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. Classic signal detection theory revealed trends in the underlying score distribution across subgroups. Specifically, we show variations in the optimal scoring threshold varies for face-pairs across different subgroups. Thus, the conventional approach of learning a global threshold for all pairs resulting in performance gaps among subgroups. By learning subgroup-specific thresholds, we not only mitigate problems in performance gaps but also show a notable boost in the overall performance. Furthermore, we do a human evaluation to measure the bias in humans, which supports the hypothesis that such a bias exists in human perception. To download the BFW database, source code, and more, visit github.com/visionjo/facerec-bias-bfw**

I. Introduction

As more of society becomes integrated with machine learning (ML), bias, fairness, and the formalization of ML standards are topics of high interest [**?**], [**?**], [**?**]. One effect of the growing dependency on technology is the increasing concern regarding biased and unfair algorithms. For instance, facial recognition (FR) systems can be untrustworthy and racist in some cases [**?**], [**?**]. .

Typically, convolutional neural networks (CNNs) are trained on faces identified by a detection system. Specifically, for FR, the goal is to project faces to an N-dimensional space with minimal distances between samples of the same identity and maximize the gap separating those that are different. Thus, the overarching goal is to discriminate between subjects with face encodings. We can deploy such a CNN to encode faces to compare via a similarity score: genuine pair if the score is high enough; and rejected as an imposter.

A fixed threshold acts as a decision boundary in similarity score space. Thus, the face features in the same class must satisfy a criterion in the form of a single value [**?**], [**?**], [**?**], [**?**]. However, a single, global threshold is a crude measure that leads to FR errors. Furthermore, the held-out set used to determine the global threshold tends to share the same distribution with test data, which makes the results skewed in favor of specific demographics that make up a majority in both sets. That skew, the difference in the performance of an algorithm of certain ethnic groups, is our definition of bias. A key question is: *is FR too biased, or not?*

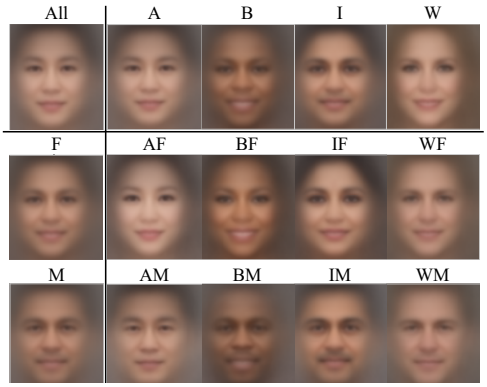

Fig. 1: **Balanced Faces In the Wild (BFW).** The average face of its different subsets: *top-left*: the entire BFW; *top-row* per race; *left-column*: per gender. The others represent the ethnicity and gender of the race and gender, respectfully. Table **??** defines the acronyms of subgroups.

The adverse effects of a global threshold are two-fold: **(1)** the mapping produced by a CNN is nonuniform. Therefore, distances between pairs of faces in different demographics vary in distribution of similarity scores (Fig **??**); **(2)** the evaluation set is imbalanced as well. Particular demographics making up a majority of the population will carry most weight on reported performance ratings. Reported results skew away from common traits to the underrepresented subgroups. Alas, demo-

TABLE I: **The proposed BFW compared to related datasets.** BFW is exactly balanced across ID, gender, and ethnicity (Table **??**). Compared with Demographic Pairs (DemogPairs), BFW provides more samples per subject and subgroups per set. Also, BFW uses a single resource, VGG2. Racial Faces in-the-Wild: (RFW), on the other hand, supports a different task (*i.e.*, domain adaptation). Furthermore, RFW focuses on race-distribution across of pairs, while not considering the distribution of identities.

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

graphics like gender, ethnicity, race, and age are underrepresented in most public datasets.

Making matters more challenging is that race and ethnicity are loosely defined. For example, the US Census Bureau allows an individual to self-identify race (https://www.census.gov). We define it as a group of people having facial characteristics similar to those found in a region. The result is various types of biases in FR systems in favor of or against particular demographics remain a question.

