Supplemental Material for Unsupervised and Semi-supervised Bias Benchmarking in Face Recognition

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

AWS AI Labs

We structure the appendices of the main text in the follow order:

- Appendix A provides further experiments demonstrating SPE-FR works for models trained with different data or configurations.
- Appendix B shows results of semi-supervised SPE-FR where an ablation study on $N_{\mathcal{L}}$ is introduced to compare the unsupervised setting with the semi-supervised ones.
- Appendix C discusses parametric modeling and demonstrates that two-piece distributions provide good approximations while other standard parametric families do not.
- Appendix D presents our procedure for estimating the proportion of true matching pairs from data with no identity annotations.
- Appendix E provides the details of our Bayesian inference strategy, including MCMC configuration.
- Appendix F provides more details on the training procedure for the face verification models.
- Appendix G provides supporting information on the datasets and face verification benchmark protocols used during training and evaluation.
- Appendix H describes how we adapted the Bayesian Calibration [16] method to the face verification application.

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

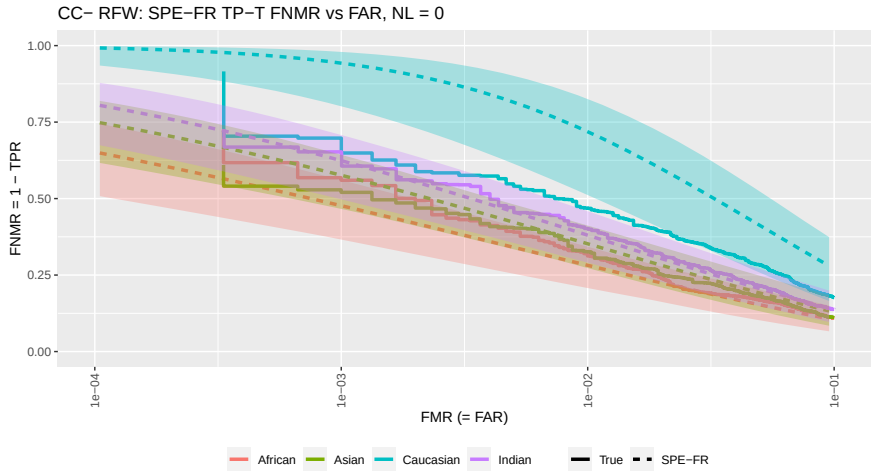
A More results on Unsupervised FR Bias Evaluation

A.1 Results on Models Trained with Controlled Biases

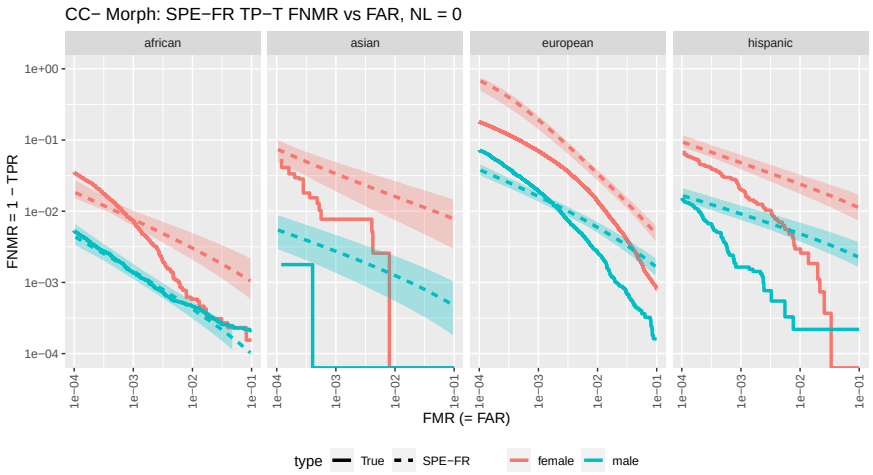
In the main text Figure 1 and 5, we presented unsupervised SPE-FR results for the AA model. We provide here corresponding Figures for the other "ablated" models (main text Table 1). Interpretations are provided in the respective Figure captions. Figure 1 shows results for CC- model, Figure 2 shows results for EA-model, Figure 3 shows results for M- model, and Figure 4 shows results for RT model.

A.2 Results on the Generalization Test of SPE-FR

In the main text we introduced that we evaluate the effectiveness of SPE-FR for performance and bias estimation of face recognition models trained under 5 settings detailed in main text Table 2. We have shown results from two models in main text Figure 7. Here we share results of the rest in Figure 5, Figure 6 and Figure 7.

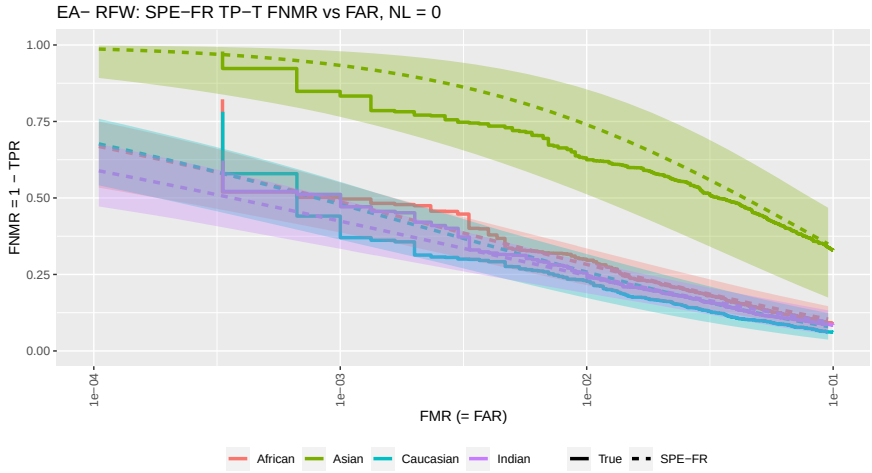


(a) Unsupervised SPE-FR applied to the **RFW** dataset. SPE-FR correctly estimates performance on all groups except for Caucasian group. SPE-FR incorrectly suggests significant underperformance for Caucasians.

