Unsupervised and Semi-supervised Bias Benchmarking in Face Recognition

Alexandra Chouldechova[†*] Siqi Deng[†] Yongxin Wang
Wei Xia[*] Pietro Perona

AWS AI Labs

Abstract. We introduce Semi-supervised Performance Evaluation for Face Recognition (SPE-FR). SPE-FR is a statistical method for evaluating the performance and algorithmic bias of face verification systems when identity labels are unavailable or incomplete. The method is based on parametric Bayesian modeling of the face embedding similarity scores. SPE-FR produces point estimates, performance curves, and confidence bands that reflect uncertainty in the estimation procedure. Focusing on the unsupervised setting wherein no identity labels are available, we validate our method through experiments on a wide range of face embedding models and two publicly available evaluation datasets. Experiments show that SPE-FR can accurately assess performance on data with no identity labels, and confidently reveal demographic biases in system performance.

Keywords: Algorithmic bias, Semi-supervised evaluation, Face verification, Bayesian inference

1 Introduction

Measuring a system's accuracy and its algorithmic bias prior to deployment is a cornerstone of responsible AI [22, 43, 54, 31, 8]. This is especially important in the context of computer vision applications, such as face analysis and recognition [8, 5, 53, 23, 45]. Assessing system performance and bias is not a one-off affair. There is no guarantee that a model that is found to perform equally across ethnic groups and genders, say, on a given benchmarking dataset will continue to do so in a different use case. This is because system operating characteristics depend on the statistics of data, which generally differ across use cases. Ideally, each organization that plans to adopt AI technology would conduct a performance and bias evaluation of the proposed system in each use case prior to deployment.

In practice, however, benchmarking vision algorithms is a tall order for most organizations. A key problem is acquiring appropriately annotated test data that mirror the statistics of the use case. In particular, annotating accurate test data is expensive and time-consuming. *This pain is particularly acute in face recognition*, which relies on identity annotations. Reliable identity ground truth for

[†] Equal contribution. Corresponding author: Siqi Deng, Email: siqideng@amazon.com.
[*] Work done when at Amazon.

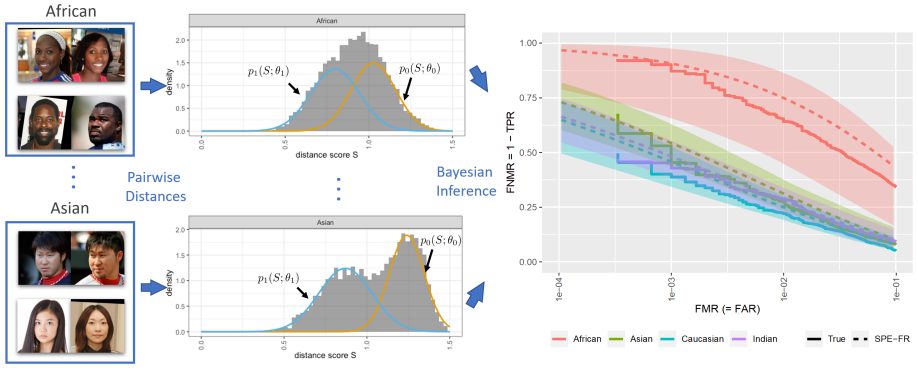


Fig. 1: **SPE-FR methodology and sample results.** SPE-FR models the similarity scores S_{ij} in each group as coming from a mixture of two parametric distributions, $p_1(S;\theta_1)$ for matching identities and $p_0(S;\theta_0)$ for impostors. Bayesian inference estimates parameters θ_0, θ_1 and proportion of true matches, π (middle). The right panel shows corresponding unsupervised error estimates, where SPE-FR was applied to the Racial Faces in the Wild (RFW) [57] dataset. SPE-FR estimates using no id labels are shown as dashed lines with credible confidence bands overlaid. Ground truth performance from fully labeled data is shown in solid lines ("True"). The face recognition model ("AA") being tested was trained on BUPT-BalancedFace [58] while leaving out the "African" group in order to artificially introduce bias (Sec. 4.1). SPE-FR correctly and confidently reveals racial bias, i.e. accuracy on the "African" group is worse than the other groups.

faces is extremely difficult to obtain. Typically it is obtained by government organizations through access to identity documents [23], or through the subjective judgment of human annotators [37]. The first approach is not available to most organizations, and the second is fraught and highly unreliable because people are not accurate and often biased in recognizing the faces of strangers [42, 2, 40, 52].

This raises the question: *What's the best an organization can do with limited resources?* We believe that the most practical scenario for face recognition is the following: data (images of faces) is plentiful, but collecting a large number of high-quality identity annotations may not be feasible. Instead, it is feasible to collect annotations as to membership in demographic or morphological groups, e.g. perceived ethnicity, gender, body shape, etc. Indeed, prior studies have found that human annotators provide reasonably reliable and consistent annotations of these types of characteristics [5, 34]. Thus, we pose the following challenge: Given a large data set of face images, where each image is annotated for group membership *but not identity*, estimate the overall performance and algorithmic bias of a face recognition system on the data. We focus here on the question of whether this is feasible at all, and how to do it.

To summarize our contributions, we introduce in this paper a method we call SPE-FR (Semi-supervised Performance Evaluation for Face Recognition), which enables one to accomplish precisely this task in the setting of face verification. We believe SPE-FR is the first semi- and un-supervised method for evaluat-

ing bias in face recognition algorithms. SPE-FR produces point estimates and uncertainty bands (specifically, Bayesian posterior credible bands) for common performance metrics used in face recognition settings, and enables the metrics to be compared across different subgroups of the population as part of an algorithmic bias assessment. Fig. 1 previews the kinds of results we are able to obtain with no access to identity information. Our experiments demonstrate the surprising result that it is possible to reliably estimate the performance and bias of face verification systems *even when no test data identity annotations are available* (what we term the *unsupervised* setting). While the methodology is presented in the more general semi-supervised setting where partial identity annotations may be available, the experimental results presented in the main paper focus on what we believe to be the more interesting and realistic unsupervised setting.

2 Related work

Algorithmic bias in face recognition. Prior work has studied demographic bias in the performance of face recognition systems, exploring whether systems have disparate error rates across different demographic groups [53, 36, 48, 1, 51, 45]. Of particular note is NIST's Face Recognition Vendor Test (FRVT), which is conducted on US government data [23]. Findings of bias have also inspired a body of work on bias mitigation strategies [33, 28, 59, 38, 32, 35]. To meet the many calls for more thorough bias benchmarking of face recognition systems, a number of public data sets have become available that contain demographic information alongside identity information [57, 39, 47]. Meanwhile, it has become clear that accuracy and bias measurements on a given dataset may not generalize to new domains and use cases [6, 5, 61, 17]. Therefore, we believe that, ultimately, accuracy and bias measurements may need to be carried out per use case.

