FAIRFACE: A NOVEL FACE ATTRIBUTE DATASET FOR BIAS MEASUREMENT AND MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing public face image datasets are strongly biased toward Caucasian faces, and other races (e.g., Latino) are significantly underrepresented. The models trained from such datasets suffer from inconsistent classification accuracy, which limits the applicability of face analytic systems to non-White race groups. To mitigate the race bias problem in these datasets, we constructed a novel face image dataset containing 108,501 images which is balanced on race. We define 7 race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. Images were collected from the YFCC-100M Flickr dataset and labeled with race, gender, and age groups. Evaluations were performed on existing face attribute datasets as well as novel image datasets to measure the generalization performance. We find that the model trained from our dataset is substantially more accurate on novel datasets and the accuracy is consistent across race and gender groups. We also compare several commercial computer vision APIs and report their balanced accuracy across gender, race, and age groups.

1 INTRODUCTION

To date, numerous large scale face image datasets (Huang et al., 2007; Kumar et al., 2011; Escalera et al., 2016; Yi et al., 2014; Liu et al., 2015; Joo et al., 2015; Parkhi et al., 2015; Yang et al., 2016; Guo et al., 2016; Kemelmacher-Shlizerman et al., 2016; Rothe et al., 2016; Cao et al., 2018; Merler et al., 2019) have been proposed and fostered research and development for automated face detection (Li et al., 2015b; Hu & Ramanan, 2017), alignment (Xiong & De la Torre, 2013; Ren et al., 2014), recognition (Taigman et al., 2014; Schroff et al., 2015), generation (Yan et al., 2016; Bao et al., 2017; Karras et al., 2018; Thomas & Kovashka, 2018), modification (Antipov et al., 2017; Lample et al., 2017; He et al., 2017), and attribute classification (Kumar et al., 2011; Liu et al., 2015). These systems have been successfully translated into many areas including security, medicine, education, and social sciences.

Despite the sheer amount of available data, existing public face datasets are strongly biased toward Caucasian faces, and other races (e.g., Latino) are significantly underrepresented. A recent study shows that most existing large scale face databases are biased towards "lighter skin" faces (around 80%), e.g. White, compared to "darker" faces, e.g. Black (Merler et al., 2019). This means the model may not apply to some subpopulations and its results may not be compared across different groups without calibration. Biased data will produce biased models trained from it. This will raise ethical concerns about fairness of automated systems, which has emerged as a critical topic of study in the recent machine learning and AI literature (Hardt et al., 2016; Corbett-Davies et al., 2017).

For example, several commercial computer vision systems (Microsoft, IBM, Face++) have been criticized due to their asymmetric accuracy across sub-demographics in recent studies (Buolamwini & Gebru, 2018; Raji & Buolamwini, 2019). These studies found that the commercial face gender classification systems all perform better on male and on light faces. This can be caused by the biases in their training data. Various unwanted biases in image datasets can easily occur due to biased selection, capture, and negative sets (Torralba & Efros, 2011). Most public large scale face datasets have been collected from popular online media – newspapers, Wikipedia, or web search– and these platforms are more frequently used by or showing White people.

To mitigate the race bias in the existing face datasets, we propose a novel face dataset with an emphasis on balanced race composition. Our dataset contains 108,501 facial images collected primarily

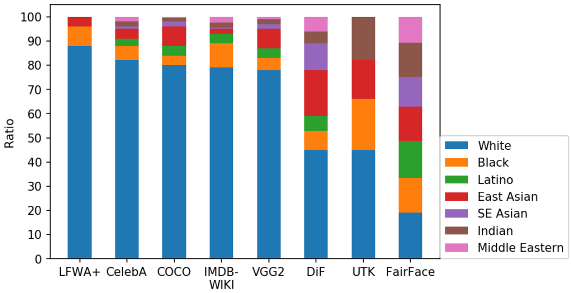
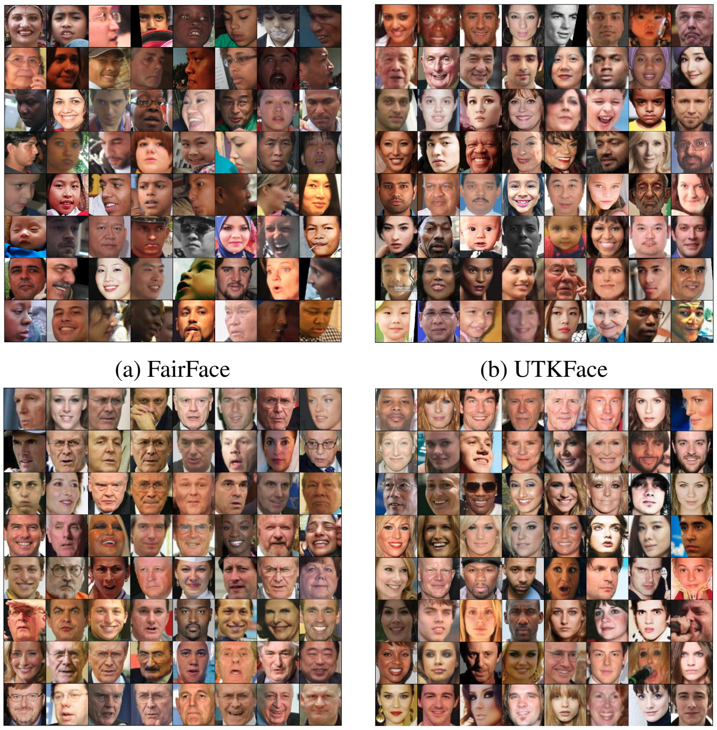


Figure 1: Racial compositions in face datasets.