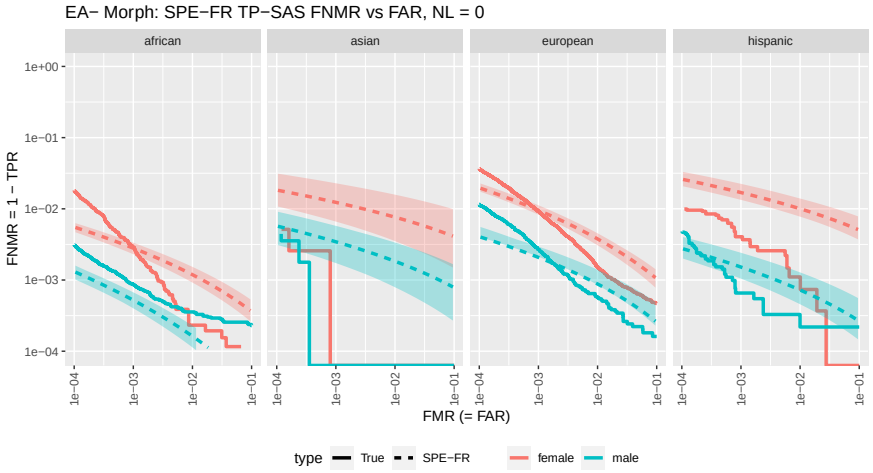


(b) Unsupervised SPE-FR applied to the **Morph** dataset restricted to a maximum of 40K samples per gender-ethnicity-match bin (detailed in Table 4). Overall SPE-FR over-estimates the system FNMR, but correctly and confidently reveals gender bias in the system that persists across all ethnicity groups.

Fig. 1: **(CC Model) Unsupervised SPE-FR estimates of the FNMR vs FMR curve. CC** model (no European faces in training data) applied to RFW (Top) and MORPH (Bottom). SPE-FRE estimates are shown as dashed lines, with 89% posterior credible confidence bands overlaid.

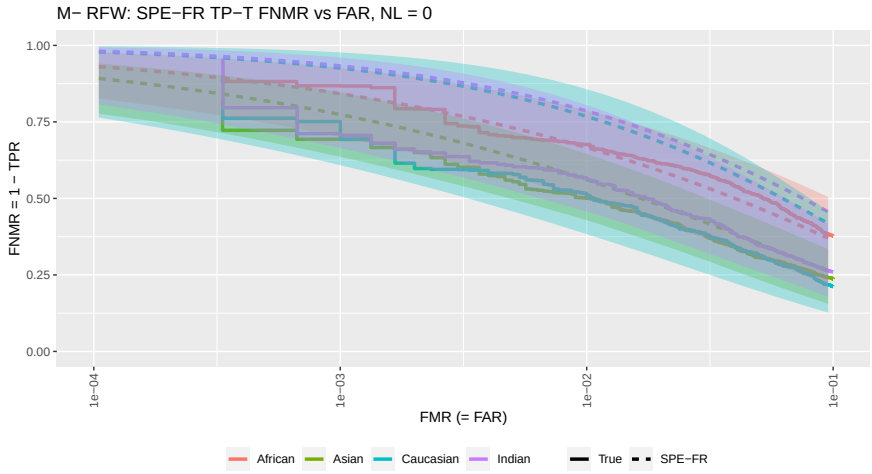


(a) Unsupervised SPE-FR applied to **RFW**. The SPE-FR estimated performance curves and confidence bands capture the True FNMR vs FMR curves, and correctly and confidently reveal poor performance on Asian faces compared to the other racial groups.

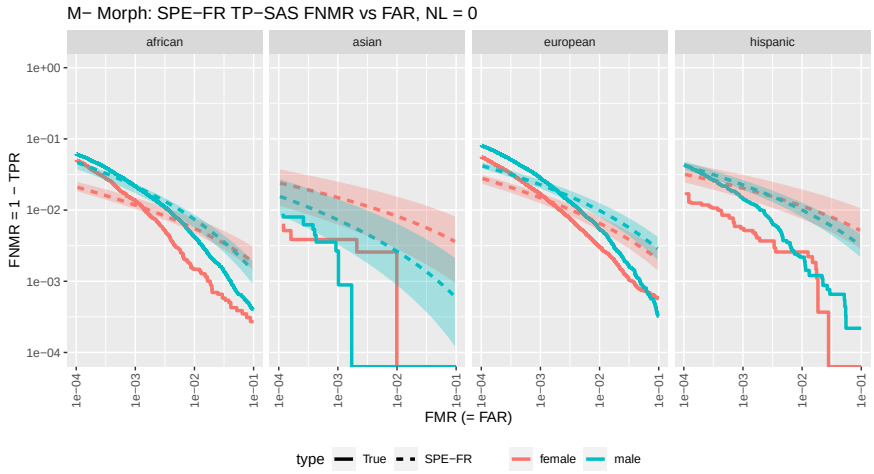


(b) Unsupervised SPE-FR applied to **Morph** data restricted to a maximum of 40K samples per gender-ethnicity-match bin (as detailed in Table 4). Overall SPE-FR over-estimates the system FNMR, but correctly and confidently reveals gender bias in the system that persists across all ethnicity groups.

Fig. 2: **(EA model) Unsupervised SPE-FR estimates of the FNMR vs FMR curve. EA** model (no East Asian faces in training data) applied to RFW (Top) and MORPH (Bottom). SPE-FRE estimates are shown as dashed lines, with 89% posterior credible confidence bands overlaid.

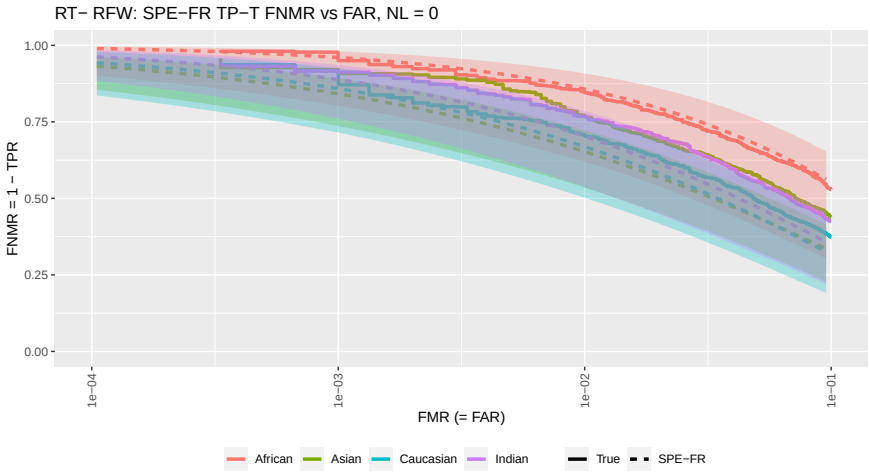


(a) Unsupervised SPE-FR applied to **RFW**. The SPE-FR estimated performance curves and confidence bands capture the True FNMR vs FMR curves, and correctly indicate overall poor but similar performance across race/ethnicity groups.

