What Should Be Balanced in a "Balanced" Face Recognition Dataset?

Haiyu Wu
hwu6@nd.edu
Kevin W. Bowyer
kwb@nd.edu

University of Notre Dame
Notre Dame, USA

Abstract

The issue of demographic disparities in face recognition accuracy has attracted increasing attention in recent years. Various face image datasets have been proposed as 'fair' or 'balanced' to assess the accuracy of face recognition algorithms across demographics. These datasets typically balance the number of identities and images across demographics. It is important to note that the number of identities and images in an evaluation dataset are *not* driving factors for 1-to-1 face matching accuracy. Moreover, balancing the number of identities and images does not ensure balance in other factors known to impact accuracy, such as head pose, brightness, and image quality. We demonstrate these issues using several recently proposed datasets. To improve the ability to perform less biased evaluations, we propose a bias-aware toolkit that facilitates creation of cross-demographic evaluation datasets balanced on factors mentioned in this paper. The dataset is at https://github.com/HaiyuWu/BA-test-dataset.

1 Introduction

Demographic disparity in face recognition accuracy has emerged as a significant and contentious issue [15, 16, 19, 27, 34, 38]. Researchers have explored approaches to uncover the causes of observed accuracy differences [4, 5, 8, 10, 33, 36, 37, 52, 57]. Datasets have been proposed as 'fair' or 'balanced' for evaluating accuracy across groups [29, 48, 54]. In this paper, we argue that merely balancing the number of identities and images is insufficient for establishing a fair evaluation. Instead, we posit that creating a fair evaluation necessitates balancing factors known to impact accuracy, such as image quality [51], head pose [18], brightness [57], and age [4]. We assembled a Bias Aware test set (BA-test) that balances multiple known accuracy-related factors, enabling cross-demographic evaluations with minimal inherent bias. Contributions of this work include:

- We demonstrate that datasets previously deemed "fair" or "balanced" for evaluation across demographics are not balanced on factors known to drive accuracy difference.
- We introduce the BA-test dataset, designed to support demographic accuracy disparity evaluations based on a better-balanced test set.
- We offer a toolkit to balance a given dataset based on all factors balanced in BA-test.
- We provide an accuracy-disparity-focused benchmark, revealing that current state-of-the-art models exhibit lowest accuracy on Asian females and highest on White males.

2 Literature Review

AI fairness is currently a topic of significant interest, with datasets [25, 44] proposed to measure fairness in audio, vision, and speech domains. Numerous datasets have been introduced to evaluate facial recognition robustness in various ways. CFP [49] and CPLFW [61] concentrate on the head pose factor, while CALFW [62] and AgeDB [41] examine accuracy across diverse ages and age gaps. LFW [28], IJB-C [39], and MegaFace [31] pose challenges to models in unconstrained environments with variations in head pose, brightness, expression, age, and image quality. On the other hand, MORPH [46], CMU-PIE [24, 50], and Multi-PIE [18] include images captured under more controlled conditions.

To support study of demographic accuracy disparity, Hupont et al. [29] introduced DemogPairs, which is balanced with 10.8K images from 600 identities across six demographics. Robbins et al.[48] developed Balanced Faces in-the-Wild (BFW), which includes an Indian subgroup and comprises 20K images from 800 identities. Wang et al.[54] proposed Racial Faces in-the-wild (RFW) with 20K images from 12K subjects across (White, Black, Asian, Indian). DemogPairs, BFW, and RFW focus on balancing the number of identities or the number of images per identity. BUPT-BalancedFace[55] is a training set containing 1.3M images from 28K celebrities across four ethnicities, with 7K identities per ethnicity. However, a fair evaluation of accuracy differences requires more than just balancing the number of subjects and images. Our proposed BA-test dataset balances image quality, head pose, and brightness as examples of controlling factors known to impact recognition accuracy.

Various tools exist to sub-sample datasets to improve performance [63, 64], clean the dataset [14], correct imbalance [35], and provide a test set with specified traits [42, 47, 53, 56]. However, none of these tools specifically address factors that directly impact demographic disparity in face recognition accuracy. Therefore, we propose the BA-toolkit to balance selected factors (e.g., brightness, head pose, image quality, age, amount of visible face) within a dataset, enabling a more controlled evaluation of accuracy across demographics.

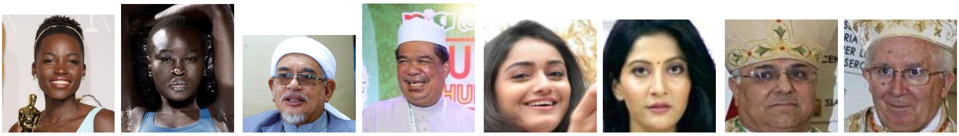
3 What Matters for a Balanced Test Set?