To address **(2)**, the lack of a balanced data, we introduce a new benchmark for FR, called BFW (Table **??**, **??**). BFW serves as a platform to fairly evaluate FR systems and enable demographic-specific ratings to be reported. We use BFW to gain a deeper understanding of the extent of bias present in facial embeddings extracted from a state-of-the-art (SOTA) CNN model. We then suggest a mechanism to counter the biased feature space to mitigate problems of bias with more balanced performance ratings for different demographics, while improving the overall accuracy. Specifically, we propose using an adaptive threshold that varies depending on the characteristics of detected facial attributes (*i.e.*, gender and ethnicity, Fig **??**). We show an increase in accuracy with a balanced performance for different subgroups of people. Similarly, we show the 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-based research, while providing a method easily adoptable in practice.

The contributions in this paper are as follows:

1) We built a balanced dataset as a proxy to measure verification performance per subgroup.
2) We analyzed an unwanted bias in scores of face pairs while showing that optimal thresholds determined per subgroup significantly boost and balances performances.
3) We showed bias causes inconsistencies in ratings across demographics– the typical use of a global threshold unfavorable. We mitigate the problem via adaptive thresholds.
4) We conducted human-evaluations to demonstrate bias in human perception.[1]

II. Background Information

A. Bias in ML

Although the community is excited about the progress and the commercial value of ML, the trust between the society and techniques will take time to build due to the bias in algorithms. The exact definitions and implications of bias tend to vary between sources, as do its sources and types. One common theme is the tendency of bias to hinder performance ratings in ways that skew in favor of a particular subpopulation. At a high-level, bias can be introduced by humans [**?**], data and label types [**?**], ML models [**?**], [**?**], and during evaluation [**?**]. For instance, a vehicle-detection model might miss cars if training set was mostly trucks. In practice, many ML systems learn on biased datasets, which could harm our society.

B. Bias in FR

Problems of bias in FR have been driven by different motivations and have therefore been solved by different means. To name a few: problems of data augmentation [**?**], one-shot learning [**?**], demographic parity and fairness with priority on privacy [**?**], domain adaptation [**?**], differences in

[1]NIH-certified, *Protect Humans in Research*, IRB 19-09-08.

TABLE II: **Database stats and nomenclature, optimal thresholds (t_o), and accuracy scores.** *Header:* Subgroup definitions. *Top-row:* Statistics of BFW. *Middle-row:* Number of pairs for each partition. *Bottom-row:* Accuracy when applying a global threshold t_g, the optimal threshold t_o, and accuracy with t_o per subgroup. Columns grouped by race and then further split by gender. Out of millions of pairs, accuracy is inconsistent across subgroups. Furthermore, *F* tend to perform inferior to that of *M* (*i.e.*, up to 8%).

	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
Acc@t_g	0.876	0.944	0.934	0.942	0.922	0.949	0.916	0.918	0.925±0.022
t_o	0.235	0.274	0.267	0.254	0.299	0.295	0.242	0.222	0.261
Acc@t_o	0.916	0.964	0.955	0.971	0.933	0.958	0.969	0.973	0.955 ± 0.018

face-based attributes across demographics [**?**], and even exploratory data analysis [**?**]. Yin et al. [**?**] proposed to augment the feature space of underrepresented classes using other classes with a diverse collection of samples to encourage distributions of underrepresented classes to more closely resemble that of the others. Similarly, [**?**] formulated the imbalanced class as one-shot learning and trained a generative adversarial network (GAN) to generate face features as a means of augmenting class with as few as one sample. [**?**] proposed Generative Adversarial Privacy and Fairness (GAPF) to create fair representations of the data in a quantifiable way, allowing for the finding of a decorrelation scheme from the data without access to its statistics. Wang et al. [**?**] defined subgroups at a finer-level (*i.e.*, Chinese, Japanese, Korean) to determine the familiarity of faces across these subgroups. Genders have also been used as subgroups for work in bias, whether for efforts of analysis and understanding of gender-based face encodings [**?**]. Most recently, [**?**] proposed adapting domains to bridge the gap between races by knowledge transfer, which was supported by a novel labeled data collection, RFW. The release of RFW came after the completion of BFW. Although similar in terms of demographics (*i.e.*, Asian, Black, Indian, and White), RFW uses faces from MSCeleb [**?**] based on the assumption CASIA-Face [**?**] and VGG2 [**?**] was used for training. In contrast, our BFW samples from VGG2 and assumes MSCeleb was the training set. As shown in [**?**], MSCeleb is highly imbalanced, primarily consisting of images of white individuals. For this, we expect bias from data, for this is the training set. Furthermore, our experiments allow for the parsing of results based on gender demographics (*i.e.*, although RFW does separate males and females in each subgroup, no consideration for splits in just gender or just races are incorporated into the experimental design). Thus, RFW and BFW are complementary, with both adding metadata on demographics for subjects with faces in renowned, large-scale FR datasets.