(a) FairFace

(b) UTKFace

(c) LFWA+

(d) CelebA

Figure 2: Random samples from face attribute datasets.

from the YFCC-100M Flickr dataset (Thomee et al.), which can be freely shared for a research purpose, and also includes examples from other sources such as Twitter and online newspaper outlets. We define 7 race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. Our dataset is well-balanced on these 7 groups (See Figures 1 and 2)

Our paper makes three main contributions. First, we emprically show that existing face attribute datasets and models learned from them do not generalize well to unseen data in which more non-White faces are present. Second, we show that our new dataset performs better on novel data, not only on average, but also across racial groups, i.e. more consistently. Third, to the best of our knowledge, our dataset is the first large scale face attribute dataset in the wild which includes Latino and Middle Eastern and differentiates East Asian and Southeast Asian. Computer vision has been rapidly transferred into other fields such as economics or social sciences, where researchers want to analyze different demographics using image data. The inclusion of major racial groups, which have been missing in existing datasets, therefore significantly enlarges the applicability of computer vision methods to these fields.

2 Related Work

2.1 Face Attribute Recognition

The goal of face attribute recognition is to classify various human attributes such as gender, race, age, emotions, expressions or other facial traits from facial appearance (Kumar et al., 2011; Joo et al., 2013; Zhang et al., 2015; Liu et al., 2015). Table 1 summarizes the statistics of existing large scale **public** and **in-the-wild** face attribute datasets including our new dataset. As stated earlier, most of these datasets were constructed from online sources and are typically dominated by the White race.

Table 1: Statistics of Face Attribute Datasets

						Race Annotation							
Name	Source	# of faces	In-the-wild?	Age	Gender	White*		Asian*		Black	Indian	Latino	Balanced?
						W	ME	E	SE				
PPB (Buolamwini & Gebru, 2018)	Gov. Official Profiles	1K		✓	✓	**Skin color prediction							
MORPH (Ricanek & Tesafaye, 2006)	Public Data	55K		✓	✓	merged				✓		✓	no
PubFig (Kumar et al., 2011)	Celebrity	13K	✓	Model generated predictions									no
IMDB-WIKI (Rothe et al., 2016)	IMDB, WIKI	500K	✓	✓	✓								no
FotW (Escalera et al., 2016)	Flickr	25K	✓	✓	✓								yes
CACD (Chen et al., 2015)	celebrity	160K	✓	✓									no
DiF (Merler et al., 2019)	Flickr	1M	✓	✓	✓	**Skin color prediction							
†CelebA (Liu et al., 2015)	CelebFace LFW	200K	✓	✓	✓								no
LFW+ (Han et al., 2018)	LFW (Newspapers)	15K	✓	✓	✓	merged		merged					no
†LFWA+ (Liu et al., 2015)	LFW (Newspapers)	13K	✓		✓	merged		merged		✓	✓		no
†UTKFace (Zhang et al., 2017)	MORPH, CACD Web	20K	✓	✓	✓	merged		merged		✓	✓		yes
FairFace (Ours)	Flickr, Twitter Newspapers, Web	108K	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	yes

*FairFace (Ours) also defines East (E) Asian, Southeast (SE) Asian, Middle Eastern (ME), and Western (W) White.
**PPB and DiF do not provide race annotations but skin color annotated or automatically computed as a proxy to race.
†denotes datasets used in our experiments.

Face attribute recognition has been applied as a sub-component to other computer vision tasks such as face verification (Kumar et al., 2011) and person re-idenfication (Layne et al., 2012; Li et al., 2015a; Su et al., 2018). It is imperative to ensure that these systems perform evenly well on different gender and race groups. Failing to do so can be detrimental to the reputations of individual service providers and the public trust about the machine learning and computer vision research community. Most notable incidents regarding the racial bias include Google Photos recognizing African American faces as Gorilla and Nikon's digital cameras prompting a message asking "did someone blink?" to Asian users (Zhang, 2015). These incidents, regardless of whether the models were trained improperly or how much they actually affected the users, often result in the termination of the service or features (e.g. dropping sensitive output categories). For this reason, most commercial service providers have stopped providing a race classifier.

Face attribute recognition is also used for demographic surveys performed in marketing or social science research, aimed at understanding human social behaviors and their relations to demographic backgrounds of individuals. Using off-the-shelf tools (Amos et al., 2016; Baltrusaitis et al., 2018) and commercial services, social scientists have begun to use images of people to infer their demographic attributes and analyze their behaviors. Notable examples are demographic analyses of social media users using their photographs (Chakraborty et al., 2017; Reis et al., 2017; Won et al., 2017; Xi et al., 2019; Wang et al., 2017). The cost of unfair classification is huge as it can over- or under-estimate specific sub-populations in their analysis, which may have policy implications.

2.2 Fair Classification and Dataset Bias

AI and machine learning communities have increasingly paid attention to algorithmic fairness and dataset and model biases (Zemel et al., 2013; Corbett-Davies et al., 2017; Zou & Schiebinger, 2018;

Zhang et al., 2018). There exist many different definitions of fairness used in the literature (Verma & Rubin, 2018). In this paper, we focus on balanced accuracy–whether the attribute classification accuracy is independent of race and gender. More generally, research in fairness is concerned with a model's ability to produce fair outcomes (e.g. loan approval) independent of protected or sensitive attributes such as race or gender.

Studies in algorithmic fairness have focused on either 1) discovering (auditing) existing bias in datasets or systems (Shankar et al., 2017; Buolamwini & Gebru, 2018; Kiritchenko & Mohammad, 2018; McDuff et al., 2019), 2) making a better dataset (Merler et al., 2019; Alvi et al., 2018), or 3) designing a better algorithm or model (Das et al., 2018; Alvi et al., 2018; Ryu et al., 2017; Zemel et al., 2013; Zafar et al., 2017). Our paper falls into the first two categories.

The main task of interest in our paper is (balanced) gender classification from facial images. Buolamwini & Gebru (2018) demonstrated many commercial gender classification systems are biased and least accurate on dark-skinned females. The biased results may be caused by biased datasets, such as skewed image origins (45% of images are from the U.S. in Imagenet) (Suresh et al., 2018) or biased underlying associations between scene and race in images (Stock & Cisse, 2018). It is, however, "infeasible to balance across all possible co-occurrences" of attributes (Hendricks et al., 2018), except in a lab-controlled setting.

