Evaluation of targeted dataset collection on racial equity in face recognition

Rachel Hong
hongrach@cs.washington.edu
University of Washington
Seattle, Washington, USA

Tadayoshi Kohno
yoshi@cs.washington.edu
University of Washington
Seattle, Washington, USA

Jamie Morgenstern
jamiemmt@cs.washington.edu
University of Washington
Seattle, Washington, USA

ABSTRACT

Algorithmic audits of industry face recognition models have recently incentivized companies to diversify their data collection methods, which in turn has reduced error disparities along demographic lines, such as gender or race. We argue that it is important to understand exactly how various forms of targeted data collection mitigate performance disparities in these updated face recognition models. We propose an empirical framework to assess the impact of additional dataset collection targeted towards various racial groups. We apply our framework to three racially-annotated benchmark datasets using three standard face recognition models. Our findings empirically validate the notion that the introduction of data from the demographic group with the initially-lowest performance improves performance on that group significantly more than adding from other groups. We also observe that in all settings, the introduction of data from a previously omitted group does not harm the performance of other groups. Furthermore, investigation of feature embeddings reveals that performance increases are associated with a larger separation among images of different identities. Despite the commonalities we observe across datasets, we also find key differences: for example, in one dataset, training on one racial group generalizes well across all groups. These differences speak to the criticality of re-applying empirical evaluation methods, such as the methods in this work, when introducing new datasets or models.

CCS CONCEPTS

• **Computing methodologies** → Neural networks; **Biometrics**; **Object recognition**; **Matching**; • **Social and professional topics** → **Race and ethnicity**; • **Information systems** → *Data mining*.

KEYWORDS

Algorithmic audit, data collection, face recognition, racial bias in computer vision

ACM Reference Format:
Rachel Hong, Tadayoshi Kohno, and Jamie Morgenstern. 2023. Evaluation of targeted dataset collection on racial equity in face recognition. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3600211.3604662

AIES '23, August 08–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08.
https://doi.org/10.1145/3600211.3604662

1 INTRODUCTION

In the last decade, extensive research studies have demonstrated the prevalence of demographic biases in machine learning systems, due to a lack of representation in training datasets [29]. Most notably, in the domain of face analysis, standard face datasets include very few images of individuals with darker skin types, and researchers have determined that commercial gender classification models have much higher error rates for women with darker skin types [8]. However, facial recognition continues to be used widely: from identity verification in mobile devices to public surveillance in certain countries, many people interact with these systems in their day-to-day lives [22]. While some argue for the complete removal of facial recognition techologies [7], the use of these technologies may not disappear. As such, opponents of face recognition along with the developers of these systems may both benefit from a careful analysis of how the demographic makeup of training datasets may impact a model's performance on various demographic groups.

In order to remedy past data representation bias, researchers have developed several new benchmark face recognition datasets that are more balanced along demographic attributes such as gender or race [38, 44]. While these balanced datasets have improved model performance, accuracy disparities still persist [45]. For example, the optimal allocation of training data by demographic group is not always the equally-balanced allocation: Gwilliam et al. [19] find that a balanced training set (with equal number of samples per racial group) obtains a higher accuracy variance across groups but the same overall accuracy compared to another training data allocation.

Additionally, curating new datasets requires time and resources, and can intrude upon the subpopulation being studied [33]. It is also incredibly time-consuming to train models on all possible allocations of demographic groups in order to find some "optimal" allocation. Rather than searching for the best subgroup allocation for a training set of a fixed size, companies may prefer a greedy solution — a solution in which new data is added in an add-only manner. Hence, we focus on the following goal: to examine *additional data collection* and its impacts on the performance of various racial groups.

Consider the following scenario: an entity (e.g., a company or a group of researchers) trains a face recognition model using some initial training dataset which lacks data from some particular racial group. Upon evaluation on held-out test data or due to an external bias audit, the company realizes their performance lags on that group, and now wishes to collect more data from the omitted group. They have the budget to collect only a fixed number of samples and have limited resources to train additional models (and, perhaps, can only train one other model). This process closely follows several

corporations' past responses detailed in Raji and Buolamwini [35] and allows us to pose these research questions:

(1) How does additional data from the underrepresented group change the test performance for that particular group, as well as the test performance for other groups?
(2) How does data collection targeted towards the group with the initially-lowest performance impact that group's test performance and overall group differences, in comparison to introducing data from other groups?
(3) Are our results consistent across racial groups, datasets, and models?

To answer these questions, we develop an empirical framework to evaluate the performance impact of data augmentation by demographic subgroup. For our framework and analyses, we focus on *one-to-one facial recognition*: given two images of faces, a one-to-one facial recognition system is designed to determine whether or not those two images are of the same person. We implement this framework for three racially-annotated datasets (BFW [38], BUPT [43, 44], and VMER [16]) and three state-of-the-art face recognition models (SE ResNet [9], CenterLoss [46], and SphereFace [27]). We summarize our main empirical findings below:

(1) The introduction of samples from any racial group X improves the performance for every group that we tested. (Different datasets use different terms. Using the terms in the source datasets, e.g., for BUPT [43, 44], we considered images labeled as *African, Asian, Caucasian, or Indian.*)
(2) The addition of data from the lowest-performing group improves that group's performance the most and closes performance gaps across racial groups.
(3) Increasing data from the highest-performing group X widens performance disparities, regardless of whether the initial training dataset contained images from group X, a specific counter to the notion that more data or more representation reduces discrimination.
(4) The above findings are *consistent* across all datasets and models we examined, while some findings are *different* across different datasets.

That some findings are *not* generalizable from the analysis of only a single dataset — speaks to the criticality of assessing various datasets. While the academic benchmark datasets we examine do not reach the commercial scale, such as Clearview AI's training data of 30 billion images [26], we find that our framework is still useful to understand how various datasets behave and how pre-conceived assumptions of additional representation do not always hold.

Thus, based on our findings, we encourage future work that introduces new datasets to re-apply our methodology (and others) as benchmarks to evaluate those datasets with known face recognition models. To facilitate this process, we publish our source code online at https://github.com/hongrachel/representation-disparities.

2 BACKGROUND AND RELATED WORK

In computer vision, researchers have extensively examined data representation biases and how models trained on datasets unrepresentative of the general population perform poorly on underrepresented groups.

For example, Pahl et al. [32] annotate several facial expression datasets and observe that these datasets skew heavily towards younger Euro-American subjects. In addition, Wilson et al. [47] find that a standard pedestrian detection dataset contains more data from individuals with lighter skin tones, and resulting models obtain higher accuracy for detecting individuals with lighter skin tones. Albiero et al. [2] investigate the source of gender bias in standard face recognition systems and determine that the test accuracy gap is attributed to models mapping images of women closer together.