Most similar to our work are the recent efforts in [**?**], [**?**], [**?**], [**?**]. A common factor of these efforts are claims of insufficient data supply to support studies on bias in FR, and the introduction of new metadata on demographics for image sets parsed out of existing collections. Specifically, [**?**] curated a set of faces based on racial demographics (*i.e.*, Asian, Black, and White) called DemogPairs, while [**?**] honed in on adults versus children called glsitwcc. Like the proposed BFW, both were built by sampling existing databases, but with the addition of tags for the respective subgroups of interest. Besides, the additional data of BFW (*i.e.*, added an additional subgroup *Indian* (I), along with additional 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.

C. Human bias in ML

Bias is not unique to ML, as humans also susceptible to a perceived bias across demographics: it exists across different races, genders, and even ages [**?**], [**?**], [**?**], [**?**]. [**?**] showed machines surpass human performance in classifying faces as Japanese, Chinese, or Korean by nearly 150%. Precisely, humans barely pass random with 38.89% accuracy (*i.e.*, $\frac{1}{3}$ is random).

We expect human bias to skew results in favor of their genders and races. For this, we measure the human perception of faces across demographics in a controlled experiment. In the end, the results concur [**?**]– we too observed an overall average below random (*i.e.*, <50%). Furthermore, we provide details on settings and counts on the number of submissions per demographics while analyzing per demographic.

III. The BFW Benchmark and Dataset

We now discuss the BFW dataset, and protocols to evaluate ML-based FR. We conclude this section by reviewing the human evaluation conducted as part of this work, to detect bias in humans.

A. The data

Problems of bias in FR motivated the design of BFW. The data evenly represent various subgroups partitioned by demographics. Inspired by DemogPairs [**?**], the specification of BFW follows in suit. BFW includes additional subgroups (*i.e.*, IF and IM), an increased in the number subjects per subgroup, with many more pairs (Table **??**).

Compiling subject list. All subjects were sampled from VGG2 [**?**]. Thus, unlike others that depend on multiple sources, BFW has less potential conflicts in train/test overlap for existing models. We used pre-trained ethnicity [**?**] and gender [**?**] classifiers to find candidates for the different subgroups.

Detecting faces. Faces were detected using *multi-task CNN* (MTCNN) [**?**].[2] Then, assigned into 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 Arcface [**?**] (see the next step). Also, coordinates of the BB and the five landmarks per

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

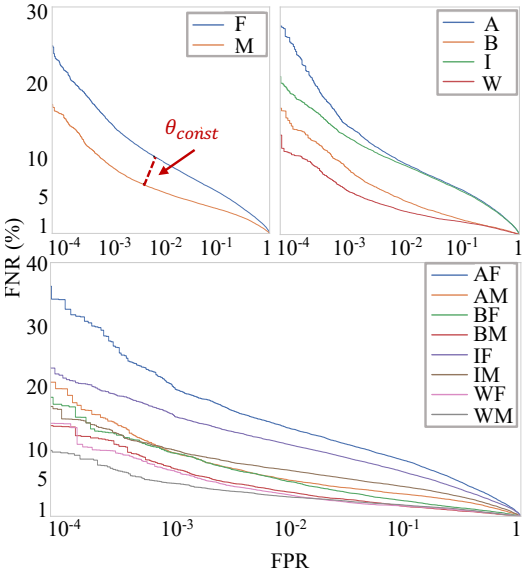

Fig. 2: **Detection Error Trade-off (DET) curves.** *Top-left*: per gender. *Top-right*: per ethnicity. *Bottom*: per subgroup (*i.e.*, combined). A dashed lined shows a difference by about 2× in FPR for the same threshold θ_{const}. FNR is the number of match errors, so closer a curve is to the bottom better.

MTCNN were stored as part of the static, raw data. Samples with multiple face detections had the BB area times the confidence score of the MTCNN to determine the instance most likely to be the face, with others set aside and labeled *miss-detection*.