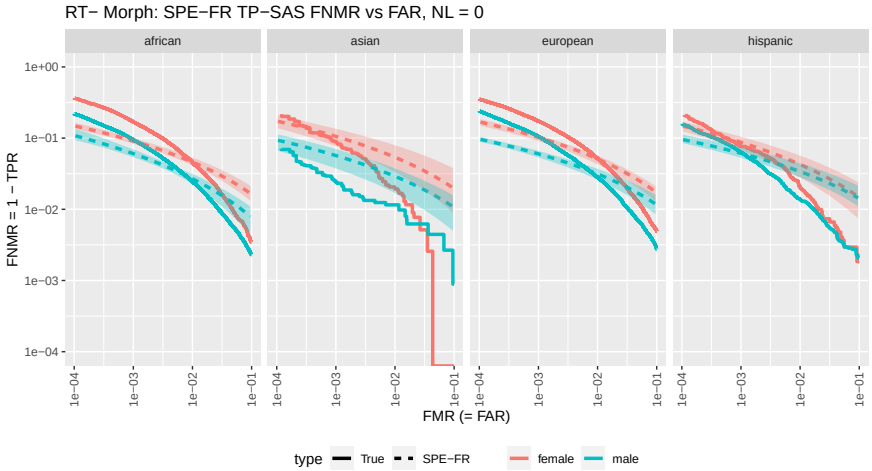


(b) Unsupervised SPE-FR applied to **Morph** data restricted to a maximum of 40K samples per gender-ethnicity-match bin (as detailed in Table 4). Overall SPE-FR over-estimates the system FNMR, but correctly infers that there is little, if any, significant gender bias in system performance. The small distance between the estimated FNMR-FMR curves for men and women within each race/ethnicity group are overall similar to the true observed gap, except in the hispanic group where the true gap is greater (but may be imprecisely estimated in the ground truth).

Fig. 3: **(M model) Unsupervised SPE-FR estimates of the FNMR vs FMR curve. EA** model (no Male faces in training data) applied to RFW (Top) and MORPH (Bottom). SPE-FRE estimates are shown as dashed lines, with 89% posterior credible confidence bands overlaid.



(a) Unsupervised SPE-FR applied to **RFW**. The SPE-FR estimated performance curves and confidence bands capture the True FNMR vs FMR curves except for the Asian group, and correctly indicate overall poor performance across race/ethnicity groups, with somewhat worse performance correctly indicated in the African group.



(b) Unsupervised SPE-FR applied to **Morph** data restricted to a maximum of 40K samples per gender-ethnicity-match bin (as detailed in Table 4). Overall SPE-FR slightly mis-estimates the system FNMR, but currectly infers that there is little, if any, significant gender bias in system performance. The small distance between the estimated FNMR-FMR curves for men and women within each race/ethnicity group are overall similar to the true observed gap, correctly indicating no significant gender bias in the hispanic group, but some gender bias in the other 3 race/ethnicity groups.

Fig. 4: **(RT model) Unsupervised SPE-FR estimates of the FNMR vs FMR curve. RT** model (only 10% of available training data is used in model training) applied to RFW (Top) and MORPH (Bottom). SPE-FRE estimates are shown as dashed lines, with 89% posterior credible confidence bands overlaid.

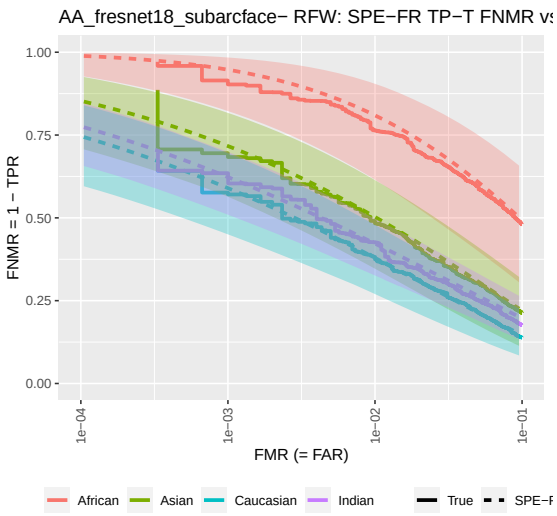


Fig. 5: **Unsupervised SPE-FR estimates of the FNMR vs FMR curve.** Model bias evaluated on RFW: FR model trained on BUPT-BalancedFace dataset (ablation "AA" applied, leaving out the African group) with Sub-Center Arcface loss and Res18 backbone.

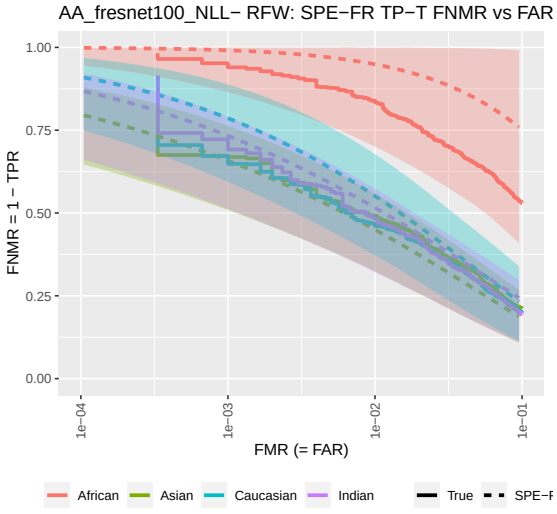


Fig. 6: **Unsupervised SPE-FR estimates of the FNMR vs FMR curve.** Model bias evaluated on RFW: FR model trained on BUPT-BalancedFace dataset (ablation "AA" applied, leaving out the African group) with Softmax loss and Res101 backbone, training sets ablation "AA" applied (see Table 1, left-out set is African.

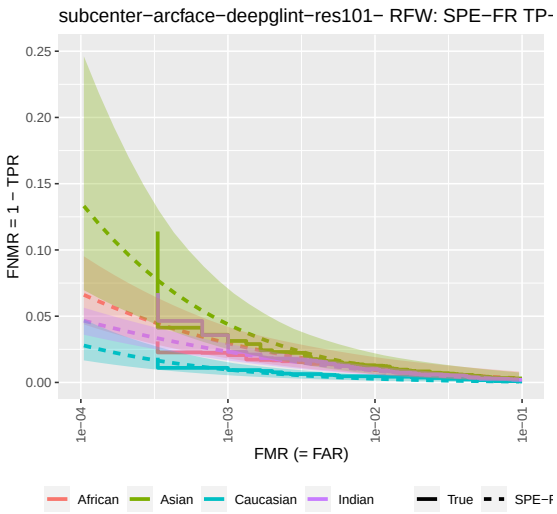


Fig. 7: **Unsupervised SPE-FR estimates of the FNMR vs FMR curve.** Model bias evaluated on RFW: FR model trained on DeepGlint dataset with Sub-Center Arcface loss and Res101 backbone.

