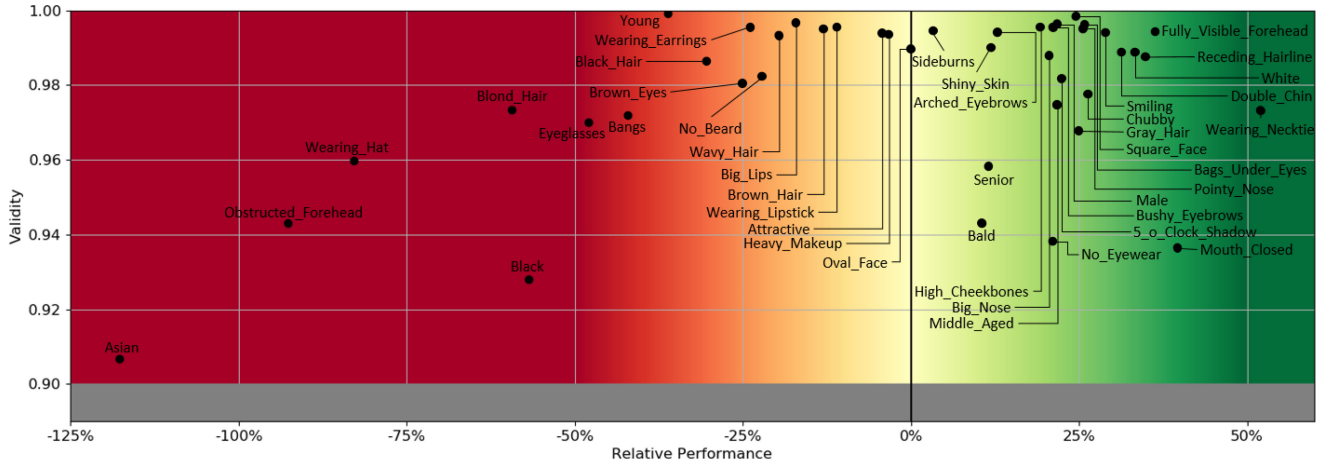
C. Performance Analysis

To provide an overview of the findings, Fig. 2 shows the relative performance differences on FaceNet and ArcFace features based on the investigated attributes. The shown relative performance is based on the FMR at 10^{-3} FNMR as recommended by the European Boarder Guard Agency Frontex [23]. The validity describes the performance difference between the positive and negative attribute-related control groups as shown in (1). An attribute performance with a validity of less than 90% is considered as *not valid* (gray area) since the unbalanced data annotations might affect the reported performance. The red area indicates that recognition performance of the positive attribute class is significantly weaker than the performance of the negative class. In contrast, the green area indicates a significant improvement of recognition performance of the positive attribute class over the negative class. If an attribute has only a minor effect on recognition performance, the relative performance is close to 0% (yellow area).

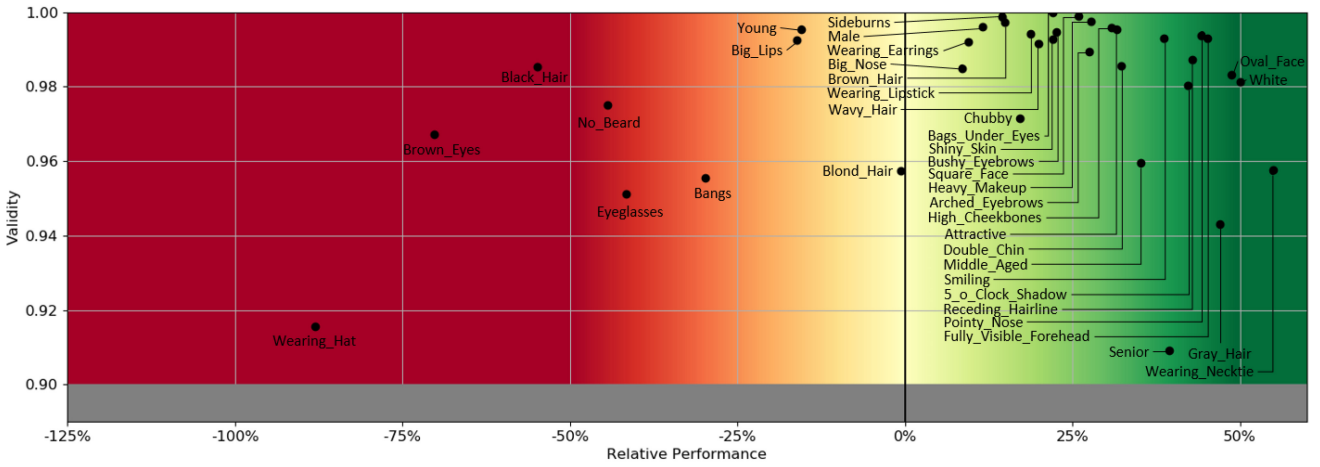
1) FaceNet Versus ArcFace: The main difference between FaceNet and ArcFace is the underlying training principles. FaceNet uses triplet-loss learning [58] that aims solely at minimizing the intraclass variations while maximizing the interclass variations. In contrast, ArcFace introduces an angular large-margin principle [17] that additionally aims at enhancing the robustness of the recognition model. The utilized training principle together with the used network structure and the training data determines the recognition behavior. This includes the effect of differential outcomes appearing when certain attributes of the face are present. Since the used FaceNet and ArcFace models share the same network structure and training data, the observed differential outcome might arise from the training principles.

