Can the accuracy bias by facial hairstyle be reduced through balancing the training data?

Kagan Ozturk, Haiyu Wu, Kevin W. Bowyer
University of Notre Dame
{kztrk, hwu6, kwb}@nd.edu

Abstract

Appearance of a face can be greatly altered by growing a beard and mustache. The facial hairstyles in a pair of images can cause marked changes to the impostor distribution and the genuine distribution. Also, different distributions of facial hairstyle across demographics could cause a false impression of relative accuracy across demographics. We first show that, even though larger training sets boost the recognition accuracy on all facial hairstyles, accuracy variations caused by facial hairstyles persist regardless of the size of the training set. Then, we analyze the impact of having different fractions of the training data represent facial hairstyles. We created balanced training sets using a set of identities available in Webface42M that both have clean-shaven and facial hair images. We find that, even when a face recognition model is trained with a balanced clean-shaven / facial hair training set, accuracy variation on the test data does not diminish. Next, data augmentation is employed to further investigate the effect of facial hair distribution in training data by manipulating facial hair pixels with the help of facial landmark points and a facial hair segmentation model. Our results show facial hair causes an accuracy gap between clean-shaven and facial hair images, and this impact can be significantly different between African-Americans and Caucasians.

1. Introduction

While deep CNN face matchers have achieved great success [9, 19, 22, 25, 32, 36, 38], researchers are also investigating concerns about fairness or bias in accuracy [10, 13, 15, 16, 30, 31, 43, 45]. Because deep CNN face matchers are trained with huge web-scraped datasets and the decision-making of the final model is not transparent, the imbalance of demographic groups in the training data is often the "knee-jerk" first suspect for the cause of any bias.

However, it has been shown that having a training set that is explicitly balanced on the number of subjects and im-
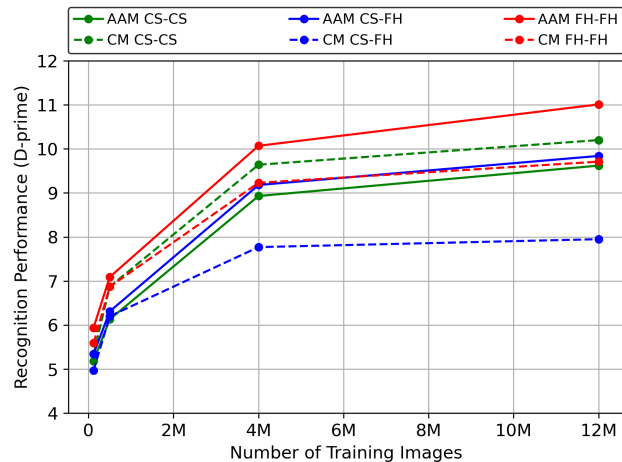

Figure 1. The effect of training size on recognition is given for African-American (AAM) and Caucasian (CM) males on MORPH using 4 training sets: a facial hair-balanced subset of WebFace42M (120K images Sec. 4.2), Casia-WebFace (500K images), WebFace4M (4M images) and WebFace12M (12M images). Higher d-prime between genuine and impostor distribution means better recognition accuracy. While d-prime values for CS-CS (clean-shaven v. clean-shaven), CS-FH (clean-shaven v. facial hair) and FH-FH (facial hair v. facial hair) image pairs consistently increase as training data gets larger, the d-prime gap across facial hairstyles also increases. AdaFace [19] loss is used to train the models.

ages is not sufficient to mitigate, for example, the accuracy difference across gender [2]. The correlation of various facial attributes with recognition accuracy differences has also been extensively investigated [37]. Moreover, bias measurement methods are analyzed in recent works [6, 11, 33].

Among the facial attributes, facial hairstyle (e.g., beard and mustache) can be easily changed to alter one's appearance. Facial hairstyle choices are influenced by a myriad of factors, including cultural norms, genetic factors and fashion trends. Facial hairstyle can occlude a substantial portion of the face and thereby affect the feature vector and similarity score produced by a face recognition model. This work

seeks to understand how the facial hairstyles present in a pair of images affect the impostor and genuine distribution, and whether the training data can be engineered to minimize accuracy differences across facial hairstyles. While the effect of facial hair on test set accuracy is analyzed in previous works (Section 2), this is the first comprehensive study of the influence of facial hair during training. First, effects of various amounts of facial hair in training set are measured. A controlled experiment is conducted by changing the facial hair percentage in a training set while maintaining the same set of identities. Then, data augmentation methods for manipulating pixels in the chin and mustache area are investigated. The accuracy discrepancy between image pairs (clean-shaven against clean-shaven), (clean-shaven against facial hair) and (facial hair against facial hair) is evaluated. Results show that facial hair can significantly impact the recognition performance on African-Americans and Caucasians even when a perfectly balanced training set is utilized. Furthermore, this impact is observed to be different across races, potentially leading to a false judgement of demographic bias in facial recognition.

This paper is organized as follows. Section 2 reviews previous works on fairness in machine learning. Also, effect of facial hair presence during evaluation is discussed for face recognition. Section 3 presents the problem statement and methodology to address the accuracy bias caused by facial hair. Our dataset and evaluation protocol is also explained. Section 4 describes our controlled training sets to measure the effect of facial hair during training. A data augmentation strategy on beard and mustache area is also investigated. Then, Section 5 discusses the effect of the data augmentation procedure on a web-crawled dataset for a comparison with our controlled training set experiment. Finally, we conclude with a summary of the work and suggestions for future research.