B Semi-supervised SPE-FR

In the main text, we presented results for unsupervised SPE-FR, which does not have access to any identity annotations. Here we present some results for semi-supervised SPE-FR where we vary the number of labeled image pairs $N_{\mathcal{L}}$ from 0 to 256. We present results for the AA model applied to RFW. Figure 8 shows how the SPE-FR point estimates and confidence intervals for FNMR vary with the number of annotated image pairs. FNMR values are calculated at a threshold chosen to achieve FMR 0.005 on the full RFW data. We see that unsupervised SPE-FR ($N_{\mathcal{L}} = 0$) performs just about as well as semi-supervised SPE-FR, except in the case of the African group, where we see significant gains coming from a small number of annotations. This is likely due to the fact that in the AA model there is very poor separation between the score distributions in the true match and non-match classes. A small number of labeled examples helps resolve the classes in ways that the fully unsupervised SPE-FR model can struggle to do.

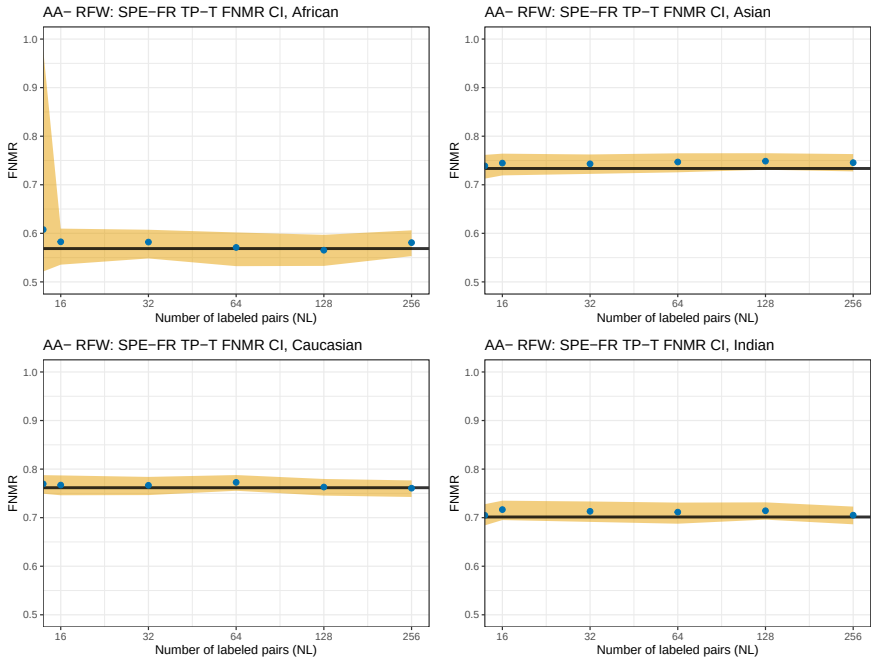


Fig. 8: **Semi-supervised SPE-FR FNMR point estimates at overall target FMR 0.005 as number of labelled pairs $N_{\mathcal{L}}$ varies.** This plot shows semi-supervised SPE-FR point estimates (blue points) and confidence intervals (orange bars) for different numbers of labelled pairs ($N_{\mathcal{L}}$). Ground truth FNMR is shown as a horizontal line. Plots show FNMR estimates for the four race/ethnicity groups in the RFW data: African (top-left), Asian (top-right), Caucasian (bottom-left), and Indian (bottom-right). We see that except for the African group, adding labels does not significantly affect SPE-FR estimates. The unsupervised SPE-FR results are mostly just as reliable. This is not the case for the African group. We see a significant decrease in the width of the confidence interval with just a small number of annotated pairs. This is likely due to the fact that in the AA model there is very poor separation between the score distributions in the true match and non-match classes. A small number of labeled examples helps resolve the classes in ways that the fully unsupervised SPE-FR model can struggle to do.

C Parametric Modeling

As discussed in the main text, we rely on two-piece distributions to model the class-conditional distributions of face recognition system similarity and distance scores. In the main text, we presented histograms of the distance scores S for the AA model on the RFW data, along with true FNMR vs FMR curves compared to parametric approximations obtained using the fully labelled data. Here we present Normal QQ plots to support our assertion that, while the densities shown in the main text may look close to normal, they are in fact significantly skewed for non-matching pairs. This is shown in Figure 9.

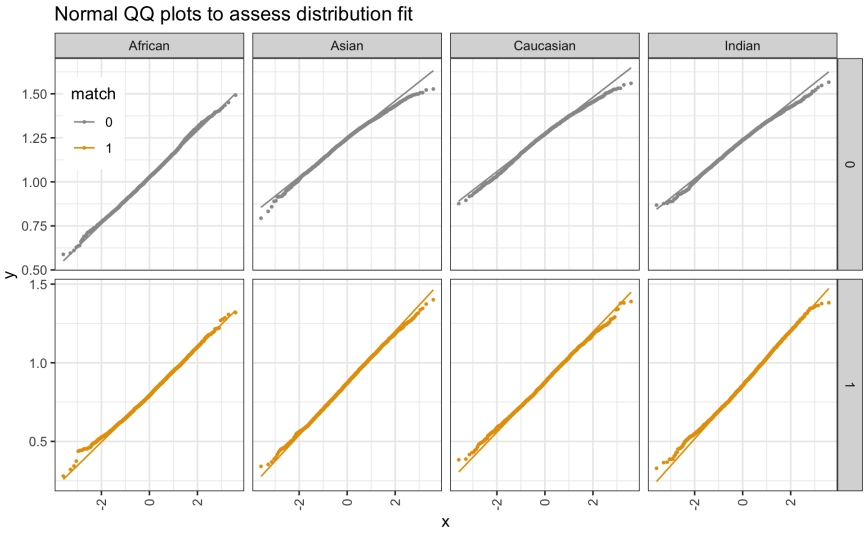


Fig. 9: **Normal QQ plots showing match-conditional distributions of distance scores for the AA model on RFW.** Departure from the diagonal lines, particularly in the lower and upper tail, indicates that the observed data are non-normally distributed. With the exception of the African subgroup, we see that the score distribution for non-matching pairs (top row, grey figures) deviates from normality.