Existing training [55] and testing [29, 48, 54] datasets balance number of subjects and number of images per subject. However, number of subjects and images per subject in a test set does not drive differences in 1-to-1 matching accuracy; see Section 3.2. Factors such as image quality [51], head pose [18], brightness [57], hairstyle [8, 10, 58], and facial morphology [3, 8] can cause accuracy differences across demographics. Moreover, [6, 23] concluded that gender and race balance in training data does not translate into gender and race balance in test accuracy. We use the well-known VGGFace2 [12] dataset to investigate and analyze factors known to drive accuracy differences.

3.1 Bias Aware Toolkit

VGGFace2 consists of in-the-wild, uncropped, unaligned images without demographic metadata and contains noise in identity labels. To support demographic analysis, we propose a Bias Aware toolkit (BA-toolkit) that integrates the function of predicting gender, race, age labels and balancing factors that matter for face recognition accuracy, including brightness, head pose, image quality, and visible face area. The details of the processes are:

Data preparation: Images are cropped and aligned by img2pose [7] and resized to 224×224. Then the measured 3D head poses are converted from 6DoF to degree in Pitch, Yaw, Roll. To prepare for face brightness measurement and balancing face area across gender, BiSeNet [2,



(a) Black female (b) Asian male (c) Indian female (d) White male

Figure 1: Identity noise examples – Each pair is labeled as same identity in VGGFace2.

60] is used to segment a face image. For image quality, FaceQnet [26] and MagFace [40] are used. FairFace [30] is used to predict the demographic and age labels of each image, and we make demographic labels consistent for a given identity by voting the race and gender within each identity. To reduce identity noise, we used ArcFace and MagFace to extract features from the selected images and use DBSCAN [45] to clean label noise.

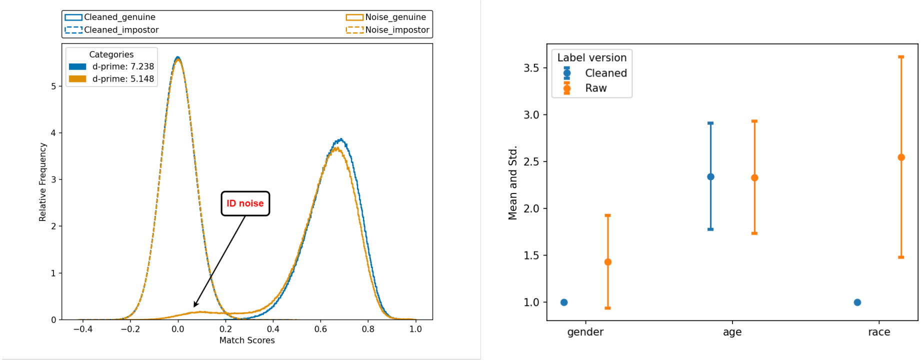
Image selection: We want to balance image quality, head pose, and brightness. First, the FaceQnet quality [26] is used to select images ($Q_{im} > 0.3$). To cross-check quality, MagFace is used the drop images with MagFace quality < 20. For frontal pose, only images with $max\{Pitch, Yaw, Roll\} \in [-20^o, 20^o]$, are selected. Filtering the image set for the brightness range recommended by [57] would reduce the number of images too much. Hence, we use the middle-exposed range $[115.86, 198.75]$ to filter images. To reduce identity noise, we use DBSCAN to get cosine distance between features in one identity folder and drop outliers.

3.2 Preliminary Data Preparation

Quality-driven selection: We first use VGGFace2 FaceQnet scores [1] to drop around 1 million low quality images, a blur classifier to drop 124,632 blurry images, and img2pose to crop and align the the most frontal of the remaining images (dropping 562,036 images). This leaves 1,286,240 images for next steps.

Intra-class noise: Existing datasets [29, 48, 54] were assembled from resources such as VGGFace2 and MS-Celeb-1M, which have identity noise. Identity noise that varies across demographics can lead to incorrect conclusions about accuracy differences, but previous 'balanced' datasets have not cleaned the identity noise. To analyze identity label noise in VGGFace2 we randomly select 200 identities and calculate the cosine similarities. Figure 2a shows that the genuine distribution before identity cleaning has an impostor-like peak from -0.2 to 0.3. Algorithms exist [14, 45] to clean this noise. We implement DBSCAN [45], on the features extracted by ArcFace [13, 21] and MagFace [40] in order to minimze the level of identity noise. The genuine distribution of the randomly-selected identities after cleaning (Figure 1) indicates the identity noise is reduced; 49,468 images are dropped in this step.