Shortly after the publication of Buolamwini and Gebru [8], which demonstrated how several commercial face recognition systems discriminate by skin tone, these corporations updated their face recognition APIs to mitigate performance disparities. In their released statements, they explicitly cited new dataset collection efforts in order to ensure diverse representation in their training sets [35]. These newly updated models significantly decreased (previously high) error rates for individuals with darker skin and attributed their improvement to the targeted collection of additonal data along the lines of skin tone, gender, and age [37]. Diverse data collection is a promising method to address bias [23], but there has been little work investigating cases when the new data is composed of some explicitly-chosen demographic group that was previously underrepresented or omitted in the initial training set.

As a result, the lack of diverse data has spurred the creation of balanced training datasets, which have shown marked improvements in classification accuracy rates for previously underrepresented groups, even when trained with the same model architecture. Specifically, much recent work has focused on the collection of diverse face image datasets, along dimensions such as race, gender, age, lighting, pose, and expression, in order to allow models to generalize well on real-world variations [9, 24, 28]. These datasets have also been used to evaluate proposed face recognition models that reduce bias, which incorporate novel loss functions or model architectures. For instance, Serna et al. [41] show that a sensitive triplet loss function improves both accuracy and fairness across racial groups.

Recently, several studies examine how demographic subgroup distribution in training plays a role in accuracy disparities. In the case of gender bias in face recognition, Albiero et al. [3] observe that training datasets equally-balanced by gender lowers the prediction accuracy gap between groups, but the equally-balanced allocation does not minimize the accuracy gap. Similarly, Gwilliam et al. [19] vary the racial group makeup of the training set and also observe that the equally-balanced allocation is not the most optimal or fair one. Our work builds off their research and extends this investigation by analyzing the impact of *adding data* from different racial group distributions, rather than holding the training size fixed.

There are also several recent works in fairness literature that formally explore data collection processes. Most notably, Rolf et al. [39] form a theoretical framework to model subgroup allocations in training for a fixed training set size. They find that dataset composition impacts performance more than upweighting samples from minority groups. Chen et al. [10] provide a procedure to estimate the value of collecting additional samples and empirically validate the notion that additional data collection can mitigate discrimination without an accuracy tradeoff. Their work focuses on introducing data drawn from the same sampling distribution rather than data

collection targeted by demographic group. Abernethy et al. [1] propose several adaptive sampling algorithms for achieving min-max fairness, which minimizes the loss of the group that is worst off, to update the model over a series of iterations. Finally, Gong et al. [15] survey several definitions of input diversity in training data, through various sampling processes that upweight diverse batches in training. Our focus complements these works by assessing the empirical impact of targeted data collection on performance inequities.

3 METHODOLOGY

3.1 Problem setup

We now describe our task setting; we focus on *face verification*, or 1-to-1 face-matching, due to its ability to handle identities outside of the training distribution. We follow the standard face recognition training process in deep learning literature [42]: given a dataset $\mathcal{D}$ with face images $\{x\}$ and identity labels $\{y\}$, we train a model that takes an image as input and outputs a vector corresponding to the image's predicted identity. This training minimizes the empirical risk with respect to a particular loss function. The model is then used to perform inference (or prediction) by removing the final output layer. The result is a model that takes an image as input and produces a feature embedding with some fixed size established during training. This output feature embedding can be thought of as a lower-dimensional representation of an individual face image.

We evaluate the performance of a given model on the task of face verification: given two images (x, x'), $x \neq x'$, do the two images belong to the same identity or not? This evaluation is performed on *pairs* of images from a held-out test set, where the images and identities beloging to the test set are disjoint from those in the training set.

To convert the model from one which produces embeddings to one which predicts whether pairs of images are of the same identity, we do the following. For a particularly fixed threshold t, the face verification system predicts that the test pair are of the same individual if the cosine similarity score of the two images' feature embeddings is at least t. As such, ground truth labels of a pair are separated into a *genuine* pair (label 1) or an *impostor* pair (label 0), following the terminology in existing literature on face verification [11]. In this manner, the verification process evaluates the differences between the genuine and impostor score distributions. This methodology does not explicitly assume that the test and training data collection processes are the same or even similar, though conceptual frameworks often assume the two are the same.

3.2 Experiment design

Given a model trained on a dataset $\mathcal{D}$, we study a method of data collection motivated by our scenario of interest, where a face recognition system developer might respond to bias audits by collecting more training data from some target demographic group. As such, we focus on benchmark datasets with each image belonging to some racial group.

We define our method, *single-group augmentation*, as the incremental addition of samples from a fixed racial group to some initial training set consisting of a single racial group. This enables us to compare the performance of re-trained models by adding data from various groups, in order to determine whether the model improves more by training on an unseen group versus the initial group. We give the formal definition of single-group augmentation below.

We stress that we are *not* arguing that this data augmentation method should be used in practice, nor does this precisely say that a facial recognition system might only train on a single demographic group in practice. Rather, our experimental methodology distills the core essence of a targeted data collection approach, such that the impacts of data augmentation can be isolated and empirically analyzed.

3.2.1 Procedure for single-group augmentation. We train our models across a variety of training set configurations to understand how the group-specific performance of a model changes with the introduction of data targeted towards a specific demographic group. We follow a very similar setup and build off of the codebase from Gwilliam et al. [19]. Unlike their work, however, we do not maintain a fixed size training set and change proportions, but instead augment the dataset with additional data, and we empirically analyze three datasets rather than one. The training configurations are defined as follows:

For each group A, the *initial training configuration* consists of images from N randomly-chosen identities from group A, where N is fixed dependent on the size of the benchmark dataset $\mathcal{D}$. Here we refer to group A as the *initial group*. To obtain subsequent training configurations, we iteratively augment the initial training configuration with n randomly-sampled identities from another group B, where n is also decided based on $\mathcal{D}$. We refer to group B as the *target group*. As an example, an initial training configuration may consist of images from 200 identities from the *African-American* group, and we incrementally add images from 50 identities from the *East-Asian* group to obtain the rest of the training configurations.

Note that in some settings, the initial group A may be equivalent to the target group B. This enables our empirical analysis to compare continually adding data from the same group to continually adding the same amount of data from a previously unrepresented group. In other words, we can assess the impact of increasing demographic representation in the training data.

The design of these training sets replicates the motivating scenario of training data collection targeted on a particular demographic group in a simple setting of moving from one group in training to two. This empirical framework therefore simulates an existing face recognition system's possible response to bias audits.

