2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Recognition Performance Variation Across Demographic Groups through the Eyes of Explainable Face Recognition

Marco Huber[1,2], Anh Thi Luu[1], Naser Damer[1,2]
[1]Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2]Department of Computer Science, TU Darmstadt, Darmstadt, Germany

***Abstract*— Face recognition systems are susceptible to differences in performance across demographic or non-demographic groups. However, the understanding of the behavior of face recognition models given such biases is still very limited and based mainly on observing model performance indicators when training/testing data is varied. On the other hand, very recently, face recognition explainability has gained increasing attention enabling the spatial explanation of face matching processes between two face images. This overcame the inapplicability of existing visual explainability methods to explain face matching decisions as they are designed for pure classification tasks. In this paper, and for the first time, we investigate the inner behavior of face recognition models with respect to bias using face recognition explainability tools. Using two state-of-the-art explainability tools, five models with different bias patterns, and a set of visualization tools, our investigation led to a set of interesting observations. This included noticing the tendency of more biased models to have more distributed attention on the facial image in comparison to focusing on the main facial features for the less biased models, all when considering the most discriminated demographic group.**

I. INTRODUCTION

Face recognition (FR) systems are integrated into our everyday lives and are used by a large number of users worldwide. Such users are diverse in terms of genders, ethnicities, and age groups, posing particular challenges for the technology, which should guarantee the same usability and security regardless of the demographic and non-demographic attributes of an individual user. However, recent works show that FR systems are biased towards demographic attributes (e.g. gender, ethnicity, age, ...) [1], [9], [27] and non-demographic attributes (e.g. facial hair style, illumination, headwear, ...) [54], [61], [10], [23]. This leads to recognition performance disparities depending on these attributes. This performance variation motivated a variety of works investigating possible sources [8], [62], [10], [3], [7], [6], measuring approaches [1], [14], ways to visualize [20] and investigate the bias problem further by e.g. asking experts [44].

Recently, motivated by the lack of transparency of the behavior of the highly complex deep learning-based FR models [41], explainability methods gained increasing interest. The lack of transparency and understandability of the inner workings of the FR models reduces trust in these systems, especially as they are often used in security-critical and sensitive areas such as passport control. Current explainability methods in FR aim at capturing uncertainties due to lack of identity information or ambiguities in presented identity information [49], [28], [43], [32], [29] or provide additional visualizations of the inner workings and the models' behavior by highlighting important areas [32], [34], [27], [37], [36].

Given the great recent advances in explaining face verifications, it is essential to gain insights into the inner implications of bias in FR models. Explainable face verification approaches are usually used to provide explanations of why two face images are matched or not matched. To do this, these approaches provide explanation maps, highlighting similar or dissimilar areas in a face image pair that support the verification decision [32], [34], [27]. In this work, we propose to use explainable face verification approaches to investigate and provide explanations of the recognition performance variations regarding ethnicity in FR models.

A first step towards gaining insights on how bias affects models' decisions was done in [20], which proposed to utilize Class Activation Maps (CAMs) to explain bias. They applied Score-CAM [55] on the feature-level which was developed to explain classification decisions. CAMs as they are designed to explain classifications, are not able to capture the whole face recognition process in their explanation as they do not utilize the feature-matching and decision-making process [36], [37], [27]. As using such classical visual explanations inherently does not fit the face matching process, we opted to utilize approaches specially designed for explainable FR rather than classification-based approaches like Score-CAM. This is thus the first work to investigate how explainable FR allows the understanding of performance variations across demographic groups.

In our investigation, we utilize two state-of-the-art explainable face verification approaches [32], [27] on five different FR models with varying ethnicity biases. The utilized explainable FR approaches were selected to include two different established paradigms, a black-box and a white-box approach. Regarding the FR models, we trained one reference model and four intentionally ethnicity-biased models using specific ethnicity subsets of a training dataset. To analyze the explanations, we provide the mean explanation maps based on the decision and compare the explanations of the more biased models with the explanations of the baseline model. In the next step, we also provide a magnitude-spatial and a spatial-variation analysis to quantify the obtained

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. This work has been partially funded by the German Federal Ministry of Education and Research through the Software Campus Project.

979-8-3503-9494-8/24/$31.00

results.

II. RELATED WORKS

A. Bias in Face Recognition

Exact definitions of bias, its implications, and its causes vary between different sources. A common interpretation is that it refers to performance variations that are influenced by a particular sub-population [45]. The ISO/IEC DIS 19795-10 standard provides an interpretation as "differential performance", which is defined as "difference in biometric system metrics across different demographics groups" [31]. Various demographic attributes were found to be vulnerable to bias and affecting the recognition performance. Several studies showed that FR systems are performing worse on female faces than on male faces [6], [5] and investigated possible sources (such as facial hair) [6], [7], [2]. A similar impact on the recognition performance can also be observed for ethnicity [1], [47], [12] and age [4], [15]. Regarding the impact of non-demographic attributes, [54] investigated 47 different attributes and their impact, including attributes related to hair color or face geometry, observing different degrees of impact. Besides the analysis of bias, several approaches have been proposed to mitigate bias [52], [35]. Nevertheless, performance variations remain an ongoing problem [46], also when using synthetic face data [26].

However, utilizing the FR model to investigate performance differences instead of investigating the impact of the data on the different performances has not been well-researched and has received little attention.

In [20] it is proposed to utilize Class Activation maps (CAMs) to gain a deeper insight into gender and ethnicity bias of trained FR models and their results aligned with human judgment on anthropometric differences. However, CAMs are not suitable to explain face verification decisions as they are designed to explain classification decisions which is not the case for face verification systems. Face verification systems include an embedding-matching and decision-making process that is not captured by CAMs. Therefore, we utilize explainable FR approaches specially designed to increase the transparency of FR systems to investigate and analyze their behavior regarding different ethnicity groups.

It should also be mentioned that bias can be observed in a wide range of other biometric scenarios then FR, such as presentation attack detection [17], [18], biometric sample quality [51], [21], face detection [38], synthetic face data [27] or keystroke dynamics [50].

B. Explainable Face Recognition

Increasing the understandability of face matching systems has recently gained attention in the biometric community [41]. The general idea is to increase the transparency of automatic decisions to ensure and increase trust in these systems. In Computer Vision, heatmaps or class activation maps (CAMs) and their variations [64], [55] have been a recent trend in explaining visual decisions to explain classification decisions. These maps highlight the most important areas that were crucial for the system's decision. These have been, for example, applied to face image quality estimation [19], [53], masked face recognition [39], [40] and face presentation attack detection [13], [48], [42], [25].