2. Related Work

Fairness is currently a topic of great interest in machine learning research. A comprehensive survey can be found in [21]. Bias in learning algorithms can be broadly analyzed in two dimensions: algorithmic bias and data bias. In the first case, it is assumed that the training set is balanced, representing all subpopulations equally, and learning bias is considered to come from the learning procedure. Bias towards low complexity solutions in deep networks are investigated in recent works [5, 27, 28, 35, 46, 50].

While the lack of theoretical foundation in deep learning complicates the study of bias, data is often the root cause of biased decisions in machine learning [7, 17, 18, 40, 43]. Particularly, bias in face recognition models is extensively studied in many works [8, 10, 23, 30, 31, 37, 41, 45, 49]. Balancing the number of images in training across demographics is initially considered as a simple solution to the bias

problem, but previous works [2, 39] show that gender gap continues to exist even with a perfectly balanced training set. In addition to the effort on constructing a "balanced" training set, bias measurement strategies on the test set are also investigated in previous studies [6, 17, 33, 34, 49].

While many of the aforementioned works investigate demographic biases across races, the effect of protected attributes obstruct the evaluation of fairness in face recognition [37]. In [41], effects of illumination on African-American and Caucasian images are investigated. A pre-trained BiSeNet [48] model is employed to segment a face image into 13 regions. Then, an image brightness metric is proposed to measure the face skin brightness. They show that comparison of two over-exposed (too bright) and comparison of two under-exposed (too dark) images can increase the match scores, resulting in higher FMR.

Impact of facial hair on recognition bias. Givens *et al.* [12] analyze the effect of facial hair together with other attributes on non-deep learning face recognition approaches. 2,144 images from the FERET [26] dataset are used in their experiments. A binary label is used to mark images as facial hair or clean-shaven. Their results suggest that when facial hair is present in one image and not the other, face recognition accuracy improves. This can be interpreted as increased dissimilarity in an impostor pair of images reducing the similarity score, and therefore reducing the chance of a false match.

In [20], the impacts of seven covariates with facial hair are examined through the utilization of five Deep Neural Networks (DNNs). The assessment of facial hair is based on four binary labels: "no facial hair", "mustache", "goatee", and "beard". Their findings indicate that state-of-the-art (SOTA) deep learning models can accommodate variations in facial hair. However, an uncontrolled dataset is used in their evaluation.

Terhörst *et al.* [37] presents a comprehensive analysis of facial attributes, including facial hair, with use of deep CNN face matchers. In one particular result investigating the effect of 5 o'clock shadow on face images, they report better recognition performance with 5 o'clock shadow compared to clean-shaven images.

The examination of the impact of both scalp hair and facial hair on face recognition is analyzed in [3]. Scalp hair is detected by a segmentation model [48] and combination of Microsoft Face API and Amazon Rekognition predictions are used to classify images as clean-shaven and facial hair. They report that the accuracy of clean-shaven prediction is lower for African-Americans than for Caucasians, suggesting development of a better classifier for a more accurate analysis.

In [42], a facial hair dataset is presented and a facial hair attribute classifier is proposed. Facial hair information about the area and length can be obtained using a set of 22
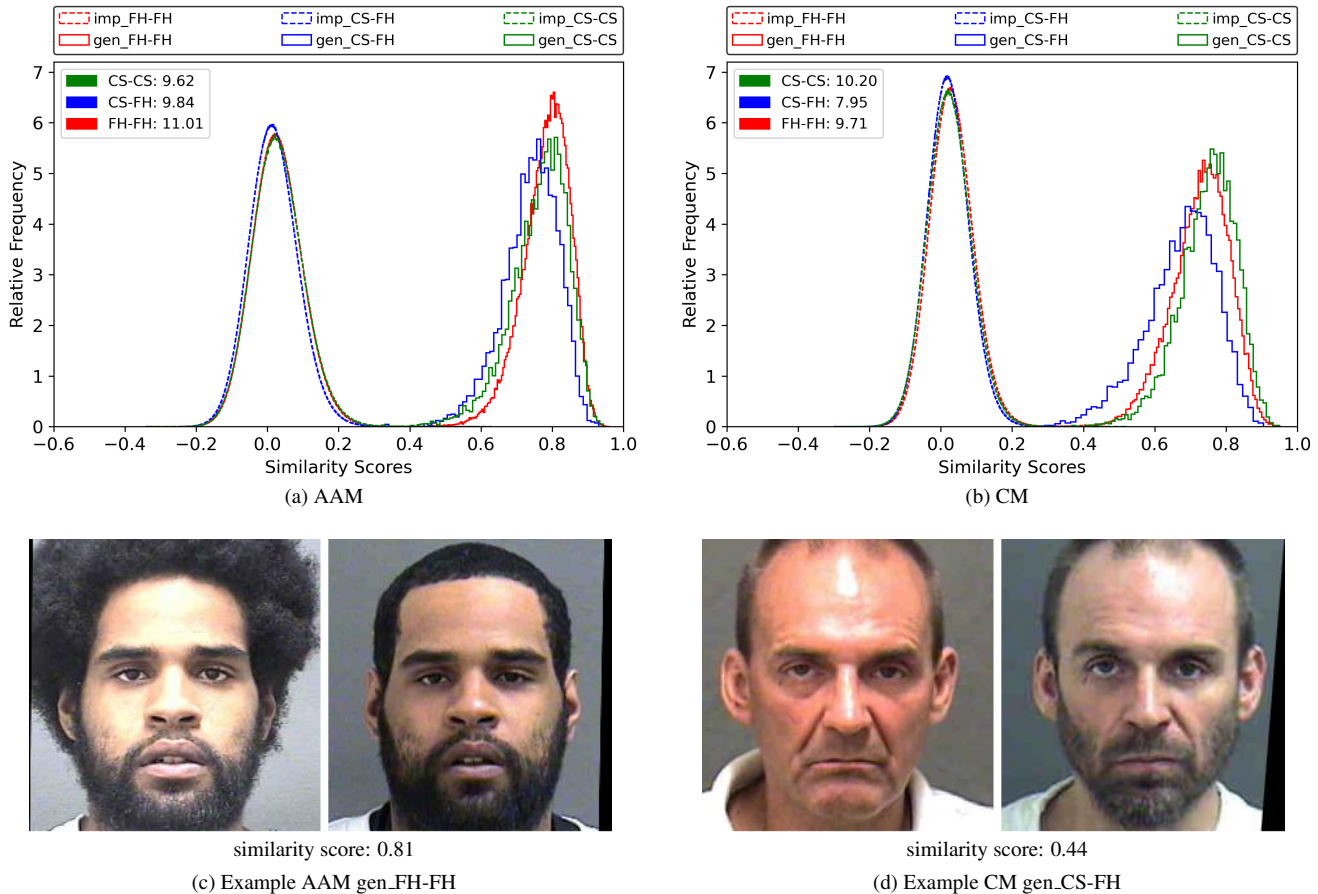
Figure 2. Impostor and genuine distributions for CS-CS, CS-FH, and FH-FH image pairs. Similarity scores are obtained using a pretrained AdaFace model on WebFace12M. D-prime values are given at the upper-left of the plots. While the recognition performance is better for Caucasian males on CS-CS pairs (b), d-prime values are greater for African-American males on CS-FH and FH-FH pairs (a). Examples of AAM FH-FH (c) and CM CS-FH (d) genuine pairs are shown.