3.3 Datasets

We conduct experiments on three existing racially-annotated datasets that we present in order of dataset size: BUPT [43, 44] (the largest dataset), VMER [16], and BFW [38], all of which have been used in face recognition model evaluations of racial bias [14, 19]. Other datasets we considered lacked sufficient images per subject to adequately train a model [34, 40], or were designed for other face-related analysis tasks [24]. Table 1 gives a breakdown of the groups in each dataset we examine. We observe that each dataset names racial categories differently from each other, and some refer to ethnicity rather than race [25]. In our results, we refer to the terminology used in the evaluated dataset in italics, but also recognize

Dataset	Categories	Subjects per category	Images per subject	Test subjects per category
BFW [38]	*Asian, Black, Indian, White*	180	25	20
BUPT / RFW [43, 44]	*African, Asian, Caucasian, Indian*	5000	18	3000
VMER [16]	*African American, East Asian, Caucasian Latin, Asian Indian*	400	108	24

Table 1: A summary composition of datasets in training and test folds, subsampled to ensure equal number of images per subject. Here, a subject refers to an identity, of which there are some number of images. It is assumed each subject belongs to exactly one category.

there are both overlaps and key distinctions between each dataset's group definitions, which is discussed further in Section 5.3.

To form the test image pairs from a given test set, we follow standard methodology as Wang et al. [44]. In every dataset, we generate all possible pairs of distinct test images (x, x'), $x \neq x'$ from the same group, assigning label 1 if the images share the same identity and 0 otherwise.

BUPT-BalancedFace (BUPT) contains a total of 1.3 million images from 28,000 individuals and is equally broken down into 4 demographic groups: *African*, *Asian*, *Caucasian*, and *Indian* [43]. Images are collected from the benchmark MS-Celeb-1M dataset [18] and augmented via Google search for additional celebrities in particular categories. The subjects are categorized by racial group using their nationality as a proxy, as well as via the Face++ API. Using nationality and race prediction are not robust methods for race categorization [25]; however, this is one of the only large-scale face datasets to consist of at least 7 thousand subjects per group. To ensure at least 18 images per subject, we constrict to 5 thousand subjects per group, which matches the setup in Gwilliam et al. [19].

The accompanying test dataset Racial Faces in the Wild (RFW) consists of fifty million test pairs and uses the same racial annotation method as BUPT. RFW is also from MS-Celeb-1M [18], but does not have any overlap with any subject from BUPT. For simplicity, we refer to the BUPT training and RFW test dataset as "BUPT."

VGGFace2 Mivia Ethnicity Recognition (VMER) dataset adds group annotations (*African American*, *Asian Indian*, *Caucasian Latin*, and *East Asian*) to the entire VGGFace2 training and test sets, which is one of the largest academic face recognition datasets [16]. VMER uses manual annotations across three million images to categorize subjects into four racial groups. Greco et al. [16] intentionally choose this annotation procedure rather than pre-trained models, in response to critiques that ethnicity classifiers fail to generalize well on racially-diverse datasets [24]. This dataset also consists of many more images per subject. To conduct our experiments with equal training set size per group, we randomly sample 440 individuals per group with 108 images per individual, which allows us to evaluate models trained on significantly more images for a given subject.

Balanced Faces in the Wild (BFW) is another dataset with an equal number of images and subjects from each racial category, but is also balanced by subgroups *Male* and *Female* within each racial group [38]. Each category consists of five thousand images from two hundred subjects with an equal number of faces per subject. BFW also samples from VGGFace2 [9], but instead uses pre-trained ethnicity classifiers to categorize subjects into the following groups: *Asian*, *Black*, *Indian*, and *White*. As with BUPT, pre-trained ethnicity classifiers, even if well-designed, may have inaccuracies [25]. To form the test set, we randomly select a hold-out fold of twenty individuals per group. Since the test sets for BUPT and VMER are fixed, for consistency of analysis, we similarly create a static test set for BFW as well.

3.4 Models

We perform these experiments on three state-of-the-art face verification architectures defined below. In each experiment, we train a model from scratch on the training configurations defined in Section 3.2.1. The models each use cross-entropy loss as the base classification loss function, stochastic gradient descent as the optimization function, and train for 50 epochs. We define the explicit hyperparameters used for each model in Appendix A.5.

The **SE-ResNet** model uses ResNet-50, a standard convolutional neural network with 50 layers [20], as a backbone and attaches Squeeze-and-Excitation blocks, which dynamically recalibrate channel wise feature responses [21]. Cao et al. [9] implement SE-ResNet to train on their proposed VGGFace2 dataset to demonstrate their improved performance in comparison to prior benchmarks. The **CenterLoss** model learns a center vector for each identity, in order to incorporate a loss penalty between feature embeddings and the identity's center, along with the base cross entropy loss function [46]. This minimizes the within-identity feature embedding distance and separates identities within the feature space. The **SphereFace** model introduces a multiplicative angular margin to the model's output, which maximizes the variance between feature embeddings of different identities.

3.5 Evaluation

To empirically measure model performance, we consider several evaluation metrics and in this section briefly describe the tradeoffs between them.

3.5.1 Global threshold. In face verification tasks, the model once trained depends on some chosen threshold to form binary predictions. We find, however, that the model evaluation of a global threshold does not sufficiently capture a model's behavior. Robinson et al. [38] demonstrate that using a singular threshold across demographic groups results in accuracy gaps, and that group-specific thresholds can strictly improve test accuracy across groups. In addition, many commercial face recognition systems, such as Amazon's Rekognition, allow users to set thresholds according to some application objective, i.e., to maintain a certain false positive rate [5]. Therefore, it is important to examine the model performance across a range of thresholds, rather than evaluation of a single one.

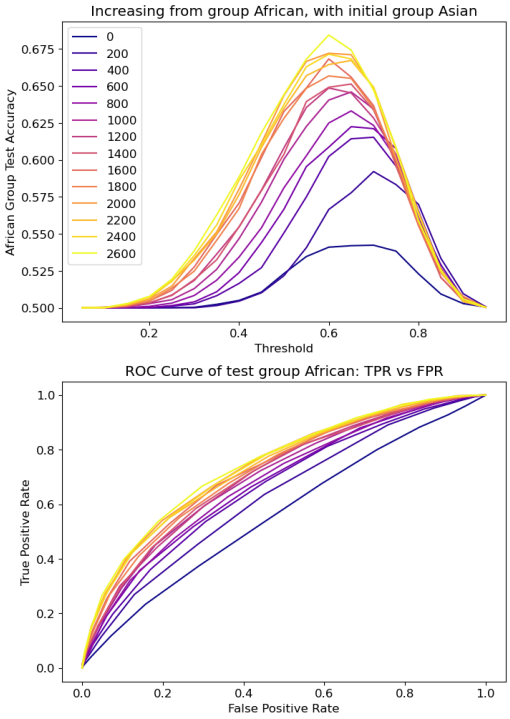


Figure 1: Initial exploration of impact of threshold on test accuracy, for an example initial group, target group pair on the BUPT dataset and SE-ResNet model. Color denotes size of target group, while initial group stays fixed at 2000.