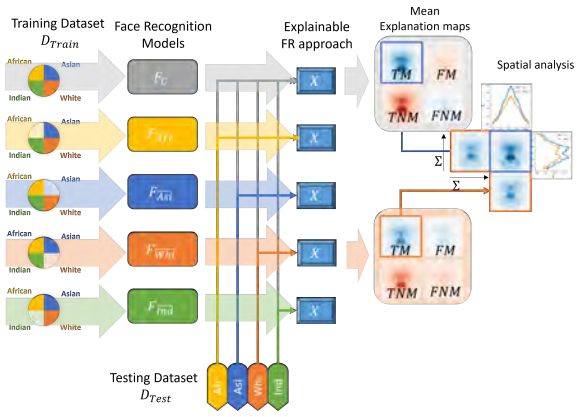

Fig. 1. Overview of the Proposed Investigation Methodology: Five FR models are trained on different (sub)sets of ethnicity-splitted data, creating four biased models and one baseline model (F_C). In the next step, ethnicity-split test data is processed by an explainable FR approach, creating mean explanation maps split by the models' decision for each ethnicity. In the final step, the obtained mean explanation maps are visually and with a spatial analysis investigated.

However, given the nature of face verification which includes a feature-matching and decision-making process, these CAMs are not suitable for explaining verification decisions. CAMs are designed to visualize why a certain class has been predicted by a classification system and not why two feature representations are matched.

Recently, inspired by this drawback, several approaches have been developed specifically to increase the transparency and interpretability of FR models [32], [34], [27], [37], [36]. These approaches provide explanation maps for a pair of face images but not based on a classification decision but on the similarity of the pair of faces or the verification decision, determined by the FR model.

In general, there are two different paradigms, either black-box approaches or white-box approaches, where black-box do not utilize the model's internals, while white-box approaches assume access to the model weights. Black-box approaches often manipulate the input images with different masks while then observing the output to estimate the impact of different facial regions on the system's decision [32], [37], [36], [34]. In contrast to this, white-box models may utilize the gradients of the input images to derive some meaningful gradient map that highlights the importance of different facial areas [27], [33]. Explainable FR approaches have proven to be applicable to different FR models, detect and highlight meaningful areas [32], [27], [34] and misguiding artificially added patches [27].

III. INVESTIGATION METHODOLOGY

In this section, we describe the proposed investigation methodology in more detail. We start with describing the approach to obtain variations in ethnicity-biased models, then provide a more detailed description of how we utilize

explainable FR methods and finalize with the proposed analysis approach.

A. Variations of Ethnicity-Biased FR Models

Although FR models are known to have performance variations across ethnicity groups [1], [47], [12] without specific attention on introducing bias, we aim at amplifying the performance variation to obtain more observable and analyzable results. To achieve this, we propose to intentionally train ethnicity-biased FR models with each model focusing on discriminating one ethnicity group. We achieve this by neglecting the discriminated ethnicity group during training.

To formalize, given a set E of ethnicity groups $e \in E$ with a face image training dataset D_{train} that provides $e-$based subsets, we train n different FR models $F_{\overline{e}}$ on $D_{train/\overline{e}}$, where the $e-$based subset was not part of the training data and n denotes the number of different ethnic groups e in E.

The hypothesis is that omitting an ethnicity subgroup during the training process leads to a relative reduction in performance compared to a baseline model trained on D_{train} without neglecting any subgroup.

B. Utilizing Explainable Face Recognition

Since FR models are not transparent to humans due to their high complexity and high number of parameters [41], explainable FR methods have been recently proposed. We propose to use these methods on different models with different degrees of performance variations across ethnicity groups to investigate and analyze these differences. The explainable FR methods provide a visualization in the form of an explanation map that indicates the most and least important face regions for the made FR decision.

Therefore, given an arbitrary explainable FR method X that is applied on an FR model F which processes two face images i, j, we can obtain an explanation map $h_{i,j}$ as:

$$h_{i,j} = X(F(i,j)) \tag{1}$$

Since the explainable FR method X is designed to capture the internal behavior of the model (to make its behavior more transparent to humans), it should capture differences due to the models' F bias and represent this in the obtained explanation map $h_{i,j}$.

In the experiments, we utilize two state-of-the-art explainable FR approaches, one black-box approach (xFace [32]) and one white-box approach (xSSAB [27]). For both approaches, we utilize the code provided in the corresponding official repositories.

xFace [32] is based on visualizing the difference or variation between an occluded and a non-occluded version of an image. If the similarity score is decreasing, the occluded area is considered dissimilar and vice versa. Since the three proposed methods in [32] only vary slightly in their performance [32], [27], we select Method 1. It calculates the cosine distance between all features of an embedded occluded and non-occluded image pair calculating the average influence of the occluded images on the distance. The obtained distance sets are then compared with the unaltered distance sets and weighted to obtain a similarity map. Since it does not require access to the models' architecture and weights, it is classified as a black-box approach.

xSSAB [27] is based on the idea of back-propagating the comparison score of an image pair based on the matching decision. To do this, the FR system is modeled in a Siamese setup and the feature dimensions of the face embeddings of the image pair are divided based on the fact if they increase or decrease the comparison score. Depending on whether they are contributing positively or negatively to the comparison score, gradients are back-propagated to obtain similar or dissimilar face regions. Since the obtained explanation maps are threshold-dependent for xSSAB, the threshold at the Equal Error Rate is chosen (False Match Rate = False Non-Match Rate) [27]. Since it requires access to the models' architecture and weights, it is considered a white-box approach.

C. Analysis Scheme

To analyze differences in the obtained explanation maps, we propose two different approaches. The first approach is based on comparing mean explanation maps between the more biased models and a baseline model. The second approach is based on creating spatial-magnitude and spatial-contribution plots that visualize the spatial differences in explanation maps of two models with different degrees of performance variations.

To investigate the performance variations with the mean explanation maps, we apply an FR explainability approach X on both, the baseline model F_C and the biased model $F_{\overline{e}}$ with image pairs of the ethnicity subset e of testing dataset D_{Test} to obtain two sets of explanation maps, $H_{C,e}$ and $H_{\overline{e},e}$. As we are interested in the differences between the F_C and the biased model $F_{\overline{e}}$, we only process image pairs that are part of ethnicity e. Since X aims at explaining the decision made by an FR model [27], [32], and genuine (matching) and imposter (non-matching) pairs are a different scenario and because previous work showed that bias affect the false matches differently than the false non-matches [30], we split the obtained set of explanation maps based on the decision before calculating the mean map. To be more precise, we choose a model-specific threshold, $t_{\overline{e}}$ for each $F_{\overline{e}}$ and split the obtained explanation map in *true match maps* (h^{TM}), *false match maps* (h^{FM}), *true non-match maps* (h^{TNM}), and *false non-match maps* (h^{FNM}) based on the comparison score of the image pair i, j used to calculate the specific explanation map $h_{i,j}$. To ensure that the mean map is calculated over the same samples, we only utilize the threshold $t_{\overline{e}}$ and the corresponding comparison score of the biased models and apply the same split on the maps $H_{C,e}$. By comparing the mean explanation maps based on the decision, we can investigate how the explainable FR approach X explains the performance variations on the same data for the different ethnicity subgroups.