Semi- and un-supervised performance evaluation on new domains. The absence of high-quality annotated target domain test data poses a persistent obstacle to the thorough performance and bias evaluation of AI systems prior to new deployments. In classification, methods have been introduced to estimate model performance on unlabelled or partly labelled test data. [60] developed the *Semi-supervised Performance Evaluation* (SPE) method for estimating the performance of classification systems, [18] proposed training a model to predict system error across different target data sets, and [24] proposed a method based on differences in model scores. [21] learns a confidence threshold such that the proportion of unlabelled examples exceeding the threshold is a reasonable estimate of model accuracy. [32] introduced the *Bayesian Calibration* (BC) method that learns a calibration function using a small sample of labelled data and then applies it to unlabelled examples to estimate model performance and bias.

Our proposed method, SPE-FR, differs in several ways from existing work. (i) None of these methods were developed for face recognition settings or consider tasks such as 1:1 verification, which is not inherently a classification task; and only [32] directly considers bias assessment. (ii) Except for [32], the methods do not output uncertainty assessments (e.g., confidence intervals or posterior credi-

ble regions) alongside point estimates of performance. (iii) None of the methods have been assessed in high-performance regimes of the kind that are relevant in face verification. That is, the methods are developed primarily for estimating overall accuracy and are not tested in challenging settings such as estimating false non-match rates at false match rates well below 0.01. In our experimental evaluation of BC [32], we found that BC performs well at estimating overall accuracy, and yet it is not suitable for estimating metrics relevant to benchmarking face recognition systems.

SPE-FR takes inspiration from the SPE [60] method. In particular, we adopt a similar approach to modeling model scores (in our case similarity or distance scores) as following user-specified parametric distributions. Our work improves upon SPE [60] in several important ways. SPE does not consider the unsupervised setting with no labels, face recognition systems, or bias evaluation. Whereas they focused on estimating Precision-Recall curves across the full range of recall (TPR), we focus on low False Match Rate (FMR$<$ 0.01) and low False Non-Match Rate (FNMR$<$ 0.01) regimes that are not considered in this or other work. We discuss further innovations and adaptations that went into SPE-FR in our methodology section below.

3 Semi-Supervised Bias Evaluation Methodology

In a typical modern face recognition system, a face embedding model ϕ is applied to extract identity information from a pre-processed (e.g., cropped and aligned) face image $x_i, i \in \{1, \ldots, I\}$ to produce a feature embedding vector $\mathbf{z}_i = \phi(x_i) \in \mathbb{R}^d$, for a choice of feature dimension, d. A common use case for face recognition models is 1:1 *face verification*. Face verification (FV) aims to determine whether two face images, x_i and x_j, belong to the same person. This is often done by applying a similarity or distance function to the embedding vectors $\mathbf{z}_i, \mathbf{z}_j$, and calling the pair a "match" if the similarity exceeds a pre-specified threshold (equivalently, if the distance falls below some threshold). Common functions include the *cosine similarity* $S_{cos}(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i \cdot \mathbf{z}_j / \|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2$, and the *Euclidean distance* $D_2(\mathbf{z}_i, \mathbf{z}_j) = \|\mathbf{z}_i - \mathbf{z}_j\|_2$.

The most common metrics used in evaluating the performance of a face verification system are the False Non-Match Rate (FNMR) and False Match Rate (FMR) [23]. Let $S_{ij} = S(\mathbf{z}_i, \mathbf{z}_j)$ denote the similarity score of images i and j and $Y_{ij} \in \{0, 1\}$ denote the ground truth indicator that i and j correspond to the same person. The FNMR and FMR at threshold τ are then defined as:

$$\text{FNMR}(\tau) = \mathbb{P}(S_{ij} < \tau \mid Y_{ij} = 1), \qquad \text{FMR}(\tau) = \mathbb{P}(S_{ij} \geq \tau \mid Y_{ij} = 0). \quad (1)$$

The FNMR-FMR curve of FNMR plotted as a function of FMR is a common summary of the accuracy of the FV system and its bias across different demographics. A common one-number summary is the FNMR at a particular FMR level (e.g., FNMR@FMR=10^{-3}). If ground truth identity labels are available (i.e., Y_{ij} is known for all pairs) then these quantities can be estimated through empirical proportions. E.g., $\widehat{\text{FNMR}}(\tau) = \frac{1}{N_1} \sum_{i<j} \mathbb{1}(S_{ij} < \tau) Y_{ij}$, where

$N_1 = \sum_{i<j} Y_{ij}$ is the number of true matches among all image pairs and $\mathbb{1}(x)$ is the indicator function. Our proposed method (Sec. 3.1) enables the estimation of such metrics even when Y_{ij} is unknown for most, or even *all*, pairs.

3.1 Semi-Supervised Performance Evaluation for Face Verification

We now proceed to present the formalism of the SPE-FR method. The intuition behind SPE-FR is twofold. First, any performance metric that involves the scores S and matches indicator Y can be computed from the joint distribution of (S, Y), as we will detail in Sec. 3.4. Second, empirical evidence suggests that the match-conditional distributions of $S \mid Y$ are well-behaved across a range of face embedding model architectures and test datasets, and may be estimated through mixture modeling *even when* Y *is unknown*.

More formally, in SPE-FR we model the similarity scores S within match class $Y = y$ as following some user-specified distribution $p_y(s \mid \theta)$, parameterized by $\theta \in \mathbb{R}^p$. In Sec. 3.3 we describe a specific parametric family that, we find, works well for face recognition systems. Let n denote the number of images and let N denote the total number of *image pairs*. Let $\mathcal{L} = \{(i,j) : i < j, Y_{ij} \text{ is known}\}$ denote the subset of image pairs for which identity match is known, and let $Y_{\mathcal{L}}$ denote the set of known Y's. In the unsupervised setting that we focus on in our experiments, where we assume no identity annotations are available, $\mathcal{L} = \emptyset$. Let $\mathcal{U}$ denote all image pairs for which identity match is *unknown*—in the unsupervised setting, $\mathcal{U}$ is *all* image pairs. Lastly, let $\pi = \mathbb{P}(Y_{ij} = 1)$ denote the proportion of true matches among all image pairs. In this notation, we can write down the likelihood of the observed data given unknown parameters π and θ as,

