This ICCV workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Rethinking Common Assumptions to Mitigate Racial Bias in Face Recognition Datasets

Matthew Gwilliam[1] Srinidhi Hegde[1] Lade Tinubu[1,2] Alex Hanson[1]
[1]University of Maryland [2]University of Chicago
{mgwillia, srihegde}@umd.edu lade@uchicago.edu hanson@cs.umd.edu

Abstract

Many existing works have made great strides towards reducing racial bias in face recognition. However, most of these methods attempt to rectify bias that manifests in models during training instead of directly addressing a major source of the bias, the dataset itself. Exceptions to this are BUPT-Balancedface/RFW [34] and Fairface [14], but these works assume that primarily training on a single race or not racially balancing the dataset are inherently disadvantageous. We demonstrate that these assumptions are not necessarily valid. In our experiments, training on only African faces induced less bias than training on a balanced distribution of faces and distributions skewed to include more African faces produced more equitable models. We additionally notice that adding more images of existing identities to a dataset in place of adding new identities can lead to accuracy boosts across racial categories. Our code is available at `https://github.com/j-alex-hanson/rethinking-race-face-datasets`.

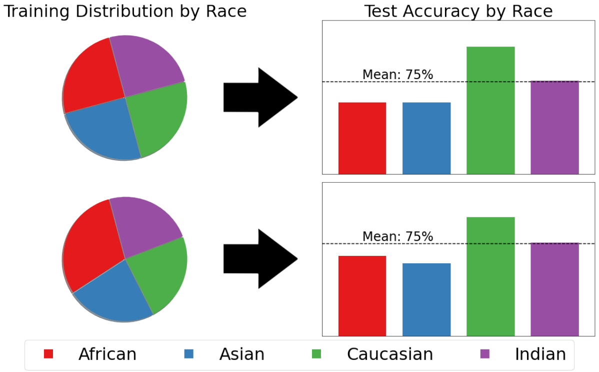

Figure 1: **Balanced vs skewed training dataset.** A training distribution balancing race (all races: 25%; top) produces a larger variance on the test accuracies of different races than a skewed distribution (African: 30%, other races: 23.3%; bottom). The overall accuracy remains at 75%, demonstrating that the skewed training distribution is the preferable choice. Data for this figure is from Section 4.2. See Supplementary Material for additional distribution variations.

1. Introduction

Since before the advent of deep learning, face recognition has been one of the most popular human-centric applications of computer vision. The introduction of deep learning has only accelerated progress in this area. The effectiveness of face recognition algorithms makes them a compelling candidate for real-world, industry-level applications. In fact, face recognition is currently used to unlock phones, validate identities at ATMs, verify drivers licenses, and aid in forensic investigations.

We consider two key attributes of face recognition applications: performance, which is often measured in the literature in terms of accuracy or a similar metric, and fairness, which is sometimes not measured at all. The computer vision community has a responsibility to ensure it delivers research, algorithms, and solutions which are not only highly performant, but also very fair. Therefore, given that existing face recognition applications have high accuracy, we ought to ensure that they are also fair. One key fairness issue is equitable performance across race. Inconsistent performance across race may lead to further disparagement of certain groups, motivating the need for research in this space.

Numerous recent works have attempted to address this concern and show progress towards equability, but this is usually done by modifying the model architecture or data sampling strategy to explicitly induce fairness. It is understood that these modifications are necessary because the data used for training these models will inherently cause the models to be biased without some intervention. What is it about this data that is causing the models to become biased? Two datasets specially created to address racial bias, BUPT-Balancedface (BUPT) / Racial Faces in the Wild (RFW) [34] and Fairface [14], posit that the reason bias is introduced is because racial categories are not evenly represented in the data. This stems from the observation that most commonly used face datasets are primarily comprised of Caucasian individuals.

But does training on a single race necessarily lead to a biased model? And is balancing the dataset across race an optimal choice for mitigating racial bias? Figure 1 offers evidence that, contrary to popular assumption, balanced datasets do not always lead to the most balanced results. The figure shows that for the BUPT dataset – which groups images into race categories African, Asian, Caucasian, and Indian – results can be better balanced in terms of accuracy variance, without affecting the mean accuracy, by allocating slightly more images to African that to other race categories. This example is merely the tip of the iceberg; more thorough analysis follows in the rest of this paper.

There are many more assumptions that are necessary for achieving fairness in face recognition datasets that we do not address in this work. We do not address the issue of the actual composition and labelling of these datasets. BUPT, which we use in this work, has 4 racial categories and every image is placed into one of those 4 categories, sometimes in spite of the represented individual's actual or self-determined race/ethnicity. Another issue that we do not cover is ensuring that the dataset is capturing a true distribution of faces. Face datasets are typically sourced from online images, and images which appear on the internet are in some sense curated – they were uploaded because the subject or a third party wanted the content to exist online.

Instead, in this work, we analyze assumptions about racial bias in face recognition by analyzing results from the BUPT training dataset and RFW testing dataset. Our contributions are the following:

- We show that, when training state-of-the-art models on a single race from BUPT, **some races generalize across all races differently** than others (Section 4.1).
- We sample subset training datasets of BUPT from many possible race distributions, and demonstrate that some **racially skewed datasets mitigate racial bias better than racially balanced datasets**; sometimes by a wide margin (Section 4.2).
- We observe that **adding more images of existing subjects in face datasets over adding new subjects can lead to greater accuracy boosts** across racial categories (Section 4.3).