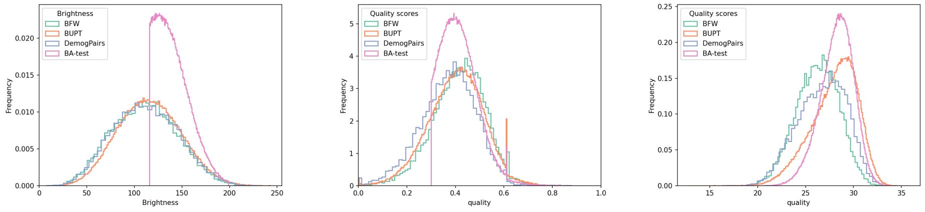
Race, Age, gender labels: The FairFace [30] classifier, trained on race-balanced data, provides confidences on four and seven races, nine age intervals, and two genders. To avoid over-classification and ambiguity between race groups (e.g., SE Asian and East Asian), we choose to group the identities in four groups (i.e. White, Black, Indian, Asian). For age, we follow previous works [4, 32], using Young (10-29 since FairFace predicts age chunks), Middle-Aged (30-49) and Senior (50+) groups. Figure 2b shows the variation in the number of race, gender, and age within the identity folders. For the gender and race attributes, if a given identity has more than one predicted value, the values were made consistent using the identity's most frequently occurring value. Age can vary across the images of the same identity, so we tried to compare the result with the age classifier F_{age} in [6]. These two age predictors disagree on 70K images, where 53K cases are between Young (Fairface) and Middle_Aged (F_{age}). The others are in group (Middle_Aged, Young), (Middle_Aged,



(a) Identity de-noising

(b) Label cleaning

Figure 2: a) Genuine / impostor distributions of random 200 VGGFace2 identities before and after cleaning identity noise. A “fair” or “balanced” dataset should have the same level of clean identity labels across demographics. b) Mean and std. dev. of the number of race, age, gender within each identity before and after cleaning; identity label cleaning results in a single gender and race across the images for a given identity.



(a) FSB brightness measurement

(b) FaceQnet

(c) MagFace (quality)

Figure 3: Brightness and quality for BFW, DemogPairs, BUPT-BalancedFace, BA-test.

Senior), (Young, Senior). We manually inspected 2000 random samples with different age predictions from both classifiers and found that FairFace has more accurate age predictions. Consequently, we only use FairFace results and manually adjust the annotations of some evidently incorrect predictions. 146,842 samples underwent label adjustments in this step.

Is balanced number of IDs and images/ID important for test set? Figure 4 shows the genuine and impostor distributions of eight demographic groups with randomly-selected (200 IDs, 15 images/ID), (100 IDs, 20 images/ID), and (50 IDs, 25 images/ID) from our prepared subset of VGGFace2. Across all groups, there is no significant difference in the impostor or genuine distribution for the different numbers of identities and images. This shows that balancing the number of identities and images across demographic groups in a test set is not relevant to a fair comparison of 1-to-1 matching accuracy. However, the frequency of difficult images (i.e., profile head pose, bad image quality, bad brightness, etc.) and identity noise are major factors impacting the accuracy.

3.3 Analysis of Factors Known to Impact Accuracy

PIE [18, 24, 50] and image quality [51] are well-known factors that affect the accuracy of facial matching. To reduce their impact, we implement the method in [7] to select the most frontal images. Face Skin Brightness (FSB) metric and the middle-exposed brightness range

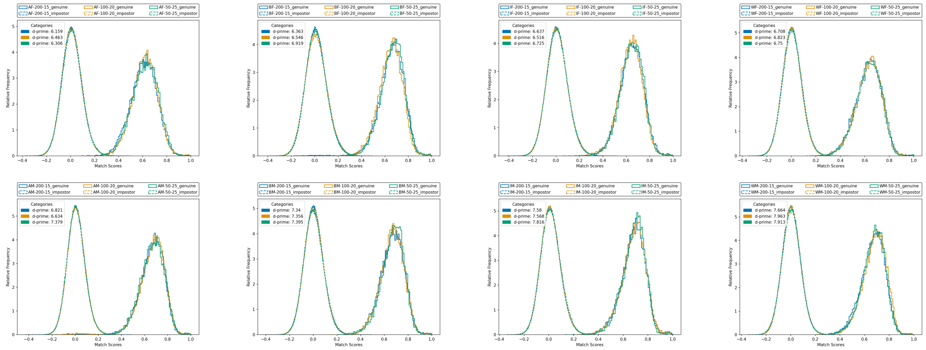
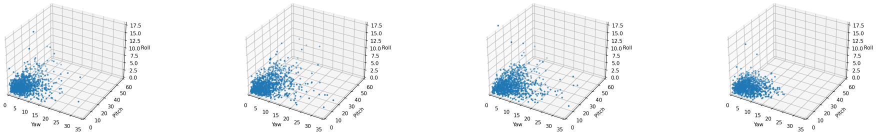


Figure 4: Similarity distributions with the varying number of identities and image per identity for 8 demographics. Top row has distributions of female groups. Bottom row has distribution of male groups. For labels, "AF-200-15" means randomly picking 200 identities with 15 images per identity from Asian Female group.