We now display the corresponding plots for MORPH. Figure 10 shows histograms and Normal QQ-plots of the distance scores S and similarity scores $\tilde{S} = \frac{1}{1+S}$ obtained by transforming the original distances. We see that among true matches, the distribution of similarity scores $\tilde{S}$ is approximately Normal for all gender-ethnicity subgroups. The distributions among non-matching scores is highly non-normal. Figure 11 shows QQ plots for the scores of non-matching pairs to assess fit of the Normal distribution and Gamma, Lognormal, Weibull and Logistic. None of these distributions provide a particularly good approx-

imation, with the Logistic distribution generally being the closest across all ethnicity-gender groups.

Figure 12 shows true FNMR vs FMR curves compared to 4 parametric approximations. We find that the Two-piece t (TP-T) and Two-piece sinh-arsinh (TP-SAS) distributions do a good job of approximating the true performance curves. Our experiments for MORPH all operate on the similarity scores $\tilde{S}$ and use TP-SAS class-conditional distributions.

In Figure 4 and 6 of the main paper, we saw that unsupervised SPE-FR tended to over-estimate error rates (FNMR) on the MORPH data. This may be somewhat surprising given that Figure 12 shows the TP-SAS model being capable of approximating the true FNMR vs FMR curve well. The challenge is that unsupervised SPE-FR does not use any identity annotations to estimate the TP-SAS model parameters. So while Figure 12 shows that there exists a TP-SAS model that produces a good approximation to the true performance curves, unsupervised SPE-FR does not necessarily find that optimal model. We can see this more directly in Figure 13, which compares the true FNMR vs FMR curve to the optimal TP-SAS approximation and to the unsupervised SPE-FR estimates for 4 different gender-ethnicity groups using the AA model applied to MORPH. We see that for some groups (e.g., african male), the optimal parametric model lies inside SPE-FR confidence band, but for others (e.g., asian and hispanic female), it does not. This confirms that unsupervised SPE-FR cannot always find the optimal model within the specified parametric class.

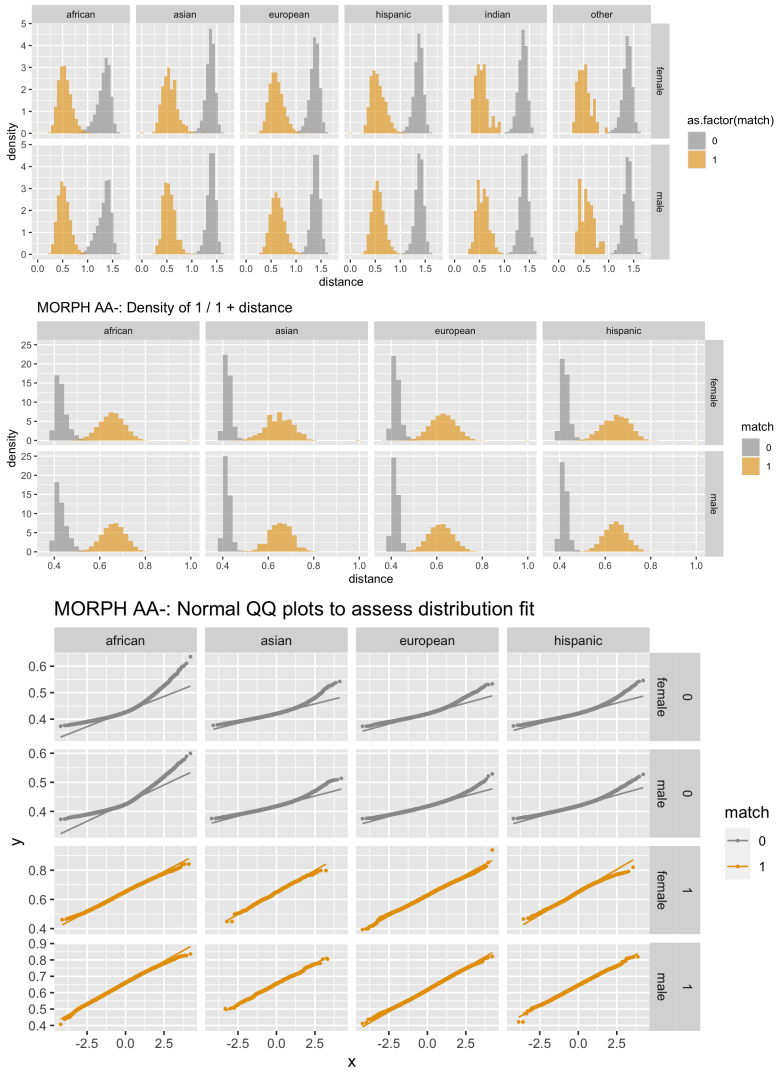


Fig. 10: **Class-conditional distributions of distance/similarity scores for the AA model on MORPH. Top**: Histograms of class-conditional *distance* scores for all race/ethnicity groups in the MORPH data. **Middle**: Histograms of class-conditional *similarity* scores (indian and other groups omitted) obtained by transforming distance scores S via $\tilde{S} = \frac{1}{1+S}$. We see that this transformation produces true match scores that are symmetric and turn out to be well-modelled by a normal distribution. **Bottom:** Orange curves confirm that $\tilde{S} \mid Y = 1$ (the similarity distribution among true matches) is approximately normally distributed within each ethnicity and gender group. The similarity distributions for non-matching pairs are clearly skewed.