For completeness, we define a few terms used throughout our work here. We define *face recognition* as the mapping of an image of a face to an identity. The implementation of face recognition we perform in this work, *pair matching*, requires a trained model (BUPT) to identify if a pair of face images (RFW) belong to the same individual or different individuals; these individuals and their corresponding images are typically unavailable during training. We define and quantify *racial bias* in face recognition systems as the variance of test race accuracies. For our work, this is the variance in accuracy on the African, Asian, Caucasian, and Indian test splits of RFW.

2. Related Work

Face Recognition Face recognition models have benefited from the inclusion of deep learning techniques [28]. In a typical deep face recognition model, the backbone architecture, discriminative loss function, and a deep feature based face matching method form the three critical components of the system [33]. A single DNN backbone architecture is the most common choice for the extraction of face features [23, 25]. To exploit these features, face recognition models employ a variety of discriminative loss functions for training face recognition models such as contrastive loss [30], triplet loss [25], angular/cosine loss [7, 31], and softmax loss variants [18, 24]. Deep feature based matching methods vary with the requirements of the face recognition application. Face verification can require a more fine-grained approach [8] and face identification necessitates discriminative features [37]. In our work, we analyze recent generic face recognition models that are trained with several loss functions and have a single backbone network.

Bias in Computer Vision for Faces In recent years, many studies have confirmed the presence of bias in deep neural networks [4, 20] which may result in undesired consequences especially for face recognition [22, 27]. Several works focus on specifically mitigating bias, either by explicitly changing the models or incorporating data sampling strategies distinctly for this purpose. Regarding model modifications, [34] propound a model to balance representations of face data from other datasets, in addition to introducing a balance dataset. [10] use adaptive convolution kernels and attention mechanisms to mitigate bias in the model. [26] incorporate triplet loss to prevent discriminatory effects. And [21] propose privacy preserving and learning agnostic representation to mask sensitive information, including race and gender. Regarding sampling strategies, [32] uses reinforcement learning dataset sampling to mitigate bias. [3] introduces the use of sampling strategies to mitigate geographic performance differences on photo ID documents. [9] uses demographic classifiers and adversarial learning to make face representations more robust. [39] adapts Cycle-GAN to racially balance training per individual in the dataset. And [35] implements a balanced sampling strategy while training to mitigate bias, though the focus of their work is primarily on gender.

Datasets Face recognition is a problem that has garnered a sustained interest and as such many datasets exist for various subproblems in this space. A few popular datasets include: CASIA-WebFace [38], VGGFace2 [6], WebFace260M [41], MS-Celeb-1M [11], LFW [13], and IJB-C [19]. Naturally, works also exist that explore potential biases these datasets exhibit [36, 5]. Addressing racial bias, datasets such as RFW [34] and FairFace [14] point out that most of the popular face datasets primarily consist of Caucasian face images and propose a racially balanced face dataset. In addition to RFW and FairFace, CASIA-SURF CeFA [17] is another racially balanced face dataset that addresses racial bias in anti-spoofing. Data augmentation techniques used to mitigate bias are discussed by [29, 32, 39]. Related to our analysis are [15], which finds no evidence that darker skin tone causes higher false match rate on the pair matching problem, and [1], which identifies accuracy differences between genders persist after balancing the dataset. [40] specifically argues for racially balancing datasets to mitigate bias. Our work refutes this claim by demonstrating that some racially skewed training datasets can result in less racially biased models than racially balanced training datasets.

3. Experimental Setup

3.1. Models

We run our experiments on four recent face recognition models: VGGFace2, CenterLoss, SphereFace, and ArcFace. **VGGFace2** [6] is simply an implementation of a squeeze-and-excitation network (a modified ResNet architecture) [12], trained with a standard cross-entropy loss where the classes are the training dataset identities. The outputs of the penultimate layer are used as feature embeddings at test time for pair-matching (see below for more details). Because VGGFace2 is a standard network implementation, we view this as a baseline method. In addition to cross-entropy loss, **CenterLoss** [37] maintains an additional center vector per identity and imposes a squared L2 distance loss between the feature vectors of the training samples and their corresponding centers. The centers are initialized randomly and are incrementally updated by the mean features of each identity as training progresses. **SphereFace** [18] modifies the softmax function on the outputs of the network to impose a multiplicative margin loss on the angle between L2 normalized feature embeddings. **ArcFace** [7] extends SphereFace by including an additive margin loss on this angle between L2 normalized feature embeddings.

Additionally, for fair comparisons between methods we used a ResNet50 as the backbone of each model and trained for 50 epochs. All inputs are resized to 128 × 128.