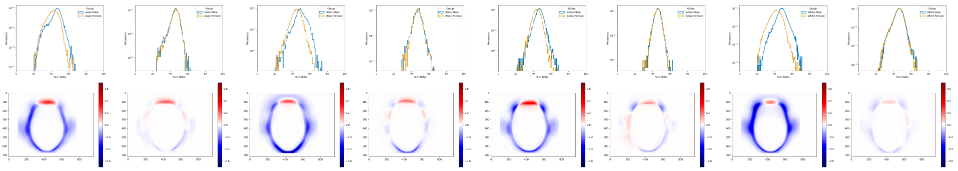


(a) BFW (b) BUPT-Balancedface (c) DemogPairs (d) BA-test

Figure 5: Head pose distribution of BFW, BUPT-Balancedface, DemogPairs, and BA-test.

in [57] are used to select images with good brightness. However, unlike controlled acquisition datasets, the face segmentation model does not perform well on in-the-wild images. To ensure accurate FSB measurement, we drop images whose face area predicted by the model is less than 20% of the image pixels or which have no nose segmentation prediction. For image quality, FaceQnet [26] and MagFace [40] are used to select good quality images and cross-check the results. Meanwhile, the distributions of brightness, 3D head pose, and image quality for BFW, BUPT-Balancedface, DemogPairs and the proposed BA-test are shown in Figure 3a, Figure 5, and Figure 3c and Figure 3b.

Figure 3a shows that the average brightness of the images in BFW, BUPT-Balancedface, and DemogPairs ranges from less than 10 to over 220. Within this range, both underexposed [19, 57] and overexposed [57] images hurt the similarity of image pairs, reducing reliability of analysis made on these datasets. The head pose in BFW, BUPT-Balancedface, and DemogPairs is not controlled. As a training set, variation in head pose in BUPT-Balancedface benefits the performance of the face matcher. However, without controlling head pose across demographics in a test set, conclusions about accuracy across demographics may not be reliable. For image quality, FaceQnet [26] is trained on the dataset that involves human perception. MagFace [40] is trained with a magnitude-based loss function, where the magnitude of the feature represents the quality that the face matcher "thinks" the image is. Figure 3b and Figure 3c show that BA-test has fewer low quality images and less quality variation than the other three datasets, for both quality assessment methods, which should minimize impact of varying image quality on cross-demographic comparisons.



(a) Asian (b) Black (c) Indian (d) White

Figure 6: Top: distributions of the visible face area ratio between genders for each race before (left) and after (right) face area balancing. Bottom: heatmaps for difference of average face area between genders before (left) and after (right) balancing visible face area.

$\Delta d'$	Asian	Black	Indian	White
Original Gen	0.5336	0.3146	0.5094	0.4115
	0.5008	0.3393	0.5030	0.4164
Balanced Gen	0.2678	0.2447	0.5028	0.284
	0.298	0.2824	0.5127	0.2856
Balanced Gen NFH	0.3901	0.6813	0.8819	0.3719
	0.4028	0.7288	0.8901	0.4095
Original Imp	0.0211	0.1511	0.0745	0.0241
	0.0125	0.0917	0.0981	0.042
Balanced Imp	0.015	0.1262	0.179	0.0398
	0.0528	0.0505	0.2455	0.0847
Balanced Imp NFH	0.004	0.0377	0.2491	0.0732
	0.0448	0.0733	0.304	0.1156

Table 1: The gender gap of genuine and impostor distributions for each race before and after balancing the face morphology measured by $\Delta d'$. Gen and Imp are genuine and impostor. NFH means no facial hair. For the $\Delta d'$ of each category, top number is ArcFace model, bottom is MagFace. Red and green represents the largest and smallest gender gap.

3.4 Further Analysis On Gender Bias

Accuracy differences have been reported across gender, age, and race. The first row in Figure 6 shows that the gender difference, where males have larger visible face area than females, also exists in the proposed dataset. Some researches [3, 8] report that balancing facial morphology between sexes can reduce the gender gap. Their conclusion is based on a controlled-acquisition dataset, MORPH [46]. *Can the observations/conclusions be transferred to in-the-wild data?* is what we investigate in this subsection. Unlike MORPH, images in BA-test are in-the-wild. We have added balance in brightness, image quality, and head pose. Therefore, our accuracy-factor-balanced version of the in-the-wild dataset may represent images captured under the real-world cases better than MORPH.