Previous work on demographic group allocation in training have studied the accuracy rates obtained from a particular threshold [3, 19]. In our initial exploration of single-group augmentation, for every test group we plot test accuracy against threshold values across each training configuration, as shown in Figure 1. While we find trends in increasing the size of the target group, it is difficult to capture how test accuracy increases, given that the optimal threshold changes at each line. If we hope to understand the different forms of single-group augmentation, we find that distilling the ROC curve to a single metric enables comparisons among many training configurations.

3.5.2 Overall accuracy. Regardless of the threshold selection problem, we also find that studying overall accuracy has its limitations: there are many cases when equalization of accuracy rates by group still allows for disparate treatment [13]. For example, a face verification system may obtain a high false positive rate and a low false negative rate for one group, but still maintain equal accuracy across all groups. If this system is used for biometric authentication, this disparity in false positive rates could result in disproportionate security vulnerabilities for one demographic group. As a result, in our evaluation, we avoid studying accuracy as a comparison metric. Moreover, this prompts us to also examine the impact of targeted data collection on the group with the lowest performance, instead of using equal performance as the primary objective.

3.5.3 Area under the curve. As a consequence of the above disadvantages, we shift our attention to the *area under the curve* (AUC) calculated by the receiver operating characteristic curve (ROC) curve, an evaluation metric that has been used in prior face recognition literature [6]. The AUC is the probability that a positive test pair has a higher similarity score than a negative test pair, which enables our analysis to capture the distance distributions of feature embeddings, rather than merely considering the accuracy (or false positive or negative rates) of a binary classification task for a fixed threshold.

We note that AUC is a single numerical value which describes the functional relationship between true positives and false positives of a classification model derived from thresholding a regression model. It therefore is an incomplete description of the ROC curve, and two regression models might have equal AUC values but very different behavior in terms of this tradeoff.

3.6 Broader contexts and limitations

In addition to the previously-mentioned assumptions of demographic group fairness, we find certain limitations to the ability to generalize beyond our datasets, which are clarified below (in Section 5.2 we discuss how the limited ability to generalize from our results to other datasets is a strength for some of our other conclusions). In this section, we also situate our methodology in relation to the broader context of machine learning research.

3.6.1 Group fairness. In our work, we examine the task of face verification from a group fairness lens because we find that the main demographic information attached to standard face benchmarks is group membership. The datasets we study partition identities into only four racial groups, which excludes and merges many racial categories. Moreover, each dataset implicitly assumes that each individual belongs to a single category. This inherently ignores individuals with multi-racial identities, and the lack of additional demographic information may prevent analysis of intersectional differences along other dimensions, such as gender or age. We believe that this is an important topic for future study, especially as adding a single training sample can often increase representation across multiple demographic groups. At the same time, it is still beneficial to understand existing differences in performance among these groups, given the limitations of real-world data containing demographic information in the first place. In Section 5.3, we elaborate upon the implications of group-level annotations based on our results.

3.6.2 Image variations by racial group. Specific to the BUPT dataset, prior research has shown that the average face-to-image ratio is much lower for images from the *Caucasian* and *African* groups [19]. We obtain similar findings even when we control for face-to-image ratios, but this discrepancy indicates that other image variations by racial group, such as lighting or pose, may factor into our results. Previous work on performance gaps in group-balanced datasets has extrapolated that learning for a particular demographic is inherently more difficult [44]. We caution against making the broad claim that performance is capped for a certain sociodemographic group, as image quality and inter-group image variations can often also explain these gaps. For instance, many face image datasets are scraped

from celebrity photographs online; as a result, researchers have pointed out distinct differences of celebrity photography by race or gender, such as higher proportions of women wearing makeup than men, which may in turn affect performance disparities by demographic group [2, 4].

3.6.3 Underrepresented versus unrepresented groups. We also highlight that the single-group augmentation framework narrows the problem space we consider: from our motivation of racial groups that are *underrepresented* in training, to our experiments on racial groups completely *unrepresented* in training. We make this simplification intentionally to isolate the impact of augmenting one group with another group — underrepresentation on the extreme end.

The specific task of evaluating a model on an unseen group relates to domain generalization, a well-studied subfield of machine learning. While domain generalization techniques can be applied to this problem, Gulrajani and Lopez-Paz [17] show that model selection may not be straightforward when evaluated on a variety of datasets; this is an important area for further study. As a result, we recognize that our study examines only one piece of the puzzle: dataset representation bias does not encapsulate demographic bias across the entire face recognition system. Due to the variations in image quality by demographic group as described above, model interventions may be needed to ensure some chosen fairness criteria or generalization property. In this work though, we center our focus on the racial composition of training datasets, instead of a specific machine learning algorithm.

3.6.4 Generalizability of datasets. Finally, the dataset-specific artifacts highlight the difficulty of making generalizations of our observed trends to apply to all future forms of data collection. We limit our study to face recognition models and benchmark datasets available for academic study. If we hope to understand how corporations should best respond to bias audits, it is unclear whether our findings extend to systems training on datasets with sizes at a much larger magnitude. Moreover, we recognize that commercial face recognition systems may rely on vast pre-trained models that are not publicly available. We therefore acknowledge that our work may not align with the training procedures and large-scale datasets that industry face recognition systems may follow — this prompts the need for the release of commercial datasets and practices to the research community.

However, the fact that differences between datasets exist is itself an important contribution, especially as BUPT, BFW, and VMER continue to be used as benchmarks in face recognition literature to evaluate racial bias [14, 19]. In Section 5.2, we explore how our methodology may inform how future work can use these benchmark datasets, in addition to new ones.

4 RESULTS

We now present some representative findings in the figures below. For brevity, we show results for the SE-ResNet model, though the relative comparisons and general trends are consistent for CenterLoss and SphereFace. In general, we focus on the BUPT dataset to demonstrate key results due to its large size, but clarify otherwise when there are distinct dataset differences. For more details and accompanying results, please refer to Appendix A.

4.1 Differences among datasets

First, we observe in Figure 2 that the group-specific performance impact of single-group augmentation differs across datasets. Training on data from some racial group may not impact performance in the same manner across various datasets. As such, evaluation of a single benchmark dataset may not be sufficient; we elaborate on this further in Section 5.2.