$$p(S, Y_{\mathcal{L}} \mid \pi, \theta) = \prod_{(i,j)\in\mathcal{L}} \pi^{y_{ij}}(1-\pi)^{1-y_{ij}} p_{y_{ij}}(s_{ij} \mid \theta_{y_{ij}}) \times \prod_{(i,j)\in\mathcal{U}} \left((1-\pi)p_0(s_{ij} \mid \theta_0) + \pi p_1(s_{ij} \mid \theta_1)\right), \quad (2)$$

where we think of $\theta = (\theta_0, \theta_1)$ as parameterizing the two distinct class-conditional densities. There are many approaches one can take to estimating the parameters (π, θ). For instance, one can attempt maximum likelihood estimation on Eq. (2) through methods such as the EM algorithm [41]. In this work, we take a Bayesian inference approach. More precisely, given a prior distribution $p(\pi, \theta)$, we base our inference on the posterior distribution,

$$p(\pi, \theta \mid S, Y_{\mathcal{L}}) \propto p(S, Y_t \mid \pi, \theta)\, p(\pi, \theta). \quad (3)$$

The posterior distribution on (π, θ) then implies a posterior for any performance metrics that can be calculated from the joint distribution of (S, Y).

This approach is based on the same philosophy as the Semi-supervised Performance Evaluation (SPE) method introduced in [60] for evaluating binary classification models. SPE-FR includes a number of innovations, which we outline in the next few subsections. To begin with SPE [60] does not consider the *unsupervised* setting, does not consider *Face Recognition (FR) systems*, nor *groups* (i.e., no bias evaluation). We introduce and study the unsupervised case, i.e. that **performance and bias evaluation** are possible with **no** Y **labels** (this is a surprising result), and focus on FR. Moreover, as we now briefly discuss, the FR

regime required more sophisticated statistical methods to address acute challenges. (i) Estimating accuracy at the low FNMR and FMR regime differs from the accuracy estimation when errors are relatively frequent. FR accuracy is often much higher than in typical classification problems, so correct approximation of the tails of the class-conditionals is crucial. Existing methods operate in much higher error settings. E.g., [60] showed results for Precision-Recall curves (not ROC or FNMR-FMR) across the whole $[0, 1]$ recall range, which can hide poor performance at the edges. (ii) Distribution tail behavior. In our extensive empirical analysis (Supplemental C) we found that the parametric families considered in [60] fail to model distance/similarity scores output by FR systems. We use instead the "two-piece" (TP) family of scale-location-shape distributions from the statistics literature on heavy-tailed distributions, which we demonstrate do a good job of approximating FR system scores across different models and different data sets (MORPH and RFW). We also provide a tailored prior specification. Our two-piece distributions approximate well the ground truth across different datasets (Supplemental C). (iii) We consider highly imbalanced data: [60], [32], and other methods do not handle extreme class imbalance. In real world FR studies, non-matches ($Y = 0$) far outnumber true matches ($Y = 1$). To adapt SPE to this highly imbalanced setting we found that an informative prior on π, the proportion of true matches, is often necessary. We show how techniques from false discovery rate control in statistical genetics can be used to estimate π and inform the prior (Supplemental D).

3.2 Bias Evaluation in Face Recognition Systems

We are interested in evaluating not only overall performance but bias as well. The most common way to assess a face verification system for *bias* is to compare performance metrics across different groups. For the purpose of this paper we assume that for each image i we have a *known* or *inferred* group membership variable $A_i \in \{1, ..., K\}$ (e.g., gender, race, combinations thereof, etc). In cases where A_i is inferred, such as through the use of a classifier, we will think of our method as estimating performance for the *inferred* rather than the true groups.*

There are two principal ways of extending SPE to perform bias evaluation. In the "stratified" approach, one can perform SPE-FR separately within each group, and then assess differences in the resulting group-level performance estimates. Alternatively, one can apply Bayesian hierarchical modeling to pool information across data from different groups in estimating parameters. Specifically, one can introduce parameters ν_Y to form the hierarchical specification:

* Some have developed methods for estimating group fairness metrics in the presence of noisy or inferred group membership labels [10, 44, 4, 9]. Understanding how SPE-FR performs with respect to the true unknown groups using inferred group information is an interesting and important question, but beyond the scope of the present work.

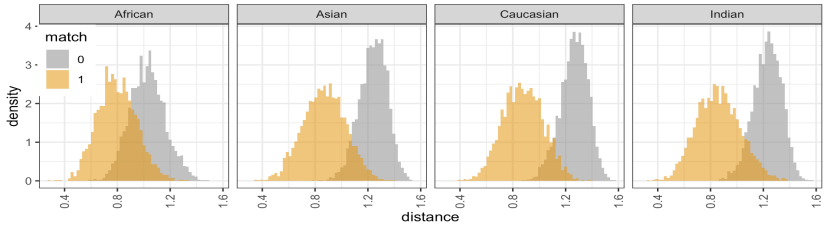


Fig. 2: **Normal model.** (Top) Empirical distribution of the class-conditional distances computed by the AA model (see Tab. 1) on the RFW dataset (each face pair is from the same ethnicity, "match" indicates a genuine identity match).

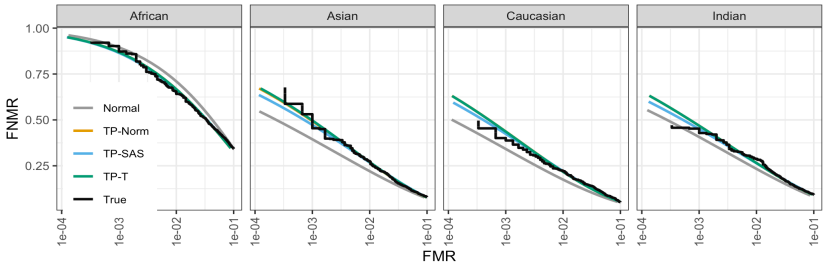


Fig. 3: **Parametric estimates of algorithmic error compared with ground truth (Black).** Comparison of Normal, Two-piece Normal (TP-Norm), Two-piece sinh-arcsinh (TP-SAS), and Two-piece Student-t (TP-T) parametric approximations of the AA model on the RFW (Sec.3.2). The TP distributions capture skewness and kurtosis (i.e., heavy-tailedness) in the class-conditional score distributions, producing better approximations of the true performance curves. TP-Norm and TP-T curves are visually indistinguishable from this data.

$$\pi_A \mid A \sim \text{Beta}(\alpha, \beta) \qquad Y \mid A \sim \text{Bernoulli}(\pi_A) \tag{4a}$$

$$\nu_Y \mid Y \sim p(\nu_Y) \qquad \theta_{Y,A} \mid Y, A, \nu_Y \sim p(\theta_{Y,A} \mid \nu_Y) \tag{4b}$$

$$S \mid Y, A, \theta \sim p_{Y,A}(s \mid \theta_{Y,A}) \tag{4c}$$

We provide details and experiments on the model specification in Sec. 3.3 and Bayesian inference strategy in Sec. 3.4 for the stratified approach.