3.2. Datasets

BUPT-Balancedface (BUPT) [34], also known as EqualizedFace in the literature, is the source of our training data. This dataset, in addition to offering subject/class labels for each image, also groups images into 1 of 4 possible race categories. Unlike standard datasets, the distribution of these races in BUPT is balanced, with each race having the same number of subjects, and roughly the same number of images. For testing, we do a pair-matching task on the pairs in Racial Faces in the Wild (RFW)[34], which is set up such

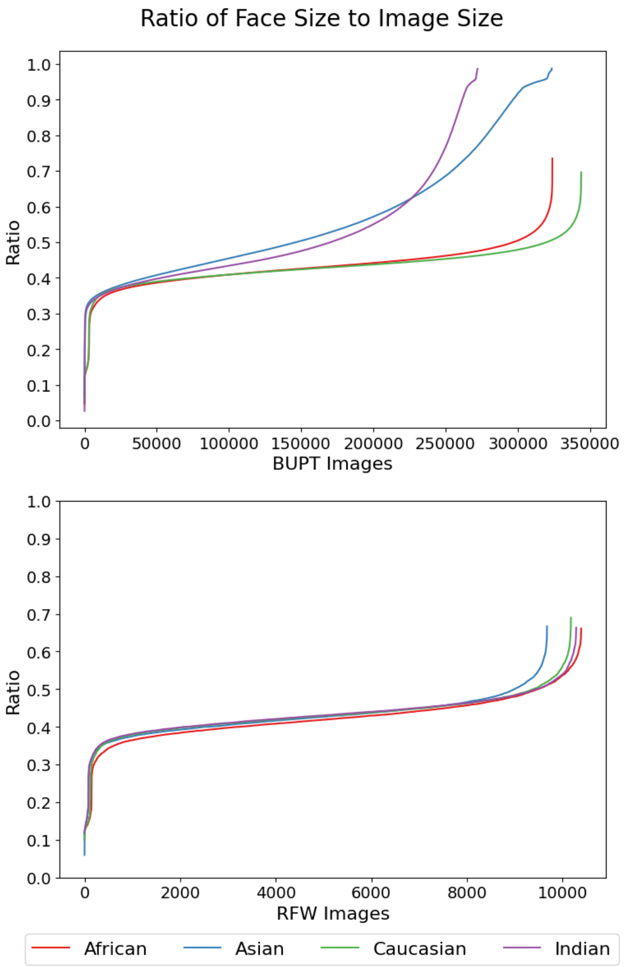

Figure 2: **Face size statistics** for training data (top) and test data (bottom). Both plots show the ratio of the face bounding box area to the area of the containing image, sorted for each race category according to the ratio.

Dataset		African	Asian	Caucasian	Indian
BUPT (train)	# sbjct	7000	7000	7000	7000
	# img	324376	325475	326484	275095
RFW (test)	# pair	6000	6000	6000	6000

Table 1: **Data Sources.** Relevant statistics pertaining to the composition of BUPT-Balancedface (BUPT), which is the source for our training datasets, and Racial Faces in the Wild (RFW), which is the source for our testing datasets.