Validating labels. Faces of BFW were encoded using the original implementation of the SOTA Arcface [**?**]. A matrix of cosine similarity scores was then generated for each subject, and removed samples (*i.e.*, rows) with median scores below threshold $\theta = 0.2$. Mathematically, the j^{th} subject with N_j faces is removed if the ordinal rank of its score $n = \frac{P \times N}{100} \geq \theta$, with the percentile $P = 50$, $\theta = 0.2$ set manually, and an ordered list of scores. This allowed us to quickly prune false-positive (FP) face detections. Following [**?**], [**?**], we built a JAVA tool to visually validate the remaining faces. The faces were first ordered from most-to-least 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

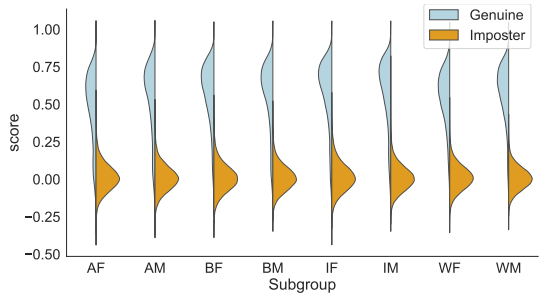

Fig. 3: **Signal detection model (SDM) across subgroups.** The scores for *imposters* have medians about 0.3 but with variation in upper percentiles, while *genuine* pairs vary in both (*e.g.*, AF has more area of overlap). A varying threshold varying across different subgroups yields a constant FPR.

subject-by-subject and flagging faces of *imposters*.

Sampling faces and creating folds. We created lists of pairs in five-folds with subjects split evenly per subject and without overlap across folds. Furthermore, balance in the number of faces per was obtained by sampling twenty-five faces at random from each. Next, we generated a list of all 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 **??**). 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 total count with negative pairs from across subgroups but in the same fold.

B. 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*), respectfully. 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 subjects is infeasible. 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:

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

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

Cosine similarity is used as the *matcher* in Eq **??** 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.

C. Human Assessment

We evaluated human on face pairs focusing on two racial groups: Chinese and Caucasians. To focus on the experiment, we honed-in on two groups, white Americans (W) and Chinese from China (C). The purpose was to the minimize variability by only analyzing the subsets of the broader groups of whites and Asians.

Samples were collected by recruiting subjects from multiple sources (*e.g.*, social media, email lists, and family/friends)– a total of 120 participants were sampled at random from all the submissions that were (1) complete and (2) from 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 (WA) and 20 (C) pairs (male and female split evenly). The other 10 pairs are of others (*e.g.*, Hispanic/ Latino, Japanese, African). Survey was created, distributed, and recorded via PaperForm.

IV. Results and Analysis

An *off-the-shelf* CNN was used throughout to control the model across all experiments. For this, Sphereface [**?**] trained on CASIA-Web [**?**], and evaluated on Labeled Faces in the Wild (LFW) [**?**] (%99.22 accuracy), encoded all of the faces.[3] Following [**?**], upon applying an affine transformation to align faces according to pre-defined eye locations, each face got fed through the network twice, the original and horizontally flipped. The two features were fused by concatenation.

[3] https:github.comclcarwinsphereface_pytorch

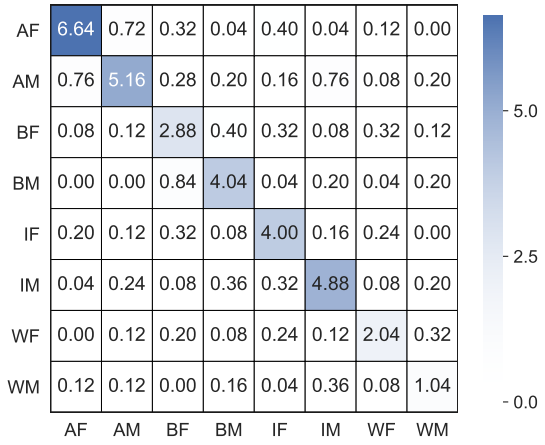

Fig. 4: **Confusion matrix.** Percent error (Rank 1) for all faces of BFW queried against all others. Notice errors concentrate within a subgroup, consistent with the SDM in Fig. **??** (*i.e.*, AF show worst performance, and is mostly confused with faces of the same demographic). This plot is evidence that while race/ethnicity may be challenging to define, the subgroups are meaningful.