binary labels. Using this classifier, the effect of beard region is analyzed [44]. They find mustache has a significant impact on recognition compared to facial hair on chin area, even though size of the mustache region is smaller.

A facial hair segmentation model is presented in [24], trained using a hand-annotated facial hair dataset. Images are categorized according to the number of facial hair pixels. It is shown that a greater amount of facial hair can cause a significant recognition bias across demographics.

3. Does recognition bias due to facial hairstyle decrease with larger training data?

We start our analysis by comparing the performance of face recognition models trained on different training sizes. Three publicly-available [1] models, trained on Casia-WebFace [47], WebFace4M and WebFace12M [52] to minimize AdaFace loss [19], are used for obtaining face representations. To investigate the recognition bias between clean-shaven and facial hair images, a facial hair classifier is used [42]. The classifier can predict attributes about the area and length of facial hair. Beard area attributes: *"no beard"*, *"chin"*, *"side to side"*. Mustache attributes: *"no mustache"*, *"mustache connected to beard"*, *"mustache isolated"*. Length attributes: *"5 o'clock shadow"*, *"short"*, *"medium"*, *"long"*. We categorize images as clean-shaven (CS) and facial hair (FH) according to predictions of the model using a threshold value of 0.7:

$$Images = \begin{cases} CS, \textbf{ if} \qquad \text{"no beard" \textbf{and} "no mustache"} \\[2ex] FH, \textbf{ if not} \quad \text{("no beard" \textbf{or} "no mustache"} \\ \qquad\qquad \textbf{or} \text{ "5 o'clock shadow")} \end{cases}$$

Dataset. MORPH [29] is a widely used dataset to perform age, gender and race analyses [2, 3, 41, 44]. MORPH images are captured in controlled conditions, as observed in passports, ID cards, etc., allowing a more detailed and ac-

curate analysis of facial hair impact. The dataset consists of 56,245 images of 8,839 African-American males (AAM) and 35,276 images of 8,835 Caucasian males (CM). We use 5,123 CS and 24,968 FH images for AAM and 8,497 CS and 12,578 FH images CM to report results across demographics. Mean face images are shown in Figure 3. The distribution of number of facial hair pixels, detected by a segmentation model [24] is given in Figure 4. AAM have more images with a smaller fraction of facial hair pixels. Distributions may vary for other datasets.

Evaluation. First, image representations are obtained using a face recognition model. Then, cosine distance between images is measured to obtain similarity score distributions. We use d-prime to calculate the distance between genuine and impostor distributions.

$$d' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}} \quad (1)$$

Higher d-prime values mean better separation of these distributions, indicating better recognition accuracy. Image pair group, CS-CS (clean-shaven against clean-shaven), CS-FH (clean-shaven against facial hair) and FH-FH (facial hair against facial hair) are created for evaluating performance discrepancy caused by facial hairstyle. We follow the same evaluation protocol for all the experiments throughout the paper.

Effects of training set size on the recognition performance for CS-CS, CS-FH and FH-FH is shown in Figure 1 across African-American (AAM) and Caucasian (CM) males. We evaluate the performance using 4 models, trained with 120K, 500K, 4M, 12M images. Even though d-prime values are increasing with larger training set size, there is a significant gap between particular groups (d-prime of AAM FH-FH is 11.01 while d-prime of CM CS-FH 7.95 using the model trained with 12M images) showing the impact of facial hairstyle in an image pair.