Second, we measure the difference in terms of spatial variation along the horizontal and vertical axes of the calculated mean explanation maps. This allows us to quantify

the spatial differences. We propose two different versions of the spatial analysis, the spatial-magnitude analysis and the spatial-contribution analysis. In the spatial-magnitude analysis, we sum the values of the mean explanation maps over the x and y axis to also include differences in the magnitude of the values. This allows an analysis of the difference in attention magnitudes. In the spatial-contribution analysis, we first scale the values in the mean explanation maps to the range of (0,1) using min-max normalization and then calculate the percentage of contribution of each pixel to the total energy of the image by dividing each pixel by the total sum in the map. This neglects the magnitude of the values and allows us to investigate shifts in the attention distribution in the spatial domain.

IV. EXPERIMENTAL SETUP

A. Training & Testing Datasets

To train the different biased FR models ($F_{\overline{Whi}}, F_{\overline{Afr}}, F_{\overline{Asi}}, F_{\overline{Ind}}$) as well as the baseline model (F_C) as described in more detail in Subsection III-A, we utilized the BUPT-Balanceface (BUPT) dataset [57], [60]. The BUPT dataset provides four ethnicity-based subsets, covering the ethnicities *White*, *Indian*, *Asian*, and *African*. Each ethnicity subset consists of around 7k different subjects and around 300k images per ethnicity. The different biased models are then trained by leaving one ethnicity subset out, leading to 21k identities and 900k training images. The baseline model F_C is trained on all 28k subjects and 1.2M images.

To evaluate the models' bias, we utilize the Racial Faces-in-the-Wild (RFW) dataset [59]. The RFW dataset provides *African*, *Asian*, *White*, and *Indian* ethnicity subsets for evaluation. Each subset consists of around 10k images of 3k individuals and the protocol provides 6000 pre-defined pairs with 3000 genuine (matching) pairs and 3000 imposter (non-matching) pairs. All pairs are intra-ethnic (Black-Black, Asian-Asian,...). In total, there are 24k evaluation pairs. For the evaluation of RFW regarding verification performance, we follow the protocol and report the accuracy (in $\%$) for each intra-ethnic pair list.

B. Face Recognition Models - Training Setup

All the models are based on the ResNet-34 architecture [22] and utilize the state-of-the-art ElasticFace-Arc [11] loss function. We follow the recommendations in [11] and use scale parameter $s = 64$, margin $m = 0.5$ and standard deviation $\sigma = 0.05$ for the loss parameters. We set the mini-batch size to 128 and use Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 1e-1 [11]. We set the momentum to 0.9 and the weight decay to 5e-4 [11]. The learning rate is divided by 10 at 80k, 140k, 210k, and 280k training iterations [11]. The total number of epochs is 50 [11]. We keep this the same for all the trained models. During the training, we apply random horizontal flipping with a probability of 0.5 [11]. The images are aligned and cropped to $112 \times 112 \times 3$ using Multi-task Cascaded Convolutional Networks (MTCNN) [63] following [16] and also normalized

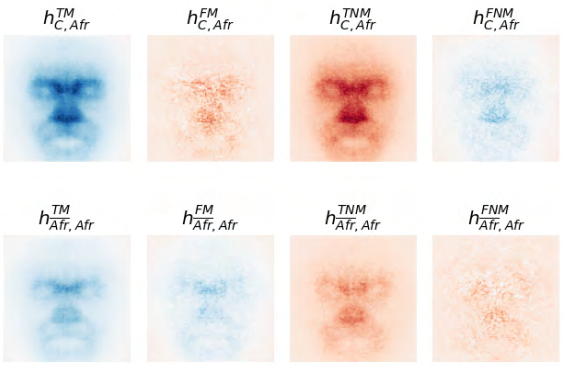

Fig. 2. Mean activation maps of xSSAB [27] for African, using the baseline Model F_C (top row) and the African-biased model $F_{\overline{Afr}}$ (bottom row).

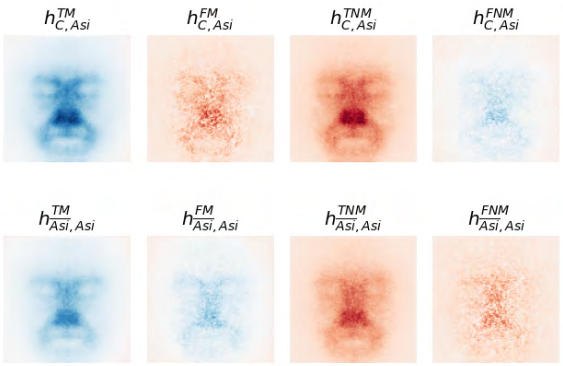

Fig. 3. Mean activation maps of xSSAB [27] for Asian, using the baseline Model F_C (top row) and the Asian-biased model $F_{\overline{Asi}}$ (bottom row).

to have pixel values between -1 and 1. The created face feature embeddings are 512-dimensional.

C. Evaluation Metric

To measure the bias of the intentionally biased models as well as the baseline model, we follow recent works and report the mean accuracy over the different ethnicities, the standard deviation (STD) [56], [24], [58] as well as the Skewed Error Ratio (SER) [56]. Error skewness is computed as the ratio of the highest error rate to the lowest error rate among different attributes (in our case ethnicities):

$$SER = \frac{max_a Err(a)}{min_b Err(b)} \tag{2}$$

where a, b are different ethnicities.

To analyze not only the overall fairness of the verification performance, we also provide the difference in ethnicity-specific accuracy ($\Delta \downarrow$) between the baseline model F_C and the biased models.

D. Explanation Map Processing & Visualization

To obtain the set of explanation maps H, we apply the explainable FR approaches on the pre-defined RFW pairs, obtaining one explanation map for each image of the pair. To be more robust against possible outliers and to increase the comparability, we normalize over all gradients of the maps obtained using the baseline models (H_C) and the maps obtained using the biased model ($H_{\overline{e}}$) to have a mean of

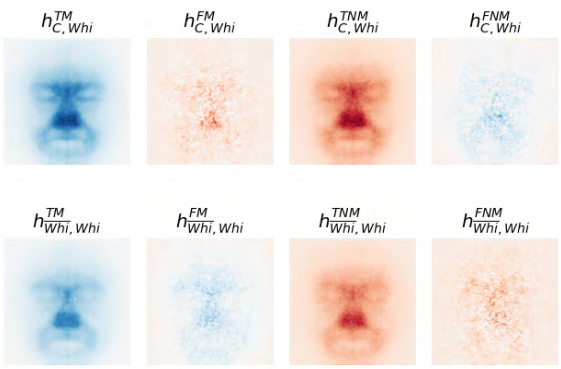

Fig. 4. Mean activation maps of xSSAB [27] for White, using the baseline model F_C (top row) and the White-biased model $F_{\overline{Whi}}$ (bottom row).

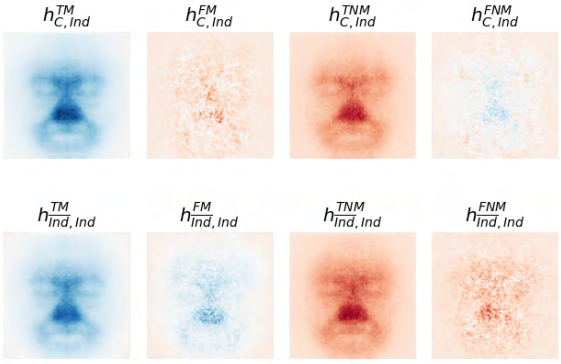

Fig. 5. Mean activation maps of xSSAB [27] for Indian, using the baseline model F_C (top row) and the Indian-biased model $F_{\overline{Ind}}$ (bottom row).

