Face Recognition: Too Bias, or Not Too Bias?

Joseph P Robinson[1], Gennady Livitz[2], Yann Henon[2], Can Qin[1],
Yun Fu[1], and Samson Timoner[2]
[1]Northeastern University [2]ISM Connect

Abstract

We reveal critical insights into problems of bias in state-of-the-art facial recognition (FR) systems using a novel Balanced Faces In the Wild (BFW) dataset: data balanced for gender and ethnic groups. We show variations in the optimal scoring threshold for face-pairs across different subgroups. Thus, the conventional approach of learning a global threshold for all pairs results in performance gaps between subgroups. By learning subgroup-specific thresholds, we reduce performance gaps, and also show a notable boost in overall performance. Furthermore, we do a human evaluation to measure bias in humans, which supports the hypothesis that an analogous bias exists in human perception. For the BFW database, source code, and more, visit https://github.com/visionjo/facerec-bias-bfw.

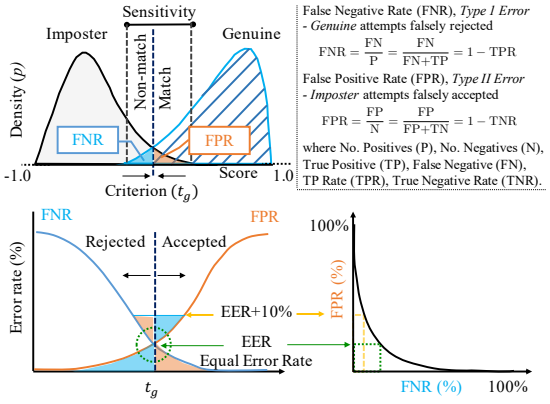


Figure 1: **Depiction of bio-metrics.** The signal detection model (SDM) (*top-left*) shows the sensitivity related to a single threshold t_g. Translated to error rate (*bottom-left*), a direct trade-off between FNP and FPR. In practice, this is entirely application dependent. Specifically, the specific chose in desired FPR (%) (*bottom-right*).

1. Introduction

As more of society becomes integrated with machine learning (ML), topics related to bias, fairness, and the formalization of ML standards attract more attention [1, 4, 38]. Thus, an effect of the growing dependency on ML is an ever-increasing concern about the biased and unfair algorithms, *e.g.*, untrustworthy and prejudiced facial recognition (FR) [26, 32].

Nowadays, convolutional neural networks (CNNs) are trained on faces identified by a detection system. Specifically, for FR, the goal is to encode faces in an N-dimensional space such that samples of the same identity are neighbors, while different persons are farther apart. Thus, we can deploy a CNN to encode faces, and then compare via a similarity score (*i.e.*, pairs with a high enough score classified as *genuine*; else, *imposter*).

Typically, a fixed threshold sets the decision boundary by which to compare scores (Fig. 1). As such, features of the same identity must satisfy a criterion based on a single value [9, 21, 36, 37]. However, we found that an individual (*i.e.*, global) threshold is a crude measure that leads to skewed errors. Furthermore, the held-out set used to determine the threshold tends to share the same distribution with the test set, favoring specific demo-

graphics that are the majority. That skew (*i.e.*, the difference in the performance of an algorithm of particular demographics) is our definition of bias. A key question is: *is FR too biased, or not?*

Making matters more challenging, precise definition of race and ethnicity vary from source-to-source. For example, the US Census Bureau allows an individual to self-identify race.[1] Even gender, our attempt to encapsulate the complexities of the sex of a human as one of two labels. Others have addressed the oversimplified class labels by representing gender as a continuous value between 0 and 1 - rarely is a person entirely *M* or *F*, but most are somewhere in between [24]. For this work, we define subgroups as specific sub-populations with face characteristics similar to others in a region. Specifically, we focus on 8 subgroups (Fig. 2).

The adverse effects of a global threshold are two-fold: **(1)** mappings produced by CNNs are nonuniform. Therefore, distances between pairs of faces in different demographics vary in distribution of similarity scores (Fig 4); **(2)** evaluation set is imbalanced. Subgroups that make up a majority of the population will carry most weight on the reported performance ratings. Reported results favor the common traits over the underrepresented. Demographics like gender, ethnicity, race, and age are underrepresented in most public datasets [24, 38].