4.1.1 VMER: Increasing representation improves AUC of unrepresented group more than addition from other groups. In Figure 2a, we show the impact of single-group augmentation on the AUC of each test group. We find that setting the target group as the test group results in the highest growth in AUC for the ranges in training size that we examine. In other words, if we were to update a face verification model by introducing samples from a single racial group, in VMER, the best choice to improve group X's performance is to add more data from group X.

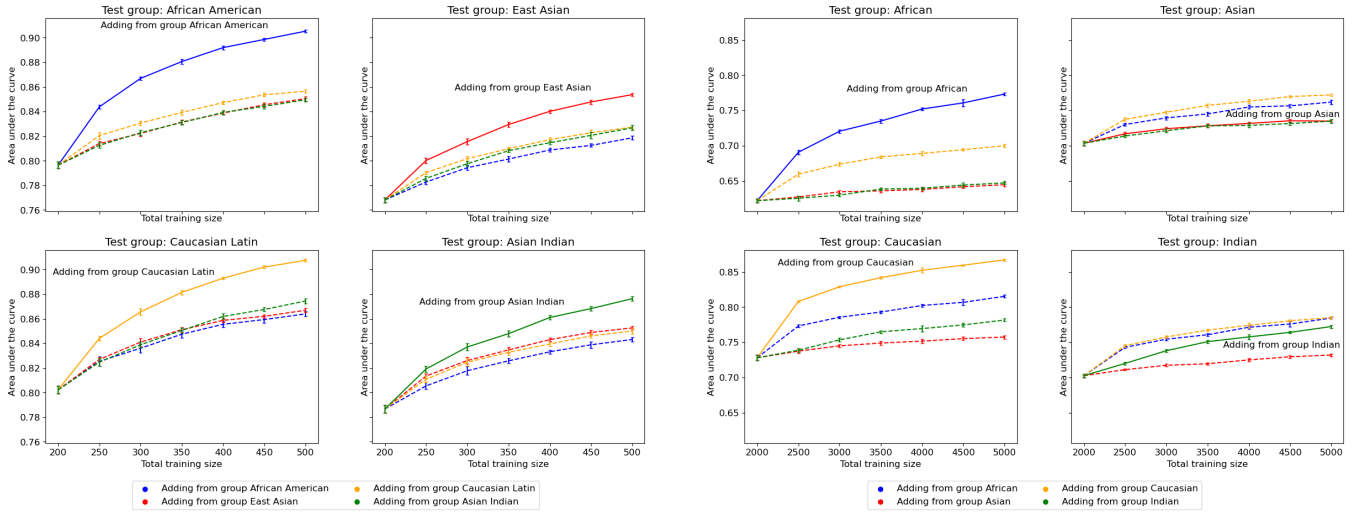
The same relative comparisons can be made when broken down by initial training configuration (see Appendix A.1 for details). Given an initial training set without group X, in the VMER dataset, the re-trained model's performance on unrepresented group X increases the most when increasing representation from group X in training. Even if the model initially trains on X, we find that continuing to augment the training set with samples from group X outperforms augmentation from any other group.

This case illustrates an example where out of all forms of single-group augmentation, improving demographic representation in training datasets increases the unrepresented group's performance the most. This matches existing intuition behind the development of training datasets that are balanced along demographics or more diverse in face composition, in response to prior face datasets that lacked representation along these dimensions [9, 24, 28].

4.1.2 BUPT: Training on some racial groups generalizes across all groups more than the addition of unrepresented groups. Figure 2b demonstrates the change in AUC for each group in the BUPT dataset. We observe that introducing data from the *African* and *Caucasian* groups improves the AUC for all groups regardless of the initial training configuration (Appendix A.1). Introducing data from the *Asian* and *Indian* groups does improve group-specific performance, but not as much as adding from the other groups, even when evaluated on the *Asian* and *Indian* test groups.

Compared to VMER, this result demonstrates that in the BUPT dataset, data from *African* and *Caucasian* groups generalizes strongly across all four groups. Gwilliam et al. [19] also confirm this trend since they find that when training on data from a single group, training on data from the *African* and *Caucasian* groups obtains the highest test accuracy for each group. A potential explanation may be that a significant proportion of images from *Asian* and *Indian* groups in training have much larger face-to-image ratios than in test [19]. We show that the same relative comparisons hold even when controlling for face-to-image ratios in Appendix A.1.1, but note that the shift from training to test sets might look different between demographic groups along other relevant dimensions.

Figure 2b shows that the addition of data from an unrepresented group is not always the best way to improve the performance for that same unrepresented group, unlike our findings in Figure 2a.



(a) VMER dataset **(b) BUPT dataset**

Figure 2: AUC for each test group under single-group augmentation averaged across all initial training configurations for both VMER and BUPT datasets. The solid line continually adds data matching the test group, and the dashed line continually adds data from a different group. Evaluated on SE-ResNet model across 5 trials with consistent results across models (per dataset).

These findings complicate the idea that the most data-efficient way to improve performance for a population X (excluded in training) is to increase representation of population X in the training set.

4.2 Similarities across datasets

In addition to the differences in results across datasets we described above, our analysis methodology revealed several trends which hold across the three datasets and three models. We highlight these trends and assess whether our analyses confirm the intuitions and findings in prior literature.

4.2.1 AUC on all test groups increases with additional training data, regardless of the group being introduced. In Figures 2a and 2b, we observe that with any form of data addition, the AUC values across all test groups increase regardless of the group being introduced and the initial training group. We find the same trend for every dataset-model pair single-group augmentation experiment we perform. Particularly, in our experiments, a model that retrains on additional training data from some target group does not sacrifice performance on the initial group in order to account for the target group. This demonstrates the notion that large neural networks have extensive capacity to capture arbitrarily complex functions [48], which also applies to new samples from distinct demographic groups.

4.2.2 No performance tradeoff among groups: Introducing data from racial groups distinct from the initial group does not harm the initial group. In various studied settings with group fairness objectives, researchers have demonstrated the existence of fairness-accuracy tradeoffs, especially in low-parametrized models, such as linear regression [12]. In our face verification experiments, we find that the introduction of data from groups distinct from the initial group does not harm the initial group; instead, the retrained model strictly increases performance across all groups.