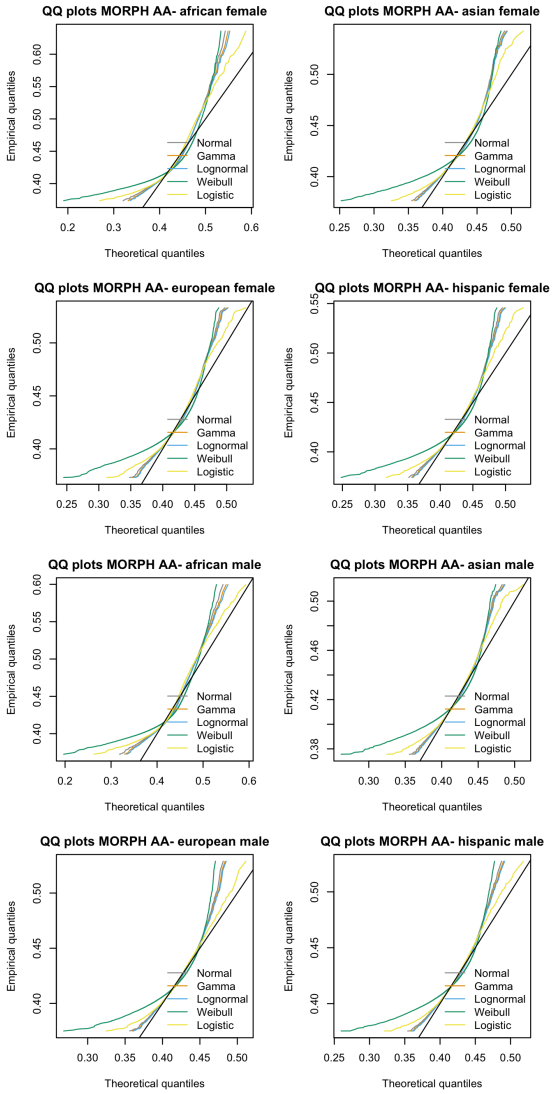


Fig. 11: **QQ plots showing distributions of similarity scores $\frac{1}{1+d}$ for *non-matching pairs* using the AA model on MORPH.** QQ plots are shown for Normal, Gamma, Lognormal, Weibull and Logistic distribution families. Best fitting parametric models are identified via maximum likelihood estimation. Departure of the from the diagonal lines indicates that the observed data are not well approximated by the given parametric distribution family. We see that these standard parametric families are generally poor fits to the observed score data.

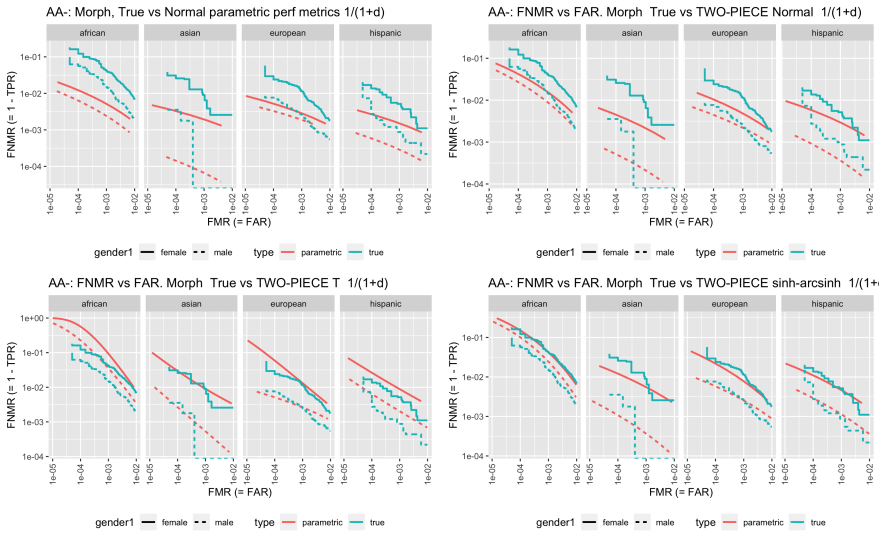


Fig. 12: **True FNMR vs FMR curves vs. parametric approximations for the AA model on MORPH.** The Figures show the true FNMR vs FMR curves as computed on the fully labelled data compared to different parametric approximations, with parameters computed via maximum likelihood on the fully annotated data. Parametric approximations shown are Normal (**top-left**), Two-piece Normal (**top-right**), Two-piece t (**bottom-left**), and Two-piece sinh-arcsinh (**bottom-right**). All plots are based on parametric modelling of the *similarity* scores $\tilde{S} = \frac{1}{1+S}$. We see that the TP-T and TP-SAS parametric models do a very good job of approximating the true FNMR vs FMR curves. Our MORPH experiments all rely on TP-SAS parametric models for the class-conditional scores.

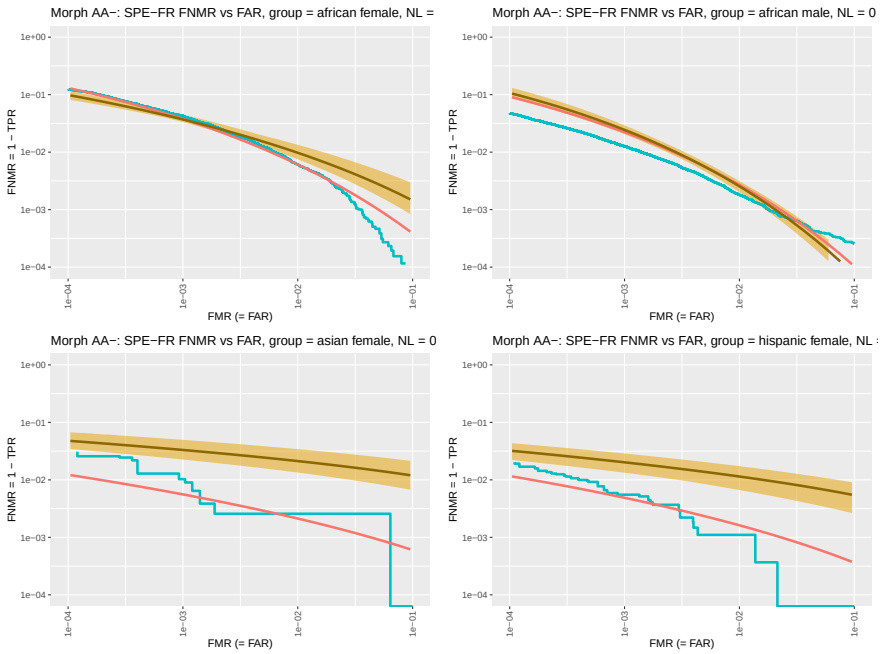


Fig. 13: **True FNMR vs FMR curves vs. parametric approximations and SPE-FR for the AA model on MORPH.** The Figures show the true FNMR vs FMR curves as computed on the fully labelled data (blue) compared to the optimal TP-SAS approximation computed via maximum likelihood estimation (red) and unsupervised SPE-FR estimates (brown curve and orange confidence band). Results are shown for 4 gender-ethnicity groups: african female (**top-left**), african male (**top-right**), asian female t (**bottom-left**), and hispanic female (**bottom-right**). We see that the unsupervised SPE-FR confidence band sometimes does contain the best parametric approximation (red curve), such as for the african male subgroup, but sometimes it does not, such as for the asian and hispanic female subgroups. Thus unsupervised SPE-FR is not guaranteed to approximate the true performance curve as well as the optimal curve in the assumed parametric family.

D Estimating Proportion of True Matches

We observed experimentally that SPE-FR performed better when informed by a good estimate of $\pi = P(Y = 1)$, the proportion of true matches in the data. To obtain initial estimates of π we rely on kernel density estimation methods for mode estimation [13] and methods for estimating the proportion of nulls in large scale hypothesis testing [19].