Impostor and genuine distributions for the AdaFace model trained on WebFace12M are given in Figure 2. D-prime values are also given in the upper left. It can be seen that, when two images have different facial hairstyle (CS-FH), both impostor and genuine similarity scores decrease. However, the genuine distribution of CS-FH Caucasian males is significantly shifted to lower scores, resulting in the worst d-prime value among all pair groups. Note that, while d-prime of CS-CS pairs is higher for Caucasian subjects, African-American subjects have better d-prime values for CS-FH and FH-FH image pairs. These results suggest that evaluation of recognition bias across races can be significantly influenced by the amount of clean-shaven and facial hair images in the test data.

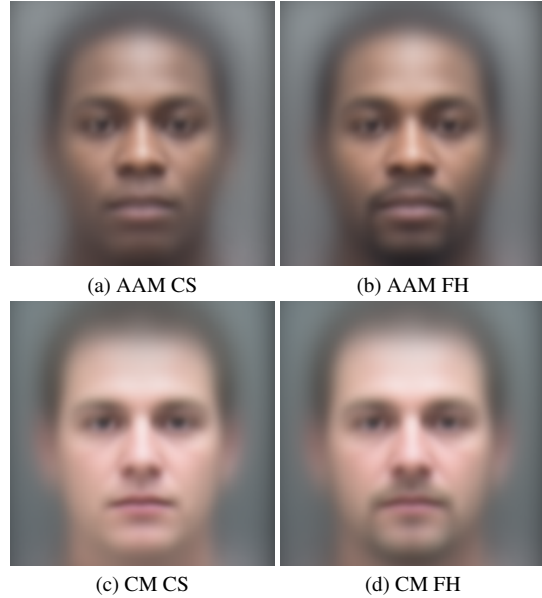| (a) AAM CS | (b) AAM FH |
| (c) CM CS | (d) CM FH |

Figure 3. Mean faces of clean-shaven (CS) and facial hair (FH) image sets for African-American (AAM) and Caucasian (CM) males on MORPH.
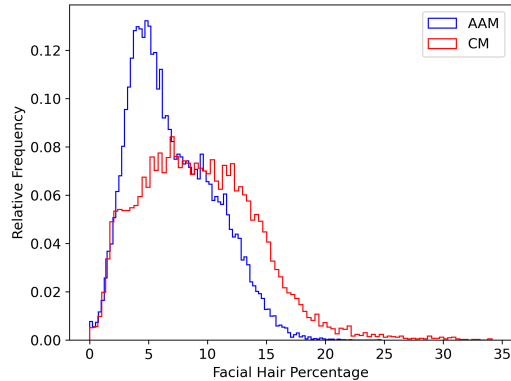

Figure 4. Percentage of number of facial hair pixels in a FH image for Caucasian and African-American males on MORPH.

4. How does the facial hair distribution in training data affect the performance?

To investigate the effect of facial hair distribution in training set, our goal is to have a set of male subjects that have both clean-shaven and facial hair images available to be used in the experiments. By doing so, we greatly control the other factors while changing the facial hair frequency in the training set; i.e., all training sets with different frequency of facial hair representation have the same set of identities and the same number of images per subject. The WebFace42M [52] dataset is used to find such subjects. Clean-shaven and facial hair images are detected using the facial hair classifier [42] (see Section 3).

Training data. Initially, we aim to have a training set size similar to Casia-WebFace [47] for comparison (10,575 subjects and 494,414 images). On WebFace42M, we found 5,000 subjects with 12 CS and 12 FH images (12-CS 12-FH). Subsets from this set of images are chosen to build our training sets. Facial hair frequency is controlled by selecting 12 images of these 5,000 male subjects. Additionally, we use the predictions of a gender classifier [17] to select 5,000 female subjects with 12 images to balance the number of images of male and female subjects. (Female subjects and images are chosen randomly from WebFace42M.) The same 60,000 female images are used in each training set.

Training sets with different facial hair frequency are constructed by two approaches: (1) selecting **x** number of males with only CS images and **5,000-x** male subjects with only FH images. (2) selecting **x** CS images and **12-x** FH images for all 5,000 male subjects. In addition to the selected 60,000 male images, the same 60,000 images of 5,000 female subjects are included in the training set, resulting in 10,000 subjects and 120,000 images for all training sets.

Implementation Details. A ResNet-50 [14] architecture is used to train models. Training lasts 30 epochs using 120,000 images with a batch size of 128 on a single GPU. Polynomial learning scheduler is employed with 3 warmup epochs. Learning rate is set to 0.1. AdaFace [19] loss is used for training.

4.1. Impact of facial hair variation across subjects

For the first experiment, we created training subsets in which each male subject has either only clean-shaven or only facial hair images. We select {0, 500, 1,000, 1,500, 2,000, 2,500, 3,000, 3,500, 4,000, 4,500, 5,000} males with 12 CS images and {5,000, 4,500, 4,000, 3,500, 3,000, 2,500, 2,000, 1,500, 1,000, 500, 0} males with 12 FH images respectively to form 11 different sets. Training is repeated 5 times, selecting a random set of subjects, from our superset (12-CS and 12-FH), with only clean-shaven and subjects with only facial hair images. In the two cases, all images are clean-shaven ("5000-CS 0-FH") and all images are facial hair ("0-CS 5000-FH"), the same 120,000 images are used in all repetitions with random initialization.

4.2. Impact of facial hair variation within subjects

In the second experiment, the frequency of facial hair is controlled by selecting {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12} CS images and {12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0} FH images for each male subject to construct 13 different training subsets. Training is repeated 5 times, selecting a random set of images, from our superset (12-CS and 12-FH), for each subject. Note that, in the case of all images are CS ("12-CS 0-FH"), the training set is the same as with 5,000-CS 0-FH (Section 4.1). Similarly, "0-CS 12-FH" corresponds to "0-CS 5,000-FH".