3.3 Parametric conditional distributions and priors

Distribution and prior selection for fitting. To have SPE-FR perform effectively, we need to have good parametric models for the match-conditional distributions $S \mid Y, \pi, \theta$. In Fig. 2 we show class-conditional score densities for the 4 most prevalent race/ethnicity groups in the RFW data. The densities are unimodal and most are close to symmetric. However, a closer analysis of the data using Normal-QQ plots (Supplemental C) reveals that only the scores for

the African group approximately follow a Normal distribution. The rest are significantly skewed. It is therefore important that our choice of parametric family $p_Y(s \mid \theta_Y)$ be able to capture at least skewness.

While there are classical skewed parametric families of distributions such as the Gamma, log-Normal and Weibull, and SPE-FR can be used with any parametric families, we found through extensive experimentation that these models were often a poor fit to the observed data. Our experiments revealed that so-called *two-piece* distributions [49] provided a much better approximation, a finding that appears robust to the choice of model architecture and data set. These distributions, which have not previously been considered in the computer vision literature, allow flexible control of skewness and tail behavior in a single simple parametric family[19, 20]. Given a symmetric unimodal density, f, centered at 0, the parametric family of *two-piece distributions generated by* f is given by:

$$g(x;\mu,\sigma_1,\sigma_2,\delta) = \begin{cases} \frac{2}{\sigma_1+\sigma_2} f\left(\frac{x-\mu}{\sigma_1};\delta\right) & x < \mu \\ \frac{2}{\sigma_1+\sigma_2} f\left(\frac{x-\mu}{\sigma_2};\delta\right) & x \geq \mu \end{cases}. \tag{5}$$

Here μ denotes the mode of the distribution, the σ_k control skewness, and δ is a shape parameter that controls kurtosis ("heavy-tailedness"). For instance, when f is chosen to be the Student-t distribution, δ denotes the degrees of freedom. Figure 3 shows that two-piece distributions provide a much closer approximation than the standard Normal to the true FNMR-FMR curves on the RFW data. The corresponding plot for Morph (Supplemental C) shows even greater improvements in approximation accuracy.

Prior specification We now outline the model specification used in our experiments in the stratified setting for a given group (and hence omit the group membership A to simplify notation). Let $\tau_{jk} = 1/\sigma^2_{jk}$ denote the precision parameter in match class $Y = j \in \{0,1\}$ of the $k \in \{1,2\}$ component of the two-piece distribution. Here $k = 1$ denotes the parameters for the left half of the two-piece distribution and $k = 2$ denotes the right. Our model specification is:

$$\pi \sim \text{Unif(L, U)} \qquad Y \sim \text{Bernoulli}(\pi) \tag{6a}$$

$$\mu_j \sim N(\eta_j, 0.25^2) \qquad \tau_{jk} \sim \text{Gamma}(10, \beta^{\tau}_{\text{jk}}) \tag{6b}$$

$$\delta_j \sim \text{Gamma}(\alpha^{\delta}_{\text{j}}, \beta^{\delta}_{\text{j}}) \qquad S \mid Y, \theta \sim TP(\mu_Y, \tau_{1Y}, \tau_{2Y}, \delta_Y), \tag{6c}$$

where TP denotes a two-piece distribution, and the Gamma distribution is shape-rate parameterized, so that $\mathbb{E}[\text{Gamma}(\alpha,\beta)] = \alpha/\beta$. For the experiments in the main paper, we show results using a two-piece Student-t (TP-T) for RFW and a Two-piece sinh-arcsinh (TP-SAS) for MORPH. The TP-SAS has previously been used specifically to model skewed heavy-tailed data [50], which is precisely the kind of distribution we expect to see for non-match scores.

Table 1: **Training sets derived from BUPT-BalancedFace.** Starting from the full training (FT) set, we obtained a number of training sets by removing selectively some identities. These "ablated" sets are used to produce corresponding face recognition models. We hypothesize that these bias-controlled models contain different types and degrees of biases, thus suitable for inspection purposes (as verified in Fig. 4). *: IDs counted from majority voting of image predictions.

Name	FT	RT	M	EA	CC	AA
Left-out set	None	Random 90%	Male	East Asian	Caucasian	African
# Identities	28000	2800	7541*	21000	21000	21000

3.4 Bayesian Inference Strategy

Markov Chain Monte Carlo (MCMC) inference. Because the posterior distribution of the parameters $(\pi, \theta) = (\pi, \mu_j, \tau_{jk}, \delta_j)$ in model specification (6) is analytically intractable, we rely on MCMC methods to obtain a sample from the posterior distribution. For our experiments we used the BayesianTools [27] implementation of the Differential Evolution MCMC sampler (DEzs) originally proposed in [7]. Given a posterior sample $\{(\pi_i, \theta_i)\}_{i=1}^T$, we can calculate values of FNMR and FMR at a given threshold τ by evaluating the tail probabilities:

$$\mathrm{FMR}_i(\tau;\theta) = \int_\tau^\infty p_0(s;\theta_{i0})ds \qquad \mathrm{FNMR}_i(\tau;\theta) = \int_{-\infty}^\tau p_1(s;\theta_{i1})ds$$

For a given metric $M(\tau;\pi,\theta)$, e.g. FNMR at a threshold τ, we obtain point estimates of M using the posterior mean $\frac{1}{T}\sum_1^T M(\tau;\pi_i,\theta_i)$. To obtain $(1-\alpha)\%$ posterior credible intervals, we take the $\alpha/2$ and $1-\alpha/2$ quantiles of $\{M(\tau;\pi_i,\theta_i)\}$.
Hierarchical clustering procedure for constructing confidence bands for performance curves. We can also obtain posterior credible confidence bands for entire FNMR-FMR curves. The procedure we present here is specifically tailored to produce non-trivial confidence bands at low FMR values. Standard approaches will generally produce FNMR confidence intervals of the form $[0, u)$: i.e., the lower endpoint of the FNMR band will be 0. Let $\mathrm{FNMR}(\zeta;\theta_i)$ denote the FNMR curve as a function of $\zeta = \mathrm{FMR}$ at parameter values θ_i. To construct the confidence band, we first apply agglomerative hierarchical single-linkage clustering with Canberra distance to the (logged) $\mathrm{FNMR}(\zeta;\theta_i)$ curves at a grid of ζ values evenly spaced on the logarithmic scale. We then cut the cluster tree at a level such that at least $(1-\alpha)\%$ of the curves are contained in the largest component. The envelope generated by the curves in that component provides a $(1-\alpha)\%$ confidence band: if the parametric assumptions of the model are met, it provides uniform coverage for the entire FNMR-FMR curve, not simply pointwise coverage.