We apply the same approach for face area extraction as described in [3]. (Note that images with a predicted face area of less than 20% were discarded.) As depicted in Figure 6, males generally have a larger visible face area than females. The heatmaps were generated by calculating the difference between the average face areas of females and males, with blue representing pixels more frequently labeled as face for males than females, and red indicating the opposite.

Since images are aligned based on eye position, our findings suggest females have greater distance between the top of their face and their eyes, while males have a longer distance between the face center and other regions. Previous research attributes this to gendered

Statistic information						
Datasets	Data sources	IDs	Images	Subgroups	Age	ID denoise
DemogPairs	[12, 43, 59]	600	10,800	6	✗	✗
RFW	[22]	12,000	80,000	4	✗	✗
BFW	[12]	800	20,000	8	✗	✗
BA-test (ours)	[12]	8,321	177,227	8	2	✓

Balanced factors						
Datasets	Head pose	Race	Quality	Brightness	ID	Gender
DemogPairs	✗	✓	✗	✗	✓	✓
RFW	✗	✓	✗	✗	✗	✗
BFW	✗	✓	✗	✗	✓	✓
BA-test (ours)	✓	✗	✓	✓	✗	✗

Table 2: Existing demographically-balanced test datasets. Upper table gives source of data, number of identities, images, demographic groups, ages, and whether identity labels have been denoised. Bottom table shows factors balanced in each dataset.

hairstyles, but this does not explain the more pronounced distance from face center to the jawline in males. We speculate this could be due to facial hair, as the face segmentation model classifies facial hair as “skin”, causing a bearded chin to appear longer.

To balance face areas, we selected image pairs with a face area intersection-over-union (IoU) greater than 0.9. The face ratio distributions and heatmaps reveal a significant reduction in face area differences. Table 1 presents the gender gap, measured by $\Delta d'$, before and after balancing face areas. For genuine pairs, the gender gap decreases by an average of 24.11% after balancing face areas for both face matchers, with the largest decrease of 45.3% observed in Asians, and a slight increase for Indians. For impostor pairs, the gender gap decreases by 27.22% for Blacks, but increases by 151.5% for Indians and 88.35% for Whites on average. The two face matchers exhibit different trends for Asians.

Studies by [58] and [10] report that facial hair significantly impacts face recognition accuracy. By removing all male samples with facial hair and balancing face areas (shown in the third and sixth rows of Table 1), we observe that gender gaps for genuine pairs increase compared to the balanced version without considering facial hair. For Blacks and Indians, the gender gap averages a 116% and 75% increase, respectively, compared to the original gender gap. For impostor pairs, the gender gap for Asians and Blacks is similar to or smaller than the other two groups. However, for Indians and Whites, the gender gap averages a 222% and 189% increase, respectively, compared to the original gender gap.

In conclusion, balancing facial morphology decreases the gender gap for genuine pairs across all four races. For impostor pairs, this approach reduces the gender gap for Blacks but increases the gap for Indians and Whites. It is important to note that the two face matchers perform differently for Asian impostor pairs, suggesting that different matchers may be sensitive to varying factors. Moreover, facial hair have a significant impact on gender accuracy disparity and more fine-grain analysis on facial hair is a topic for future work.

4 Bias-Aware Test Dataset

The proposed dataset, BA-test, contains 177K images from 8K identities, which is larger than the existing related datasets. Since it is assembled from a single data source, the re-

FMR	BM	BF	IM	IF	WM	WF	AM	AF
BFW	0.0161	0.0357	0.0323	0.0334	0.001	0.0080	0.0225	0.0533
VGGFace2	0.0059	0.0314	0.0174	0.0273	0.001	0.0026	0.0143	0.0267
BA-test	**0.005**	**0.0302**	**0.0141**	**0.0251**	0.001	**0.0024**	**0.0100**	**0.0160**

Table 3: False match rate comparisons with 1-in-10,000 threshold value of White Males in each dataset. Note that both BFW and BA-test are purely assembled on VGGFace2.

liability of identity classification should be higher than that of multiple sources. It has 8 demographic groups with sufficient images per group: 45,642 images of 3,631 White males, 53,245 images of 2,865 White females, 13,311 images of 288 Asian males, 19,454 images of 277 Asian females, 10,610 images of 577 Black males, 6,190 images of 188 Black females, 11,091 images of 244 Indian males, 17,684 images of 251 Indian females. In addition, the images are classified in three age groups - Young, Middle_Aged, Senior. However, due to the small number of seniors in the original dataset (e.g., 77 senior black women, 56 senior Asian females), only Young and Middle_Aged images are selected.