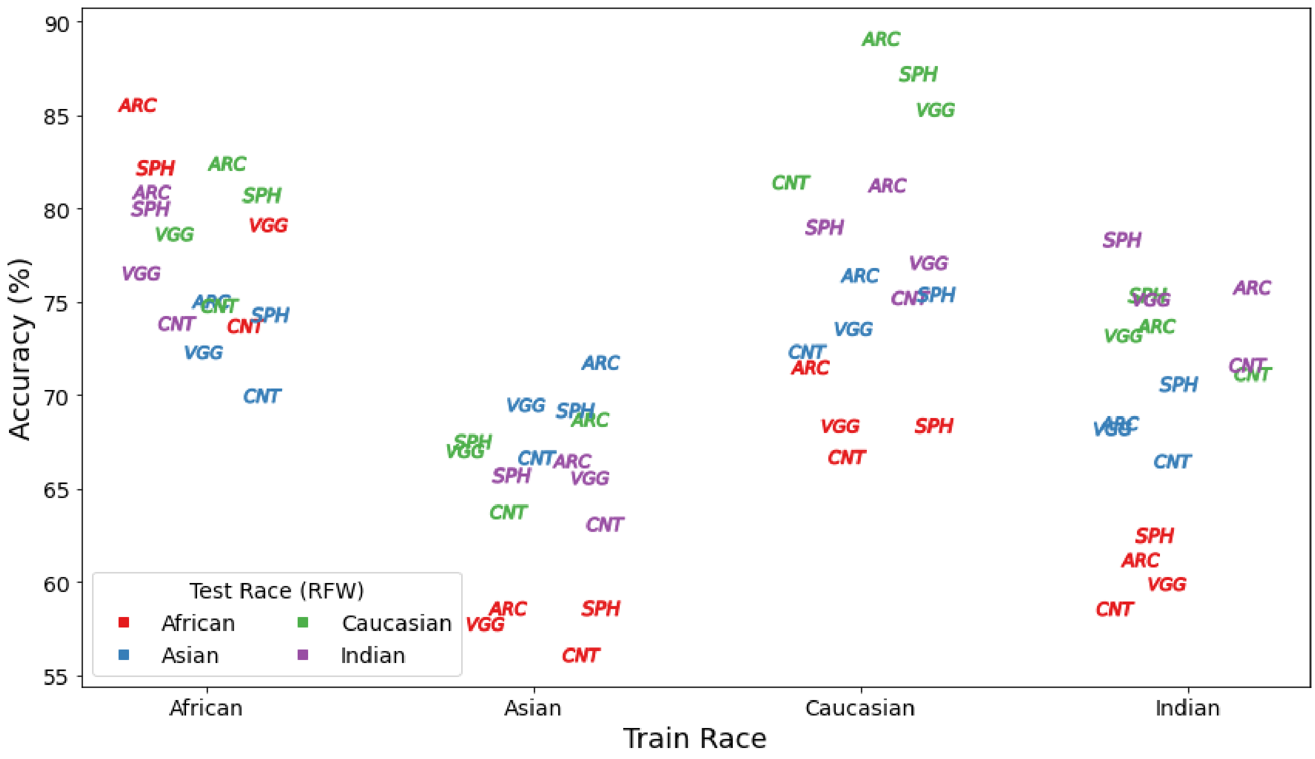

Figure 3: **Results for single race distributions.** Models from ArcFace (ARC), CenterLoss (CNT), SphereFace (SPH), and VGGFace2 (VGG) were trained on each race, and then tested on each race. Results were averaged across 5 trials.

that each race has the same number of corresponding pairs. Relevant statistics for both datasets can be found in Table 1.

We perform additional analysis regarding the contents of these two datasets to produce additional statistics. In Figure 2, we show the size distributions of the faces of both datasets, grouped by race. We calculated the ratios for each image by dividing the bounding box area of the face, calculated via OpenCV's pre-trained face detector [2], by the area of the containing image. Notice that while the African and Caucasian size statistics of BUPT are similar to those of all 4 races for RFW, the sizes of the Indian and Asian faces are distributed quite differently. Specifically, both races are represented by an abundance of larger faces (relative to the images) in the training data. This could contribute to discrepancies in accuracy and representational power that we observe in Section 4.3. Moreover, if the balance hypothesis was valid but perturbed by these inconsistent ratios, we would expect an optimally fair distribution to contain an even split between African and Caucasian. However, our experiments demonstrate this is not the case.

3.2.1 Single Race Setup

For the experiments in Section 4.1, we construct our training and testing subsets from the two datasets described above. Specifically, we take training data from BUPT-Balancedface and testing data from RFW. Rather than training on the entire BUPT dataset, for the Single Race experiments we train a given model only on a single race. We repeat this for each race, such that we have different models corresponding to the training data associated with each of the 4 race groups in BUPT. We evaluate each model on all 4 races in the testing data, separately, such that each model is associated with 4 different test accuracy results, 1 per race.

3.2.2 Race Distribution Setup

Prior work on these datasets deals primarily with the uniform (balanced) dataset which consists of equal amounts of persons representing each racial category. In Section 4.2, we expand our analysis from Section 4.1 to a representative collection of non-uniform distributions of data. Thus, we explore the space of imbalanced datasets.

In order to conduct meaningful exploration, we impose the following constraints:

- Each dataset must have the same number of images.
- Each dataset must have the same number of subjects.

To abide by these constraints, we only consider the 5000 persons corresponding to each race for whom the most data

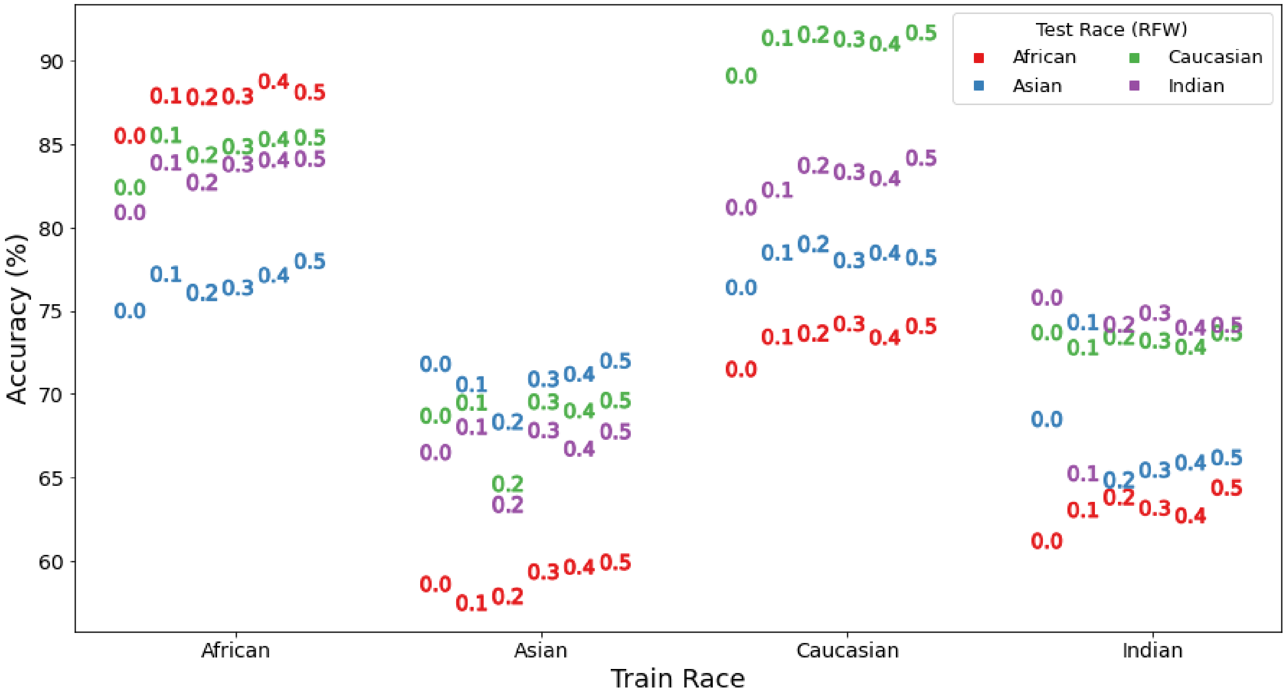

Figure 4: **Effect of noise injection** in training data on model performance. 0.1, 0.2, 0.3, 0.4, and 0.5 are the different noise percentage probabilities.

is available. Since the least-represented of these persons has 18 corresponding images, we choose to use 18 images to represent each person. Furthermore, since we only have 5000 persons available for each race, every dataset considered has exactly 5000 persons, selected from the different races as described below.