For **(1)**, we propose subgroup-specific (*i.e.*, optimal) thresholds while addressing **(2)** with a new benchmark dataset to measure bias in FR, BFW (Table 1 and 2). BFW serves as a proxy for fair evaluations for FR while enabling per subgroup ratings to be reported. We use BFW to gain an understanding of the extent to which bias is present in state-of-the-art (SOTA) CNNs used FR. Then, we suggest a mechanism to mitigate problems of bias with more balanced performance ratings for different demographics. Specifically, we propose using an adaptive threshold that varies depending on the characteristics of detected facial attributes (*i.e.*, gender and ethnicity). We show an increase in accuracy with a balanced performance for different subgroups. Similarly, we show a positive effect of adjusting the similarity threshold based on the facial features of matched faces. Thus, selective use of similarity thresholds in current SOTA FR systems provides more intuition in FR research with a method easy to adopt in practice.

The contributions of this work are 3-fold. (1) We built a balanced dataset as a proxy to measure verification performance per subgroup for studies of bias in FR. (2) We revealed an unwanted bias in scores of face pairs - a bias that causes ratings to skew across demographics. For this, we showed that an adaptive threshold per subgroup balances performance (*i.e.*, the typical use of a global threshold unfavorable, which we address via optimal thresholds). (3) We surveyed humans to demonstrate bias in human perception (NIH-certified, *Protect Humans in Research*).

2. Background Information

2.1. Bias in ML

The progress and commercial value of ML are exciting. However, due to inherent biases in ML, society is not readily able to trust completely in its widespread use. The exact definitions and implications of bias vary between sources, as do its causes and types. A common theme is that bias hinders performance ratings in ways that skew to a particular sub-population. In essence, the source varies, whether from humans [40], data or label types [35], ML models [3, 19], or evaluation protocols [34]. For instance, a vehicle-detection model might miss cars if training data were mostly trucks. In practice, many ML systems learn biased data, which could be detrimental to society.

2.2. Bias in FR

Different motivations have driven problems of bias in FR. Instances can be in issues of data augmentation [42], one-shot learning [10], demographic parity and fairness with priority on privacy [16], domain adaptation [38], differences in face-based attributes across demographics [39], data exploration [25], and even characterizing different commercial systems [5].

Yin *et al*. proposed to augment the feature space of underrepresented classes using different classes

[1] www.census.gov/mso/www/training/pdf/race-ethnicity-onepager

Table 1: **Database stats and nomenclature.** *Header:* Subgroup definitions. *Top:* Statistics of Balanced Faces In the Wild (BFW). *Bottom:* Number of pairs for each partition. Columns grouped by ethnicity and then further split by gender.

	Asian (A)		Black (B)		Indian (I)		White (W)		
	Female (AF)	Male (AM)	BF	BM	IF	IM	WF	WM	Aggregated
# Faces	2,500	2,500	2,500	2,500	2,500	2,500	2,500	2,500	20,000
# Subjects	100	100	100	100	100	100	100	100	800
# Faces / Subject	25	25	25	25	25	25	25	25	25
# Positive Pairs	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	240,000
# Negative Pairs	85,135	85,232	85,016	85,141	85,287	85,152	85,223	85,193	681,379
# Pairs (Total)	115,135	115,232	115016	115,141	115287	115,152	115,223	115193	921,379

with a diverse collection of samples [42]. This was to encourage distributions of underrepresented classes to resemble the others more closely. Similarly, others formulated the imbalanced class problem as one-shot learning, where a generative adversarial network (GAN) was trained to generate face features to augment classes with fewer samples [10]. Generative Adversarial Privacy and Fairness (GAPF) was proposed to create fair representations of the data in a quantifiable way, allowing for the finding of a de-correlation scheme from the data without access to its statistics [16]. Wang *et al*. defined subgroups at a finer level (*i.e*., Chinese, Japanese, Korean), and determined the familiarity of faces inter-subgroup [39]. Genders have also been used to make subgroups (*e.g*., for analysis of gender-based face encodings [25]). Most recently, [38] proposed to adapt domains to bridge the bias gap by knowledge transfer, which was supported by a novel data collection, Racial Faces in-the-Wild: (RFW). The release of RFW occurred after BFW was built - although similar in terms of demographics, RFW uses faces from MSCeleb [13] for testing, and assumes CASIA-Face [41] and VGG2 [6] were used to train. In contrast, our BFW assumes VGG2 as the test set. Furthermore, BFW balance subgroups: RFW splits subgroups by gender and race, while BFW has gender, race, or both).