Additional CNN-based models demonstrate the same phenomena: proportional to the overall model performance, exact in which the ordering subgroups in sensitivity in scores space (Appendix **??**).

A. Score Analysis

Fig. **??** shows score distributions for faces of the same (*i.e.*, *Genuine*) and different (*i.e.*, *Imposter*) identity, with a subgroup per SDM plot. Notice that score distributions for imposters tends 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). Asian Fem Fig. **??** shows the confusion matrix of the subgroups. A vast majority of errors occurs in 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 in the context of FR.

B. Detection Error Trade-off (DET) analysis

DET curves averaged across 5-folds show per-subgroup trade-offs (Fig. **??**). Note that M perform

TABLE III: **TAR at intervals of FAR**. For each subgroup, and overall average FAR, listed are the TAR scores for a global threshold (top) and the proposed category-based threshold (bottom). Higher is better. The proposed shows improvement in all cases.

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

better than F, precisely as one would expect from the tails of score-distributions for *genuine* pairs (Fig. **??**). AF and IF perform the worst.

The gender-based DET curve shows a difference in performances between M and F with a fixed threshold (dashed-line). Similar effects exist for the other curves as well (lines omitted to declutter). For many FR applications, systems are operated at the highest FPR allowed, so the line of constant threshold indicates that a single threshold produces different operating points (*i.e.*, FPR), which is undesirable. The difference in FPR is approximately a factor of two– quite large indeed. If this is the case in an industrial system, one would expect a difference in about double the false positives to be reported on based on subgroup alone. The potential ramifications of such a bias should not be overlooked, which it has not as of lately– gaining interest of even main-stream media [**?**], [**?**].

C. Verification threshold

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 test. For this, the ethnicity of the query image is assumed. We

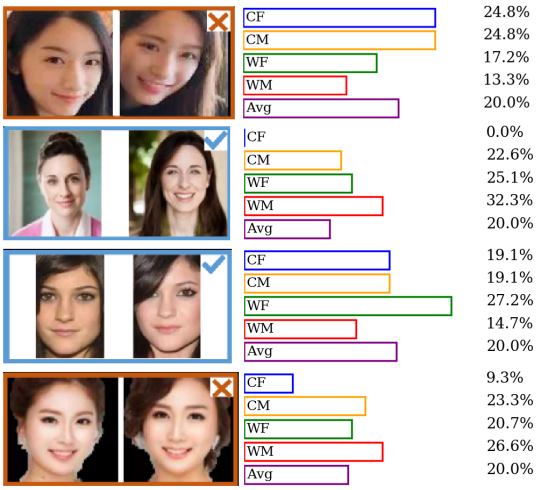

Fig. 5: **Qualitative results of human assessment.** ✓ for *match*; × for *non-match*, which scores plotted next to each. Humans tend to be more successful at recognizing their own subgroup, with few exceptions (bottom).

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 **??**, 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 **??**. 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 further experimented using ethnicity estimators and using the ethnicity of both query and test image, which yielded similar results to those reported here (results not shown).

D. Human evaluation

Quantitative and qualitative results are in Table **??** and Fig. **??**, respectfully. One might expect that the most exposure to others would be within the same subgroup, and, therefore, would be best at labeling their own. Secondarily, they would be best at labeling images of the same ethnicity, but opposite 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-to-best having the same ethnicity and opposite gender. On average, subgroups are comparable at labeling images.

TABLE IV: **Quantitative of human assessment.** Different human subgroups listed per row. Each column is the subgroup labeled. Note that each people are best within their subgroup, and second-best within the same subgroup but different gender. CF shows the least variation, but with the lowest accuracy. CM shows the best accuracy, but second to WM in deviation from the mean. Thus, scores for males vary more than females.

	CF	CM	WF	WM	Avg
CF	**0.529**	0.480	0.438	0.447	0.474±0.041
CM	0.456	**0.504**	0.444	0.362	0.441±0.059
WF	0.447	0.438	**0.573**	0.480	0.485±0.062
WM	0.301	0.474	0.453	**0.561**	0.447±0.108
Avg	0.433	0.474	0.477	0.463	0.462±0.020

V. CONCLUSION

We introduce a new data set Balanced Faces In the Wild (BFW) 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 facial recognition (FR) algorithm rarely makes mistakes across subgroups. We trained Arc-Face net on MSCeleb, expecting that the results would suffer 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 false-positive rate (FPR) up to a factor of two, which is seemingly the cause of the frenzy about bias in FR in main-stream media. 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 grown amongst more than their own demographic and, therefore, effectively learn from imbalanced datasets– a human evaluation validated that humans are biased, as most recognized their personal demographic best. The focused research findings presented here, along with the public database included, are extendable in vast ways. Thus, we see this as the

slither to a much larger problem of bias in ML.