4 Experiments

We mimic the setting in which an organization has a large collection of face images x_i, many of which are of the same people. We assume that we *do not*

Table 2: **Models trained under various settings.** We trained face embedding models on various popular training datasets, model architectures and loss functions and verified the effectiveness of SPE-FR (Sec. 4.2).

Training data	IMDB [55]	DeepGlint [11]	BUPT-BalancedFace [58]		
Loss function	Sub-Center Arc	Sub-Center Arc	Sub-Center Arc	CosFace [56]	ℓ_2-Softmax [46]
Architecture	R101	R101	R18	R101	R101 [29]
AA ablation	✗	✗	✓	✓	✓

have access to identity indicators Y_{ij}, and that we have access to true or inferred group label identities A_i for the purpose of assessing bias. For privacy reasons, organizations wishing to assess the performance of a third-party system on their data may be unwilling or prohibited from sharing image data with system developers, who are in many cases, commercial vendors. Since SPE-FR relies only on the scores S_{ij}, it is sufficient for owners of the test data to provide just the scores from running the 1:1 verification system on their data. Sharing of the raw face images x_i is not required; the resulting S_{ij} and A_i are sufficient to run SPE-FR. *The goal of our experiments is to evaluate whether the SPE-FR algorithm can assess group-level performance well enough to reflect existing trends and to reveal demographic bias in system performance.* We compare the results of SPE-FR to ground truth assessed using fully identity-annotated data. SPE-FR is applied in the *unsupervised* setting where $N_{\mathcal{L}} = 0$, i.e. *no identity labels are available.* "Ground truth" values are calculated from known Y_{ij} for all image pairs, to which SPE-FR does not have access.

Our experiments are performed on the RFW [57] and MORPH [47] datasets. Details about the test sets and protocols, specifics on the Bayesian inference MCMC configurations and hyperparameters estimation process (Sec. 3.4), as well as ablation study on $N_{\mathcal{L}}$ comparing the unsupervised setting ($N_{\mathcal{L}} = 0$) with the semi-supervised ones ($N_{\mathcal{L}} > 0$), may be found in the Supplemental B.

4.1 Face Embedding Model Training for Bias Analysis

We apply our method to two sets of face recognition (FR) models: (i) demographically "biased" models trained using a common architecture to assess whether SPE-FR can reveal bias in the performance of 1:1 verification systems; and (ii) models trained with various datasets, network architectures, and loss functions that allow us to examine whether SPE-FR performs well across different settings.

Model Training with Controlled Biases. We employ BUPT-BalancedFace dataset [58] as our training set. This dataset provides images across 4 race/ethnicity groups (African, Asian, Caucasian, and Indian), each with 7,000 identities. We adopt the popular state-of-the-art Sub-center ArcFace [12] method in our face recognition model and employ a variant of ResNet [30] as our feature extractor.

In order to explore bias, we train a set of models by using all but one particular race or gender group from the BUPT-BalancedFace dataset. In this way, we can obtain several different training sets with different types or degrees of

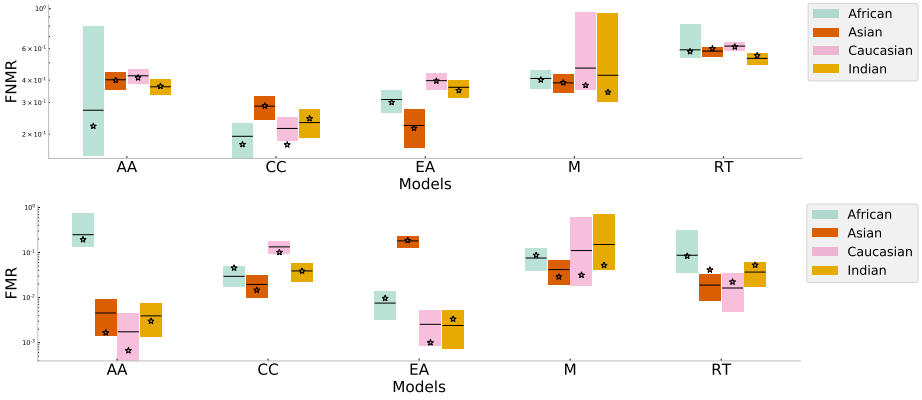
bias. Configurations of the training sets is detailed in Tab. 1*. We evaluate each model's performance and bias on the test sets and use ground truth labels to compute the *true* performance (in the form of FNMR-FMR curves) for comparing with SPE-FR estimates.

Model Training for Generalization Validation. To validate if SPE-FR is applicable to face embedding models trained across different settings, we test on the second set of FR models trained to represent state-of-the-art for popular training datasets, model architectures, and loss functions. See Tab. 2 for details. We share more details on model training in Supplemental F.