Results. Figure 5 shows the recognition performance for training subsets with different facial hair frequency. Dashed lines represent facial hair variation across subjects (Section 4.1), and solid lines are for facial hair variation within subjects (Section 4.2) in the training set. *It can be seen that, models trained with subjects that have both CS and FH images always perform better than the models trained with subjects that have only one type of facial hairstyle.* Overall, across 3 image pair groups and 2 races, d-prime values are higher when the facial hair frequency is between 30% and 70%. However, a performance gap still exits among CS-CS, CS-FH and FH-FH pairs, even when a balanced set (6-CS 6-FH) is used to train a face recognition model. In this experiment, we do not find a frequency of facial hairstyle in the training data that makes a major change in the hairstyle-related accuracy difference in the test data.

To further investigate the effects of facial hair representation in the training set, we explore two data augmentation approaches (Alg. 1). The first approach is copying facial hair pixels, detected by a facial hair segmentation model [24], from the images in Casia-WebFace. We use 2,500 images that have more than 15% of the pixels detected as facial hair. During training, a facial hair image is picked randomly and facial hair pixels are copied to a clean-shaven image. Facial landmark points [4] are used to warp the facial hair mask to match points between two images. In the second approach, instead of transferring real facial

---

Algorithm 1: Augment clean-shaven image

Data: CS_image, FH_image
Result: Augmented_image

1 FH_mask ← segmentFacialHair(FH_image);
2 CS_landmarks ← predictLandmarks(CS_image);
3 FH_landmarks ← predictLandmarks(FH_image);
4 Warped_FH_mask ← warpImage(FH_mask, CS_landmarks, FH_landmarks);
5 Augmented_image ← replacePixels(CS_image, Warped_FH_mask);
6 **return** Augmented_image;

---

hair pixels from a different image, we apply random pixel values to a target image [51]. Examples are shown in Figure 6.

The impact of data augmentation on recognition performance for CS-CS, CS-FH and FH-FH pair groups is reported in Table 1. Augmentation is applied to every image in the clean-shaven training set (12-CS 0-FH) with the probability of 0.2, 0.4, 0.6, 0.8 and 1.0. In addition, results of augmentation on only male images against augmenting all images (male and female) is compared. We find that, setting random pixel values in the beard area can be useful to increase recognition accuracy not only on facial hair im-
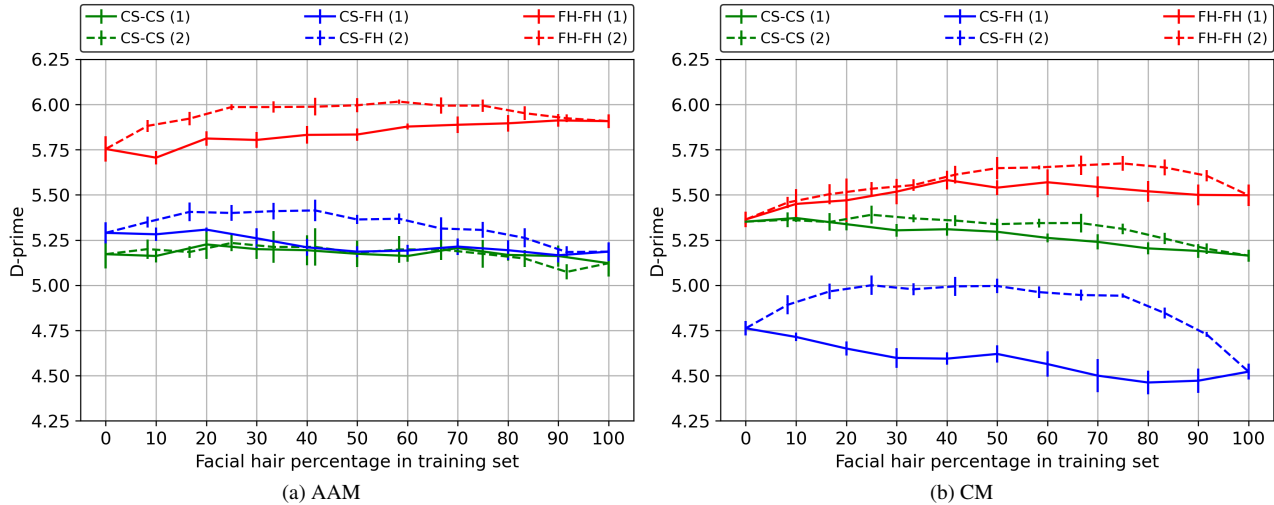
(a) AAM

(b) CM

Figure 5. Effect of facial hair distribution in training set. Recognition performance is measured on African-American (a) and Caucasian (b) males on MORPH. D-prime values are shown for CS-CS (clean-shaven versus clean-shaven), CS-FH (clean-shaven versus facial hair) and FH-FH (facial hair versus facial hair) image pairs. Dashed lines show the effect of facial hair ratio variation within subjects (Section 4.2) and solid lines shows the variation across subjects (Section 4.1). Vertical bars show the standard deviation of 5 repetition. Lower d-prime values are observed in most cases as facial hair percentage exceed 50% in training data.