We describe each dataset as a tuple consisting of 4 values, (w_1, x_1, y_1, z_1), where w_1 gives the portion of the dataset comprised of African persons (as a percent), x_1 gives the portion of the dataset that represent Asian persons, and so forth. Since the portions are given as percents, the values of each tuple always sum to 100. From the set of all possible tuples, we consider the uniform distribution, $(25, 25, 25, 25)$ and 88 other points. To choose the 88 points, we consider 4 nested 3-simplexes. The corners of the outermost simplex are given by permutations of $(100, 0, 0, 0)$. The 3 inner simplexes have corners given by permutations of $(60, 13.3, 13.3, 13.3)$, $(40, 20, 20, 20)$, $(30, 23.3, 23.3, 23.3)$ such that we examine more distributions that lie closer to the center (closer to uniform). Note that the corners only account for 16 points (4 corners each for 4 simplexes). The remaining points are derived by taking 3 equidistant points from each edge on each simplex. With 6 edges per simplex, and 4 simplexes, this accounts for the remaining 72 points. We believe this method gives reasonable coverage of non-uniform race distributions.

3.3. Evaluation

3.3.1 Test Accuracy

We compute accuracy for the pair-matching task on RFW. This is a binary task, a pair is either a match or not a match. Test accuracy is thus the percent of all pairs for a race that are correctly identified as matches or non-matches.

3.3.2 Cluster Analysis

In order to better understand why different racial image distributions produce different results, we analyze the relationships between features. We consider the 4 races as clusters, and assess the image representations by computing metrics that approximate the tightness and spread of the clusters. We compute 2 metrics to accomplish this, *Intra-race Cosine Distance* and *Race Cluster Membership*.

Intra-race Cosine Distance The first form of “clustering analysis” we perform attempts to measure the compactness of the representations corresponding to each race. We represent images from RFW using the same features used in the pair matching task. We first compute the mean vector of all the features for the images of a given race. Then, we compute the cosine distances between the image features for a race and that race’s mean vector. We thus report the average of the cosine distances (cluster compactness) from the mean vector (cluster center).

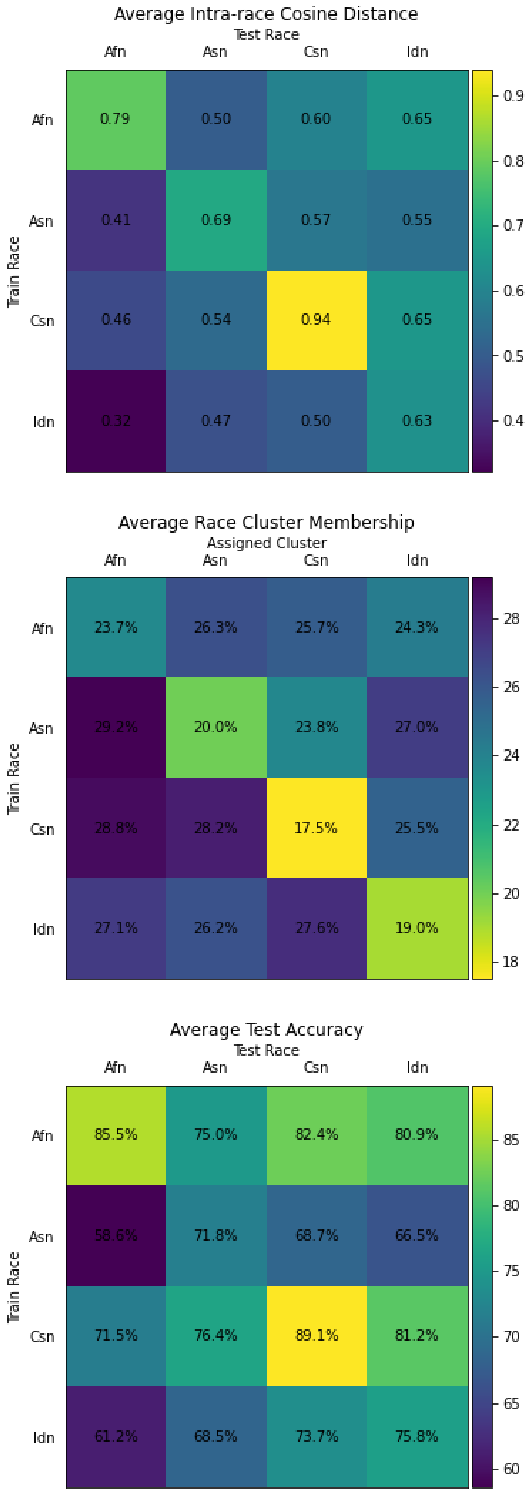

Figure 5: **Cluster vs. accuracy** for the single race experiments using Arcface.