Therefore, the contribution of our paper is to mitigate, not entirely solve, the current limitations and biases of existing databases by collecting more diverse face images from non-White race groups. We empirically show this significantly improves the generalization performance to novel image datasets whose racial compositions are not dominated by the White race. Furthermore, as shown in Table 1, our dataset is the first large scale in-the-wild face image dataset which includes Southeast Asian and Middle Eastern races. While their faces share similarity with East Asian and White groups, we argue that not having these major race groups in datasets is a strong form of discrimination.

3 Dataset Construction

3.1 Race Taxonomy

Our dataset defines 7 race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. Race and ethnicity are different categorizations of humans. Race is defined based on physical traits and ethnicity is based on cultural similarities (Schaefer, 2008). For example, Asian immigrants in Latin America can be of Latino ethnicity. In practice, these two terms are often used interchangeably.

We first adopted a commonly accepted race classification from the U.S. Census Bureau (White, Black, Asian, Hawaiian and Pacific Islanders, Native Americans, and Latino). Latino is often treated as an ethnicity, but we consider Latino a race, which can be judged from the facial appearance. We then further divided subgroups such as Middle Eastern, East Asian, Southeast Asian, and Indian, as they look clearly distinct. During the data collection, we found very few examples for Hawaiian and Pacific Islanders and Native Americans and discarded these categories. All the experiments conducted in this paper were therefore based on 7 race classification.

An important criterion to measure dataset bias is on which basis the bias should be measured: **skin color or race?** A few recent studies (Buolamwini & Gebru, 2018; Merler et al., 2019) use skin color as a proxy to racial or ethnicity grouping. While skin color can be easily computed without subjective annotations, it has limitations. First, skin color is heavily affected by illumination and light conditions. The Pilot Parliaments Benchmark (PPB) dataset (Buolamwini & Gebru, 2018) only used profile photographs of government officials taken in well controlled lighting, which makes it non-in-the-wild. Second, within-group variations of skin color are huge. Even same individuals can show different skin colors over time. Third, most importantly, race is a multidimensional concept whereas skin color (i.e. brightness) is one dimensional. Figure 5 in Appendix shows the distributions of the skin color of multiple race groups, measured by Individual Typology Angle (ITA) (Wilkes et al., 2015). As shown here, the skin color provides no information to differentiate many groups such as East Asian and White. Therefore, we explicitly use race and annotate the physical race by human annotators' judgments. To complement the limits of race categorization, however, we also use skin color, measured by ITA, following the same procedure used by Merler et al. (2019).

3.2 Image Collection and Annotation

Many existing face datasets have been sourced from photographs of public figures such as politicians or celebrities (Kumar et al., 2011; Huang et al., 2007; Joo et al., 2015; Rothe et al., 2016; Liu et al., 2015). Despite the easiness of collecting images and ground truth attributes, the selection of these populations may be biased. For example, politicians may be older and actors may be more attractive than typical faces. Their images are usually taken by professional photographers in limited situations, leading to the quality bias. Some datasets were collected via web search using keywords such as "Asian boy" (Zhang et al., 2017). These queries may return only stereotypical faces or prioritize celebrities in those categories rather than diverse individuals among general public.

Our goal is to minimize the selection bias introduced by such filtering and maximize the diversity and coverage of the dataset. We started from a huge public image dataset, Yahoo YFCC100M dataset (Thomee et al.), and detected faces from the images without any preselection. A recent work also used the same dataset to construct a huge unfiltered face dataset (Diversity in Faces, DiF) (Merler et al., 2019). Our dataset is smaller but more balanced on race (See Figure 1).

For an efficient collection, we incrementally increased the dataset size. We first detected and annotated 7,125 faces randomly sampled from the entire YFCC100M dataset ignoring the locations of images. After obtaining annotations on this initial set, we estimated demographic compositions of each country. Based on this statistic, we adaptively adjusted the number of images for each country sampled from the dataset such that the dataset is not dominated by the White race. Consequently, we excluded the U.S. and European countries in the later stage of data collection after we sampled enough White faces from those countries. The minimum size of a detected face was set to 50 by 50 pixels. This is a relatively smaller size compared to other datasets, but we find the attributes are still recognizable and these examples can actually make the classifiers more robust against noisy data. We only used images with "Attribution" and "Share Alike" Creative Commons licenses, which allow derivative work and commercial usages.

We used Amazon Mechanical Turk to annotate the race, gender and age group for each face. We assigned three workers for each image. If two or three workers agreed on their judgements, we took the values as ground-truth. If all three workers produced different responses, we republished the image to another 3 workers and subsequently discarded the image if the new annotators did not agree. These annotations at this stage were still noisy. We further refined the annotations by training a model from the initial ground truth annotations and applying back to the dataset. We then manually re-verified the annotations for images whose annotations differed from model predictions.

4 Experiments

4.1 Measuring Bias in Datasets

We first measure how skewed each dataset is in terms of its race composition. For the datasets with race annotations, we use the reported statistics. For the other datasets, we annotated the race labels for 3,000 random samples drawn from each dataset. See Figure 1 for the result. As expected, most existing face attribute datasets, especially the ones focusing on celebrities or politicians, are biased toward the White race. Unlike race, we find that most datasets are relatively more balanced on gender ranging from 40%-60% male ratio.

4.2 Model and Cross-Dataset Performance

To compare model performance of different datasets, we used an identical model architecture, ResNet-34 (He et al., 2016), to be trained from each dataset. We used ADAM optimization (Kingma & Ba, 2014) with a learning rate of 0.0001. Given an image, we detected faces using the dlib's (dlib.net) CNN-based face detector (King, 2015) and ran the attribute classifier on each face. The experiment was done in PyTorch.

Throughout the evaluations, we compare our dataset with three other datasets: UTKFace (Zhang et al., 2017), LFWA+, and CelebA (Liu et al., 2015). Both UTKFace and LFWA+ have race annotations, and thus, are suitable for comparison with our dataset. CelebA does not have race annotations, so we only use it for gender classification. See Table 1 for more detailed dataset characteristics.

Table 2: Cross-Dataset Classification Accuracy on White Race.