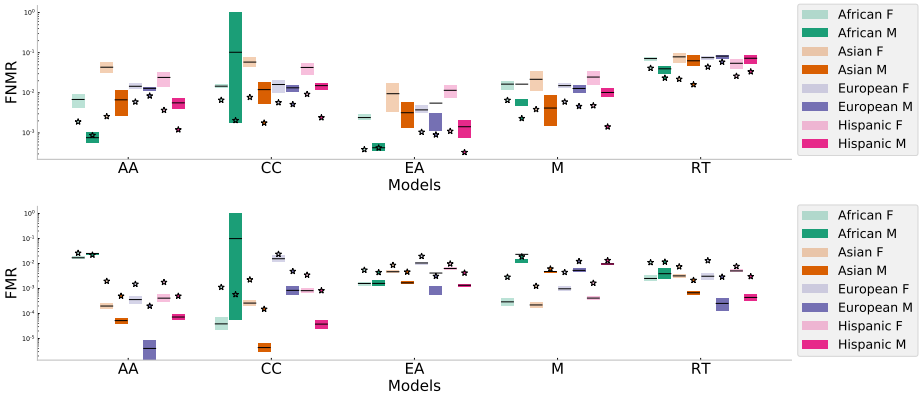
4.2 Result and Analysis

Results on Models Trained with Controlled Biases. To validate the effectiveness of SPE-FR, we evaluate the models trained with controlled biases (Tab. 1) on several test datasets where face verification performance is compared across cohorts of face pairs representing different genders and races. **Fig. 1** shows the results of applying unsupervised SPE-FR to estimate the RFW data performance of the system trained excluding African faces from the training data. SPE-FR produces estimates (dashed lines) and 89% posterior credible regions (shaded bands) for the full FNMR vs FMR curves for each of the 4 racial/ethnic groups coded in the data. While the confidence bands are wide, in part due to the relatively small dataset, SPE-FR has good coverage of the true (solid) curves and correctly reveals that the system under-performs on African faces. **Fig. 5** shows the unsupervised SPE-FR results for MORPH disaggregated by race/ethnicity and gender. Even though the confidence bands fail to capture the true (solid) curves, they are in the right ballpark, and SPE-FR still confidently and correctly identifies significant gender bias across all race/ethnicity groups. As we show in the Supplemental C), poor confidence band prediction on MORPH is not due to the TP-SAS parametric model being an overall poor fit to the score distributions within each group and match class. In particular, if all labels Y_{ij} are made available, the TP-SAS distribution results in good approximations of the FNMR-FMR curve on the MORPH data. Confidence band prediction of SPE-FR on MORPH can in principle be improved by using SPE-FR in the semi-supervised setting and obtaining Y_{ij} for pairs with borderline similarity scores. However, this requires obtaining true identity annotations for pairs that are somewhat difficult for the system to distinguish, and which therefore may be difficult for humans to correctly annotate. **Fig. 4a and Fig. 4b** show estimated FNMR at an overall target FMR for five different models (Tab. 1) on the RFW and MORPH datasets. The overall target FMR is set to be 0.001 on MORPH and 0.005 on RFW. The SPE-FR estimates are highly accurate for RFW but slightly overestimate the error on MORPH across the board. The results are nevertheless useful: They reflect trends in the ground truth, such as the poor performance of the RT model compared to others.

* BUPT-BalancedFace does not provide gender annotations, we generated pseudo labels from open-source face analysis repository Insightface [25, 3, 13, 15, 26, 16, 14].



(a) **Ground truth bias vs Unsupervised SPE-FR estimates on RFW.**



(b) **Ground truth model vs Unsupervised SPE-FR estimates on MORPH.**

Fig. 4: **Unsupervised SPE-FR model bias estimates vs ground truth.** We show the error rates of models that were trained using five different training settings (Tab. 1) derived from BUPT [58]. SPE-FR was used to assess the error rates of each model on RFW (4a) and MORPH (4b). We show False Match Rate (FMR) and False Non-Match Rate (FNMR). We set the number of labeled instances $N_{\mathcal{L}} = 0$, for unsupervised estimates. Here we apply one decision threshold per model, across all demographic groups. The threshold is selected so that FMR=0.05 for RFW and FMR=0.001 for MOPRH over the entire dataset containing all demographic groups. Then FMR and FNMR at each model's selected threshold are produced on each group-specific benchmark. The $\star$ indicates the ground truth performance (FMR or FNMR) measured with fully labeled data. The horizontal lines and colored bars are the corresponding performance point estimates and 89% confidence bands. On RFW, unsupervised estimates of SPE-FR are on target (stars within the confidence intervals) for both FNMR and FMR, and the error rate differences across groups are predicted correctly. On the MORPH dataset, SPE-FR overestimates the ground truth FNMR and underestimates the ground truth FMR. Since the overestimate and underestimate are consistent across settings, the error rate differences across groups are predicted correctly (as confirmed in Fig. 5).

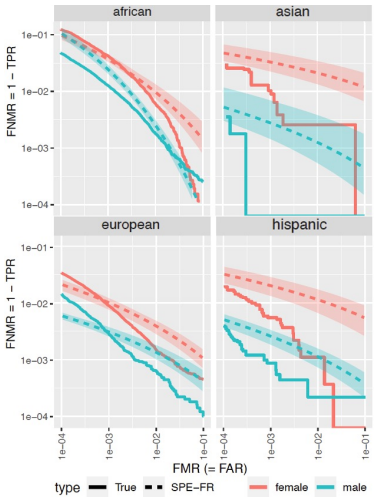


Fig. 5: **Unsupervised SPE-FR estimates of the FNMR vs FMR curve.** AA model applied to MORPH, SPE-FRE estimates shown as dashed lines with credible confidence bands overlaid (see Sec 3.4 for details), ground truth ("True") performance from fully labeled data shown in solid lines. The jagged shape in the true performance curve is caused by insufficient sample size at the operating ranges. SPE-FR correctly reveals gender bias within each ethnic group, where females are recognized less accurately.

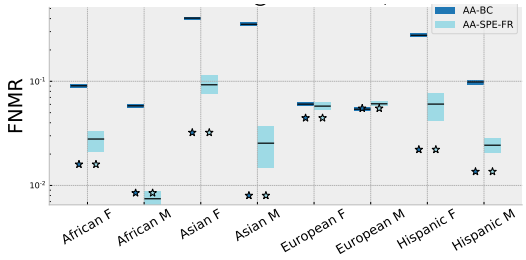


Fig. 6: **Comparison with BC adapted to face verification systems.** Evaluation of AA model on MORPH. Results show Bayesian Calibration [32] (left) and SPE-FR (right) estimates for each group, with corresponding Bayesian confidence intervals. True performance from full label is marked with stars. Both BC and SPE-FR over-estimate FNMR, but our estimates are significantly more close to the ground truth.

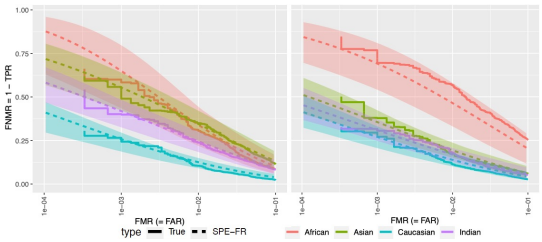


Fig. 7: **Unsupervised SPE-FR estimates of the FNMR vs FMR curve.** Two models are evaluated on RFW: (left) FR model trained on IMDB with Sub-Center Arcface loss and Res101 backbone; (right) FR model trained on BUPT with CosFace loss and Res101 backbone.