Race Cluster Membership To complement the cosine distance metric, which is an absolute measure of cluster compactness (treating races as clusters), we perform clustering by taking the 20 nearest neighbors in feature space for each of 5000 randomly selected images from each race of RFW (20000 images in total). The neighbors then vote, with the vote weighted according to the inverse of the inner product distance between the image's feature vector and the neighbor's feature vector. Each images is then assigned the race that receives the maximum votes as the "cluster" label. We repeat this process for each trained model and report the average percent of the RFW images that are assigned to each race. The resulting cluster membership statistics serve as a relative measure of cluster compactness.

4. Experiments and Analysis

As Figure 1 establishes, a balanced dataset doesn't actually lend itself to the most balanced results. In this section, we explore how training set composition affects results. In Section 4.1, we begin by examining the accuracy results that models trained on a single race achieve when transferred to other races. Then, in Section 4.2, we compare models which are trained exclusively on a single race to models which are trained on an even distribution of data to every race, along with 84 distributions in between those extremes, as explained in Section 3.3.2. Finally, in Section 4.3, we compare the two different ways to increase the size of a dataset (overall number of images)- gathering more images from existing subjects, and acquiring new subjects.

4.1. Generalizing from Single Race

We train each model (ArcFace, CenterLoss, SphereFace, and VGGFace2) on each race of BUPT (African, Asian, Caucasian, Indian) for 5 trials. This gives 16 models per trial and 80 models in total. We test each model on each race of RFW. Results are averaged across trials.

4.1.1 Test Accuracy

Figure 3 gives average accuracy results. Note the major role played by the train race, as it dominates the range of possible accuracy scores. Also note that, for a given train race, results tend to be grouped not by method, but by train race. For a given train race and test race, the performance of the various models are relatively similar, with ArcFace tending to have the best results, followed by SphereFace, VGGFace2, and CenterLoss, respectively. Data is clearly the determining factor in terms of both mean accuracy (across 4 races) as well as fairness (variance in test race accuracies).

Robust Augmentation For understanding the effects of data augmentation on model performance we inject noise in the BUPT train images. To introduce noise in the image samples, we, first, localize the face using template matching method with Haar features [16]. Then we subdivide the localized face bounding box into a 4×4 grid and pick a square