2) Effect of Attributes on Recognition: It turns out that the majority of the investigated attributes strongly affect recognition performance of both FaceNet and ArcFace. For FaceNet, many faces that are perceived as *Attractive* or make use of *Heavy Makeup* do not show to alter recognition performance unlike previously reported in [54]. The same goes for *Oval Faces* and faces with *Sideburns*. For ArcFace, *Blond Hair*, *Big Nose*, *Big Lips*, *Wearing Earrings*, and *Young* faces show only a minor effect on recognition performance. For both recognition models, the majority of the investigated attributes strongly affect recognition performance. Some of the observations might be explainable.

1) *Demographics:* Recent works [6], [29], [34], [57] extensively discussed the impact of demographic attributes on FR. Our results support the findings from previous works. We observe an improved recognition performance for the attributes *Middle Aged*, *Senior*, *White*, and *Male*. On the contrary, a degraded recognition performance is observed for *Young*, *Asian*, *Black*, and *Female* faces. For FaceNet, the observed differential outcome is stronger for ArcFace. Moreover, we could



(a)



(b)

Fig. 2. Visual summary on the differential outcome affected by each attribute. (a) visualizes the results for FaceNet, while (b) visualizes the results for ArcFace. The relative performance is based on recognition performance on the positively labeled data versus the performance of negatively labeled data. The validity is based on the performance differences of the control groups. Validity values below 0.9 (more than 10% performance differences between the control groups) are considered as *not valid* (gray area) and are not shown in this figure. The red areas indicate an attribute-related bias that leads to a degraded FR performance for faces with the specific attribute. Green areas indicate that faces possessing a specific attribute enhances recognition performance. It can be observed that the majority of the investigated attributes strongly affect recognition performance.

not show that *Asian* or *Black* faces perform weaker than *White* faces on ArcFace, since the data unbalance lead to a low validity for our results.

2) *Visibility-Related Attributes:* We observe that attributes that indicate a fully visible face lead to an improved FR performance. This includes the attributes *Fully Visible Forehead*, *Receding Hairline*, *No Eyewear*, and *Bald*. In contrast, attributes that might lead to small partial occlusions of the face lead to significantly degraded recognition performances [75]. For FaceNet, this includes faces with an *Obstructed Forehead*, *Bangs*, and *Wavy Hair*. For ArcFace, this includes samples with *Eyeglasses* or *Bangs*.
3) *Temporary Attributes:* For faces with temporary attributes, such as for accessories, degraded FR performance can be observed. This includes *Wearing Hat*, *Wearing Earrings*, *Wearing Lipstick*, and *Eyeglasses*. Besides a partial occlusion of small parts of the face, these attributes are nonpermanent and can quickly change the appearance of the face.
4) *Anomalous Characteristics:* It turns out that conspicuous characteristics that are only possessed by a small proportion of the population lead to strongly enhanced recognition performances. This includes *Arched Eyebrows*, *Big Nose*, *Pointy Nose*, *Bushy Eyebrows*, *Double Chin*, and *High Cheekbones* [69].
5) *Facial Expressions:* Faces that are *Smiling* or that have their *Mouth Closed* perform above average for FR. However, faces with other expressions lead to degraded FR performances. This bias might come from the data utilized for training that usually contains neutral or smiling faces and was discussed in more detail by previous works [10], [11].