A. Other convolutional neural network (CNN) models.

Variations in optimal threshold exists across different models (Fig. **??**). Like in Fig. **??**, the Detection Error Trade-off (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. **??**). For example, the plots generated with ArcFace and VggFace2 all have the *White-Male* (WM) curve at the bottom (*i.e.*, best) and *Asian-Female* (AF) on top (*i.e.*, worst).

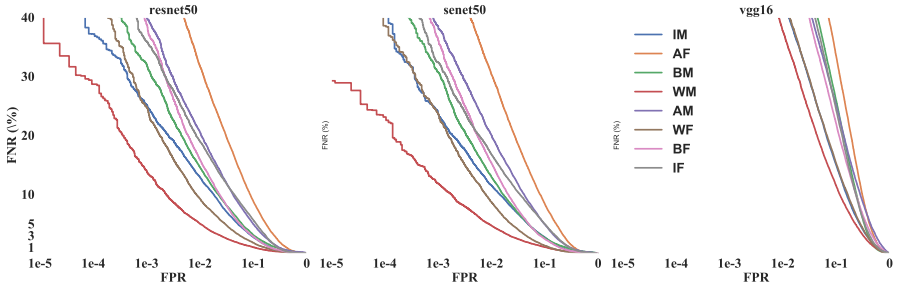

Fig. 6: **DET curves for different CNN models**. false-negative rate (FNR) (%) (vertical) vs (log-scale) (horizontal) for VGG2 [**?**] models with different backbones (vgg16, Resnet50 [**?**], SEnet50 [**?**]). Lower is better. For each plot, WM is the best performing curve, AF is the worst. The the ordering of the curves is roughly the same for each backbone.

Sample faces per subgroup of the Balanced Faces In the Wild (BFW) dataset are shown in Fig. **??**.

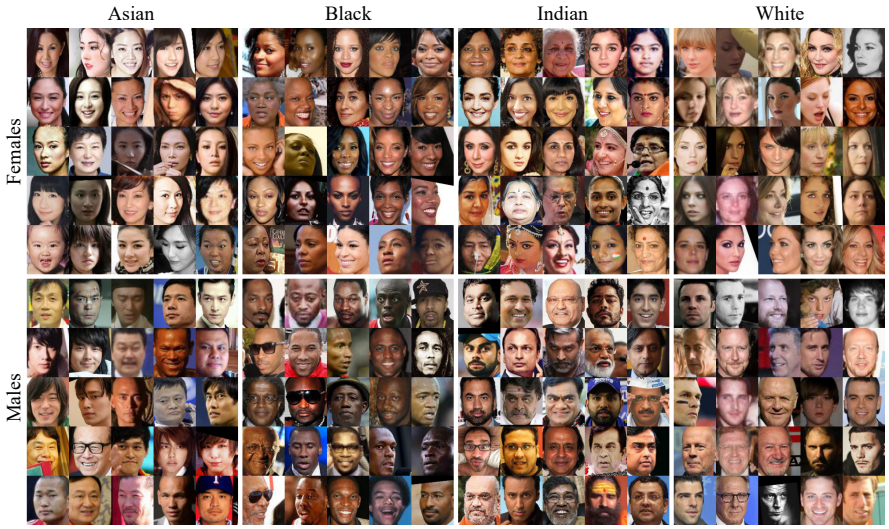

Fig. 7: **Sample of BFW**. Each row depicts a different gender, *Female* (F) (top) and *Male* (M) (bottom). Columns are grouped by ethnicity (*i.e.*, *Asian* (A), *Black* (B), *Indian* (I), and *White* (W), respectfully).

References

[1] Cao, Qiong, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. "Vggface2: A dataset for recognising faces across pose and age." In *IEEE International Conference on Automatic Face Gesture Recognition.* 2018.

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

[3] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." In *IEEE Conference on Computer Vision and Pattern Recognition.* 2018.

[4] Used pre-trained models and public Github, https://github.com/rcmalli/keras-vggface