		Tested on								
		Race			Gender				Age	
		FairFace	UTKFace	LFWA+	FairFace	UTKFace	LFWA+	CelebA*	FairFace	UTKFace
Trained on	FairFace	**.937**	.936	**.970**	**.942**	**.940**	.920	**.981**	**.597**	.565
	UTKFace	.800	.918	.925	.860	.935	.916	.962	.413	**.576**
	LFWA+	.879	**.947**	.961	.761	.842	**.930**	.940	-	-
	CelebA	-	-	-	.812	.880	.905	.971	-	-

* CelebA doesn't provide race annotations. The result was obtained from the whole set (white and non-white).

Table 3: Cross-Dataset Classification Accuracy on non-White Races.

		Tested on								
		Race†			Gender				Age	
		FairFace	UTKFace	LFWA+	FairFace	UTKFace	LFWA+	CelebA*	FairFace	UTKFace
Trained on	FairFace	**.754**	.801	**.960**	**.944**	**.939**	**.930**	**.981**	**.607**	.616
	UTKFace	.693	**.839**	.887	.823	.925	.908	.962	.418	**.617**
	LFWA+	.541	.380	.866	.738	.833	.894	.940	-	-
	CelebA	-	-	-	.781	.886	.901	.971	-	-

* CelebA doesn't provide race annotations. The result was obtained from the whole set (white and non-white).
† FairFace defines 7 race categories but only 4 races (White, Black, Asian, and Indian) were used in this result to make it comparable to UTKFace.

Using models trained from these datasets, we first performed cross-dataset classifications, by alternating training sets and test sets. Note that FairFace is the only dataset with 7 races. To make it compatible with other datasets, we merged our fine racial groups when tested on other datasets. CelebA does not have race annotations but was included for gender classification.

Tables 2 and 3 show the classification results for race, gender, and age on the datasets across subpopulations. As expected, each model tends to perform better on the same dataset on which it was trained. However, the accuracy of our model was highest on some variables on the LFWA+ dataset and also very close to the leader in other cases. This is partly because LFWA+ is the most biased dataset and ours is the most diverse, and thus more generalizable dataset.

4.3 Generalization Performance

4.3.1 Datasets

To test the generalization performance of the models, we consider three novel datasets. Note that these datasets were collected from completely different sources than our data from Flickr and not used in training. Since we want to measure the effectiveness of the model on diverse races, we chose the test datasets that contain people in different locations as follows.

Geo-tagged Tweets. First we consider images uploaded by Twitter users whose locations are identified by geo-tags (longitude and latitude), provided by (Steinert-Threlkeld, 2018). From this set, we chose four countries (France, Iraq, Philippines, and Venezuela) and randomly sampled 5,000 faces.

Media Photographs. Next, we also use photographs posted by 500 online professional media outlets. Specifically, we use a public dataset of tweet IDs (Littman et al., 2017) posted by 4,000 known media accounts, e.g. @nytimes. Note that although we use Twitter to access the photographs, these tweets are simply external links to pages in the main newspaper sites. Therefore this data is considered as media photographs and different from general tweet images mostly uploaded by ordinary users. We randomly sampled 8,000 faces from the set.

Protest Dataset. Lastly, we also use a public image dataset collected for a recent protest activity study (Won et al., 2017). The authors collected the majority of data from Google Image search by using keywords such as "Venezuela protest" or "football game" (for hard negatives). The dataset exhibits a wide range of diverse race and gender groups engaging in different activities in various countries. We randomly sampled 8,000 faces from the set.

These faces were annotated for gender, race, and age by Amazon Mechanical Turk workers.

Race	White		Black		East Asian		SE Asian		Latino		Indian		Middle Eastern						
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F	Max	Min	AVG	STDV	ϵ
FairFace	.967	.954	.958	.917	.873	.939	.909	.906	.977	.960	.966	.947	.991	.946	**.991**	**.873**	**.944**	**.032**	**.055**
UTK	.926	.864	.909	.795	.841	.824	.906	.795	.939	.821	.978	.742	.949	.730	.978	.730	.859	.078	.127
LFWA+	.946	.680	.974	.432	.826	.684	.938	.574	.951	.613	.968	.518	.988	.635	.988	.432	.766	.196	.359
CelebA	.829	.958	.819	.919	.653	.939	.768	.923	.843	.955	.866	.856	.924	.874	.958	.653	.866	.083	.166

Table 4: Gender classification accuracy measured on external validation datasets across gender-race groups.

4.3.2 Result

Table 7 shows the classification accuracy of different models. Because our dataset is larger than LFWA+ and UTKFace, we report the three variants of the FairFace model by limiting the size of a training set (9k, 18k, and Full) for fair comparisons.

Improved Accuracy. As clearly shown in the result, the model trained by FairFace outperforms all the other models for race, gender, and age, on the novel datasets, which have never been used in training and also come from different data sources. The models trained with fewer training images (9k and 18k) still outperform other datasets including CelebA which is larger than FairFace. This suggests that the dataset size is not the only reason for the performance improvement.

Balanced Accuracy. Our model also produces more consistent results – for race, gender, age classification – across different race groups compared to other datasets. We measure the model consistency by standard deviations of classification accuracy measured on different sub-populations, as shown in Table 5. More formally, one can consider conditional use accuracy equality (Berk et al.) or equalized odds (Hardt et al., 2016) as the measure of fair classification. For gender classification:

$$\begin{aligned} P(\widehat{Y} = i|Y = i, A = j) = P(\widehat{Y} = i|Y = i, A = k), \\ i \in \{\text{male, female}\}, \forall j, k \in \mathrm{D}, \end{aligned} \tag{1}$$

where $\widehat{Y}$ is the predicted gender, Y is the true gender, A refers to the demographic group, and D is the set of different demographic groups being considered (race). When we consider different gender groups for A, this needs to be modified to measure accuracy equality Berk et al.:

$$P(\widehat{Y} = Y|A = j) = P(\widehat{Y} = Y|A = k), \forall j, k \in \mathrm{D}. \tag{2}$$

We therefore define the maximum accuracy disparity of a classifier as follows:

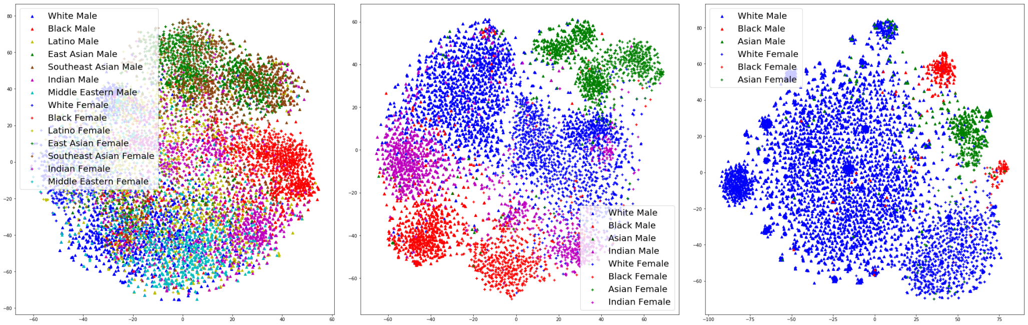
$$\epsilon(\widehat{Y}) = \max_{\forall j,k \in \mathrm{D}} \left(\log \frac{P(\widehat{Y} = Y|A = j)}{P(\widehat{Y} = Y|A = k)} \right). \tag{3}$$

Table 4 shows the gender classification accuracy of different models measured on the external validation datasets for each race and gender group. The FairFace model achieves the lowest maximum accuracy disparity. The LFWA+ model yields the highest disparity, strongly biased toward the male category. The CelebA model tends to exhibit a bias toward the female category as the dataset contains more female images than male.

The FairFace model achieves less than 1% accuracy discrepancy between male $\leftrightarrow$ female and White $\leftrightarrow$ non-White for gender classification (Table 7). All the other models show a strong bias toward the male class, yielding much lower accuracy on the female group, and perform more inaccurately on the non-White group. The gender performance gap was the biggest in LFWA+ (32%), which is the smallest among the datasets used in the experiment. Recent work has also reported asymmetric gender biases in commercial computer vision services (Buolamwini & Gebru, 2018), and our result further suggests the cause is likely due to the unbalanced representation in training data.

Data Coverage and Diversity. We further investigate dataset characteristics to measure the data diversity in our dataset. We first visualize randomly sampled faces in 2D space using t-SNE (Maaten & Hinton, 2008) as shown in Figure 3. We used the facial embedding based on ResNet-34 from dlib, which was trained from the FaceScrub dataset (Ng & Winkler, 2014), the VGG-Face dataset (Parkhi et al., 2015) and other online sources, which are likely dominated by the White faces. The faces in FairFace are well spread in the space, and the race groups are loosely separated from each other.

This is in part because the embedding was trained from biased datasets, but it also suggests that the dataset contains many non-typical examples. LFWA+ was derived from LFW, which was developed for face recognition, and therefore contains multiple images of the same individuals, i.e. clusters. UTKFace also tends to focus more on local clusters compared to FairFace.



(a) FairFace

(b) UTKFace

(c) LFWA+

Figure 3: t-SNE visualizations (Maaten & Hinton, 2008) of faces in datasets.

To explicitly measure the diversity of faces in these datasets, we examine the distributions of pair-wise distance between faces (Figure 4). On the random subsets, we first obtained the same 128-dimensional facial embedding from dlib and measured pair-wise distance. Figure 4 shows the CDF functions for 3 datasets. As conjectured, UTKFace had more faces that are tightly clustered together and very similar to each other, compared to our dataset. Surprisingly, the faces in LFWA+ were shown very diverse and far from each other, even though the majority of the examples contained a white face. We believe this is mostly due to the fact that the face embedding was also trained on a very similar white-oriented dataset which will be effective in separating white faces, not because the appearance of their faces is actually diverse. (See Figure 2)

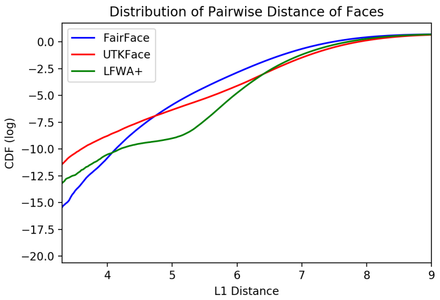


Figure 4: Distribution of pairwise distances of faces in 3 datasets measured by L1 distance on face embedding.

Table 5: Gender classification accuracy on external validation datasets, across race and age groups.

		Mean across races	SD across races	Mean across ages	SD across ages
Model trained on	FairFace	**94.89%**	**3.03%**	**92.95%**	**6.63%**
	UTKFace	89.54%	3.34%	84.23%	12.83%
	LFWA+	82.46%	5.60%	78.50%	11.51%
	CelebA	86.03%	4.57%	79.53%	17.96%

4.4 Evaluating Commercial Face Gender Classifiers

Previous studies have reported that popular commercial face analytic models show inconsistent classification accuracies across different demographic groups (Buolamwini & Gebru, 2018; Raji & Buolamwini, 2019). We used the FairFace images to test several online APIs for gender classification: Microsoft Face API, Amazon Rekognition, IBM Watson Visual Recognition, and Face++. Compared to prior work using politicians' faces, our dataset is much more diverse in terms of race, age, expressions, head orientation, and photographic conditions, and thus serves as a much better benchmark for bias measurement. We used 7,476 random samples from FairFace such that it contains an equal number of faces from each race, gender, and age group. We left out children under the age of 20, as these pictures were often ambiguous and the gender could not be determined for certain. The experiments were conducted on August 13th - 16th, 2019.

Table 6: Classification accuracy of commercial services on FairFace dataset. (*Microsoft, *Face++, *IBM indicate accuracies only on the detected faces, ignoring mis-detections.)