Most similar to us is [8, 18, 22, 33] - each was motivated by insufficient paired data for studying bias in FR. Then, problems were addressed using labeled data from existing image collections. Uniquely, Hupont *et al*. curated a set of faces based on racial demographics (*i.e*., *Asian* (A), *Black* (B), and *White* (W)) called Demographic Pairs (DemogPairs) [18]. In contrast, [33] honed in on adults versus children called Wild Child Celebrity (ITWCC). Like the proposed BFW, both were built by sampling existing databases, but with the addition of tags for the respective subgroups of interest. Aside from the additional data of BFW (*i.e*., added subgroup *Indian* (I), along with other subjects with more faces for all subgroups), we also further split subgroups by gender. Furthermore, we focus on the problem of facial verification and the different levels of sensitivity in cosine similarity scores per subgroup.

2.3. Human bias in ML

Bias is not unique to ML - humans are also susceptible to a perceived bias. Biases exist across race, gender, and even age [1, 7, 23, 27]. Wang *et al*. showed machines surpass human performance in discriminating between Japanese, Chinese, or Korean faces by nearly 150% [39], as humans just pass random (*i.e*., 38.89% accuracy).

We expect the human bias to skew to their genders and races. For this, we measure human perception with face pairs of different subgroups (Section 3.3). The results concur with [39], as we also recorded overall averages below random (<50%).

3. The BFW Benchmark and Dataset

We now discuss the acquisition of BFW, protocols, and settings used to survey bias in humans.

3.1. The data

Problems of bias in FR motivated us to build BFW. Inspired by DemogPairs [18], the data is

Table 2: **BFW and related datasets.** BFW is balanced across ID, gender, and ethnicity (Table 1). Compared with DemogPairs, BFW provides more samples per subject and subgroups per set, while using a single resource, VGG2. RFW, on the other hand, supports domain adaptation, and focuses on race-distribution - not the distribution of identities.

Database		Number of			Balanced Labels		
Name	Source Data	Faces	IDs	Subgroups	ID	Ethnicity	Gender
DemogPairs [18]	CASIA-W [41], VGG [30] &VGG2 [6]	10,800	600	6	✓	✓	✓
RFW [38]	MS-Celeb-1M	≈80,000	≈12,000	4	✘	✓	✘
BFW (ours)	VGG2	20,000	800	8	✓	✓	✓

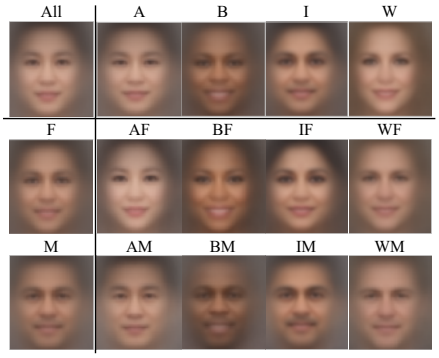


Figure 2: **BFW.** Average face per subgroup: *top-left*: the entire BFW; *top-row* per ethnicity; *left-column*: per gender. The others represent the ethnicity and gender, respectively. Table 1 defines the acronyms of subgroups.

made up of evenly split subgroups, but with an increase in subgroups (*i.e.*, IF and IM), subjects per subgroup, and face pairs (Table 1).

Compiling subject list. Subjects were sampled from VGG2 [6] - unlike others built from multiple sources, BFW has fewer potential conflicts in train and test overlap with existing models. To find candidates for the different subgroups, we first parsed the list of names using a pre-trained ethnicity model [2]. This was then further refined by processing faces using ethnicity [12] and gender [20] classifiers. This resulted in hundreds of candidates per subgroup, which allowed us to manually filter 100 subjects per the 8 subgroups.

Detecting faces. Faces were detected using MTCNN [43].[2] Then, faces were assigned to one of two sets. Faces within detected bounding box (BB) regions extended out 130% in each direction, with zero-padding as the boundary condition made-up one set. The second set were faces aligned and cropped for Sphereface [21] (see the next step). Also, coordinates of the BB and the five landmarks from *multi-task CNN* (MTCNN) were stored as part of the static, raw data. For samples with multiple face detections, we used the BB area times the confidence score of the MTCNN to determine the face most likely to be the subject of interest, with the others set aside and labeled *miss-detection*.