As discussion in Section 3.2, the number of identities and images across demographic groups is not necessary to be balanced for face verification, so this imbalanced version is proposed. However, for face identification (1-to-many matching), the number of images and identities do affect accuracy [11, 17, 20]. In the BA-test, there are 7,896 identities that have more than 2 images, 5,335 identities that have more than 10 images, and 2,969 identities that have more than 20 images. Hence, BA-test can potentially be used for face identification analysis. Moreover, the identity noise, samples in Figure 1, has been reduced. Head pose, brightness, and image quality are balanced in our dataset, but not in the others. Therefore, conclusions about cross-demographic accuracy difference based on our dataset should be more reliable.

4.1 Racial Accuracy Disparity

To illustrate the advantage of this work, we measure the false match rate of each demographic group, with the 1-in-10,000 FMR threshold of White Males in the BFW, VGGFace2, and BA-test datasets. Since BFW and BA-test are assembled based on VGGFace2, it should be fair to compare the results. However, due to the large number of images VGGFace2, we do not measure the similarity of image pairs across the whole dataset. We first run FairFace to obtain demographic labelsforof VGGFace2, then randomly select the same number of identities and images as that of the BA-test dataset. Table 3 shows that BA-test has the smallest FMRs for all demographic groups and the smallest gender gap across races.

4.2 Benchmarks on Bias

BA-test is over-large as a testing benchmark. Since image quality is balanced in a good range, it is not a strong challenge for state-of-the-art face matchers. However, this dataset is a good indicator for measuring how biased models may be on gender, age, and race. To pick challenging samples, we first normalized the two quality measurements into $[0,1]$ by Min-Max normalization where Q is a image quality vector of BA-test. After aggregating the quality values, for each demographic group, we used the image quality value at 50% percentile Q_{50th} to randomly pick 90 subjects with 5 images of relatively poor quality ($< Q_{50th}$) images per subject. Consequently, there are 3,600 images from 720 identities, where there

Loss	Model	Train	AF	AM	diff.	WF	WM	diff.
MagFace	r50	Mv2	65.00	79.78	14.78	76.67	86.22	9.56
MagFace	r100	Mv2	81.56	94.44	12.89	89.44	96.56	7.11
ArcFace	r100	Mv2	81.56	93.11	11.56	90.11	97.11	7.00
ArcFace	r50	Glint	81.22	93.00	11.78	92.67	95.78	3.11
ArcFace	r100	Glint	90.00	96.78	6.78	95.78	98.78	3.00
Loss			BF	BM	diff.	IF	IM	diff.
MagFace	r50	Mv2	85.56	86.78	1.22	86.78	90.78	4.00
MagFace	r100	Mv2	91.00	94.22	3.22	96.00	96.11	0.11
ArcFace	r100	Mv2	91.56	94.11	2.56	94.56	95.56	1.00
ArcFace	r50	Glint	93.44	93.78	0.33	94.89	93.67	-1.22
ArcFace	r100	Glint	98.00	97.67	-0.33	98.56	97.22	-1.33

Table 4: True positive rates (%) with a false match rate of 10^{-5} and the best (green) and worst (red) accuracy for each face matcher across eight demographic groups. diff. is the highest TPR - the lowest TPR in each block. Mv2 and Glint are MS1MV2 and Glint360K.

Loss	Model	Train	Y	M	diff.	F	M	diff.
MagFace	r50	Mv2	79.95	89.96	10.01	78.50	85.89	7.39
MagFace	r100	Mv2	91.44	95.49	4.05	89.50	95.33	5.83
ArcFace	r100	Mv2	91.00	95.83	4.83	89.44	94.97	5.53
ArcFace	r50	Glint	91.51	95.49	3.98	90.56	94.06	3.50
ArcFace	r100	Glint	96.12	98.16	2.04	95.58	97.61	2.03
Loss	Model	Train	W	B	A	I	diff.	Overall
MagFace	r50	Mv2	81.44	86.17	72.39	88.78	16.39	82.19
MagFace	r100	Mv2	93.00	92.61	88.00	96.06	8.06	92.42
ArcFace	r100	Mv2	93.61	92.83	87.33	95.06	7.73	92.21
ArcFace	r50	Glint	94.22	93.61	87.11	94.28	7.17	92.31
ArcFace	r100	Glint	97.28	97.83	93.39	97.89	4.50	96.60

Table 5: True positive rates (%) with a false match rate of 10^{-5} and the difference (best - worst) within age (Y,M), gender (F, M), race group (W, B, A, I). diff. is the highest TPR - the lowest TPR in each block. Red and green represent the group with highest and lowest TPR in each block.

are 2,465 images in Young, 1,135 images in Middle-Aged, 1,800 females, and 1,800 males. We evaluated pre-trained face matchers from [9, 13, 40] with ResNet50 and ResNet100 backbones trained with MS1MV2 and Glint360K.