While these attribute-dependent differential outcomes might be explainable, the reason for the impact of other attributes on recognition is currently unclear.

1) *Colors:* The results demonstrate strong differential outcome based on the user's hair color and eyecolor. For FaceNet, faces with *Blond Hair*, *Black Hair*, and *Brown Hair* show strongly degraded recognition performances. In contrast, faces with *Gray Hair* lead to improved recognition. For ArcFace, *Gray Hair* also strongly improves recognition performance while *Black Hair* decreases it. The differential outcome for *Blond Hair* and *Brown Hair* strongly varies depending on the used decision threshold. For instance, for high FMRs, *Blond Hair* has a positive effect on recognition, for a lower FMR (e.g., 10^{-4}), the same attribute changes to a negative effect. The same can be observed for eyecolors. Faces with *Brown Eyes* perform weaker than faces from the opposite group. The differential outcome of these attributes does not reflect the distribution of the training data and, thus, might arise from a different origin.
2) *Beard:* As we discussed before, attributes that might induce a partially occluded face lead to a degraded FR performance. Although beards can cover parts of the face, the results demonstrate the faces with *No Beard* perform below average, while faces with, e.g., a *5 o Clock Shadow* achieve much higher recognition rates.
3) *Wearing Necktie:* Unlike other accessories, *Wearing Necktie* improved the FR performance drastically. We assume that this might result from a data collection bias induced by the correlation with hidden factors, such as the environment. Persons who present themselves in public (e.g., celebrities) might often wear a necktie and, thus, photos are often taken with frontal poses and full lightning. However, the high validity and the strong differential outcome make it hard to argue in this direction.
4) *Antagonistic Behavior:* Some attributes might result in a differential outcome of the opposite direction depending on the used training principle (triplet versus angular margin loss). For instance, faces with *Wavy Hair* lead to a negative performance on FaceNet and to a positive performance on ArcFace. Also, the attributes *Attractiveness*, *Heavy Makeup*, and *Oval Faces* negatively affect recognition performance on FaceNet but show some strong positive impacts on recognition performance of ArcFace.

As mentioned earlier, the resulting performance of an FR model is mainly determined by its loss function, its network architecture, and the utilized training data. Since both investigated models have the last two points in common, the observed differences in the performance might arise from the underlying training principles. Generally, we observe that the large angular margin loss from ArcFace leads to a significantly stronger overall recognition performance compared to FaceNet. The loss aiming to enhance the model robustness also shows a clearly visible effect on the attribute-related differential outcome. On ArcFace, slightly fewer attributes negatively affect recognition performance than on FaceNet. However, the differential outcome that origins from the affected (biased) attributes are still of high impact. A remarkable observation is the fact that the differential outcome remains relatively constant over several decision thresholds for FaceNet, while for ArcFace, the differential outcome often significantly varies for different decision thresholds. This can be observed, for instance, for faces with *Bangs*, *Blond Hair*, or a *Double Chin*.

3) Future Challenges for Face Recognition: The observations of the experiment point out some critical issues of current FR solutions, especially in terms of annotations, robustness, fairness, and explainability.