Our analysis is informed by the empirical observation that the true match score distribution $S \mid Y = 1$ tends to be approximately normally distributed. On MORPH, we observed that transforming the distance scores via $\tilde{S} = \frac{1}{1+S}$ produced perfectly normally distributed similarity scores $\tilde{S}$. It is not surprising that (potentially transformed) distances or similarities among true matches tend to be approximately normally distributed. Such distances are often the accumulation of a number of small differences between image pairs, which is precisely the setting where the Central Limit Theorem is expected to apply.

We leverage the approximate normality of $S \mid Y = 1$ and separation between the match-conditional distributions $S \mid Y = 0$ and $S \mid Y = 1$ to estimate π as follows. We describe the procedure assuming the score S is a **similarity**. If S is a distance, the inequalities/orderings referenced below need to be flipped. There are two stages. First, we use mode estimation methods to estimate the parameters (μ, σ) in the Normal approximation of the conditional true match distribution, $S \mid Y = 1 \sim N(\mu, \sigma^2)$. Then we transform the scores S into p-

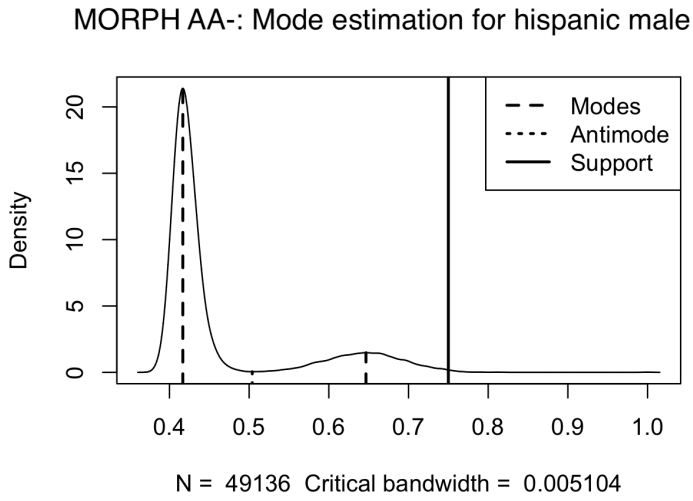


Fig. 14: **Mode estimation step for estimating proportion of true matches.** Kernel density estimate-based mode estimates for similarity score distribution of the AA model on hispanic males in the MORPH data. We estimate $\mu = \mathbb{E}(S \mid Y = 1)$ using the greater of the two estimated modes shown (the one around 0.65).

values and apply a method for estimating the proportion of nulls, which in our case translates into the proportion of true matches in the data. Note that no identity labels are used in this estimation strategy. Group labels *are* used: the procedure is carried out separately for each group. To reduce notation burden we will index the scores with a single index i as opposed to a double index ij going over all image pairs ij. We will think of there being n image *pairs*. Within each group we:

1. Since the Normal is unimodal and symmetric, the mean μ is equal to me mode. We estimate the two modes of the marginal score distribution S using the KDE-based mode estimation method of [13] as implemented in the `multimode` library in R. We then let $\hat{\mu}$ be the *greater* of the two modes. (Figure 14 shows mode estimation applied to the hispanic male score distribution.)
2. If the class-conditions are well-separated, there are no (or very few) observations from $Y = 0$ with score $S > \hat{\mu}$. We therefore estimate σ as:

$$\hat{\sigma}^2 = \frac{2}{n_U} \sum_{i=1}^{n_U} (s_i - \hat{\mu})^2 I(s_i \geq \hat{\mu}),$$

   where $n_U = |\{i : s_i > \hat{\mu}\}$. This amounts to using only the upper half of the data to estimate the variance of a symmetric distribution.
3. We then transform all of the scores into Z-scores:

$$z_i = \frac{s_i - \hat{\mu}}{\hat{\sigma}}, \tag{1}$$

   and then turn these Z-scores into p-values by taking,

$$p_i = 2\,(1 - \Phi(|z_i|)) \tag{2}$$

   where Φ is the standard normal CDF function.
4. We then use the Storey(λ) estimator [19]. Given a threshold $0 < \lambda \leq 1$, the Storey(λ) estimator for the number of true matches is then given by:

$$\hat{\pi}_{Storey}(\lambda) = \frac{\#\{i : p_i > \lambda\}}{n(1 - \lambda)}, \tag{3}$$

   where n is the total number of observations in the group.

The intuition for Step (4) is as follows. If the normal approximation to $S \mid Y = 1$ is approximately correct, then $p_i \mid Y = 1 \sim Unif(0, 1)$. For a given threshold λ, we have

$$P(p > \lambda) = \pi(1 - \lambda) + (1 - \pi)P(p > \lambda \mid Y = 0)$$

Non-match scores S_i coming from non-match pairs $Y_i = 0$ will tend to have p-values very close to 0, because they will look like unlikely observations from the $N(\hat{\mu}, \hat{\sigma}^2)$ distribution of the $S \mid Y = 1$ true match scores. This means we

Table 1: **Estimates of proportions of true matches.** Results shown for MORPH 40K subset data (as detailed in Table 4). Table shows ground truth (true π column) along with Storey(0.05) estimate that uses the fully labelled data to estimate the mean and standard deviation of the normal approximation for the true match score distribution (storey*(0.05) column). Three right-most columns are estimated from data with no identity annotations. We find that at $\lambda = 0.05$ and 0.1 the Storey procedure consistently does a good job of estimating π, the proportion of true matches in the data.

ethnicity	gender	true π	storey* (0.05)	storey(0.01)	storey(0.05)	storey(0.10)
african	female	0.392	0.395	0.441	0.416	0.425
african	male	0.500	0.502	0.524	0.524	0.541
asian	female	0.030	0.031	0.254	0.036	0.034
asian	male	0.027	0.027	0.028	0.029	0.029
european	female	0.500	0.503	0.591	0.527	0.546
european	male	0.500	0.502	0.669	0.527	0.548
hispanic	female	0.064	0.065	0.174	0.069	0.070
hispanic	male	0.186	0.187	0.201	0.195	0.203

expect the second term in the sum to be approximately 0. Rearranging to solve for π yields the estimator in(3).

Table 1 compares the ground truth proportion of true matches in each ethnicity-gender subground of the MORPH 40K data (as detailed in Table 4) to the Storey(0.05) estimate obtained by using fully labelled data to estimate the $(\hat{\mu}, \hat{\sigma})$ parameters and also to Storey(λ) estimates at $\lambda = 0.01, 0.05, 0.1$ using data with no identity annotations. We find that the procedure at $\lambda = 0.05$ and 0.1 does a good job of estimating the true proportions π across all ethnicity-gender groups. Smaller values of λ such as 0.01 are rarely used in practice, and are shown only to demonstrate how the estimation accuracy can degrade. For our SPE-FR experiments we apply Storey(0.05) to inform the prior on π for each group.