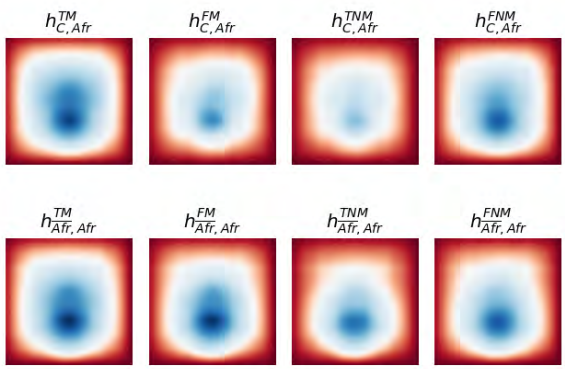

Fig. 6. Mean activation maps of xFace [32] for African, using the baseline model F_C (top row) and the African-biased model $F_{\overline{Afr}}$ (bottom row).

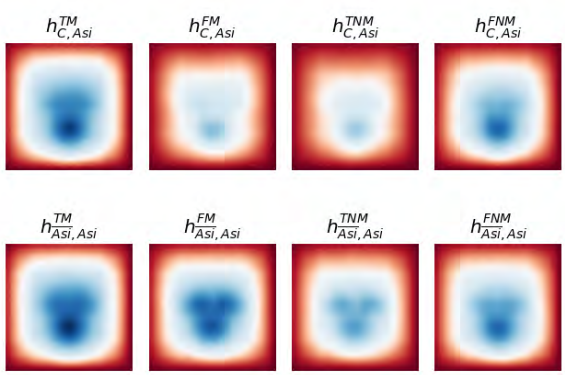

Fig. 7. Mean activation maps of xFace [32] for Asian, using the baseline model F_C (top row) and the Asian-biased model $F_{\overline{Asi}}$ (bottom row).

0 and a standard deviation of 1. To get the decision of the model $F_{\overline{e}}$, we set the decision threshold based on the Equal Error Rate (EER) and split the maps according to the decision into *True Match (TM)*, *False Match (FM)*, *True Non-Match (TNM)*, and *False Non-Match (FNM)*. To maintain the same pairs and the same number of pairs, we apply the same split on H_C.

For the visualization of the mean explanation maps, we select a two-slope color map, ranging from red (low) to blue (high), set white as 0, and normalize the values accordingly with the minimum and maximum value set based on the minimum and maximum values of H_C and $H_{\overline{e}}$ combined, which allows us to compare the intensity (magnitude) of the visualized mean maps within each figure. For the plots of the magnitude-spatial variation and the spatial variation, we follow the procedure described in Section III-C.

V. INVESTIGATION RESULTS

We summarize here the findings of our performed investigation. This includes the analysis of the mean explanation maps, the spatial-magnitude analysis, and the spatial-contribution analysis, from both FR explainability methods.

- *Fairness of the FR Models*: The fairness analysis is provided in Table I. The baseline model F_G achieves the highest mean accuracy (93.98%) as well as the highest accuracy in each ethnicity subset. Removing a certain ethnicity subset during training reduces the performance on the corresponding testing subset drastically (e.g. in the case of "African" from 92.92% to 80.25%) indicating a high performance variation. While removing a specific ethnicity affects also the performance on other ethnicity subsets, this effect is smaller. Based on the analysis above, the intentionally ethnicity-biased models are ethnicity-biased regarding the expected ethnicity subgroup and therefore suitable for the following investigations.
- *Differences in the Magnitude in the Explanation Maps:* The visual investigation in the Figures 2, 3, 4, and 5 for xSSAB and Figure 6, 7, 8, and 9 for xFace, as well as the spatial-magnitude analysis provided in Figures 10 and 11 show that both explainable FR approaches show differences in the magnitude of its explanations attentions based on the performance variations of the biased models and the baseline model. While the explanations of xSSAB show higher magnitudes in their maps (darker blue-ish and red-ish areas in the mean map visualizations) for the baseline model than for the more biased models, the explanation maps of xFace show a different behavior with slightly smaller magnitudes for the baseline model's explanation than for the more biased models' explanations. In the spatial-magnitude analysis, similar observations can be made. For example, in Figure 10, while both models focus on the eye and mouth region in the True match (TM) case for

Model	Training Data	RFW African	Δ ↓	Asian	Δ ↓	White	Δ ↓	Indian	Δ ↓	Bias Evaluation Metric Mean	STD	SER
F_C	BUPT	92.92	-	93.30	-	95.67	-	94.02	-	93.98	1.05	1.64
$F_{\overline{Afr}}$	w/o African	80.25	**-12.67**	92.33	-0.97	94.88	-0.79	93.15	-0.87	90.15	5.79	3.86
$F_{\overline{Asi}}$	w/o Asian	92.30	-0.62	83.97	**-9.33**	95.22	-0.45	93.32	-0.7	91.20	4.31	3.35
$F_{\overline{Whi}}$	w/o White	91.93	-0.99	92.78	-0.52	90.88	**-4.79**	93.10	-0.92	92.17	0.86	1.32
$F_{\overline{Ind}}$	w/o Indian	92.05	-0.87	92.73	-0.57	95.28	-0.39	90.17	**-3,85**	92.56	1.83	2.08

TABLE I

RECOGNITION PERFORMANCE [IN %] AND FAIRNESS OF THE FIVE USED FR MODELS ON THE RFW DATASET. REMOVING AN ETHNICITY SUBGROUP FROM THE TRAINING DATA DRASTICALLY REDUCES THE VERIFICATION PERFORMANCE ON THIS GROUP.

Africans (first column, first row, and second row), the baseline model (blue) shows a higher magnitude than the more African-biased model (orange). The opposite is true for the results using the xFace approach (Figure 11), where especially in the True Non-match cases (TNM, 3rd column) the magnitudes of the baseline model (blue) are smaller than the more biased models (orange). Nevertheless, differences in the magnitudes are observable in either direction (smaller values for xFace, larger values for xSSAB), capturing performance differences.

- *Investigating False Decisions:* A larger difference between the mean maps and also in the spatial-magnitude and spatial-contribution analysis is observable in the case of decision errors. In the mean maps using xSSAB on the baseline model F_C, the visual maps of false matches (FM) are visually more similar to the mean maps of true matches, while the more biased models show more similarity to the mean maps of the true matches. The opposite case is true for false non-matches (FNM). For the mean explanation maps for the errors produced by the xFace approach, it can be observed that especially for the false matches, the overall face area has higher values. This can also be observed in the spatial-magnitude analysis (Figure 11).
- *Differences in the Spatial-contribution Analysis:* The spatial-contribution analysis in Figures 12 for xSSAB and 13 for xFace shows that for xSSAB a larger difference between the baseline model (blue) and the more biased model (orange) in the spatial contributions are observable in the error cases (FM, column 2 and FNM, row column) in all different biased models. For the baseline model, the distribution of the contribution is more focused on the eye and mouth region (middle of the face) in comparison to the baseline model, where the contribution is more across the face regions. A similar observation can be made for xFace, however, with less difference in the distribution between the baseline and the biased models. This is especially true, for example, in the case of the y-axis of the Asian-biased model compared to the baseline model in the FM and TNM cases (column 2, 3, and row 4).