1) *Need for Annotations:* Most databases include nondemographic attributes. However, these are rarely labeled, which presents a barrier to further explore nondemographic biases. To the best of our knowledge, the MAAD-Face annotations database is currently the only publicly available and large-scale face database that provides demographic and nondemographic annotations a) of high quality; b) for a large variety of soft-biometrics; and c) in large numbers. To further investigate the issue of soft-biometric bias, more databases are needed that meet these requirements.
2) *Need for Robustness:* FR systems need to become more robust against partial occlusions (from accessories or hair) [43], [75], facial expressions (beyond neutral and smiling faces) [52], and temporary attributes that might change the daily appearance of a face [67], [74]. This can greatly enhance the applicability in more real-life scenarios.
3) *Need for Fairness:* FR systems need to enhance user fairness. We observed a differential outcome based on the user demographics (demographic-bias), anomalous characteristics (such as pointy noses, bushy eyebrows, and high cheekbones), beard types, and accessories. This can lead to discriminative decisions [59] of FR systems that several political regulations, such as the GDPR [77], try to prevent.
4) *Need for Explainability:* FR models need to explain themselves. Why do colors/face shapes/beards/accessories lead to the differential outcome? Why can we observe an antagonistic behavior between the two different learning principles for some attributes? In order to enhance the model transparency and to enable efficient model debugging, future work has to elaborate on the explainability [5], [50] of FR models.
5) *Need for Comprehensive Approaches and Transfer Learning:* The previous areas related to robustness, fairness, and explainability will significantly benefit from more comprehensive approaches that consider simultaneously all the elements and attributes in place [20], [62], exploiting at the same time previous or general knowledge of the problem at hand [21], [64]. Most of the research so far in biometrics bias, especially around face biometrics, has been mainly oriented to studying individual elements (e.g., gender or ethnicity) not exploiting previous models or evidence. There is a need for more comprehensive approaches like the one presented here (incorporating simultaneously 47 relevant attributes) and new schemes to easily exploit the generated knowledge.

V. Conclusion

The growing effect of FR systems on daily life, including critical decision-making processes, shows the need for nondiscriminative FR solutions. Previous works focused on estimating and mitigating demographic-bias. However, to deploy nondiscriminatory FR systems, it is necessary to know which differential outcome appears in the presence of certain facial attributes beyond demographics. Driven by this need, we analyzed the performance differences on two popular FR models concerning 47 different attributes. The experiment was conducted on the publicly available MAAD-Face database, a large-scale dataset with over 120M attribute annotations of high quality. To prevent misleading statements of attribute biases, we consider attribute correlations and minimize the effect of unbalanced testing data via control group-based validity values. We investigated the effect of two different learning principles on the differential outcome originating from facial attributes. The results show that, besides demographics, many attributes strongly affect recognition performance of both investigated FR models: 1) FaceNet and 2) ArcFace. While for FaceNet, the observed differential outcome originated by several attributes remains relatively constant, these differences strongly depend on the used decision threshold for ArcFace. We provided explanations for many observed performance differences. However, the reason for some observations remains unclear and has to be addressed by future work. The findings of this work strongly demonstrate the need for further advances in making FR systems more robust, explainable, and fair. We hope these findings lead to the development of more robust and unbiased FR solutions.

References

[1] A. Acien, A. Morales, R. Vera-Rodriguez, I. Bartolome, and J. Fierrez, "Measuring the gender and ethnicity bias in deep models for face recognition," in *Proc. IAPR Iberoamerican Congr. Pattern Recognit. (CIARP)*, vol. 11401, Nov. 2018, pp. 584–593.

[2] V. Albiero and K. W. Bowyer, "Is face recognition sexist? no, gendered hairstyles and biology are," 2020. [Online]. Available: arXiv:2008.06989.

[3] V. Albiero, K. Zhang, and K. W. Bowyer, "How does gender balance in training data affect face recognition accuracy?" 2020. [Online]. Available: arXiv:2002.02934.

[4] M. S. Alvi, A. Zisserman, and C. Nellåker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, vol. 11129. Munich, Germany, Sep. 2018, pp. 556–572.

[5] A. B. Arrieta *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[6] G. Balakrishnan, Y. Xiong, W. Xia, and P. Perona, "Towards causal benchmarking of bias in face analysis algorithms," 2020. [Online]. Available: arXiv:2007.06570.

[7] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. 1st Conf. Fairness Accountability Transparency*, vol. 81. New York, NY, USA, 2018, pp. 77–91.

[8] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, Xi'an, China, 2018, pp. 67–74.

[9] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?" 2019. [Online]. Available: arXiv:1912.07398.

[10] K. I. Chang, K. W. Bowyer, and P. J. Flynn, "Multiple nose region matching for 3D face recognition under varying facial expression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1695–1700, Oct. 2006.