(a) Clean-shaven images

(b) Facial hair images

(c) Augmentation using facial hair pixels
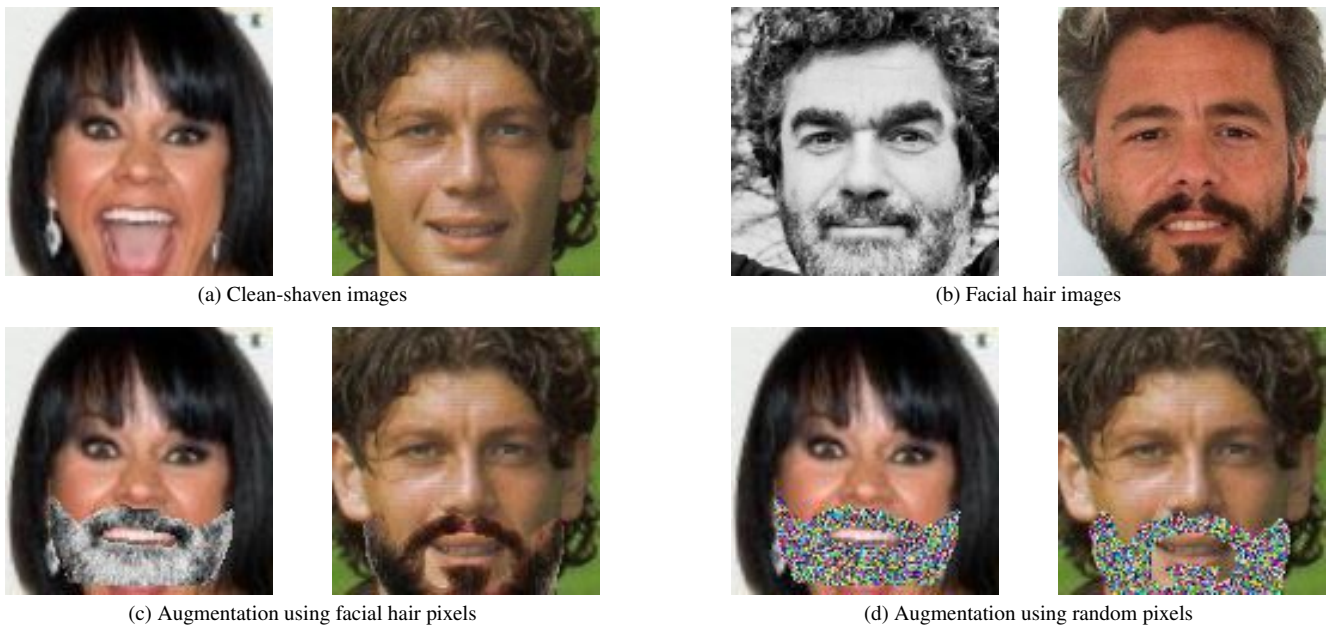
(d) Augmentation using random pixels

Figure 6. Data augmentation using facial hair and random pixels. Clean-shaven images (a) are augmented using facial hair pixels (b) from Casia-WebFace. First, a facial hair mask in obtained using a segmentation model. Then the mask is warped to match facial landmarks between two images and facial hair pixels copied to a clean-shaven image (c). Instead of facial hair pixel values, random values are applied to the corresponding area in the second approach (d).

ages but also on CS-CS pairs. It can be seen that d-prime values are increasing over the baseline (12-CS 0-FH) and even exceed the performance of perfectly balanced set (6-CS 6-FH) for AAM CS-CS, AAM CS-FH and CM CS-CS image pairs.

5. Facial hair area and Casia-WebFace

We further investigate the effect of beard area manipulation on Casia-WebFace to measure the effect on a larger training dataset. First, a face recognition model is trained without data augmentation as baseline. Training is repeated

Train	Test					
	AAM			CM		
	CS-CS	CS-FH	FH-FH	CS-CS	CS-FH	FH-FH
0-CS 12-FH	5.12 ± 0.07	5.19 ± 0.05	5.91 ± 0.04	5.16 ± 0.03	4.52 ± 0.04	5.50 ± 0.06
6-CS 6-FH	5.18 ± 0.06	5.36 ± 0.03	6.00 ± 0.04	5.34 ± 0.03	5.00 ± 0.04	5.65 ± 0.06
12-CS 0-FH	5.17 ± 0.08	5.29 ± 0.06	5.75 ± 0.07	5.35 ± 0.01	4.76 ± 0.04	5.36 ± 0.04
12-CS 0-FH w/ facial hair pixel aug. (male)						
prob. 0.2	5.19	5.30	5.85	5.33	4.79	5.40
prob. 0.4	5.16	5.37	5.87	5.28	4.85	5.45
prob. 0.6	5.08	5.36	5.83	5.25	4.93	5.45
prob. 0.8	5.07	5.30	5.79	5.16	4.80	5.36
prob. 1.0	5.13	5.27	5.82	5.09	4.73	5.43
12-CS 0-FH w/ facial hair pixel aug. (male and female)						
prob. 0.2	5.21	5.39	5.82	5.34	4.88	5.41
prob. 0.4	5.10	5.38	5.84	5.32	4.90	5.43
prob. 0.6	5.07	5.34	5.81	5.27	4.87	5.39
prob. 0.8	5.12	5.33	5.80	5.19	4.86	5.39
prob. 1.0	5.12	5.35	5.82	5.23	4.85	5.44
12-CS 0-FH w/ random pixel aug. (male)						
prob. 0.2	5.10	5.32	5.83	5.37	4.95	5.49
prob. 0.4	5.19	5.40	5.90	5.29	4.84	5.47
prob. 0.6	5.22	5.33	5.84	5.22	4.76	5.36
prob. 0.8	5.20	5.34	5.85	5.22	4.81	5.36
prob. 1.0	5.00	5.13	5.63	5.13	4.60	5.20
12-CS 0-FH w/ random pixel aug. (male and female)						
prob. 0.2	5.27	5.34	5.88	5.46	4.89	5.41
prob. 0.4	5.31	5.43	5.93	5.38	4.96	5.51
prob. 0.6	5.07	5.27	5.77	5.28	4.83	5.30
prob. 0.8	5.11	5.28	5.79	5.24	4.82	5.37
prob. 1.0	5.07	5.22	5.76	5.14	4.73	5.36