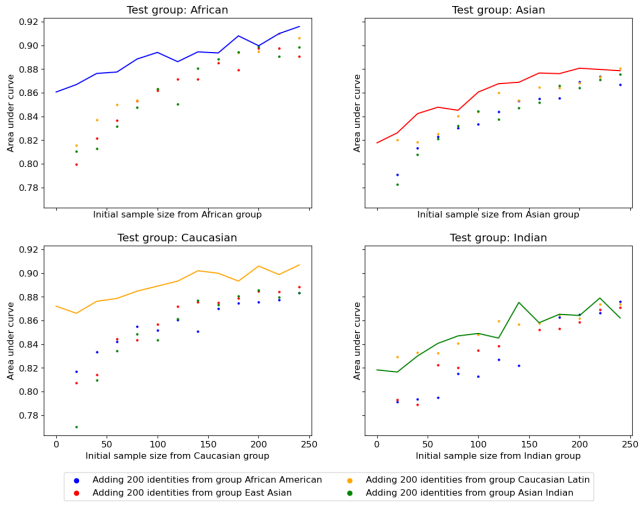


Figure 3: AUC of each test group under fixed-size data additions versus the size of the initial training set, composed of samples matching the test group. The solid line represents adding 200 identities from the test group, and the points represent adding 200 identities from a different group. Note that AUC increases as the initial training set size increases along the x-axis. Evaluated on VMER dataset with SE-ResNet model with consistent results across models.

4.2.3 Marginal performance of fixed-size data additions from the test group versus data additions from other groups shrinks as initial

training set grows. In Figure 3, we focus on the addition of a fixed number of samples from any group, instead of the continual introduction of samples from a single group as illustrated in prior figures. This form of analysis thus answers the following question: given an initial training set on group X, how does performance on group X differ between adding N samples from group X versus from another group?

In the VMER dataset, Figure 2a shows that adding data from the same group as test improves the test group's performance the most, in comparison to other forms of single-group augmentation. As such, the AUC value from adding from the test group (solid line) is higher than adding from another group, across most initial training set sizes. However, we observe that as the initial training set size gets large, the marginal benefit of introducing data from the test group compared to another group shrinks. We find this phenomenon across other models and in the BFW dataset as well. Due to the size limitations of the benchmark datasets we examined, it is unclear if adding data from a non-test group will ever obtain a higher performance than adding from the test group and requires further study.

4.3 Performance disparities for single-group augmentation

In Figure 4, we study different measures of performance disparity among racial groups with single-group augmentation.

4.3.1 Examination of group AUC disparities reveals examples that additional data can widen performance gaps. Figure 4a uses widest AUC disparity as a metric for unfairness via single-group augmentation. While equalizing performance across groups is not always desirable due to cases of sacrificing performance to satisfy parity, we have observed no decrease in performance with any form of additional data. As such, we still find it valuable to understand how performance gaps may change as a result of incorporating data from some racial group.

In Figure 4a, introducing data from the *African* group lowers the AUC disparity. This is driven by an increase in the *African* group's performance, which was originally the lowest. On the other hand, introducing data from the *Caucasian* group increases the test performance gap. This is driven by an increase in the *Caucasian* group's performance, which was originally the highest, even without inclusion of the *Caucasian* group in the initial training configuration.

4.3.2 Results contradict principle that more data reduces demographic bias. Figure 4a thus illustrates how data collection can generate various outcomes in performance disparities, and we find similar examples in other datasets (Appendix A.2). Moreover, the finding that adding data from an unrepresented group, such as the *Caucasian* group, widens performance gaps is a clear counter to the idea that more data mitigates discrimination as discussed in Chen et al. [10]. Their work proves that collecting more data from the population distribution decreases the population loss gap between groups. In our work, we consider data collection methods that may not match the test distribution, which may be realistic in cases when the test distribution is unknown. As a result, we demonstrate how the introduction of data from a group unrepresented in training may worsen performance disparities.

4.3.3 Adding data from the group with the initially-lowest AUC increases the AUC for that group significantly more than adding data from other groups. Figure 4b distinguishes different forms of single-group augmentation based on whether the target data is from the group that originally obtained the lowest AUC value. Across all models, datasets, and when separated by initial training configurations (Appendix A.3), we find that if the objective is to most improve the test performance for the group with the lowest AUC in the initial training set, adding data from that group increases performance significantly more than adding data from any other group.

4.3.4 Results connect to prior theoretical work on sampling from group with lowest performance. This validates prior theoretical analysis on active learning in group fairness. Abernethy et al. [1] find that updating the model with the samples from the current worst-off group converges to a min-max fairness solution, or minimizes the maximum classification loss across groups. In this manner, suppose a developer wishes to update their face recognition system to address concerns about a demographic group on which the model classifies poorly. Then targeted data collection on that group may improve the retrained model's performance, even if that group was already included in the initial training set.

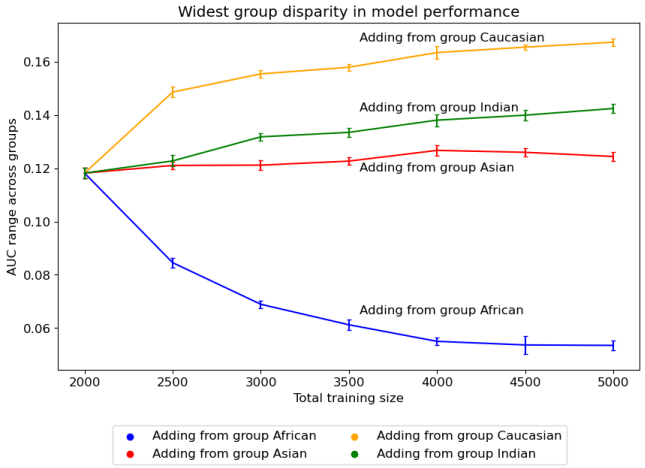
4.3.5 Lowest-performing group does not equal the least-represented group. Note the distinction between a group with the lowest performance and a group that is unrepresented in training. Although Figure 2b shows that data augmentation from some omitted group X may not significantly improve that group's AUC, this is still consistent with Figure 4b since group X did not have the lowest test performance in the initial training configuration.

4.4 Feature embedding similarity score distribution

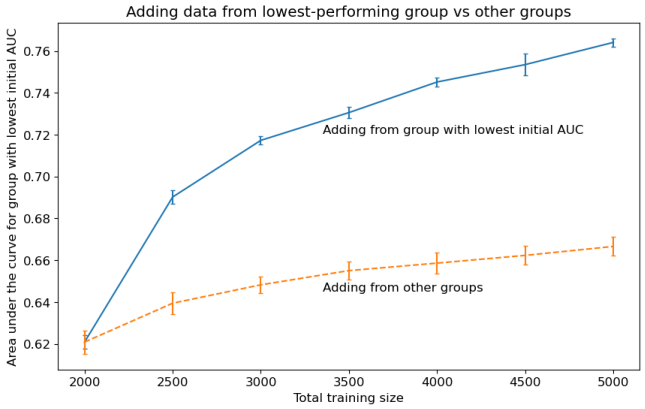
In order to explain the increase in AUC values from single-group augmentation, we investigate a model's feature embeddings from test images. Figure 5 plots the difference in average cosine similarity scores between genuine (label 1) and impostor (label 0) test pairs against the overall AUC of the test group. Each point represents a training configuration where the target group matches the test group, over all initial groups of the same size, and the color encodes the target group size at every point.