Validating labels. Faces of BFW were encoded using the original implementation of the SOTA Sphereface [21]. For this, each face was aligned to predefined eye locations via an affine transformation. Then, faces were fed through the CNN twice (*i.e.*, the original and horizontally flipped), with two features fused by average pooling (*i.e.*, 512 D). A matrix of cosine similarity scores was then generated for each subject and removed samples (*i.e.*, rows) with median scores below threshold $\theta = 0.2$ (set manually). Mathematically, the n^{th} sample for the j^{th} subject with N_j faces was removed if the ordinal rank of its score $n = \frac{P \times N}{100} \geq \theta$, where $P = 50$. In other words, the median (*i.e.*, 50 percentile) of all scores for a face with respect to all of faces for the respective subject must pass a threshold of $\theta = 0.2$; else, the face is dropped. This allowed us to quickly prune false-positive (FP) face detections. Following [28, 29], we built a JAVA tool to visually validate the remaining faces. For this, the faces were ordered by decreasing confidence, with confidence set as the average score, and then displayed as image icons on top toggling buttons arranged as a grid in a sliding pane window. Labeling then consisted of going subject-by-

[2] https://github.com/polarisZhao/mtcnn-pytorch

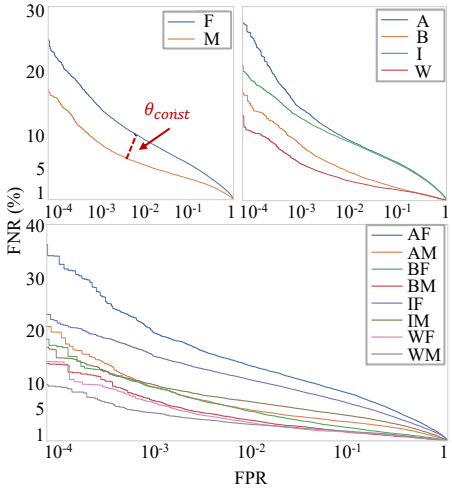


Figure 3: **Detection error trade-off (DET) curves.** *Top-left*: per gender. *Top-right*: per ethnicity. *Bottom*: per subgroup (*i.e.*, combined). Dashed line shows about $2\times$ difference in FPR for the same threshold θ_{const}. FNR is the match error count (closer to the bottom is better).

subject and flagging faces of *imposters*.

Sampling faces and creating folds. We created lists of pairs in five-folds with subjects split evenly per person and without overlap across folds. Furthermore, a balance in the number of faces per subgroup was obtained by sampling twenty-five faces at random from each. Next, we generated a list of all the face pairs per subject, resulting in $\sum_{l=1}^{L}\sum_{k=1}^{K_d}\binom{N_k}{2}$ positive pairs, where the number of faces of all K_l subjects $N_k = 25$ for each of the L subgroups (Table 1). Next, we assigned subjects to a fold. To preserve balance across folds, we sorted subjects by the number of pairs and then started assigning to alternating folds from the one with the most samples. Note, this left no overlap in identity between folds. Later, a negative set from samples within the same subgroup randomly matched until the count met that of the positive. Finally, we doubled the number with negative pairs from across subgroups but in the same fold.

3.2. Problem formulation

Facial verification (FV) is the special case of the two-class (*i.e.*, boolean) classification. Hence, pairs are labeled as the "same" or "different" *genuine* pairs (*i.e.*, *match*) or *imposter* (*i.e.*, *mismatch*), respectively. This formulation (*i.e.*, FV) is highly practical for applications like access control, re-identification, and surveillance. Typically, training a separate model for each unique subject is unfeasible. Firstly, the computational costs compound as the number of subjects increase. Secondly, such a scheme would require model retraining each time a new person is added. Instead, we train models to encode facial images in a feature space that captures the uniqueness of a face, to then determine the outcome based on the output of a scoring (or distance) function. Formally put:

$$f_{boolean}(\vec{x}_i, \vec{x}_j) = d(\vec{x}_i, \vec{x}_j) \leq \theta, \qquad (1)$$

where $f_{boolean}$ is the *matcher* of the feature vector $\vec{x}$ for the i^{th} and j^{th} sample [17].

Cosine similarity is used as the *matcher* in Eq 1 the closeness of i^{th} and j^{th} features, *i.e.*, $s_l = \frac{f_i \cdot f_j}{||f_i||_2||f_j||_2}$ is the closeness of the l^{th} pair.