VI. CONCLUSION

We investigated performance variations across ethnicity groups using explainable face recognition. To do this, we trained five different FR models with bias regarding different ethnicities and investigated how the bias influences the explanation maps produced by two state-of-the-art explainable FR approaches, xFace and xSSAB. The findings obtained by investigating mean explanation maps as well as performing a spatial-magnitude and a spatial-contribution analysis are that performance variations lead to observable differences in the magnitude of the explanations. Moreover, major differences are observable in the error cases, false matches, and false non-matches, rather than in correct decisions.

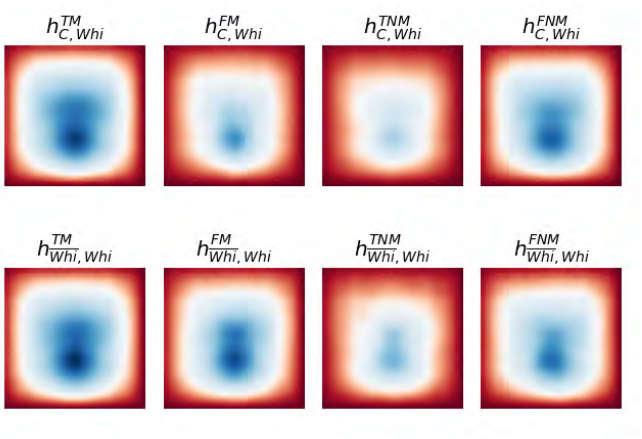

Fig. 8. Mean activation maps of xFace [32] for White, using the baseline model F_C (top row) and the White-biased model $F_{\overline{Whi}}$ (bottom row).

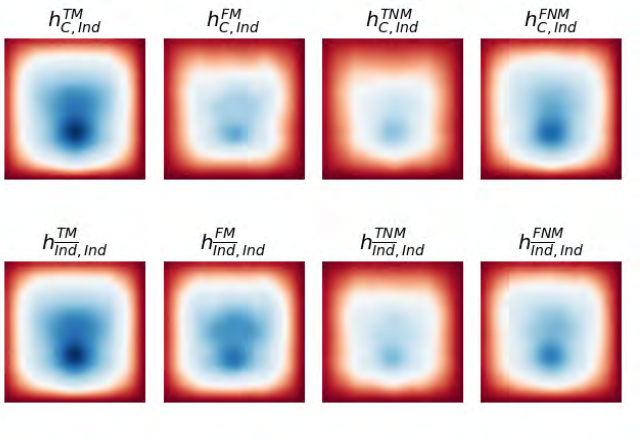

Fig. 9. Mean activation maps of xFace [32] for Indian, using the baseline model F_C (top row) and the Indian-biased model $F_{\overline{Ind}}$ (bottom row).

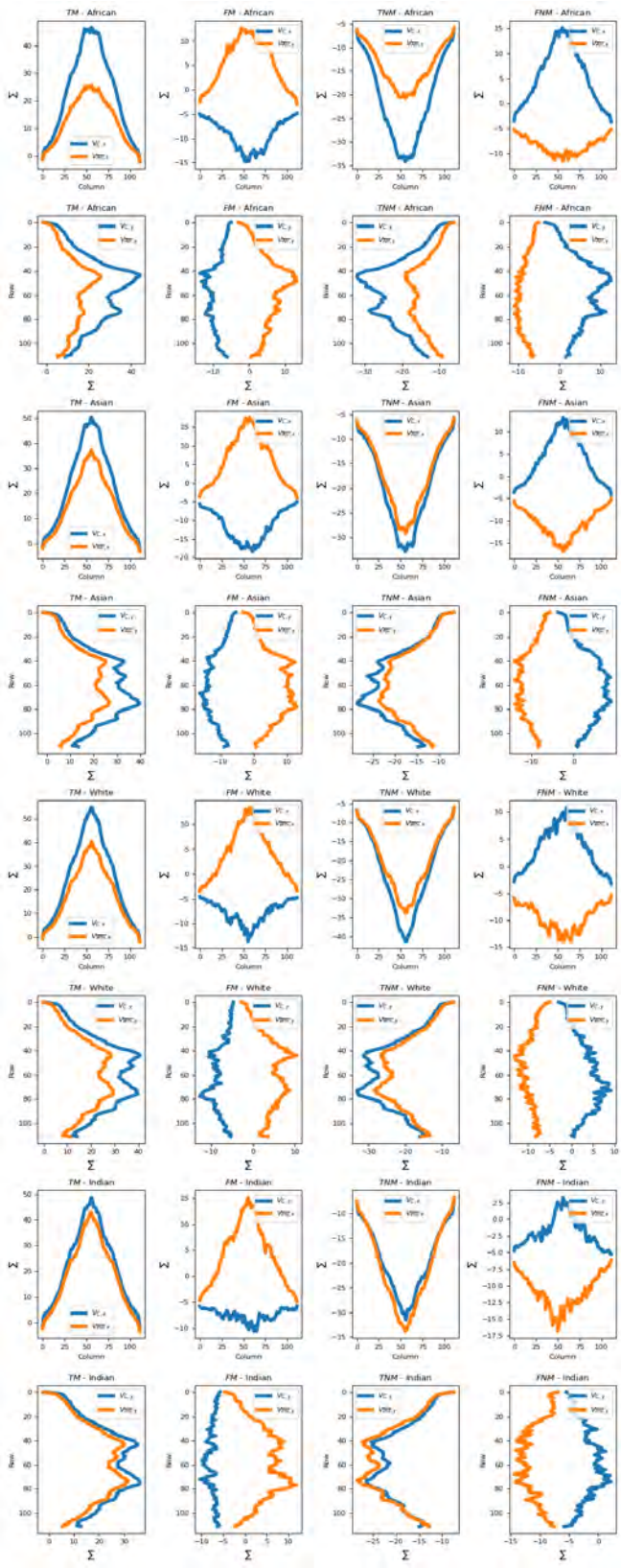

Fig. 10. Spatial-magnitude analysis for the different models and ethnicities (orange) compared with the baseline model (blue) using xSSAB [27].