Table 1. Effect of data augmentation on beard area. D-prime values between genuine and impostor distributions of image pairs (CS-CS, CS-FH and FH-FH) are given for African-American (AAM) and Caucasian (CM) males. The first three rows show the performance of models trained with: only facial hair images (0-CS 12-FH), balanced set (6-CS 6-FH) and only clean-shaven (12-CS 0-FH) images. Four data augmentation approaches are investigated on only clean-shaven set (12-CS 0-FH) with 5 different probability. Best performance gain is observed with random pixel augmentation to beard area on all male and female images.

5 times with random initiation. Mean and standard deviation is reported in Table 2. We first conduct an experiment to investigate usage of random pixel values on a beard area, instead of real facial hair pixels. A facial hair segmentation model [24] is used to segment facial hair region on 490,623 images. Facial hair masks for 93,273 images are obtained. Pixels labeled as facial hair are replaced with random pixel values with the 5 levels of probability (0.2, 0.4, 0.6, 0.8, 1.0) for each image during training. It can be seen in Table 2 (*Casia-WebFace w/ random pixel aug. instead of facial hair (male)*) that even replacing all real facial hair pixels with random values does not have a significant effect over the baseline (*Casia-WebFace wo/ aug.*). The best results for CM CS-CS image pairs (6.94) is observed with this data augmentation approach with the probability of 0.2.

Next, we apply random pixel augmentation procedure to only mustache area and both beard and mustache area to measure the importance of the particular face area to be augmented. Comparison between data augmentation on only male images against augmenting all male and females images is also reported. 68 facial landmark points [4] are used to define mustache and beard area for each image. Lower performance is observed for all cases as the probability of applying augmentation increases. The best results are achieved augmenting mustache and beard area on all images with probability of 0.2.

6. Conclusion and Discussion

We investigate the recognition bias (accuracy difference) that arises due to clean-shaven and facial hair face images.

Train	Test					
	AAM			CM		
	CS-CS	CS-FH	FH-FH	CS-CS	CS-FH	FH-FH
Casia-WebFace wo/ aug.	6.20 ± 0.05	6.29 ± 0.02	7.10 ± 0.02	6.86 ± 0.02	6.11 ± 0.09	6.82 ± 0.05
Casia-WebFace w/ random pixel aug. instead of facial hair (male)						
prob. 0.2	6.20	6.38	7.12	6.94	6.10	6.81
prob. 0.4	6.24	6.35	7.10	6.90	6.09	6.77
prob. 0.6	6.17	6.28	7.04	6.90	6.11	6.88
prob. 0.8	6.22	6.30	7.08	6.91	6.17	6.83
prob. 1.0	6.19	6.32	7.16	6.88	6.06	6.82
Casia-WebFace w/ random pixel aug. mustache area (male)						
prob. 0.2	6.20	6.26	7.14	6.87	6.11	6.89
prob. 0.4	6.29	6.35	7.11	6.84	6.08	6.81
prob. 0.6	6.17	6.21	7.04	6.75	6.13	6.77
prob. 0.8	6.08	6.29	7.05	6.70	6.04	6.79
prob. 1.0	5.87	5.80	6.44	6.30	5.60	6.19
Casia-WebFace w/ random pixel aug. mustache area (male and female)						
prob. 0.2	6.28	6.37	7.15	6.81	6.14	6.89
prob. 0.4	6.24	6.28	7.13	6.80	6.13	6.86
prob. 0.6	6.16	6.30	7.10	6.72	6.00	6.78
prob. 0.8	6.12	6.29	7.04	6.58	5.99	6.76
prob. 1.0	5.65	5.86	6.60	6.24	5.66	6.24
Casia-WebFace w/ random pixel aug. mustache and beard area (male)						
prob. 0.2	6.23	6.34	7.13	6.83	6.24	6.93
prob. 0.4	6.16	6.27	7.05	6.81	6.17	6.80
prob. 0.6	6.15	6.39	7.09	6.71	6.02	6.66
prob. 0.8	6.02	6.26	6.95	6.62	6.05	6.67
prob. 1.0	5.54	5.54	6.05	5.99	5.14	5.57
Casia-WebFace w/ random pixel aug. mustache and beard area (male and female)						
prob. 0.2	6.29	6.45	7.19	6.85	6.20	6.82
prob. 0.4	6.15	6.34	7.07	6.76	6.16	6.87
prob. 0.6	6.04	6.26	6.97	6.71	6.07	6.65
prob. 0.8	6.05	6.26	6.88	6.52	5.95	6.55
prob. 1.0	5.41	5.60	6.27	5.86	5.26	5.95

Table 2. Effect of data augmentation on beard and mustache area during training is analyzed on Casia-WebFace. Augmentation results are reported for 5 range of probability. Comparison between augmentation on only male against all images is given.

First, we show that even though larger training sets increase the performance for all image pairs, the accuracy gap caused by facial hairstyles does not decrease; see Figure 1. Going from 4M to 12M training images does not significantly improve results for CM CS-FH. To investigate this discrepancy, training sets with varying facial hair amount are constructed. Results show that the accuracy gap does not diminish even with a perfectly balanced set.