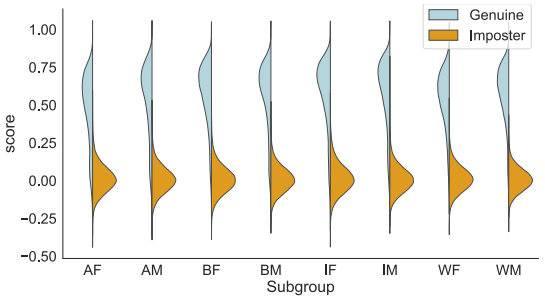


Figure 4: **SDM across subgroups.** Scores of *imposters* have medians around 0.3 but with variations in upper percentiles; *genuine* pairs vary in mean and spread (*e.g.*, AF has more area of overlap). A threshold varying across different subgroups yields a constant FPR.

3.3. Human assessment

To focus on the human evaluation experiment, we honed-in on pairs from two groups, White Americans (W) and Chinese from China (C). This minimized variability compared to the broader groups of Whites and Asians, which was thought to be best, provided only a small subset of the data was used on fewer humans than subjects in BFW.

Samples were collected from individuals recruited from multiple sources (*e.g.*, social media,

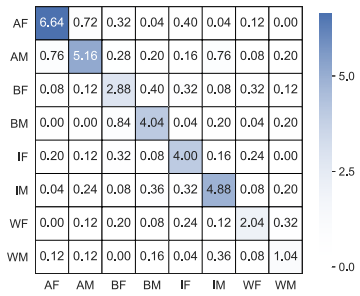


Figure 5: **Confusion matrix.** Error (Rank 1, %) for all BFW faces versus all others. Errors concentrate intra-subgroup - consistent with the SDM (Fig. 4). Although subgroups are challenging to define, this shows the ones chosen are meaningful for FR.

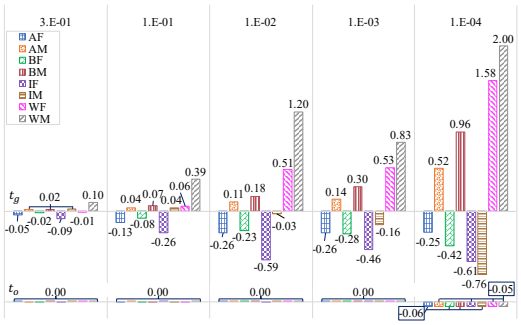


Figure 6: **Percent difference from intended FPR.** *Top:* t_g yields FPR the span as large as 2× (*i.e.*, 200%) that intended (*i.e.*, WM for 1e-4). Furthermore, F subgroups tend to perform worse than intended for all cases (while M tend to overshoot intended performance, with exception of IM in for FPR=1e-4). *Bottom:* Subgroup-specific thresholds reduces this difference to near zero, where there are small differences, the percent difference across different subgroups is fair (*i.e.*, FPR=1e-4).

email lists, family, and friends) - a total of 120 participants were sampled at random from all submissions that were completed and done by a W or C participant. Specifically, there were 60 W and 60 C, both with *Male* (M) and *Female* (F) split evenly. A total of 50 face pairs of non-famous "look-alikes" were collected from the internet, with 20 (*W*A) and 20 (*C*) pairs with, again, M and F split evenly. The other 10 were of a different demographic (*e.g.*, Hispanic/ Latino, Japanese, African). The survey was created, distributed, and recorded via *PaperForm*. It is important to note that participants were only aware of the task (*i.e.*, to verify whether or not a face-pair was a *match* or *non-match*, but with no knowledge of it being a study on the bias).

4. Results and Analysis

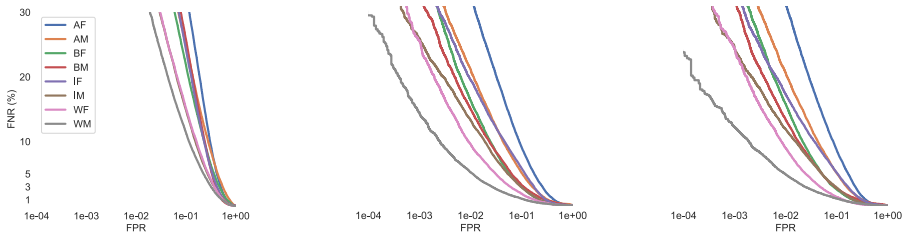
A single CNN was used as a means to control the experiments. For this, Sphereface [21] trained on CASIA-Web [41], and evaluated on Labeled Faces in the Wild (LFW) [17] (%99.22 accuracy), encoded all of the faces.[3] As reported in [**?**], LFW has about 13%, 14%, 3%, and 70% ratio in Asian, Black, Indian, and White, respectively. Furthermore, CASIA-Web is even more unbalanced (again, as reported in [38]), with about 3%, 11%, 2%, and 85% for the same subgroups.