[11] K. J. Chang, K. W. Bowyer, and P. J. Flynn, "Effects on facial expression in 3D face recognition," in *Proc. Biometric Technol. Hum. Identif. II*, vol. 5779, 2005. pp. 132–143.

[12] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 1, no. 1, pp. 32–41, Jan. 2019.

[13] N. Damer, P. Terhörst, A. Braun, and A. Kuijper, "General borda count for multi-biometric retrieval," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Denver, CO, USA, Oct. 2017, pp. 420–428.

[14] N. Damer *et al.*, "CrazyFaces: Unassisted circumvention of watchlist face identification," in *Proc. 9th IEEE Int. Conf. Biometrics Theory Appl. Syst. (BTAS)*, Redondo Beach, CA, USA, Oct. 2018, pp. 1–9.

[15] A. Das, A. Dantcheva, and F. Bremond, "Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, vol. 11129. Munich, Germany, Sep. 2018, pp. 573–585.

[16] D. Deb, N. Nain, and A. K. Jain, "Longitudinal study of child face recognition," in *Proc. Int. Conf. Biometrics (ICB)*, Gold Coast, QLD, Australia, Feb. 2018, pp. 225–232.

[17] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.

[18] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Trans. Technol. Soc.*, vol. 1, no. 2, pp. 89–103, Jun. 2020.

[19] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper, "Demographic bias in presentation attack detection of iris recognition systems," 2020. [Online]. Available: arXiv:2003.03151.

[20] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, "Multiple classifiers in biometrics. Part 1: Fundamentals and review," *Inf. Fusion*, vol. 44, pp. 57–64, Nov. 2018.

[21] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Adapted user-dependent multimodal biometric authentication exploiting general information," *Pattern Recognit. Lett.*, vol. 26, no. 16, pp. 2628–2639, Dec. 2005.

[22] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Trans. Inf. Syst.*, vol. 14, no. 3, pp. 330–347, 1996.

[23] *Best Practice Technical Guidelines for Automated Border Control (ABC) Systems*, Frontex, Warsaw, Poland, 2017.

[24] N. Furl, P. J. Phillips, and A. J. O'Toole, "Face recognition algorithms and the other-race effect: Computational mechanisms for a developmental contact hypothesis," *Cogn. Sci.*, vol. 26, pp. 797–815, Nov./Dec. 2002.

[25] C. Garvie, *The Perpetual Line-Up: Unregulated Police Face Recognition in America*. Washington, DC, USA: Georgetown Law, 2016.

[26] M. Georgopoulos, Y. Panagakis, and M. Pantic, "Investigating bias in deep face analysis: The KANFace dataset and empirical study," *Image Vis. Comput.*, vol. 102, 2020, Art. no. 103954.

[27] S. Gong, X. Liu, and A. K. Jain, "Jointly de-biasing face recognition and demographic attribute estimation," 2019. [Online]. Available: arXiv:1911.08080.

[28] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez, "Facial soft biometrics for recognition in the wild: Recent works, annotation, and COTS evaluation," *IEEE Trans. Inf. Forensics Security*, vol. 13, pp. 2001–2014, 2018.

[29] P. J. Grother, M. L. Ngan, and K. K. Hanaoka, "Face recognition vendor test part 3: Demographic effects," NIST, Gaithersburg, MD, USA, Rep. NIST Interagency/Internal Report (NISTIR) 8280, 2019.

[30] J. Guo, J. Deng, N. Xue, and S. Zafeiriou, "Stacked dense U-nets with dual transformers for robust face alignment," in *Proc. Brit. Mach. Vis. Conf. (BMVC)* Sep. 2018, p. 44.

[31] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 9907. Amsterdam, The Netherlands, Oct. 2016, pp. 87–102.

[32] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics systems under spoofing attack: An evaluation methodology and lessons learned," *IEEE Signal Process. Mag.*, vol. 32, no. 5, pp. 20–30, Sep. 2015.

[33] J. Hernandez-Ortega, J. Fierrez, A. Morales, and J. Galbally, "Introduction to face presentation attack detection," in *Handbook of Biometric Anti-Spoofing*, S. Marcel, M. Nixon, J. Fierrez, and N. Evans, Eds. Cham, Switzerland: Springer, 2019, pp. 187–206.