First, we find a clear positive relationship between distance and performance. This is consistent across datasets with more details in Appendix A.4. This observation indicates that higher AUC values for some test group are associated with genuine pairs having much higher cosine similarity scores on average than those of impostor pairs. This relationship follows from the model test pair procedure because models that further separate similarity scores between genuine and impostor pairs will obtain a higher AUC value by definition. This result matches findings in Albiero et al. [2], which examine test pair similarity distributions along gender and race.

Second, we notice that for every test group, the upwards trajectory is driven by adding samples matching the test group, regardless of the initial training configuration. This observation indicates that



(a) Widest disparity in AUC among groups when introduced with more data from each group.



(b) AUC of group with lowest performance in initial training configuration.

Figure 4: Performance disparity measures of single-group augmentation. Evaluated on BUPT dataset with SE-ResNet model across 5 trials with consistent results across models (per dataset).

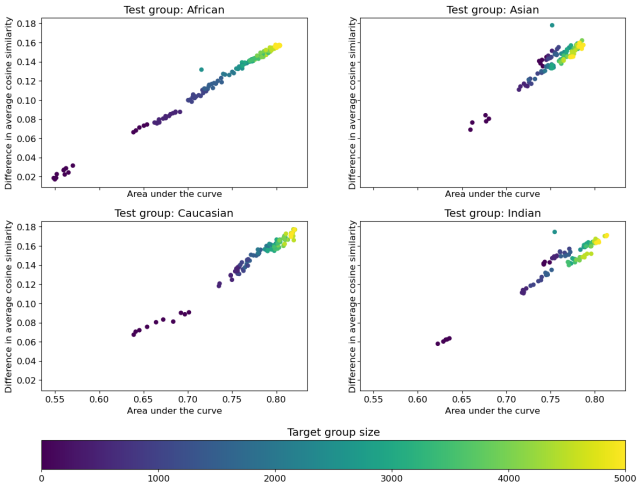


Figure 5: Difference in average cosine similarity scores between genuine (label 1) and impostor (label 0) test pairs for each training configuration run, plotted against area under curve. Color denotes the number of identities in training from the group that matches the test group, with the initial group held constant at 2000 identities. Evaluated on BUPT dataset for SE-ResNet, with consistent results for other models and datasets.

the introduction of some group X to any initial training set allows the model to better distinguish between genuine and imposter pairs from group X, which in turn, results in a higher AUC.

5 DISCUSSION

We now turn to a discussion of the broader implications of our results to (1) the addition of new training data in Section 5.1, (2) the general use of benchmark datasets in Section 5.2, and (3) the difficulty of group-level annotations in Section 5.3.

5.1 Broader implications of additional training data collection

From our analysis, we form several takeaways about the conditions and factors associated with data collection. Through simulating model retraining on the addition of new samples from a specific target group, we emphasize that we *do not* claim that this is the best method to add data, nor that data collection is the most effective way to improve a model. Instead, we aim to understand the impact of introducing data from various groups to some initial training set.

5.1.1 Results summary. Our empirical results illustrate an example in the BUPT dataset, where increasing representation from a group X initially omitted in training is not the best form of single-group augmentation to improve X's performance the most. However, across all datasets, we find that introducing data from the group that was originally worst-off obtains significant performance gains for that group. We make these performance comparisons by measuring AUC, but also recognize that AUC is an imperfect metric for capturing model behavior.

5.1.2 Importance of group annotations of both new and old data. Our results convey several implications about additional data collection. First, when augmenting training data, if we do not know the demographic group annotations of the additional samples, it is unclear how this new data will impact group-specific performance or group disparities. In other words, it is necessary to have knowledge of the demographic makeup of any additional data in

order to improve any group-specific fairness objectives. Second, our experimental analysis requires knowledge of initial performance across demographic groups. This underscores the importance of bias audits in the first place.

5.1.3 Data collection costs. At the same time, we recognize that data collection comes with various costs: Raji et al. [36] examine the ethical considerations when collecting diverse data collections, especially violations of the privacy and consent of the population being studied. As a result, targeted data collection can harm and unfairly monitor marginalized populations, as face recognition becomes used as a form of surveillance [30]. Therefore we qualify our recommendations to researchers and developers and encourage them to first assess the harms before choosing to engage in additional data collection.

5.2 Broader implications of the use of benchmark datasets

Given the limited number of publicly-available, large datasets with racial group annotations for face verification, our and other empirical findings may well be artifacts specific to particular datasets or models. For example, in Figure 2b, we observe that training on data from only the *African* or *Caucasian* group generalizes across all racial groups in the BUPT dataset, which is not replicated in the BFW and VMER datasets. The reason for this key difference between datasets is unclear and warrants further exploration. Yet because these datasets are used as benchmarks for racial bias evaluation in face recognition [14, 19], our findings are still valuable for models trained and evaluated on these same datasets.

5.2.1 Recommendations for future research on datasets and models. While the individual properties of our datasets, as discussed in Section 3.6, limits the full generalizability of our results, the unique characteristics of the datasets also leads to a strength of our study: recommendations for future research. As context for these recommendations, we observe that prior analyses on racially-balanced datasets examine one dataset instead of many. This is perhaps not surprising — and is not a criticism of past works — because these datasets are relatively new.

By studying three different datasets (across three models), we demonstrably find that there *are* important differences between datasets. Our findings here thus speak to the criticality of future work repeating evaluations like ours. For example, we recommend that future research that introduces new face recognition models to address racial bias should evaluate their models with several datasets. Similarly, we recommend that future research that introduces new datasets re-apply our methods and share the results of their analyses.

5.3 Annotations of demographic groups

For both data collection and dataset curation methods, we recognize the importance of demographic group-level annotations of data points, but also are aware of its limitations. Recent work, for instance, demonstrates that curators in each dataset follow different racial group annotation methods. Khan and Fu [25] point out that racially-annotated face recognition datasets define racial categories inconsistently, in spite of similarly named categories, and also encode stereotypes by excluding minority ethnic groups. From their evaluated datasets, the authors note that BUPT and BFW are the most consistent, due to having more images per individual.