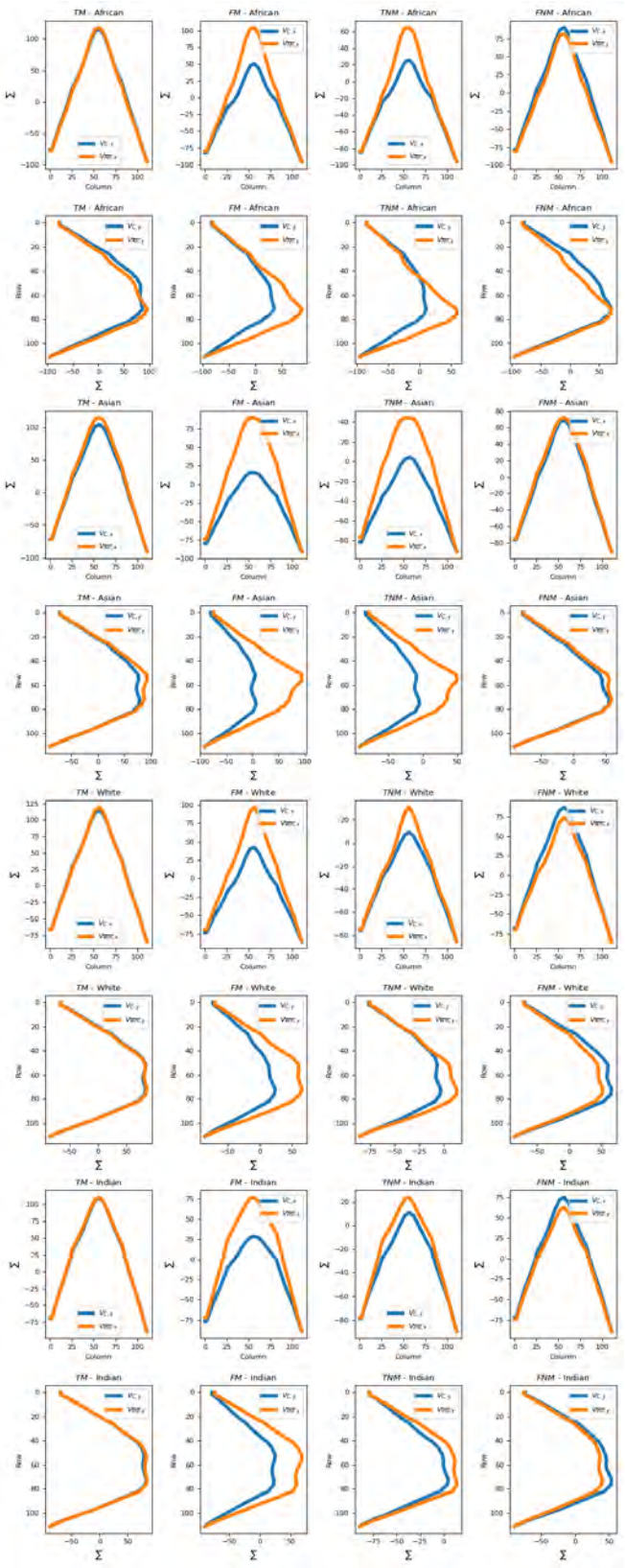

Fig. 11. Spatial-magnitude analysis for the different models and ethnicities (orange) compared with the baseline model (blue) using xFace [32].

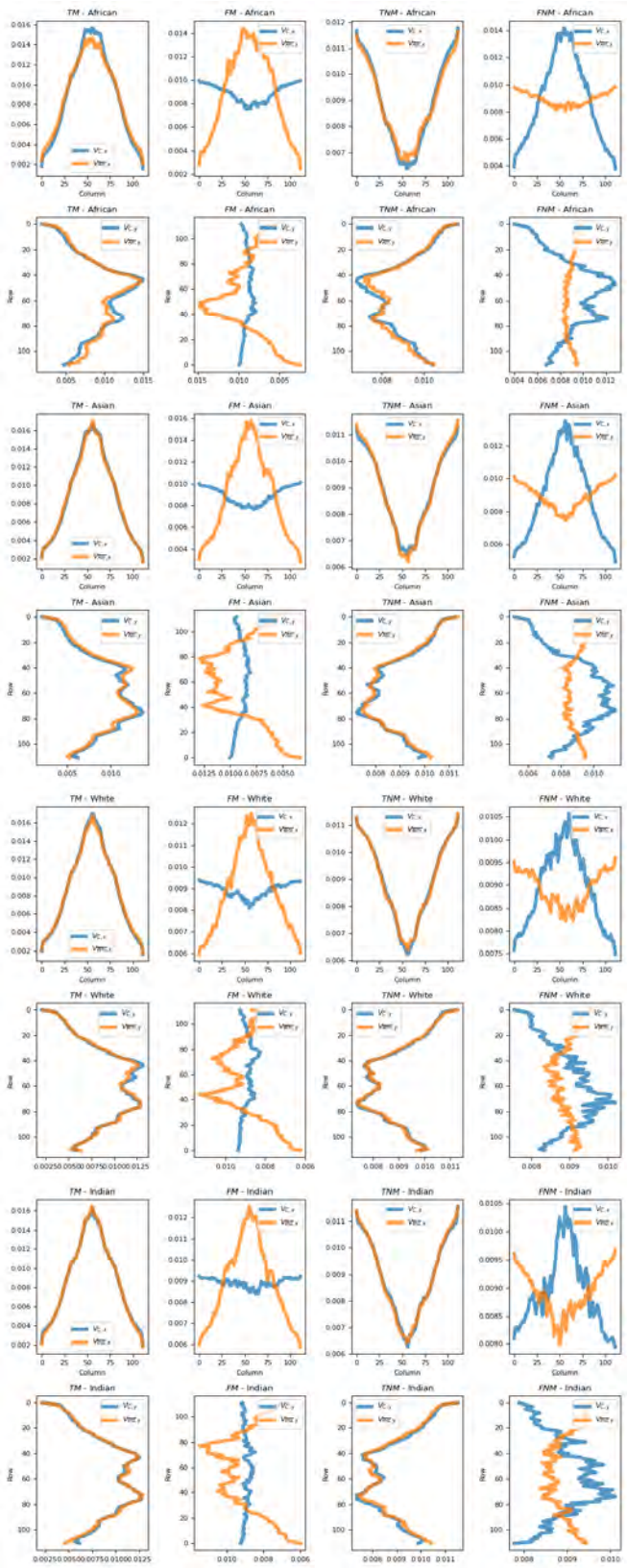

Fig. 12. Spatial-contribution analysis for the different models and ethnicities (orange) compared with the baseline model (blue) using xSSAB [27].