[3] https://github.com/clcarwin/sphereface_pytorch

DET analysis. DET curves (5-fold, averaged) show per-subgroup trade-offs (Fig. 3). Note that M performs better than F, precisely as one would expect from the tails of score-distributions for *genuine* pairs (Fig. 4). AF and IF perform the worst.

Score analysis. Fig. 4 shows score distributions for faces of the same (*i.e.*, *genuine*) and different (*i.e.*, *imposter*) identity, with a subgroup per SDM graph. Notice that score distributions for imposters tend to peak about zero for all subgroups, and with minimal deviation comparing modes of the different plots. On the other hand, the score distribution of the *genuine* pairs varies across subgroups in location (*i.e.*, score value) and spread (*i.e.*, overall shape). Fig. 5 shows the confusion matrix of the subgroups. A vast majority of errors occur in the intra-subgroup. It is interesting to note that while the definition of each group based on ethnicity and race may not be crisply defined, the confusion matrix indicates that in practice, the CNN finds that the groups are effectively separate. The categories are, therefore, meaningful for FR.

The gender-based DET curves show performances with a fixed threshold (dashed line). Other curves relate similarity (lines omitted to declutter). For many FR applications, systems operate



(a) VGG16 [31] (b) ResNet50 [14] (c) SENet [15]

Figure 7: **DET curves for different CNNs**. FNR (%) (vertical) vs FPR (horizontal, log-scale) for VGG2 [6] models with different backbones (VGG16, Resnet50, SEnet50). Lower is better. For each plot, WM is the top-performer, while AF is the worst. The ordering of the curves is roughly the same for each backbone.

at the highest FPR allowed. The constant threshold shows that a single threshold produces different operating points (*i.e.*, FPR) across subgroups, which is undesirable. If this is the case in an industrial system, one would expect a difference in about double the FPs reported based on subgroup alone. The potential ramifications of such a bias should be considered, which it has not as of lately– even noticed in main-stream media [11, 32].

To further demonstrate the extent of the problem, we follow settings typical for systems in practice. We set the desired FPR, and then determine the percent difference, *i.e.*, desired versus actual (Fig. 6, *top*). Furthermore, we mitigate this highly skewed phenomenon by applying subgroup-specific thresholds (*bottom*) - by this, minimal error from the desired FPR. Besides where there is a small error, the offset is balanced across subgroups.

Table 3: **TAR at intervals of FAR**. FAR, listed are the TAR scores for a global threshold (top) and the proposed category-based threshold (bottom). Higher is better.

FAR	0.3	0.1	0.01	0.001	0.0001
AF	0.990	0.867	0.516	0.470	0.465
	1.000	0.882	0.524	0.478	0.474
AM	0.994	0.883	0.529	0.482	0.477
	1.000	0.890	0.533	0.486	0.482
BF	0.991	0.870	0.524	0.479	0.473
	1.000	0.879	0.530	0.484	0.480
BM	0.992	0.881	0.526	0.480	0.474
	1.000	0.891	0.532	0.485	0.480
IF	0.996	0.881	0.532	0.486	0.481
	1.000	0.884	0.534	0.488	0.484
IM	0.997	0.895	0.533	0.485	0.479
	1.000	0.898	0.535	0.486	0.481
WF	0.988	0.878	0.517	0.469	0.464
	1.000	0.894	0.526	0.478	0.474
WM	0.989	0.896	0.527	0.476	0.470
	1.000	0.910	0.535	0.483	0.478
Avg.	0.992	0.881	0.526	0.478	0.474
	1.000	0.891	0.531	0.483	0.479

Model analysis. Variations in optimal threshold exist across models (Fig. 7). Like in Fig. 3, the DET curves for three CNN-based models, each trained on VGG2 with softmax but with different backbones.[4] Notice similar trends across subgroups and models, which is consistent with Sphereface as well (Fig. 3). For example, the plots generated with Sphereface and VggFace2 all have the WM curve at the bottom (*i.e.*, best) and AF on top (*i.e.*, worst). Thus, the additional CNN-based models demonstrate the same phenomena: proportional to the overall model performance, exact in the ordering of curves for each subgroup.