Comparison to Bayesian Calibration [32] We compare SPE-FR to the Bayesian Calibration (BC) [32] method by analogously recasting verification in a manner such that BC can be applied despite originally developed for binary classification (implementation details in Supplemental H). We apply the BC method under setting $N_{\mathcal{L}} = 256$ to the "AA" model. On MORPH, BC's estimate of the *overall system accuracy* is close to the ground truth (Supplemental H). However, at the same threshold introduced in Fig. 4, we can see from Fig. 6 that BC's estimates of FNMR underperform SPE-FR by a large margin, especially considering that the BC is applied with partial annotation ($N_{\mathcal{L}} = 256$), whereas SPE-FR with none ($N_{\mathcal{L}} = 0$). There are two important takeaways from these results: (i) Methods that are effective at estimating overall performance

may fail to accurately estimate metrics that are relevant to face verification (e.g., FNMR at low FMR); (ii) SPE-FR performed with no identity annotations can outperform methods informed by partial identity annotation.

Results on Generalization Capability Test We evaluate the effectiveness of SPE-FR for performance estimation and bias detection of face recognition models trained under a wide range of settings (Tab. 2). We show two examples in Fig. 7 (the rest may be found in Supplemental A). As is seen, unsupervised SPE-FR ($N_{\mathcal{L}} = 0$) produces estimated curves close to the true (fully labeled) performance curves, and the confidence bands provide good coverage of the ground truth values. This indicates that SPE-FR generalizes well to different training pipelines. On the three models trained on BUPT-BalancedFace data with the leave-one-out setting, we again observe similar trends as Fig. 1 where the performance and bias estimates are dead-on. As is seen, SPE-FR performs well across different optimization functions model architectures. This also provides evidence that the two-piece family of distributions we propose in SPE-FR is a good choice across a range of training pipelines and test datasets.

5 Discussion and Conclusions

We have presented SPE-FR, the first unsupervised method for measuring bias in face recognition algorithms. It is based on parametric modeling of the distribution of confidence values that are assigned by the algorithm to proposed matches of faces. SPE-FR can be applied to assess the performance of face recognition systems both in the unsupervised setting where no identity annotations are available and in the semi-supervised setting where some annotations are available. SPE-FR produces Bayesian posterior intervals for any performance metric that can be evaluated from the joint distribution of the match indicator Y and the algorithmic score S. In particular, it can be used to estimate entire performance curves (such as the FNMR vs FMR curve) and produces confidence bands to communicate the uncertainty in the estimation. We validated SPE-FR with experiments on a carefully constructed set of FR models and datasets. The main observations are fourfold: First, it is effective in revealing demographic biases in model performance. Second, our method can estimate performance even when the test set is rather small, and when the ratio of true matches to non-matches is low, as is the case for certain subgroups in the MORPH data. Third, even when the confidence bands do not contain the ground truth, the degree of misestimation is found to be fairly consistent across groups, and thus SPE-FR can still provide a strong indication of demographic bias in system performance. Lastly, our experiments show that SPE-FR can be applied off-the-shelf to a wide range of face embedding models with state-of-the-art designs and trained on different datasets. Therefore, SPE-FR can be especially useful to companies and agencies prior to system adoption who may otherwise be unable to estimate system performance or detect potential biases as they cannot collect reliable identity annotations for their data.

Bibliography

[1] Albiero, V., KS, K., Vangara, K., Zhang, K., King, M.C., Bowyer, K.W.: Analysis of gender inequality in face recognition accuracy. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. pp. 81–89 (2020) 3

[2] Albright, T.D.: Why eyewitnesses fail. Proceedings of the National Academy of Sciences **114**(30), 7758–7764 (2017) 2

[3] An, X., Zhu, X., Xiao, Y., Wu, L., Zhang, M., Gao, Y., Qin, B., Zhang, D., Ying, F.: Partial fc: Training 10 million identities on a single machine. In: Arxiv 2010.05222 (2020) 11

[4] Awasthi, P., Beutel, A., Kleindessner, M., Morgenstern, J., Wang, X.: Evaluating fairness of machine learning models under uncertain and incomplete information. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 206–214 (2021) 6

[5] Balakrishnan, G., Xiong, Y., Xia, W., Perona, P.: Towards causal benchmarking of bias in face analysis algorithms. In: European Conference on Computer Vision. pp. 547–563. Springer (2020) 1, 2, 3

[6] Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European conference on computer vision (ECCV). pp. 456–473 (2018) 3

[7] ter Braak, C.J., Vrugt, J.A.: Differential evolution markov chain with snooker updater and fewer chains. Statistics and Computing **18**(4), 435–446 (2008) 9

[8] Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91. PMLR (2018) 1

[9] Chen, J., Kallus, N., Mao, X., Svacha, G., Udell, M.: Fairness under unawareness: Assessing disparity when protected class is unobserved. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 339–348 (2019) 6

[10] Coston, A., Ramamurthy, K.N., Wei, D., Varshney, K.R., Speakman, S., Mustahsan, Z., Chakraborty, S.: Fair transfer learning with missing protected attributes. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 91–98 (2019) 6

[11] Deepglint: http://trillionpairs.deepglint.com/overview, http://trillionpairs.deepglint.com/overview 10

[12] Deng, J., Guo, J., Liu, T., Gong, M., Zafeiriou, S.: Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In: European Conference on Computer Vision. pp. 741–757. Springer (2020) 10

[13] Deng, J., Guo, J., Liu, T., Gong, M., Zafeiriou, S.: Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In: Proceedings of the IEEE Conference on European Conference on Computer Vision (2020) 11