To better compare accuracy in age, gender and race, we use the 1-in-100K false match rate (FMR) of all the impostor similarities as a decision threshold to calculate the true positive rate (TPR). From Table 5, a general conclusion for age, gender and race is that accuracy disparity exists in all five matchers even though they are different algorithms and have different accuracy. For age, the TPR of Middle-Aged is 4.98% higher than Young. Males have 4.86% higher TPR than females. For race, matchers perform best on Indian and worst on Asian; the difference on average is 8.77%. White and Black do not have a general conclusion on accuracy across the matchers. Results suggest that the largest difference in cross-demographic accuracy is based on race.

Table 4 shows the TPR of each demographic with the same 1-in-100K FMR threshold. For gender bias, matchers are biased Asian > White > Black > Indian, where the average gender gap for Asian is 11.6%, for White is 6%, for Black is 1.4%, and 0.5% for Indian.

Results show that gender gap is smaller for Indian and Black, but larger for Asian and White. Furthermore, the difference between best (White Male) and worst (Asian Female) accuracy is 15% on average. Again, test samples are balanced in quality; thus, the difference in accuracy reflects the bias of the models caused by gender, race, and age. Therefore, even though current matchers have high accuracy, demographic differences are still an issue.

5 Conclusion

This work demonstrates that balancing the number of identities and images per identity is insufficient to address bias in 1-to-1 matching. Instead, factors such as head pose, brightness, image quality, and gendered characteristics play critical roles in understanding bias.

We propose a bias-aware toolkit for assembling datasets and creating bias-aware test sets (BA-test). This test set, with more identities and images, enables researchers to draw reliable conclusions about the sources of bias in real-world scenarios.

We introduce a face recognition bias benchmark dataset and evaluate three state-of-the-art models, revealing that age gap, gender gap, and demographic accuracy disparity persist.

Future research includes exploring additional bias-contributing factors, examining their impact in face identification, and developing algorithms to mitigate bias in face recognition.

6 Acknowledgement

Thanks to the help from Dr. Vítor Albiero. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via *2022-21102100003*. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] https://github.com/javier-hernandezo/FaceQnet, last accessed on April 2023.

[2] https://github.com/zllrunning/face-parsing.PyTorch, last accessed on February 2021.

[3] Vítor Albiero and Kevin W Bowyer. Is face recognition sexist? no, gendered hairstyles and biology are. In *British Machine Vision Conference (BMVC)*, 2020.

[4] Vítor Albiero, Kevin Bowyer, Kushal Vangara, and Michael King. Does face recognition accuracy get better with age? deep face matchers say no. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 261–269, 2020.

[5] Vitor Albiero, Krishnapriya KS, Kushal Vangara, Kai Zhang, Michael C King, and Kevin W Bowyer. Analysis of gender inequality in face recognition accuracy. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 81–89, 2020.

[6] Vítor Albiero, Kai Zhang, and Kevin W Bowyer. How does gender balance in training data affect face recognition accuracy? In *Proceedings of the IEEE International Joint Conference on Biometrics*, pages 1–10. IEEE, 2020.

[7] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7617–7627, 2021.

[8] Vítor Albiero, Kai Zhang, Michael C. King, and Kevin W. Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. 17:127–137, 2022. doi: 10.1109/TIFS.2021.3135750.

[9] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021.

[10] 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.

[11] Aman Bhatta, Gabriella Pangelinan, Micheal C. King, and Kevin W. Bowyer. Demographic disparities in 1-to-many facial identification. *arXiv preprint arXiv:2309.04447*, 2023.

[12] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *proceeding of the IEEE international Conference on Automatic Face and Gesture Recognition*, pages 67–74. IEEE, 2018.

[13] 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.

[14] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the European Conference on Computer Vision*, pages 741–757. Springer, 2020.

[15] C. Doctorow. NIST confirms that facial recognition is a racist, sexist dumpster-fire, 2019. https://boingboing.net/2019/12/19/demographics-v-robots.html5.

[16] 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.

[17] Pawel Drozdowski, Christian Rathgeb, and Christoph Busch. The watchlist imbalance effect in biometric face identification: Comparing theoretical estimates and empiric measurements. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3757–3765, 2021.

[18] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010.

[19] Patrick Grother. NISTIR 8429: Face recognition vendor test (FRVT) part 8: Summarizing demographic differentials. Technical report.