[34] J. J. Howard, Y. B. Sirotin, and A. R. Vemury, "The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance," in *Proc. 10th IEEE Int. Conf. Biometrics Theory Appl. Syst. (BTAS)*, Tampa, FL, USA, Sep. 2019, pp. 1–8.

[35] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Deep imbalanced learning for face recognition and attribute prediction," 2018. [Online]. Available: arXiv:1806.00194.

[36] I. Hupont and C. Fernández, "DemogPairs: Quantifying the impact of demographic imbalance in deep face recognition," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Lille, France, May 2019, pp. 1–7.

[37] *Information Technology—Biometric Performance Testing and Reporting*, Standard ISO/IEC 19795-1:2006, 2016.

[38] V. Kazemi and J. Sullivan, 'One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 1867–1874.

[39] 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 Trans. Inf. Forensics Security*, vol. 7, pp. 1789–1801, 2012.

[40] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Long Beach, CA, USA, Jun. 2019, pp. 2261–2268.

[41] K. S. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer, "Issues related to face recognition accuracy varying based on race and skin tone," *IEEE Trans. Technol. Soc.*, vol. 1, no. 1, pp. 8–20, Mar. 2020.

[42] J. Liang, Y. Cao, C. Zhang, S. Chang, K. Bai, and Z. Xu, "Additive adversarial learning for unbiased authentication," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 11420–11429.

[43] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 3730–3738.

[44] B. Lu, J.-C. Chen, C. D. Castillo, and R. Chellappa, "An experimental evaluation of covariates effects on unconstrained face verification," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 1, no. 1, pp. 42–55, Jan. 2019.

[45] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips, "A meta-analysis of face recognition covariates," in *Proc. IEEE 3rd Int. Conf. Biometrics Theory Appl. Syst.*, Washington, DC, USA, 2009, pp. 1–8.

[46] D. Michalski, S. Y. Yiu, and C. Malec, "The impact of age and threshold variation on facial recognition algorithm performance using images of children," in *Proc. Int. Conf. Biometrics (ICB)*, Gold Coast, QLD, Australia, Feb. 2018, pp. 217–224.

[47] V. Mirjalili, S. Raschka, and A. Ross, "PrivacyNet: Semi-adversarial networks for multi-attribute face privacy," *IEEE Trans. Image Process.*, vol. 29, pp. 9400–9412, 2020.

[48] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana, "SensitiveNets: Learning agnostic representations with application to face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2158–2164, Jun. 2021.

[49] M. Orcutt, *Are Face Recognition Systems Accurate? Depends on Your Race*, MIT Technol. Rev., Cambridge, MA, USA, 2016.

[50] A. Ortega, J. Fierrez, A. Morales, Z. Wang, and T. Ribeiro, "Symbolic AI for XAI: Evaluating LFIT inductive programming for fair and explainable automatic recruitment," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVw)*, Waikola, HI, USA, Jan. 2021, pp. 78–87.

[51] A. Peña, I. Serna, A. Morales, and J. Fierrez, "Bias in multimodal AI: Testbed for fair automatic recruitment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRw)*, Seattle, WA, USA, Jun. 2020, pp. 129–137.

[52] A. Peña, I. Serna, A. Morales, J. Fierrez, and À. Lapedriza, "Facial expressions as a vulnerability in face recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Anchorage, AK, USA, Sep. 2021.

[53] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole, "An other-race effect for face recognition algorithms," *ACM Trans. Appl. Percept.*, vol. 8, no. 2, pp. 1–11, Feb. 2011.

[54] C. Rathgeb, A. Dantcheva, and C. Busch, "Impact and detection of facial beautification in face recognition: An overview," *IEEE Access*, vol. 7, pp. 152667–152678, 2019.

[55] C. Rathgeb, P. Drozdowski, N. Damer, D. C. Frings, and C. Busch, "Demographic fairness in biometric systems: What do the experts say?" 2021. [Online]. Available: arXiv:2105.14844.

[56] K. Ricanek, S. Bhardwaj, and M. Sodomsky, "A review of face recognition against longitudinal child faces," in *Proc. 14th Int. Conf. Biometrics Spec. Interest Group*, vol. P-245. Darmstadt, Germany, Sep. 2015, pp. 15–26.