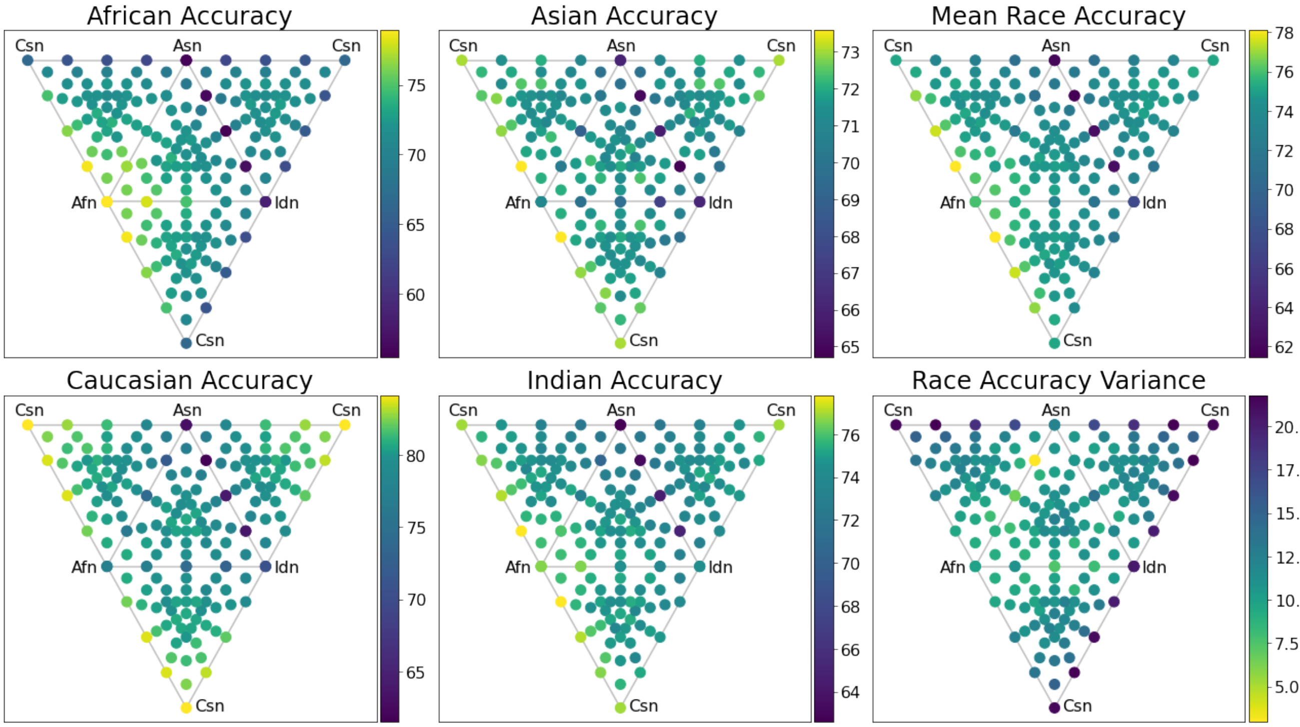

Figure 6: **Race distribution results** of ArcFace models on the 89 distributions identified in Section 3.2.2. Each plot gives percent accuracy values for a single test race. These plots roughly correspond to flattened versions of the 4 nested 3-simplexes; note that each plot consists of 4 connected equilateral triangles, where each triangle contains 4 parallel simplex faces projected into 2 dimensions, one from each simplex, as well as the center point. The corners of the outer simplex, which correspond to training distributions containing only data from a single race, and are labeled accordingly, i.e., the point labeled "Afn" is (100, 0, 0, 0). For readability, some points are therefore represented multiple times, such as the center (25, 25, 25, 25), which appears in the middle of all triangles. Thus, each plot contains 181 points, of which only 89 are unique. In addition to per-race accuracy, overall average accuracy of the races is represented, as well as the accuracy variance between the 4 per-race accuracy results. The variance plot thus represents how "balanced" the results are. For reference, the highest mean race accuracy is (75, 0, 25, 0) and the lowest race accuracy variance is (25, 75, 0, 0).

patch uniformly at random. Finally, we apply Gaussian blur on the selected square patch. For the Gaussian blur filter, we pick a Gaussian kernel whose size varies randomly between 11 and 21 pixels and with a variance of 1.5. Throughout this experiment we maintain the original number of training identities, i.e, 7000. Figure 4 represents the model performance under different noise injection probabilities across the different races which means that each image sample is assigned noise with the specified noise injection probability. We see that the introduction of noise indeed results in slight boost in accuracy for African and Caucasian races which have stronger representations compared to other races.

4.1.2 Cluster Analysis

We attempt to understand how test accuracy is related to the representations learned during training. As described in Section 3.3.2, we do this both by considering the races as clusters and obtaining a cosine distance measure for each, as well as clustering via a nearest neighbor voting method. Figure 5 shows the results of these methods, in addition to test accuracy, for ArcFace.

We observe reasonable correspondence between the values in the clustering matrices and the accuracy matrix. Caucasian, from the clustering plots, appears to have the representations which are the most spread out in embedding space, featuring the highest cosine distance as well as the lowest number of assigned image for its own cluster. African also is fairly spread out according to the cosine distance metric, although it is worth considering that it is somewhat unique in the race cluster membership analysis. Specifically, African-African is the highest value on the main diagonal, and African training seems to induce more balanced clusters than the other races. This suggests that while the African embeddings spread out in absolute terms (cosine distance), they also maintain a distinct cluster while also keeping other races clusters distinct. Overall this seems to empirically suggest that training on races that spread out more in feature space may lead to higher overall accuracy and that more distinct clusters in those spread out embeddings may yield better transfer performance.

4.2. Analyzing Race Distributions

It is clear that the distribution of subjects and images among the races affects per-race accuracy. Furthermore, it is apparent that training on a single race can achieve, in some cases, reasonable performance on other races. Nevertheless, neither training on a single race nor training on equal amounts of data from all races achieves totally balanced performance across all races. While, as discussed previously, some researchers attempt to address the remaining imbalances by adopting specific data sampling strategies during training or altering the model embeddings, we study how the dataset itself might be used to balance results. To more aptly investigate this, we carefully vary the distribution of training data among races, as explained in Section 3.2.2. In this way we gain an understanding, not only of the performance of the perfectly balanced dataset, but of a representative portion of the possible distributions of training data among the 4 race supercategories.

Figure 6 shows the results from the selected data distributions with ArcFace. Notice the importance of training race. For all races, distributions containing a mix of Caucasian and African images tend to be the highest performing. This matches some of the findings from Figure 3, where Asian accuracy is higher when trained on African or Caucasian images than when trained on Asian images. Further, the data shown here confirms that the equally balanced dataset, the central point in each plot, is not the best for either performance (mean accuracy) or fairness (variance). Instead, as stated before, datasets containing African and Caucasian images tend to give the highest accuracy scores, while various blends that contain African images are the most “fair” (lowest variance). Refer to the Supplementary Material for the race distribution analysis of VGGFace2.

4.3. Understanding Increasing Dataset Size

This section continues our exploration of the importance of data distribution by investigating how the addition of new data affects results. For these experiments, we establish a base dataset, consisting of 10 images per subject for 2500 subjects from each race. We examine a special case of data addition where data is added to only one race, and consider two ways this data can be added. The first is the introduction of new images for existing subjects; we add 5 images to each of the existing subjects for the selected race. The second way is adding new subjects; we introduce 1250 new subjects, and add 10 images to each. Both methods for adding data result in the same number of total images, enabling us to compare the two methods for adding to an existing dataset in terms of performance and fairness.

Figure 7 shows the results of this experiment. Note that, with the exception of a single train-test pair, accuracy improvements are much higher when images are added to the existing subjects. Furthermore, adding Caucasian and Indian images in this manner increases accuracy results in a far more balanced fashion than one might expect, with nearly identical improvement for African and Caucasian when adding Caucasian images, and very similar improvement for African, Caucasian, and Indian when adding Indian images. This reinforces the idea that balancing the distribution of training data can contribute to, but is neither necessary nor sufficient, for balanced results.