Even simple investigation of the racial group annotation techniques reveals that some of these datasets conflate race, nationality, and ethnicity [43, 44]. Given that racial groups are socially constructed and dependent on cultural contexts [31], it is difficult to form concrete recommendations when training machine learning models that are equitable along the lines of race. However, since face recognition models have historically underperformed for people from certain racial groups [8], it is necessary to be aware of disparate treatment across groups, in spite of these groups not being well-formed. We find that our methodology still adds value and can still be performed for future datasets with differently-defined demographic groups even outside of the face recognition task.

6 CONCLUSION

In this work, we examine the group-specific performance impact of introducing additional training data from a particular racial group, if, for instance, a developer discovers that their face recognition model underperforms for some group unrepresented in its initial training set. By studying facial recognition, we acknowledge that some of its applications may create societal harm or invasions of privacy [7]. This work does not make a normative claim on the use of face recognition technologies; instead, we focus on the role that data collection plays on the model performance across groups, if these systems were to be used.

By proposing and evaluating an empirical framework that models targeted data collection, we find differences and general trends across 3 benchmark datasets and 3 standard face verification models. Some findings confirm previous intuitions about the relationship between a model's performance and the importance of data representation, while other findings reveal exceptions to these intuitions. In addition, significant differences in datasets reveal shortcomings in racial bias evaluation that use only one benchmark. We hope that our experimental results inform future instances of targeted data collection and racial bias evaluation on existing or new datasets.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. National Science Foundation under awards CCF-2045402, CNS-2205171, and UTA20-000943, as well as a grant from the Simons Foundation. We thank Ivan Evtimov for our early discussions on the project, and Josh Gardner and Kentrell Owens for their feedback on the final manuscript.

REFERENCES

[1] Jacob D Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, and Jie Zhang. 2022. Active Sampling for Min-Max Fairness. In *International Conference on Machine Learning*. PMLR, PMLR, Online, 53–65.

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

[3] Vítor Albiero, Kai Zhang, and Kevin W Bowyer. 2020. How does gender balance in training data affect face recognition accuracy?. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, IEEE, New York, NY, USA, 1–10.

[4] Vítor Albiero, Kai Zhang, Michael C King, and Kevin W Bowyer. 2021. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. *IEEE Transactions on Information Forensics and Security* 17 (2021), 127–137.

[5] Amazon 2023. *Guidelines on face attributes, Amazon Rekognition Developer Guide*. Amazon.

[6] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. 2016. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science* 6, 2 (2016), 20.

[7] Kevin W Bowyer. 2004. Face recognition technology: security versus privacy. *IEEE Technology and Society Magazine* 23, 1 (2004), 9–19.

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 2018 ACM Conference on Fairness, Accountability, and Transparency*. PMLR, ACM, New York, NY, USA, 77–91.

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

[10] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems* 31 (2018), 3543–3554.

[11] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1. IEEE, IEEE, New York, NY, USA, 539–546.

[12] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. 2021. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 66–76.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, New York, NY, USA, 214–226.

[14] Biying Fu and Naser Damer. 2022. Towards Explaining Demographic Bias through the Eyes of Face Recognition Models. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, IEEE, New York, NY, USA, 1–10.

[15] Zhiqiang Gong, Ping Zhong, and Weidong Hu. 2019. Diversity in machine learning. *IEEE Access* 7 (2019), 64323–64350.

[16] Antonio Greco, Gennaro Percannella, Mario Vento, and Vincenzo Vigilante. 2020. Benchmarking deep network architectures for ethnicity recognition using a new large face dataset. *Machine Vision and Applications* 31 (2020), 1–13.

[17] Ishaan Gulrajani and David Lopez-Paz. 2020. In search of lost domain generalization. In *International Conference on Learning Representations*. 1–9.

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

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

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, New York, NY, USA, 770–778.

[21] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the 2018 IEEE/CVF International Conference on Computer Vision*. IEEE, New York, NY, USA, 7132–7141.

[22] Anil K Jain and Stan Z Li. 2011. *Handbook of face recognition*. Vol. 1. Springer, New York, NY, USA.

[23] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 306–316.

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

[25] Zaid Khan and Yun Fu. 2021. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 587–597.

[26] Terence Liu. 2023. How we store and search 30 billion faces.

[27] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, New York, NY, USA, 212–220.

[28] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods* 47 (2015), 1122–1135.

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

[30] Paul Mozur. 2019. One month, 500,000 face scans: How China is using AI to profile a minority. *The New York Times* 14 (2019), 2019.

[31] Brian K Obach. 1999. Demonstrating the social construction of race. *Teaching Sociology* 27, 3 (1999), 252–257.

[32] Jaspar Pahl, Ines Rieger, Anna Möller, Thomas Wittenberg, and Ute Schmid. 2022. Female, white, 27? Bias evaluation on data and algorithms for affect recognition in faces. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 973–987.

[33] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336.

[34] P Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J Rauss. 1998. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing* 16, 5 (1998), 295–306.

[35] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 429–435.

[36] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 145–151.

[37] John Roach. 2018. Microsoft improves facial recognition technology to perform well across all skin tones, genders.

[38] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. 2020. Face recognition: too bias, or not too bias?. In *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, New York, NY, USA, 0–1.

[39] Esther Rolf, Theodora T Worledge, Benjamin Recht, and Michael Jordan. 2021. Representation matters: Assessing the importance of subgroup allocations in training data. In *International Conference on Machine Learning*. PMLR, PMLR, Online, 9040–9051.

[40] Ignacio Serna, Aythami Morales, Julian Fierrez, Manuel Cebrian, Nick Obradovich, and Iyad Rahwan. 2019. Algorithmic discrimination: Formulation and exploration in deep learning-based face biometrics. *arXiv preprint arXiv:1912.01842* abs/1912.01842 (2019).

[41] Ignacio Serna, Aythami Morales, Julian Fierrez, and Nick Obradovich. 2022. Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence* 305 (2022), 103682.

[42] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation by joint identification-verification. *Advances in Neural Information Processing Systems* 27 (2014).

[43] Mei Wang and Weihong Deng. 2020. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the 2020 IEEE/CVF International Conference on Computer Vision*. IEEE, New York, NY, USA, 9322–9331.

[44] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. 2019. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*. IEEE, New York, NY, USA, 692–702.

[45] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*. IEEE, New York, NY, USA, 5310–5319.

[46] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*. Springer, Springer, New York, NY, USA, 499–515.

[47] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097* abs/1902.11097 (2019).

[48] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* 64, 3 (2021), 107–115.