	White		Black		East Asian		SE Asian		Latino		Indian		Mid-Eastern			
	F	M	F	M	F	M	F	M	F	M	F	M	F	M	**Mean**	**STD**
Amazon	.923	.966	.901	.955	.925	.949	.918	.914	.921	.987	.951	.979	.906	.983	.941	.030
Microsoft	.822	.777	.766	.717	.824	.775	.852	.794	.843	.848	.863	.790	.839	.772	.806	.042
Face++	.888	.959	.805	.944	.876	.904	.884	.897	.865	.981	.770	.968	.822	.978	.896	.066
IBM	.910	.966	.758	.927	.899	.910	.852	.919	.884	.972	.811	.957	.871	.959	.900	.061
FairFace	.987	.991	.964	.974	.966	.979	.978	.961	.991	.989	.991	.987	.972	.991	.980	.011
*Microsoft	.973	.998	.962	.967	.963	.976	.960	.957	.983	.993	.975	.991	.966	.993	.975	.014
*Face++	.893	.968	.810	.956	.878	.911	.886	.899	.870	.983	.773	.975	.827	.983	.901	.067
*IBM	.914	.981	.761	.956	.909	.920	.852	.926	.892	.977	.819	.975	.881	.979	.910	.066

Table 6 shows the gender classification accuracies of the tested APIs. These APIs first detect a face from an input image and classify its gender. Not all 7,476 faces were detected by these APIs with the exception of Amazon Rekognition which detected all of them. Table 8 in Appendix reports the detection rate.[1] We report two sets of accuracies: 1) treating mis-detections as mis-classifications and 2) excluding mis-detections. For comparison, we included a model trained with our dataset to provide an upper bound for classification accuracy. Following prior work (Merler et al., 2019), we also show the classification accuracy as a function of skin color in Figure 6.

The results suggest several findings. First, all tested gender classifiers still favor the **male** category, which is consistent with the previous report (Buolamwini & Gebru, 2018). Second, **dark-skinned females** tend to yield higher classification error rates, but there exist many exceptions. For example, Indians have darker skin tones (Figure 5), but some APIs (Amazon and MS) classified them more accurately than Whites. This suggests skin color alone, or any other individual phenotypic feature, is not a sufficient guideline to study model bias. Third, face detection can also introduce significant gender bias. Microsoft's model failed to detect many **male** faces, an opposite direction from the gender classification bias. This was not reported in previous studies which only used clean profile images of frontal faces.

5 Conclusion

This paper proposes a novel face image dataset balanced on race, gender and age. Compared to existing large-scale in-the-wild datasets, our dataset achieves much better generalization classification performance for gender, race, and age on novel image datasets collected from Twitter, international online newspapers, and web search, which contain more non-White faces than typical face datasets. We show that the model trained from our dataset produces balanced accuracy across race, whereas other datasets often lead to asymmetric accuracy on different race groups.

This dataset was derived from the Yahoo YFCC100m dataset (Thomee et al.) for the images with Creative Common Licenses by Attribution and Share Alike, which permit both academic and commercial usage. Our dataset can be used for training a new model and verifying balanced accuracy of existing classifiers.

Algorithmic fairness is an important aspect to consider in designing and developing AI systems, especially because these systems are being translated into many areas in our society and affecting our decision making. Large scale image datasets have contributed to the recent success in computer vision by improving model accuracy; yet the public and media have doubts about its transparency. The novel dataset proposed in this paper will help us discover and mitigate race and gender bias present in computer vision systems such that such systems can be more easily accepted in society.

[1] These detection rates should not be interpreted as general face detection performance because we did not measure false detection rates using non-face images.

REFERENCES

Mohsan Alvi, Andrew Zisserman, and Christoffer Nellaaker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018.

Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6, 2016.

Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 2089–2093. IEEE, 2017.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66. IEEE, 2018.

Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2745–2754, 2017.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, pp. 0049124118782533.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

Abhijnan Chakraborty, Johnnatan Messias, Fabricio Benevenuto, Saptarshi Ghosh, Niloy Ganguly, and Krishna P Gummadi. Who makes trends? understanding demographic biases in crowdsourced recommendations. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia*, 17(6):804–815, 2015.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. ACM, 2017.

Abhijit Das, Antitza Dantcheva, and Francois Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018.

Sergio Escalera, Mercedes Torres Torres, Brais Martinez, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Georgios Tzimiropoulos, Ciprian Corneou, Marc Oliu, Mohammad Ali Bagheri, et al. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–8, 2016.

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*, pp. 87–102. Springer, 2016.

Hu Han, Anil K Jain, Fang Wang, Shiguang Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 40(11): 2597–2609, 2018.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Arbitrary facial attribute editing: Only change what you want. *arXiv preprint arXiv:1711.10678*, 1(3), 2017.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pp. 793–811. Springer, 2018.

Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 951–959, 2017.

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 07-49, University of Massachusetts, Amherst, October 2007.

Jungseock Joo, Shuo Wang, and Song-Chun Zhu. Human attribute recognition by rich appearance dictionary. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 721–728, 2013.

Jungseock Joo, Francis F Steen, and Song-Chun Zhu. Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE international conference on computer vision*, pp. 3712–3720, 2015.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018.

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*, pp. 4873–4882, 2016.

Davis E King. Max-margin object detection. *arXiv preprint arXiv:1502.00046*, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Svetlana Kiritchenko and Saif M Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.

Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10): 1962–1977, 2011.

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pp. 5967–5976, 2017.

Ryan Layne, Timothy M Hospedales, Shaogang Gong, and Q Mary. Person re-identification by attributes. In *Bmvc*, volume 2, pp. 8, 2012.

Annan Li, Luoqi Liu, Kang Wang, Si Liu, and Shuicheng Yan. Clothing attributes assisted person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(5):869–878, 2015a.

Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5325–5334, 2015b.

Justin Littman, Laura Wrubel, Daniel Kerchner, and Yonah Bromberg Gaber. News Outlet Tweet Ids, 2017. URL `https://doi.org/10.7910/DVN/2FIFLH`.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Daniel McDuff, Shuang Ma, Yale Song, and Ashish Kapoor. Characterizing bias in classifiers using generative models. *arXiv preprint arXiv:1906.11891*, 2019.

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

Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 343–347. IEEE, 2014.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *bmvc*, volume 1, pp. 6, 2015.

Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AAAI/ACM Conf. on AI Ethics and Society*, volume 1, 2019.

Julio Reis, Haewoon Kwak, Jisun An, Johnnatan Messias, and Fabricio Benevenuto. Demographics of news sharing in the us twittersphere. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pp. 195–204. ACM, 2017.

Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1685–1692, 2014.

Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pp. 341–345. IEEE, 2006.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)*, July 2016.

Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. *arXiv preprint arXiv:1712.00193*, 2017.

Richard T Schaefer. *Encyclopedia of race, ethnicity, and society*, volume 1. Sage, 2008.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

Zachary C Steinert-Threlkeld. *Twitter as data*. Cambridge University Press, 2018.

Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 498–512, 2018.

Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry Steven Davis, and Wen Gao. Multi-task learning with low rank attribute embedding for multi-camera person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1167–1181, 2018.

Harini Suresh, Jen J Gong, and John V Guttag. Learning tasks for multitask learning: Heterogenous patient populations in the icu. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 802–810. ACM, 2018.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708, 2014.

Christopher Thomas and Adriana Kovashka. Persuasive faces: Generating faces in advertisements. *arXiv preprint arXiv:1807.09882*, 2018.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2): 64–73.

A Torralba and AA Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1521–1528. IEEE Computer Society, 2011.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pp. 1–7. IEEE, 2018.

Yu Wang, Yang Feng, Zhe Hong, Ryan Berger, and Jiebo Luo. How polarized have we become? a multimodal classification of trump followers and clinton followers. In *International Conference on Social Informatics*, pp. 440–456. Springer, 2017.

Marcus Wilkes, Caradee Y Wright, Johan L du Plessis, and Anthony Reeder. Fitzpatrick skin type, individual typology angle, and melanin index in an african population: steps toward universally applicable skin photosensitivity assessments. *JAMA dermatology*, 151(8):902–903, 2015.

Donghyeon Won, Zachary C Steinert-Threlkeld, and Jungseock Joo. Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 786–794. ACM, 2017.

Nan Xi, Di Ma, Marcus Liou, Zachary C Steinert-Threlkeld, Jason Anastasopoulos, and Jungseock Joo. Understanding the political ideology of legislators from social media images. *arXiv preprint arXiv:1907.09594*, 2019.

Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 532–539, 2013.

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pp. 776–791. Springer, 2016.

Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5525–5533, 2016.

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

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340. ACM, 2018.

Maggie Zhang. Google photos tags two african-americans as gorillas through facial recognition software, Jul 2015. URL `https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/#55d05821713d`.

Zhanpeng Zhang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Learning social relation traits from face images. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3631–3639, 2015.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5810–5818, 2017.

James Zou and Londa Schiebinger. Ai can be sexist and racistits time to make it fair, 2018.

A Appendix

Table 7: Classification accuracy on external validation datasets.

		Race Classification															
		All	Female	Male	White	Non-White	Black	Asian	E Asian	SE Asian	Latino	Indian	Mid-East	0-9	10-29	30-49	50+
Twitter	FairFace	**.733**	**.726**	**.737**	.899	**.548**	**.695**	**.888**	.705	.465	.305	**.492**	.743	**.756**	**.691**	**.768**	**.777**
	UTKFace	.544	.543	.544	.741	.354	.591	.476	-	-	-	.474	-	.606	.516	.574	.567
	LFWA+	.626	.596	.647	**.965**	.284	.283	.425	-	-	-	-	-	.639	.562	.705	.751
Media	FairFace	**.866**	**.874**	**.863**	.949	**.685**	**.890**	**.918**	.886	.152	.267	**.691**	.704	**.833**	**.853**	**.852**	**.893**
	UTKFace	.772	.795	.763	.883	.546	.802	.588	-	-	-	.599	-	.646	.755	.757	.804
	LFWA+	.679	.823	.835	**.978**	.393	.485	.578	-	-	-	-	-	.682	.656	.651	.722
Protest	FairFace	**.846**	**.849**	**.844**	.935	**.683**	**.859**	**.843**	.702	.510	.169	**.649**	.779	**.839**	**.821**	**.837**	**.881**
	UTKFace	.706	.723	.697	.821	.536	.714	.456	-	-	-	.591	-	.681	.658	.685	.787
	LFWA+	.747	.759	.741	**.964**	.366	.418	.488	-	-	-	-	-	.689	.645	.668	.801
Average	FairFace	**.815**	**.816**	**.815**	.928	**.639**	**.815**	**.883**	.764	.376	.247	**.611**	.742	**.809**	**.788**	**.819**	**.850**
	FairFace 18K	.800	.812	.795	.917	.588	.779	.856	.685	.355	.279	.502	.625	.786	.773	.809	.827
	FairFace 9K	.774	.788	.768	.885	.564	.756	.827	.641	.315	.281	.531	.544	.723	.757	.789	.787
	UTKFace	.674	.687	.668	.815	.479	.702	.507	-	-	-	.555	-	.644	.643	.672	.719
	LFWA+	.684	.726	.741	**.969**	.348	.395	.497	-	-	-	-	-	.670	.621	.675	.758
		Gender Classification															
		All	Female	Male	White	Non-White	Black	Asian	E Asian	SE Asian	Latino	Indian	Mid-East	0-9	10-29	30-49	50+
Twitter	FairFace	**.940**	.948	**.935**	**.949**	**.932**	**.932**	**.894**	**.864**	**.942**	**.963**	**.932**	**.976**	**.817**	**.932**	**.973**	**.959**
	UTKFace	.884	.859	.899	.897	.874	.864	.829	.803	.871	.901	.898	.947	.671	.874	.933	.912
	LFWA+	.797	.637	.899	.815	.773	.789	.724	.716	.736	.804	.728	.911	.634	.769	.857	.859
	CelebA	.829	**.955**	.750	.850	.812	.818	.764	.716	.839	.843	.831	.876	.539	.818	.889	.881
Media	FairFace	**.973**	**.957**	**.980**	**.976**	**.969**	**.953**	**.956**	**.967**	**.891**	**.980**	**.977**	**.988**	**.821**	**.952**	**.984**	**.979**
	UTKFace	.927	.841	.961	.928	.915	.907	.908	.915	.869	.928	.945	.932	.679	.917	.931	.924
	LFWA+	.887	.656	.976	.893	.871	.851	.864	.875	.804	.859	.897	.944	.688	.835	.832	.911
	CelebA	.899	.950	.880	.909	.881	.847	.858	.857	.870	.925	.884	.926	.560	.860	.908	.924
Protest	FairFace	**.957**	**.944**	**.963**	**.962**	**.951**	**.957**	.887	**.879**	**.906**	**.970**	**.973**	**.991**	**.861**	**.934**	**.967**	**.976**
	UTKFace	.901	.829	.934	.905	.873	.911	.814	.802	.843	.902	.918	.921	.611	.812	.924	.919
	LFWA+	.829	.567	.954	.841	.801	.821	.758	.782	.697	.811	.811	.929	.568	.705	.851	.908
	CelebA	.882	.935	.856	.893	.866	.876	**.892**	.750	.833	.892	.878	.956	.492	.842	.904	.927
Average	FairFace	**.957**	**.950**	**.959**	**.962**	**.951**	**.947**	**.912**	**.903**	**.913**	**.971**	**.961**	**.985**	**.833**	**.939**	**.975**	**.971**
	FairFace 18K	.941	.930	.946	.946	.934	.931	.891	.886	.895	.955	.960	.967	.803	.920	.957	.962
	FairFace 9K	.926	.921	.927	.929	.921	.922	.864	.851	.883	.942	.951	.974	.760	.901	.949	.943
	UTKFace	.904	.843	.931	.910	.887	.894	.850	.840	.861	.910	.920	.933	.654	.868	.929	.918
	LFWA+	.838	.620	.943	.850	.815	.820	.782	.791	.746	.825	.812	.928	.630	.770	.847	.893
	CelebA	.870	.947	.829	.884	.853	.847	.838	.774	.847	.887	.864	.919	.530	.840	.900	.911
		Age Classification															
		All	Female	Male	White	Non-White	Black	Asian	E Asian	SE Asian	Latino	Indian	Mid-East	0-9	10-29	30-49	50+
Twitter	FairFace	**.578**	**.586**	**.573**	**.563**	**.590**	**.557**	**.620**	**.629**	**.606**	**.581**	**.576**	**.555**	**.805**	**.666**	**.439**	**.408**
	UTKFace	.366	.355	.384	.343	.385	.338	.397	.382	.419	.411	.356	.345	.585	.499	.104	.307
Media	FairFace	**.516**	**.511**	**.517**	**.513**	**.520**	**.483**	**.557**	**.559**	**.543**	**.537**	**.532**	**.475**	**.714**	**.686**	**.447**	**.501**
	UTKFace	.275	.273	.282	.281	.267	.271	.276	.279	.261	.231	.292	.222	.511	.529	.112	.238
Protest	FairFace	**.515**	**.543**	**.502**	**.498**	**.539**	**.527**	**.584**	**.605**	**.531**	**.507**	**.581**	**.469**	**.885**	**.687**	**.395**	**.478**
	UTKFace	.302	.306	.294	.291	.319	.305	.316	.318	.312	.314	.371	.318	.516	.503	.114	.349
Average	FairFace	**.536**	**.547**	**.531**	**.525**	**.550**	**.522**	**.587**	**.598**	**.560**	**.542**	**.563**	**.500**	**.801**	**.680**	**.427**	**.462**
	FairFace 18K	.492	.508	.484	.485	.496	.463	.528	.538	.506	.510	.454	.490	.700	.646	.387	.410
	FairFace 9K	.470	.493	.459	.462	.478	.449	.506	.515	.483	.473	.458	.463	.662	.611	.361	.394
	UTKFace	.314	.311	.320	.305	.324	.305	.330	.326	.331	.319	.340	.295	.537	.510	.110	.298