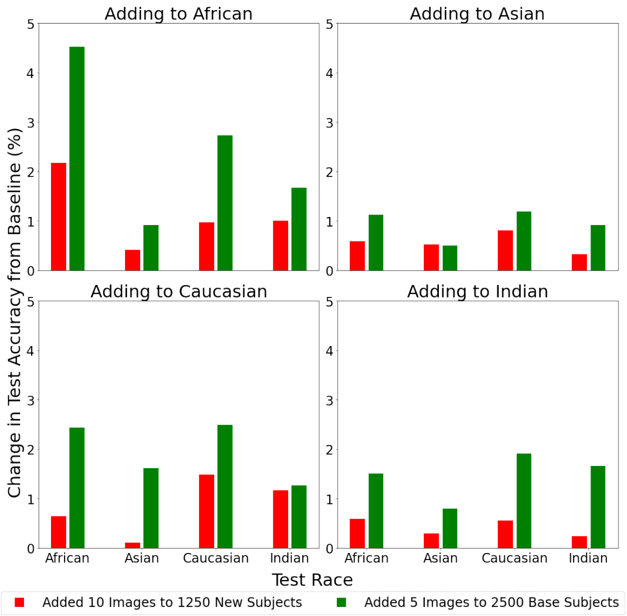

Figure 7: **Increasing dataset size** by adding new subjects vs adding data to existing subjects, for a given race.

5. Conclusion

We have shown that common assumptions to mitigate racial bias in datasets do not necessarily hold. The African BUPT data split produces more equitable models on RFW than a balanced BUPT data split. Training set distributions skewed to include more African faces also mitigate racial bias better than balanced training sets. Data augmentations appear to benefit more robust racial categories and adding more images to the base identities of a dataset can boost performance across race. We have demonstrated some ways to improve mitigation of racial bias on existing datasets, but we hope that illuminating these erroneous assumptions will ultimately assist the face recognition community in building more equitable systems.

Acknowledgments: Alex Hanson was supported by the NDSEG fellowship.

References

[1] Vitor Albiero, Krishnapriya KS, Kushal Vangara, Kai Zhang, Michael C King, and Kevin W Bowyer. Analysis of gender inequality in face recognition accuracy. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 81–89, 2020. 3

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 4

[3] Martins Bruveris, Jochem Gietema, Pouria Mortazavian, and Mohan Mahadevan. Reducing geographic performance differentials for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 98–106, 2020. 2

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. 2

[5] Dong Cao, Xiangyu Zhu, Xingyu Huang, Jianzhu Guo, and Zhen Lei. Domain balancing: Face recognition on long-tailed domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5671–5679, 2020. 3

[6] 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)*, pages 67–74. IEEE, 2018. 2, 3

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 2, 3

[8] Weihong Deng, Jiani Hu, Nanhai Zhang, Binghui Chen, and Jun Guo. Fine-grained face verification: Fglfw database, baselines, and human-dcmn partnership. *Pattern Recognition*, 66:63–73, 2017. 2

[9] Sixue Gong, Xiaoming Liu, and Anil K Jain. Jointly debiasing face recognition and demographic attribute estimation. In *European Conference on Computer Vision*, pages 330–347. Springer, 2020. 2

[10] Sixue Gong, Xiaoming Liu, and Anil K Jain. Mitigating face recognition bias via group adaptive classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3414–3424, 2021. 2

[11] 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*, pages 87–102. Springer, 2016. 3

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[13] 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. Number 07-49, October 2007. 3

[14] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 1, 3

[15] KS Krishnapriya, Vítor Albiero, Kushal Vangara, Michael C King, and Kevin W Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020. 3

[16] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *Proceedings. international conference on image processing*, volume 1, pages I–I. IEEE, 2002. 6

[17] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z. Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1179–1187, January 2021. 3

[18] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 2, 3

[19] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018. 3

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

[21] Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. Sensitivenets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2158–2164, 2020. 2

[22] Shruti Nagpal, Maneet Singh, Richa Singh, and Mayank Vatsa. Deep learning for face recognition: Pride or prejudiced? *arXiv preprint arXiv:1904.01219*, 2019. 2

[23] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015. 2

[24] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017. 2

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[26] Ignacio Serna, Aythami Morales, Julian Fierrez, Manuel Cebrian, Nick Obradovich, and Iyad Rahwan. Sensitiveloss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *arXiv preprint arXiv:2004.11246*, 2020. 2

[27] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019. 2

[28] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 2

[29] Valeska Uchôa, Kelson Aires, Rodrigo Veras, Anselmo Paiva, and Laurindo Britto. Data augmentation for face recognition with cnn transfer learning. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 143–148. IEEE, 2020. 3

[30] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017. 2

[31] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 2

[32] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9322–9331, 2020. 2, 3

[33] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. 2

[34] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 3

[35] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020. 2

[36] Xin Wen, Biying Li, Haiyun Guo, Zhiwei Liu, Guosheng Hu, Ming Tang, and Jinqiao Wang. Adaptive variance based label distribution learning for facial age estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 379–395. Springer, 2020. 3

[37] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 2, 3

[38] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 2

[39] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020. 2, 3

[40] Yaobin Zhang and Weihong Deng. Class-balanced training for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 824–825, 2020. 3

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