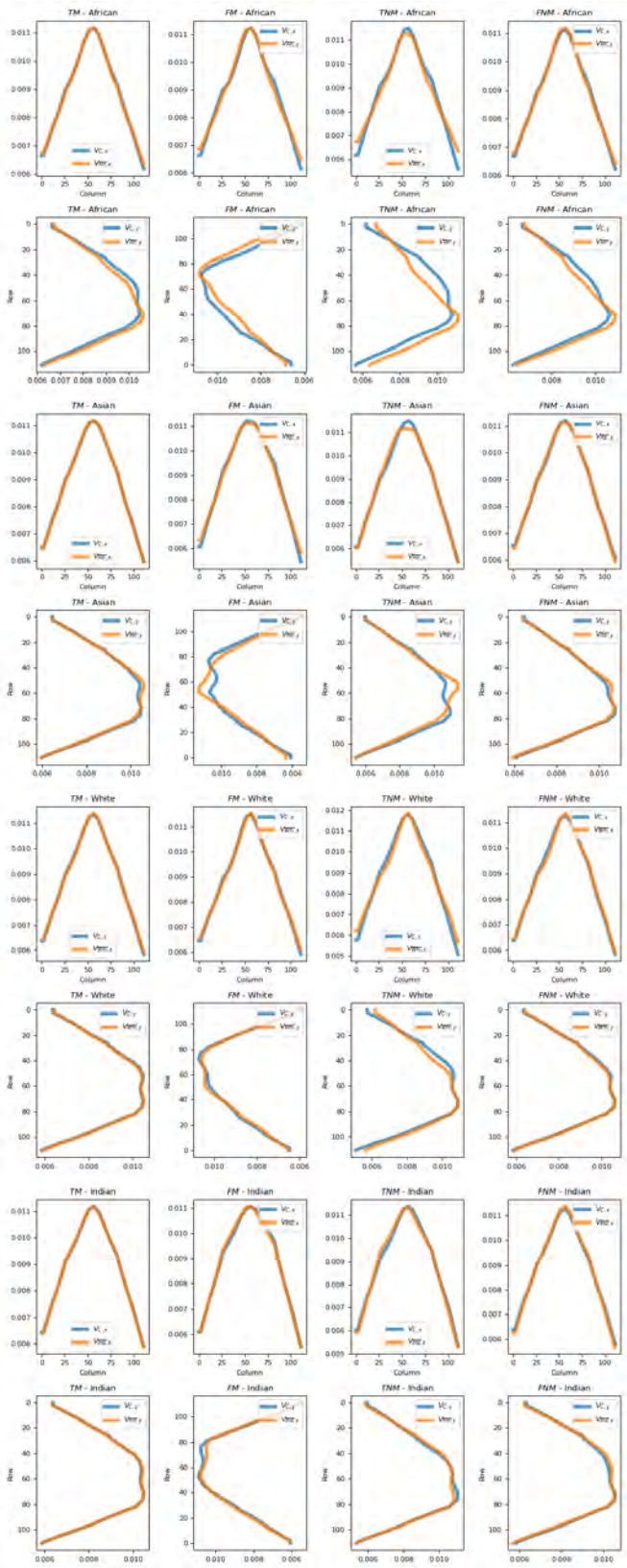

Fig. 13. Spatial-contribution analysis for the different models and ethnicities (orange) compared with the baseline model (blue) using xFace [32].

REFERENCES

[1] A. Acien, A. Morales, R. Vera-Rodríguez, I. Bartolome, and J. Fiérrez. Measuring the gender and ethnicity bias in deep models for face recognition. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 23rd Iberoamerican Congress, CIARP 2018, Madrid, Spain, November 19-22, 2018, Proceedings*, volume 11401 of *Lecture Notes in Computer Science*, pages 584–593. Springer, 2018.

[2] V. Albiero and K. W. Bowyer. Is face recognition sexist? no, gendered hairstyles and biology are. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

[3] V. Albiero, K. W. Bowyer, and M. C. King. Face regions impact recognition accuracy differently across demographics. In *IEEE International Joint Conference on Biometrics, IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022*, pages 1–9. IEEE, 2022.

[4] V. Albiero, K. W. Bowyer, K. Vangara, and M. C. King. Does face recognition accuracy get better with age? deep face matchers say no. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 250–258. IEEE, 2020.

[5] V. Albiero, K. K. S, K. Vangara, K. Zhang, M. C. King, and K. W. Bowyer. Analysis of gender inequality in face recognition accuracy. In *IEEE Winter Applications of Computer Vision Workshops, WACV Workshops 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 81–89. IEEE, 2020.

[6] V. Albiero, K. Zhang, and K. W. Bowyer. How does gender balance in training data affect face recognition accuracy? In *2020 IEEE International Joint Conference on Biometrics, IJCB 2020, Houston, TX, USA, September 28 - October 1, 2020*, pages 1–10. IEEE, 2020.

[7] V. Albiero, K. Zhang, M. C. King, and K. W. Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. *IEEE Trans. Inf. Forensics Secur.*, 17:127–137, 2022.

[8] A. Atzori, G. Fenu, and M. Marras. Explaining bias in deep face recognition via image characteristics. In *IEEE International Joint Conference on Biometrics, IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022*, pages 1–10. IEEE, 2022.

[9] A. Atzori, G. Fenu, and M. Marras. Demographic bias in low-resolution deep face recognition in the wild. *IEEE J. Sel. Top. Signal Process.*, 17(3):599–611, 2023.

[10] A. Bhatta, V. Albiero, K. W. Bowyer, and M. C. King. The gender gap in face recognition accuracy is a hairy problem. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV 2023 - Workshops, Waikoloa, HI, USA, January 3-7, 2023*, pages 1–10. IEEE, 2023.

[11] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 1577–1586. IEEE, 2022.

[12] 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? *IEEE Trans. Biom. Behav. Identity Sci.*, 3(1):101–111, 2021.

[13] A. da Silva Pinto, S. Goldenstein, A. M. Ferreira, T. J. Carvalho, H. Pedrini, and A. Rocha. Leveraging shape, reflectance and albedo from shading for face presentation attack detection. *IEEE Trans. Inf. Forensics Secur.*, 15:3347–3358, 2020.

[14] D. DeAlcala, I. Serna, A. Morales, J. Fiérrez, and J. Ortega-Garcia. Measuring bias in AI models: An statistical approach introducing n-sigma. In *47th IEEE Annual Computers, Software, and Applications Conference, COMPSAC 2023, Torino, Italy, June 26-30, 2023*, pages 1167–1172. IEEE, 2023.

[15] D. Deb, N. Nain, and A. K. Jain. Longitudinal study of child face recognition. In *2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018*, pages 225–232. IEEE, 2018.

[16] 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.

[17] 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.

[18] M. Fang, W. Yang, A. Kuijper, V. Struc, and N. Damer. Fairness in face presentation attack detection. *Pattern Recognit.*, 147:110002, 2024.

[19] 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.

[20] B. Fu and N. Damer. Towards explaining demographic bias through the eyes of face recognition models. In *IEEE International Joint Conference on Biometrics, IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022*, pages 1–10. IEEE, 2022.

[21] J. Galbally, R. Haraksim, and L. Beslay. Fingerprint quality: a lifetime story. In *2018 International Conference of the Biometrics Special Interest Group, BIOSIG 2018, Darmstadt, Germany, September 26-28, 2018*, volume P-282 of *LNI*, pages 1–5. GI / IEEE, 2018.

[22] 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.

[23] N. M. Herranz, C. Galdi, and J. Dugelay. Impact of digital face beautification in biometrics. In *10th European Workshop on Visual Information Processing, EUVIP 2022, Lisbon, Portugal, September 11-14, 2022*, pages 1–6. IEEE, 2022.

[24] L. Huang, M. Wang, J. Liang, W. Deng, H. Shi, D. Wen, Y. Zhang, and J. Zhao. Gradient attention balance network: Mitigating face recognition racial bias via gradient attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 38–47. IEEE, 2023.