[57] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, "Face recognition: Too bias, or not too bias?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Seattle, WA, USA, Jun. 2020, pp. 1–10.

[58] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 815–823.

[59] I. Serna, A. Morales, J. Fiérrez, M. Cebrián, N. Obradovich, and I. Rahwan, "Algorithmic discrimination: Formulation and exploration in deep learning-based face biometrics," in *Proc. AAAI Workshop Artif. Intell. Safety (SafeAI)*, Feb. 2020, pp. 146–152.

[60] I. Serna, A. Morales, J. Fierrez, M. Cebrian, N. Obradovich, and I. Rahwan, "SensitiveLoss: Improving accuracy and fairness of face representations with discrimination-aware deep learning," 2020. [Online]. Available: arXiv:2004.11246.

[61] I. Serna, A. Peña, A. Morales, and J. Fierrez, "InsideBias: Measuring bias in deep networks and application to face gender biometrics," in *Proc. IAPR Intl. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 3720–3727.

[62] M. Singh, R. Singh, and A. Ross, "A comprehensive overview of biometric fusion," *Inf. Fusion*, vol. 52, pp. 187–205, Dec. 2019.

[63] R. Singh, A. Agarwal, M. Singh, S. Nagpal, and M. Vatsa, "On the robustness of face recognition algorithms against attacks and bias," in *Proc. 34th AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 13583–13589.

[64] R. Singh, M. Vatsa, V. M. Patel, and N. K. Ratha, Eds., *Domain Adaptation for Visual Understanding*. Cham, Switzerland: Springer, 2020.

[65] N. Srinivas, M. Hivner, K. Gay, H. Atwal, M. King, and K. Ricanek, "Exploring automatic face recognition on match performance and gender bias for children," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Waikoloa, HI, USA, 2019, pp. 107–115.

[66] N. Srinivas, K. Ricanek, D. Michalski, D. S. Bolme, and M. King, "Face recognition algorithm bias: Performance differences on images of children and adults," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2269–2277.

[67] Y. Sun, M. Zhang, Z. Sun, and T. Tan, "Demographic analysis from biometric data: Achievements, challenges, and new frontiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 332–351, Feb. 2018.

[68] P. Terhörst, D. Fährmann, N. Damer, F. Kirchbuchner, and A. Kuijper, "Beyond identity: What information is stored in biometric face templates?" in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Houston, TX, USA, Sep. 2020, pp. 1–10.

[69] P. Terhörst, D. Fährmann, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "MAAD-face: A massively annotated attribute dataset for face images," 2020. [Online]. Available: arXiv:2012.01030.

[70] P. Terhörst *et al.*, "Reliable age and gender estimation from face images: Stating the confidence of model predictions," in *Proc. 10th IEEE Int. Conf. Biometrics Theory Appl. Syst. (BTAS)*, Tampa, FL, USA, Sep. 2019, pp. 1–8.

[71] 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 *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Houston, TX, USA, Sep./Oct. 2020, pp. 1–11.

[72] 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," 2020. [Online]. Available: arXiv:2002.03592.

[73] P. Terhörst, M. L. Tran, N. Damer, F. Kirchbuchner, and A. Kuijper, "Comparison-level mitigation of ethnic bias in face recognition," in *Proc. Int. Workshop Biometrics Forensics (IWBF)*, Porto, Portugal, Apr. 2020, pp. 1–6.

[74] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M. S. Nixon, "Soft biometrics and their application in person recognition at a distance," *IEEE Trans. Inf. Forensics Security*, vol. 9, pp. 464–475, 2014.

[75] P. Tome, J. Fierrez, R. Vera-Rodriguez, and D. Ramos, "Identification using face regions: Application and assessment in forensic scenarios," *Forensic Sci. Int.*, vol. 233, pp. 75–83, Dec. 2013.

[76] R. Vera-Rodríguez, M. Blázquez, A. Morales, E. Gonzalez-Sosa, J. C. Neves, and H. Proença, "FaceGenderID: Exploiting gender information in DCNNs face recognition systems," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2254–2260.