Verification threshold analysis. We seek to reduce the bias between subgroups. Such that an operating point (*i.e.*, FPR) is constant across subgroups. To accomplish that, we used a per subgroup threshold. In FV, we consider one image as the query, and all others as the test. For this, the ethnicity of the query image is assumed. We can then examine the DET curves and pick the best threshold per subgroup for a certain FPR.

We evaluated True Acceptance Rate (TAR) for specific False Acceptance Rate (FAR) values. As described in Section 3.2, the verification experiments were 5-fold, with no overlap in subject ID between folds. Results reported are averaged across folds in all cases and are shown in Table 3.

[4] https://github.com/rcmalli/keras-vggface

Table 4: **Human assessment (quantitative).** Subgroups listed per row (*i.e.*, human) and column (*i.e.*, image). Note, most do the best intra-subgroup (blue), and second-best (red) intra-subgroup but inter-gender. WF performs the best; WF pairs are most correctly matched.

		Image				
	(Acc, %)	CF	CM	WF	WM	Avg
Human	CF	**52.9**	**48.0**	43.8	44.7	47.4
	CM	**45.6**	**50.4**	44.4	36.2	44.1
	WF	44.7	43.8	**57.3**	**48.0**	**48.5**
	WM	30.1	**47.4**	45.3	**56.1**	44.7
	Avg	43.3	47.4	**47.7**	46.3	46.2

For each subgroup, the TAR of using a global threshold is reported (upper row), as well as using the optimal per subgroup threshold (lower row).

Even for lower FAR, there are notable improvements, often of the order of 1%, which can be challenging to achieve when FAR is near $\geq$90%. More importantly, each subgroup has the desired FPR, so that substantial differences in FPR will remain unfounded. We experimented with ethnicity estimators on both the query and test image, which yielded similar results to those reported here.

Human evaluation analysis. Subjects of a subgroup likely have mostly been exposed to others of the same (Table 4 and Fig. 8). Therefore, it is expected they would be best at labeling their own, similar to the same ethnicity, but another gender. Our findings concur. Each subgroup is best at labeling their type, and then second best at labeling the same ethnicity but opposite sex.

Interestingly, each group of images is best tagged by the corresponding subgroup, with the second-best having the same ethnicity and opposite gender. On average, subgroups are comparable at labeling images. Analogous to the FR system, performance ratings differ when analyzing within and across subgroups. In other words, performance on BFW improved with subgroup-specific thresholds. Similarly, humans tend to better recognize individuals by facial cues of the same or similar demographics. Put differently, as the recognition performances drop with a global threshold optimized for one subgroup and deployed on another, human performance tends to fall when across subgroups (*i.e.*, performances drop for less familiar subgroups).

Figure 8: **Human assessment (qualitative).** ✓ for *match*; × for *non-match*. Accuracy scores shown as bar plots. Humans are most successful at recognizing their own subgroup, with a few exceptions (*e.g.*, bottom).

5. Conclusion

We introduce the Balanced Faces In the Wild (BFW) dataset with eight subgroups balanced across gender and ethnicity. With this, and upon highlighting the challenges and shortcomings of grouping subjects as a single subset, we provide evidence that forming subgroups is meaningful, as the FR algorithm rarely makes mistakes across subgroups. We used an *off-the-shelf* Sphereface, hypothesizing this SOTA CNN suffers from bias because of the imbalanced train-set. Once established that the results do suffer from problems of bias, we observed that the same threshold across ethnic and gender subgroups leads to differences in the FPR up to a factor of two. Also, we clearly showed notable percent differences in ratings across subgroups. Furthermore, we ameliorate these differences with a per-subgroup threshold, leveling out FPR, and achieving a higher true-positive rate (TPR). We hypothesized that most humans grew among more than their own demographic and, therefore, effectively learn from imbalanced datasets. In essence, a human evaluation validated that humans are biased, as most recognized their personal demographic best. This research, along with the data and resources, are extendable in vast ways. Thus, this is just a sliver of the larger problem of bias in ML.