Our findings on the effect of facial hair show that balanced training set construction either using only real images or using a data augmentation method that targets only the facial hair area on face images does not significantly mitigate bias caused by facial hairstyle. Note that the accuracy discrepancy is observed on African-American and Caucasian images in different amounts. While better recognition performance is measured on CS-CS image pairs for Caucasian male subjects, genuine similarity scores of CS-FH and FH-FH pairs are higher for African-American resulting in better accuracy. This suggests that judgement on whether recognition bias exists between demographics is heavily influenced by the facial hair present in the evaluation set and rigorous methodologies are necessary for bias assessment.

References

[1] https://github.com/mk-minchul/AdaFace. 3
[2] Vítor Albiero, Kai Zhang, and Kevin W. Bowyer. How does

gender balance in training data affect face recognition accuracy? In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 2020. 1, 2, 3

[3] Aman Bhatta, Vítor Albiero, Kevin W Bowyer, and Michael C King. The gender gap in face recognition accuracy is a hairy problem. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 303–312, 2023. 2, 3

[4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 5, 7

[5] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. In *30th International Joint Conference on Artificial Intelligence (IJCAI 2021)*, pages 2205–2211. International Joint Conferences on Artificial Intelligence, 2021. 2

[6] Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):101–111, 2020. 1, 2

[7] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019. 2

[8] Tiago de Freitas Pereira and Sébastien Marcel. Fairness in biometrics: a figure of merit to assess biometric verification systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(1):19–29, 2021. 2

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1

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

[11] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223:103552, 2022. 1

[12] G Givens, JR Beveridge, BA Draper, P Grother, and PJ Phillips. How features of the human face affect recognition: a statistical comparison of three face recognition algorithms. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 381–389, 2004. 2

[13] Patrick J Grother, Mei L Ngan, and Kayee K Hanaoka. Face recognition vendor test part 3: Demographic effects. 2019. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[15] Isabelle Hupont and Carles Fernández. Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019. 1

[16] Anil K Jain, Debayan Deb, and Joshua J Engelsma. Biometrics: Trust, but verify. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):303–323, 2021. 1

[17] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 2, 5

[18] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019. 2

[19] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18750–18759, 2022. 1, 3, 5

[20] Boyu Lu, Jun-Cheng Chen, Carlos D Castillo, and Rama Chellappa. An experimental evaluation of covariates effects on unconstrained face verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pages 42–55, 2019. 2

[21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. 2

[22] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021. 1

[23] Pedro C Neto, Eduarda Caldeira, Jaime S Cardoso, and Ana F Sequeira. Compressed models decompress race biases: What quantized models forget for fair face recognition. In *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2023. 2

[24] Kagan Ozturk, Grace Bezold, Aman Bhatta, Haiyu Wu, and Kevin Bowyer. Beard segmentation and recognition bias. *arXiv preprint arXiv:2308.15740*, 2023. 3, 4, 5, 7

[25] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015. 1

[26] P.Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998. 2

[27] Tomaso Poggio, Kenji Kawaguchi, Qianli Liao, Brando Miranda, Lorenzo Rosasco, Xavier Boix, Jack Hidary, and Hrushikesh Mhaskar. Theory of deep learning iii: the non-overfitting puzzle. *CBMM Memo*, 73:1–38, 2018. 2

[28] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and

Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019. 2

[29] Karl Ricanek Jr and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 341–345, 2006. 3

[30] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–1, 2020. 1, 2

[31] Joseph P Robinson, Can Qin, Yann Henon, Samson Timoner, and Yun Fu. Balancing biases and preserving privacy on balanced faces in the wild. *IEEE Transactions on Image Processing*, 2023. 1, 2

[32] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 1

[33] Ignacio Serna, Alejandro Pena, Aythami Morales, and Julian Fierrez. Insidebias: Measuring bias in deep networks and application to face gender biometrics. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3720–3727. IEEE, 2021. 1, 2

[34] Richa Singh, Puspita Majumdar, Surbhi Mittal, and Mayank Vatsa. Anatomizing bias in facial analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12351–12358, 2022. 2

[35] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018. 2

[36] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. 1

[37] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, 3(1):16–30, 2021. 1, 2

[38] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 1

[39] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319, 2019. 2

[40] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference*

on *Computer Vision and Pattern Recognition*, pages 8919–8928, 2020. 2

[41] Haiyu Wu, Vítor Albiero, KS Krishnapriya, Michael C King, and Kevin W Bowyer. Face recognition accuracy across demographics: Shining a light into the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1041–1050, 2023. 2, 3

[42] Haiyu Wu, Grace Bezold, Aman Bhatta, and Kevin W Bowyer. Logical consistency and greater descriptive power for facial hair attribute learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8588–8597, 2023. 2, 3, 4

[43] Haiyu Wu and Kevin W Bowyer. What should be balanced in a" balanced" face recognition dataset. In *BMVC*, volume 1, page 2, 2023. 1, 2

[44] Haiyu Wu, Sicong Tian, Aman Bhatta, Kağan Öztürk, Karl Ricanek, and Kevin 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*, pages 1131–1140, 2024. 3

[45] Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. Consistent instance false positive improves fairness in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 578–586, 2021. 1, 2

[46] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26*, pages 264–274. Springer, 2019. 2

[47] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 3, 5

[48] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 2

[49] Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby P Breckon. Measuring hidden bias within face recognition via racial phenotypes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 995–1004, 2022. 2

[50] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 2

[51] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020. 5

[52] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021. 3, 4