[77] P. Voigt and A. V. D. Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 1st ed. Cham, Switzerland: Springer Publ. Company, Incorp., 2017.
[78] M. Wang and W. Deng, "Deep face recognition: A survey," 2018. [Online]. Available: arXiv:1804.06655.
[79] M. Wang and W. Deng, "Mitigate bias in face recognition using skewness-aware reinforcement learning," 2019. [Online]. Available: arXiv:1911.10692.
[80] P. Wang, F. Su, Z. Zhao, Y. Guo, Y. Zhao, and B. Zhuang, "Deep class-skewed learning for face recognition," *Neurocomputing*, vol. 363, pp. 35–45, Oct. 2019.
[81] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Feature transfer learning for face recognition with under-represented data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5704–5713.
[82] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5419–5428.
[83] Y. Zhang and Z.-H. Zhou, "Cost-sensitive face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1758–1769, Oct. 2010.

Philipp Terhörst received the Ph.D. degree from TU Darmstadt, Darmstadt, Germany, in 2021.

He is a Researcher with Fraunhofer IGD, Darmstadt.

Dr. Terhörst received several awards for his scientific work, such as the EAB Biometrics Industry Award 2020 from the European Association for Biometrics for his dissertation and the IJCB 2020 Qualcomm PC Chairs Choice Best Student Paper Award.

Jan Niklas Kolf received the B.Sc. degree in computer science from TU Darmstadt, Darmstadt, Germany, where he is currently pursuing the M.Sc. degree in autonomous systems and visual computing.

He is a Researcher with Fraunhofer IGD, Darmstadt, working in the area of efficient machine learning for embedded biometrics.

Marco Huber received the M.Sc. degree in computer science and the M.Sc. degree in Internet- and Web-based systems from the Technical University of Darmstadt, Darmstadt, Germany, in 2021.

He is a Research Fellow with Fraunhofer IGD, Darmstadt. He is currently working on secure identity management in biometric systems.

Florian Kirchbuchner (Member, IEEE) received the M.Sc. degree in computer science from TU Darmstadt, Darmstadt, Germany, in 2014.

He is the Head of the Department for Smart Living and Biometric Technologies, Fraunhofer IGD, Darmstadt, Germany. He is the spokesperson for the Fraunhofer Alliance Ambient Assisted Living AAL and a Principal Investigator with the National Research Center for Applied Cybersecurity ATHENE.

Naser Damer (Member, IEEE) received the Ph.D. degree from TU Darmstadt, Darmstadt, Germany, in 2018.

He is a Senior Researcher with Fraunhofer IGD, Darmstadt. He is a Principal Investigator with the National Research Center for Applied Cybersecurity ATHENE.

Dr. Damer serves as an Associate Editor for the *Visual Computer* and represents the German Institute for Standardization (DIN) in ISO/IEC SC37 Standardization Committee.

Aythami Morales Moreno received the M.Sc. degree in electrical engineering and the Ph.D. degree from the Universidad de LPGC, Las Palmas, Spain, in 2006 and 2011, respectively.

Since 2017, he has been an Associate Professor with the Universidad Autonoma de Madrid, Madrid, Spain. In his work, he combines his interests in machine learning, biometric processing, security, and privacy.

Julian Fierrez (Member, IEEE) received the M.Sc. and Ph.D. degrees from the Universidad Politecnica de Madrid, Madrid, Spain, in 2001 and 2006, respectively.

Since 2004, he has been an Associate Professor with the Universidad Autonoma de Madrid, Madrid. His research is on signal and image processing, AI fundamentals and applications, HCI, forensics, and biometrics for security and human behavior analysis.

Arjan Kuijper received the Ph.D. degree from the Department of Computer Science and Mathematics, Utrecht University, Utrecht, The Netherlands, in 2002.

He is a Member of the Management of Fraunhofer IGD, Darmstadt, Germany, responsible for scientific dissemination. He has authored over 350 peer-reviewed publications. His research interests cover all aspects of mathematics-based methods for computer vision, biometrics, graphics, imaging, pattern recognition, interaction, and visualization.

Mr. Kuijper is an Associate Editor for *Computer Vision and Image Understanding*, *Pattern Recognition*, and *The Visual Computer Journal* and a Secretary of the International Association for Pattern Recognition.