References

[1] Alejandro Acien, Aythami Morales, Ruben Vera-Rodriguez, Ivan Bartolome, and Julian Fierrez. Measuring the gender and ethnicity bias in deep models for face recognition. In Ruben Vera-Rodriguez, Julian Fierrez, and Aythami Morales, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer International Publishing, 2019. 1, 3

[2] Anurag Ambekar, Charles Ward, Jahangir Mohammed, Swapna Male, and Steven Skiena. Name-ethnicity classification from open sources. In *Proceedings of the SIGKDD conference on Knowledge Discovery and Data Mining*, 2009. 4

[3] Alexander Amini, Ava Soleimany, Wilko Schwarting, Sangeeta Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. 2019. 2

[4] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 2018. 2

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2018. 3, 4, 7

[7] Abhijit Das, Antitza Dantcheva, and Francois Bremond. Nature and nurture in own-race face processing. In *Psychological science*, 2006. 3

[8] Abhijit Das, Antitza Dantcheva, and Francois Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. *European Conference on Computer Vision*, 2018. 3

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[10] Zhengming Ding, Yandong Guo, Lei Zhang, and Yun Fu. One-shot face recognition via generative learning. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2018. 2, 3

[11] Rachel England. Facial recognition misidentified 26 california lawmakers as criminals. 2019. 7

[12] Siyao Fu, Haibo He, and Zeng-Guang Hou. Learning race from face: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014. 4

[13] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, 2016. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 7

[16] Chong Huang, Xiao Chen, Peter Kairouz, Lalitha Sankar, and Ram Rajagopal. Generative adversarial models for learning private and fair representations. 2018. 2, 3

[17] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007. 5, 6

[18] Isabelle Hupont and Carles Fernandez. Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2019. 3, 4

[19] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[20] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Computer Vision and Pattern Recognition Workshop*, pages 34–42, 2015. 4

[21] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 4, 6

[22] Eric López-López, Xosé M Pardo, Carlos V Regueiro, Roberto Iglesias, and Fernando E Casado. Dataset bias exposed in face verification. *IET Biometrics*, 2019. 3

[23] Christian A Meissner and John C Brigham. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. In *Psychology, Public Policy, and Law*, 2001. 3

[24] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv:1901.10436*, 2019. 2

[25] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush Varshney. Understanding unequal gender classification from face images. *arXiv:1812.00099*, 2018. 2, 3

[26] Shruti Nagpal, Maneet Singh, Richa Singh, Mayank Vatsa, and Nalini Ratha. Deep learning for face recognition: Pride or prejudiced? *arXiv:1904.01219*, 2019. 1

[27] Michael ER Nicholls, Owen Churches, and Tobias Loetscher. Perception of an ambiguous figure is affected by own-age social biases. In *Scientific reports*, 2018. 3

[28] Joseph P Robinson, Ming Shao, Yue Wu, and Yun Fu. Families in the wild (fiw): Large-scale kinship image database and benchmarks. In *ACM Conference on Multimedia*, 2016. 4

[29] Joseph P Robinson, Ming Shao, Yue Wu, Hongfu Liu, Timothy Gillis, and Yun Fu. Visual kinship recognition of families in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 4

[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 4

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015. 7

[32] Jacob Snow. Amazons face recognition falsely matched 28 members of congress with mugshots. 2018. 1, 7

[33] Nisha Srinivas, Karl Ricanek, Dana Michalski, David S Bolme, and Michael King. Face recognition algorithm bias: Performance differences on images of children and adults. In *Computer Vision and Pattern Recognition Workshop*, 2019. 3

[34] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *European Conference on Computer Vision*, 2018. 2

[35] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*. Springer, 2017. 2

[36] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 2018. 1

[37] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[38] Mei Wang, Weihong Deng, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang. Race faces in-the-wild: Reduce bias by deep unsupervised domain adaptation. *arXiv:1812.00194*, 2019. 1, 2, 3, 4, 6

[39] Yu Wang, Yang Feng, Haofu Liao, Jiebo Luo, and Xiangyang Xu. Do they all look the same? deciphering chinese, japanese and koreans by fine-grained deep learning. In *IEEE Conference on Multimedia Information Processing and Retrieval*, 2018. 2, 3

[40] Sabine Windmann and Thomas Krüger. Subconscious detection of threat as reflected by an enhanced response bias. *Consciousness and cognition*, 1998. 2

[41] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014. 3, 4, 6

[42] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[43] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *Signal Processing Letters*, 2016. 4