[14] Deng, J., Guo, J., Niannan, X., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019) 11
[15] Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: CVPR (2020) 11
[16] Deng, J., Roussos, A., Chrysos, G., Ververas, E., Kotsia, I., Shen, J., Zafeiriou, S.: The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. IJCV (2018) 11
[17] Deng, S., Xiong, Y., Wang, M., Xia, W., Soatto, S.: Harnessing unrecognizable faces for improving face recognition. arXiv preprint arXiv:2106.04112 (2021) 3
[18] Deng, W., Zheng, L.: Are labels always necessary for classifier accuracy evaluation? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15069–15078 (2021) 3
[19] Fechner, G.T.: Kollektivmasslehre. Engelmann (1897) 8
[20] Fernández, C., Steel, M.F.: On bayesian modeling of fat tails and skewness. Journal of the american statistical association **93**(441), 359–371 (1998) 8
[21] Garg, S., Balakrishnan, S., Lipton, Z.C., Neyshabur, B., Sedghi, H.: Leveraging unlabeled data to predict out-of-distribution performance. arXiv preprint arXiv:2201.04234 (2022) 3
[22] GoogleAI: Responsible ai practices, https://ai.google/responsibilities/responsible-ai-practices/ 1
[23] Grother, P.J., Ngan, M.L., Hanaoka, K.K., et al.: Face recognition vendor test part 3: demographic effects (2019) 1, 2, 3, 4
[24] Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., Schmidt, L.: Predicting with confidence on unseen distributions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1134–1144 (2021) 3
[25] Guo, J., Deng, J., Lattas, A., Zafeiriou, S.: Sample and computation redistribution for efficient face detection. arXiv preprint arXiv:2105.04714 (2021) 11
[26] Guo, J., Deng, J., Xue, N., Zafeiriou, S.: Stacked dense u-nets with dual transformers for robust face alignment. In: BMVC (2018) 11
[27] Hartig, F., Minunno, F., Paul, S.: BayesianTools: General-Purpose MCMC and SMC Samplers and Tools for Bayesian Statistics (2019), https://CRAN.R-project.org/package=BayesianTools, r package version 0.1.7 9
[28] Hashimoto, T., Srivastava, M., Namkoong, H., Liang, P.: Fairness without demographics in repeated loss minimization. In: International Conference on Machine Learning. pp. 1929–1938. PMLR (2018) 3
[29] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 10
[30] He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European conference on computer vision. pp. 630–645. Springer (2016) 10
[31] IBM: Trustworthy ai, https://www.ibm.com/watson/trustworthy-ai 1
[32] Ji, D., Smyth, P., Steyvers, M.: Can i trust my fairness metric? assessing fairness with unlabeled data and bayesian inference. arXiv preprint arXiv:2010.09851 (2020) 3, 4, 6, 13

[33] Kearns, M., Roth, A.: The ethical algorithm: The science of socially aware algorithm design. Oxford University Press (2019) 3

[34] Keles, U., Lin, C., Adolphs, R.: A cautionary note on predicting social judgments from faces with deep neural networks. Affective Science **2**(4), 438–454 (2021) 2

[35] Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., Vetter, T.: Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019) 3

[36] Krishnapriya, K., Albiero, V., Vangara, K., King, M.C., Bowyer, K.W.: 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

[37] Krivosheev, E., Bykau, S., Casati, F., Prabhakar, S.: Detecting and preventing confused labels in crowdsourced data. Proceedings of the VLDB Endowment **13**(12), 2522–2535 (2020) 2

[38] Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., Chi, E.H.: Fairness without demographics through adversarially reweighted learning. arXiv preprint arXiv:2006.13114 (2020) 3

[39] Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 international conference on biometrics (ICB). pp. 158–165. IEEE (2018) 3

[40] McKone, E., Dawel, A., Robbins, R.A., Shou, Y., Chen, N., Crookes, K.: Why the other-race effect matters: Poor recognition of other-race faces impacts everyday social interactions. British Journal of Psychology (2021) 2

[41] Muthén, B., Shedden, K.: Finite mixture modeling with mixture outcomes using the em algorithm. Biometrics **55**(2), 463–469 (1999) 5

[42] Phillips, P.J., Yates, A.N., Hu, Y., Hahn, C.A., Noyes, E., Jackson, K., Cavazos, J.G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., et al.: Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. Proceedings of the National Academy of Sciences **115**(24), 6171–6176 (2018) 2

[43] PricewaterhouseCoopers: Responsible ai toolkit, https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html 1

[44] Prost, F., Awasthi, P., Blumm, N., Kumthekar, A., Potter, T., Wei, L., Wang, X., Chi, E.H., Chen, J., Beutel, A.: Measuring model fairness under noisy covariates: A theoretical perspective. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. pp. 873–883 (2021) 6

[45] Raji, I.D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., Denton, E.: Saving face: Investigating the ethical concerns of facial recognition auditing. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 145–151 (2020) 1, 3

[46] Ranjan, R., Castillo, C.D., Chellappa, R.: L2-constrained softmax loss for discriminative face verification. arXiv preprint arXiv:1703.09507 (2017) 10

[47] Ricanek, K., Tesafaye, T.: Morph: A longitudinal image database of normal adult age-progression. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR06). pp. 341–345. IEEE (2006) 3, 10
[48] Robinson, J.P., Livitz, G., Henon, Y., Qin, C., Fu, Y., Timoner, S.: Face recognition: too bias, or not too bias? In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops. pp. 0–1 (2020) 3
[49] Rubio, F., Steel, M.: The family of two-piece distributions. Significance **17**, 12–13 (02 2020). https://doi.org/10.1111/j.1740-9713.2020.01352.x 8
[50] Rubio, F.J., Ogundimu, E.O., Hutton, J.L.: On modelling asymmetric data using two-piece sinh–arcsinh distributions. Brazilian Journal of Probability and Statistics pp. 485–501 (2016) 8
[51] Srinivas, N., Ricanek, K., Michalski, D., Bolme, D.S., King, M.: Face recognition algorithm bias: Performance differences on images of children and adults. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019) 3
[52] Tanaka, J.W., Kiefer, M., Bukach, C.M.: A holistic account of the own-race effect in face recognition: Evidence from a cross-cultural study. Cognition **93**(1), B1–B9 (2004) 2
[53] Vangara, K., King, M.C., Albiero, V., Bowyer, K., et al.: Characterizing the variability in face recognition accuracy relative to race. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019) 1, 3
[54] Vorvoreanu, M., Walker, K.: Advancing ai trustworthiness: Updates on responsible ai research (Feb 2022), https://www.microsoft.com/en-us/research/blog/advancing-ai-trustworthiness-updates-on-responsible-ai-research/ 1
[55] Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., Change Loy, C.: The devil of face recognition is in the noise. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 765–780 (2018) 10
[56] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018) 10
[57] Wang, M., Deng, W., Hu, J., Tao, X., Huang, Y.: Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 692–702 (2019) 2, 3, 10
[58] Wang, M., Zhang, Y., Deng, W.: Meta balanced network for fair face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 2, 10, 12
[59] Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards fairness in visual recognition: Effective strategies for bias mitigation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8919–8928 (2020) 3

[60] Welinder, P., Welling, M., Perona, P.: A lazy man's approach to benchmarking: Semisupervised classifier evaluation and recalibration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3262–3269 (2013) 3, 4, 5, 6

[61] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization in vision: A survey. arXiv preprint arXiv:2103.02503 (2021) 3