[25] M. Huber, M. Fang, F. Boutros, and N. Damer. Are explainability tools gender biased? A case study on face presentation attack detection. In *31st European Signal Processing Conference, EUSIPCO 2023, Helsinki, Finland, September 4-8, 2023*, pages 945–949. IEEE, 2023.

[26] M. Huber, A. T. Luu, F. Boutros, A. Kuijper, and N. Damer. Bias and diversity in synthetic-based face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6215–6226, January 2024.

[27] M. Huber, A. T. Luu, P. Terhörst, and N. Damer. Efficient explainable face verification based on similarity score argument backpropagation. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pages 4724–4733. IEEE, 2024.

[28] M. Huber, P. Terhörst, F. Kirchbuchner, N. Damer, and A. Kuijper. Stating comparison score uncertainty and verification decision confidence towards transparent face recognition. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 506. BMVA Press, 2022.

[29] M. Huber, P. Terhörst, F. Kirchbuchner, A. Kuijper, and N. Damer. Uncertainty-aware comparison scores for face recognition. In *11th International Workshop on Biometrics and Forensics, IWBF 2023, Barcelona, Spain, April 19-20, 2023*, pages 1–6. IEEE, 2023.

[30] Face Recognition Vendor Test (FRVT) - Part 8: Summarizing Demographic Differentials. Technical report, National Institute of Standards and Technology, 2022.

[31] ISO/IEC DIS 19795-10: Information technology -Biometric performance testing and reporting - Part 10: Quantifying biometric system performance variation across demographic groups. Standard, International Organization for Standardization, 2024.

[32] M. Knoche, T. Teepe, S. Hörmann, and G. Rigoll. Explainable model-agnostic similarity and confidence in face verification. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV 2023 - Workshops, Waikoloa, HI, USA, January 3-7, 2023*, pages 1–8. IEEE, 2023.

[33] Y. Lu, Z. Xu, and T. Ebrahimi. Explainable face verification via feature-guided gradient backpropagation, 2024.

[34] Y. Lu, Z. Xu, and T. Ebrahimi. Towards visual saliency explanations of face verification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4726–4735, January 2024.

[35] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodríguez, A. Morales, D. Lawatsch, F. Domin, and M. Schaubert. Synthetic data for the mitigation of demographic biases in face recognition. *CoRR*, abs/2402.01472, 2024.

[36] D. Mery. True black-box explanation in facial analysis. In *IEEE/CVF*

Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022, pages 1595–1604. IEEE, 2022.

[37] D. Mery and B. Morris. On black-box explanation for face verification. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 1194–1203. IEEE, 2022.

[38] S. Mittal, K. Thakral, P. Majumdar, M. Vatsa, and R. Singh. Are face detection models biased? In *17th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2023, Waikoloa Beach, HI, USA, January 5-8, 2023*, pages 1–7. IEEE, 2023.

[39] P. C. Neto, F. Boutros, J. R. Pinto, N. Damer, A. F. Sequeira, and J. S. Cardoso. Focusface: Multi-task contrastive learning for masked face recognition. In *16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021, Jodhpur, India, December 15-18, 2021*, pages 1–8. IEEE, 2021.

[40] P. C. Neto, F. Boutros, J. R. Pinto, M. Saffari, N. Damer, A. F. Sequeira, and J. S. Cardoso. My eyes are up here: Promoting focus on uncovered regions in masked face recognition. In *Proceedings of the 20th International Conference of the Biometrics Special Interest Group, BIOSIG 2021, Digital Conference, September 15-17, 2021*, volume P-315 of *LNI*, pages 21–30. Gesellschaft für Informatik e.V., 2021.

[41] P. C. Neto, T. Gonçalves, J. R. Pinto, W. Silva, A. F. Sequeira, A. Ross, and J. S. Cardoso. Explainable biometrics in the age of deep learning. *CoRR*, abs/2208.09500, 2022.

[42] P. C. Neto, A. F. Sequeira, and J. S. Cardoso. Myope models - are face presentation attack detection models short-sighted? In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV - Workshops, Waikoloa, HI, USA, January 4-8, 2022*, pages 390–399. IEEE, 2022.

[43] P. C. Neto, A. F. Sequeira, J. S. Cardoso, and P. Terhörst. Pic-score: Probabilistic interpretable comparison score for optimal matching confidence in single- and multi-biometric face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 1021–1029. IEEE, 2023.

[44] C. Rathgeb, P. Drozdowski, D. C. Frings, N. Damer, and C. Busch. Demographic fairness in biometric systems: What do the experts say? *IEEE Technol. Soc. Mag.*, 41(4):71–82, 2022.

[45] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner. Face recognition: Too bias, or not too bias? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 1–10. Computer Vision Foundation / IEEE, 2020.

[46] A. Ross, S. Banerjee, C. Chen, A. Chowdhury, V. Mirjalili, R. Sharma, T. Swearingen, and S. Yadav. Some research problems in biometrics: The future beckons. In *2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*, pages 1–8. IEEE, 2019.

[47] K. K. S, K. Vangara, M. C. King, V. Albiero, and K. W. Bowyer. Characterizing the variability in face recognition accuracy relative to race. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2278–2285. Computer Vision Foundation / IEEE, 2019.

[48] A. F. Sequeira, T. Gonçalves, W. Silva, J. R. Pinto, and J. S. Cardoso. An exploratory study of interpretability for face presentation attack detection. *IET Biom.*, 10(4):441–455, 2021.

[49] Y. Shi and A. K. Jain. Probabilistic face embeddings. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6901–6910. IEEE, 2019.

[50] G. Stragapede, R. Vera-Rodríguez, R. Tolosana, A. Morales, N. Damer, J. Fiérrez, and J. Ortega-Garcia. Keystroke verification challenge (KVC): biometric and fairness benchmark evaluation. *IEEE Access*, 12:1102–1116, 2024.

[51] 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.

[52] 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.

[53] P. Terhörst, M. Huber, N. Damer, F. Kirchbuchner, K. Raja, and A. Kuijper. Pixel-level face image quality assessment for explainable face recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pages 1–1, 2023.

[54] 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.

[55] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 111–119. Computer Vision Foundation / IEEE, 2020.

[56] M. Wang and W. Deng. Mitigate bias in face recognition using skewness-aware reinforcement learning. *CoRR*, abs/1911.10692, 2019.

[57] 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.

[58] M. Wang and W. Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9319–9328. Computer Vision Foundation / IEEE, 2020.

[59] 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 *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 692–702. IEEE, 2019.

[60] M. Wang, Y. Zhang, and W. Deng. Meta balanced network for fair face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):8433–8448, 2022.

[61] H. Wu, V. Albiero, K. S. Krishnapriya, M. C. King, and K. W. Bowyer. Face recognition accuracy across demographics: Shining a light into the problem. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 1041–1050. IEEE, 2023.

[62] H. Wu, S. Tian, A. Bhatta, K. Öztürk, K. Ricanek, and K. W. Bowyer. Facial hair area in face recognition across demographics: Small size, big effect. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 1131–1140, January 2024.

[63] 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.

[64] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2921–2929. IEEE Computer Society, 2016.