Table 8: Face detection rate of commercial services on FairFace dataset.

	White		Black		East Asian		Southeast Asian		Latino Hispanic		Indian		Middle Eastern			
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	**Mean**	**STD**
Amazon	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.000
Microsoft	.845	.779	.796	.742	.856	.794	.888	.830	.858	.854	.886	.798	.869	.777	.812	.047
Face++	.994	.991	.994	.987	.998	.993	.998	.998	.994	.998	.996	.993	.994	.994	.993	.003
IBM	.996	.985	.996	.970	.989	.989	1.000	.993	.991	.994	.991	.981	.989	.979	.991	.008

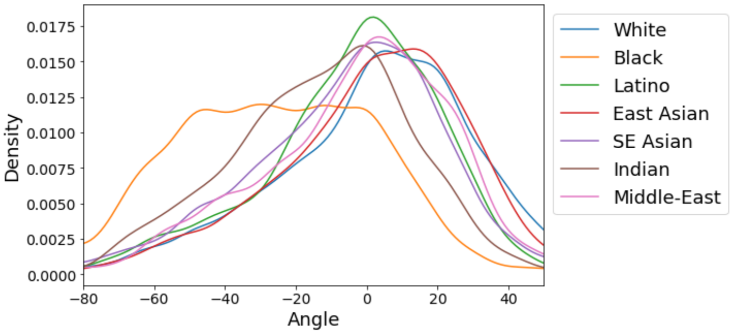


Figure 5: Individual Typology Angle (ITA), i.e. skin color, distribution of different races measured in our dataset.

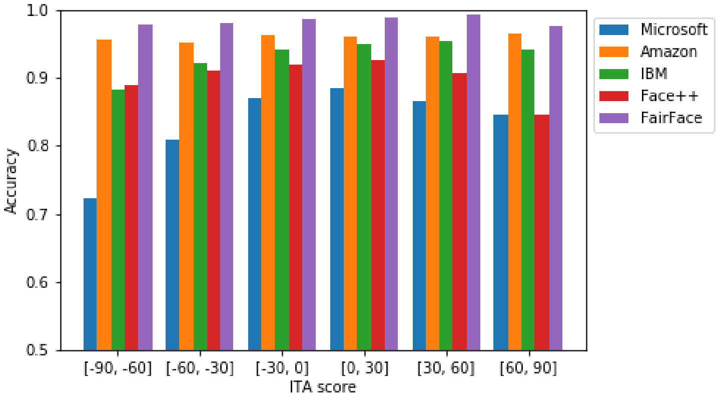


Figure 6: Classification accuracy based on Individual Typology Angle (ITA), i.e. skin color.