[20] Patrick Grother and P Jonathon Phillips. Models of large population recognition performance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–II. IEEE, 2004.

[21] Jia Guo. Insightface: 2D and 3D face analysis project. https://github.com/deepinsight/insightface, last accessed on February 2021.

[22] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision*, pages 87–102. Springer, 2016.

[23] Matthew Gwilliam, Srinidhi Hegde, Lade Tinubu, and Alex Hanson. Rethinking common assumptions to mitigate racial bias in face recognition datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4123–4132, 2021.

[24] Hu Han, Shiguang Shan, Xilin Chen, and Wen Gao. A comparative study on illumination preprocessing in face recognition. *Pattern Recognition*, 46(6):1691–1699, 2013.

[25] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):324–332, 2021.

[26] Javier Hernandez-Ortega, Javier Galbally, Julian Fiérrez, and Laurent Beslay. Biometric quality: Review and application to face recognition with faceqnet. *arXiv preprint arXiv:2006.03298*, 2020.

[27] T. Hoggins. 'racist and sexist' facial recognition cameras could lead to false arrests, Dec. 20 2019. https://www.telegraph.co.uk/technology/2019/12/20/racist-sexist-facial-recognition-cameras-could-lead-false-arrests/.

[28] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

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

[30] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

[31] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

[32] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

[33] KS Krishnapriya, Vítor Albiero, Kushal Vangara, Michael C King, and Kevin W Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020.

[34] KS Krishnendu. Analysis of recent trends in face recognition systems.

[35] Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563, 2017.

[36] Hao Liang, Josue Ortega Caro, Vikram Maheshri, Ankit B Patel, and Guha Balakrishnan. Towards causally linking architectural parametrizations to algorithmic bias in neural networks. *arXiv preprint arXiv:2302.03750*, 2023.

[37] Hao Liang, Pietro Perona, and Guha Balakrishnan. Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation. *arXiv preprint arXiv:2308.05441*, 2023.

[38] S. Lohr. Facial recognition is accurate, if you're a white guy. *The New York Times*, Feb. 9 2018.

[39] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *Proceedings of the international conference on biometrics*, pages 158–165. IEEE, 2018.

[40] 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*, 2021.

[41] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 51–59, 2017.

[42] Avanika Narayan, Piero Molino, Karan Goel, Willie Neiswanger, and Christopher Re. Personalized benchmarking with the ludwig benchmarking toolkit. *arXiv preprint arXiv:2111.04260*, 2021.

[43] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[44] Bilal Porgali, Vítor Albiero, Jordan Ryda, Cristian Canton Ferrer, and Caner Hazirbas. The casual conversations v2 dataset. *arXiv preprint arXiv:2303.04838*, 2023.

[45] Anant Ram, Sunita Jalal, Anand S Jalal, and Manoj Kumar. A density based algorithm for discovering density varied clusters in large spatial databases. *International Journal of Computer Applications*, 3(6):1–4, 2010.

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

[47] Wes Robbins, Steven Zhou, Aman Bhatta, Chad Mello, Vítor Albiero, Kevin W Bowyer, and Terrance E Boult. Cast: Conditional attribute subsampling toolkit for fine-grained evaluation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 919–929, 2023.

[48] 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.

[49] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *Proceedings of the IEEE winter conference on applications of computer vision*, pages 1–9. IEEE, 2016.

[50] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12): 1615–1618, 2003.

[51] Philipp Terhörst, Malte Ihlefeld, Marco Huber, Naser Damer, Florian Kirchbuchner, Kiran Raja, and Arjan Kuijper. Qmagface: Simple and accurate quality-aware face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3484–3494, 2023.

[52] Kushal Vangara, Michael C King, Vitor Albiero, Kevin Bowyer, et al. Characterizing the variability in face recognition accuracy relative to race. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[53] Jun Wang, Yinglu Liu, Yibo Hu, Hailin Shi, and Tao Mei. Facex-zoo: A pytorch toolbox for face recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3779–3782, 2021.

[54] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the International Conference on Computer Vision*, October 2019.

[55] Mei Wang, Yaobin Zhang, and Weihong Deng. Meta balanced network for fair face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[56] Qingzhong Wang, Pengfei Zhang, Haoyi Xiong, and Jian Zhao. Face. evolve: A high-performance face recognition library. *arXiv preprint arXiv:2107.08621*, 2021.

[57] 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.

[58] 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.

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

[60] 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*, pages 325–341, 2018.

[61] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7), 2018.

[62] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

[63] Manli Zhu and Aleix M Martínez. Optimal subclass discovery for discriminant analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, pages 97–97. IEEE, 2004.

[64] Manli Zhu and Aleix M Martinez. Subclass discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1274–1286, 2006.