E SPE-FR Experimental Details

E.1 MCMC Configurations

For the RFW experiments we ran the sampler for 17000 iterations, using the first 2000 iterations for adaptation and discarding a burn-in of 4000. For the MORPH experiments the chains were slower to mix. Through experimentation we found that running the sampler for 25000 iterations and taking a burn-in of 7000 resulted in adequate mixing. At these settings the Gelman-Rubin diagnostic [2] was typically under 1.2 for each model parameter, and was often much smaller.

E.2 SPE-FR hyperparameters estimation

We inform the analysis by estimating $\{\mu_j, \eta_j, \beta^{\tau}_{jk}, \alpha^{\delta}_j, \beta^{\delta}_j\}$ using unlabeled data. Specifically, we begin by fitting a two-component Gaussian mixture model and taking μ_j to be the means of the estimated components. When S denotes distance (rather than similarity), μ_0 is taken to be the greater of the two-component means. We set the β^{τ} parameters to make the Gamma means equal to the estimated standard deviations of the mixture components. $\alpha^{\delta}_j = 5$ when using the TP-T and 50 when using TP-SAS. β^{δ}_j is then taken to be $\alpha^{\delta}_j/20$ for TP-T and α^{δ}_j for TP-SAS. This centers the prior on δ_j at 1, at which value the TP-SAS distribution has Gaussian tails. $\delta_j < 1$ gives tails that are heavier than Gaussian; $\delta > 1$ gives tails that are lighter. Lastly, to estimate the parameter π, we adapt methods for estimating the proportion of non-nulls in large-scale hypothesis testing [9]. These methods are specifically tailored to the setting where the number of non-nulls (here, true match image pairs) is small relative to the number of nulls (here, non-match image pairs).

F Face Embedding Model Training

F.1 Model Training with Controlled Biases.

We adopt the well-known state-of-the-art Sub-center ArcFace [4] method in our face recognition model. We employ a variant of ResNet [15] as our underlying feature extractor that take as inpout a face image $x_i, i \in \{1, \ldots, I\}$, and outputs an embedding $\mathbf{z}_i \in \mathcal{R}^d$ where d is the embedding feature dimension. We use the Sub-center ArcFace loss as described in equation 4 for training the face recognition model.

$$
\begin{aligned}
L &= -\log \frac{e^{s(\cos(m_1\theta_{i,y_i}+m_2)-m_3)}}{e^{s(\cos(m_1\theta_{i,y_i}+m_2)-m_3)} + Z(\theta)} \\
\theta_{i,j} &= \arccos(\max_k(W_{jk}^T \mathbf{z}_i) \\
Z(\theta) &= \sum_{j=1, j\neq y_i}^{N} e^{s\cos(\theta_{i,j})}
\end{aligned}
\tag{4}
$$

where s and m_* are the hyper-parameters representing scale and margin respectively. N is the total number of classes and y_i is the corresponding class label of x_i. W_j denotes the class center for class j and $k \in \{1, 2, ..., K\}$ where K is the number of sub-centers for each W_j. For all our experiments, we set $s = 24$, $m_1 = 1$, $m_1 = 0.5$, $m_3 = 0.1$ and $K = 3$. We use an initial learning rate of 0.01 and a cosine learning rate schedule that decays the learning rate periodically. The batch size is set to 64, and a weight decay of 0.0005 is used. We train our model with 8 Tesla V100 GPUs for a total of 32 epochs.

Model Architecture We implemented our model in MxNet version 1.5.0. Here we take the ResNet 101 model as an example and share a summary of the architecture. There are four ResNet blocks in the model, and each block contains 3, 13, 30, 3 Basic Blocks. Each Basic Block uses a number of channels 64, 128, 256, and 512 respectively. No bottle neck is used in our architecture. We also swap the ReLU activation with PReLU. We use a Conv2D filter with 64 channels and kernal size of 3 as the first layer before the main ResNet blocks. A Dense layer with output dimension $d = 128$ is used for the feature extractor.

F.2 Model Training for Generalization Validation.

To validate if SPE-FR is applicable to face embedding models trained across different settings, we test on a second set of FR models trained to represent popular and the state-of-the-art choices. This suite of models are trained across IMDB [20], DeepGlint [3] and BUPT-BalancedFace [23] datasets. Sub-Center Arcface [4], CosFace [21] and ℓ_2-Softmax [17] loss functions are studied, as well as the ResNet-101[14] and a light-weight ResNet-18 model architectures. For models trained on the BUPT-BalancedFace dataset, we again introduce a leave-one-out ablation setting ("AA") for easy observation of controlled biases.

G Datasets and evaluation protocols

G.1 Testing datasets

We use the MORPH [18] dataset and RFW [22] dataset as our test data for the face verification performance and bias analysis. Basic statistics of the evaluation dataset MORPH [18] and RFW [22] are shown in Table 2 and Table 3.

The MORPH longitudinal Database is a large facial recognition database which contains over 400K images of nearly 70K subjects. The images are 8-bit color and of generally high image quality. MORPH provides descriptive statistics associated with the variables including age, gender, ethnicity, height, etc. Noticeably, it is also a longitudinal database that provides multiple images of a given subject over time. In our study, we make use of the gender and ethnicity annotation. Within each gender and ethnicity intersectional group, we sampled 1v1 face verification protocols consisting up to 40K face pairs (Table 4, some groups have fewer images), which we refer to as the **MORPH 40K protocol**. The 'Indian', 'Other', 'Unknown' groups are not included in our experiments due to insufficient number of images for reliable estimation of ground truth performance.

Racial Faces in-the-Wild (RFW) is a testing database for studying racial bias in face recognition. Four testing subsets, namely Caucasian, Asian, Indian and African, are constructed, and each contains about 3K individuals with 6K image pairs for face verification. They can be used to evaluate and compare the recognition ability of the algorithm on different races. For RFW, we used these officially released 1v1 verification protocol, so we have 6K pairs for each ethnicity group, and half are genuine pairs and half are imposter pairs.

Table 2: **MORPH Database Demographics.** MORPH [18] dataset image statistics breakdown by gender and ancestry.

MORPH		Gender		Total
		Female	Male	
Ethnicity	African	24898	155783	180681
	Asian	536	1150	1686
	European	109132	99093	208225
	Hispanic	1880	8908	10788
	Indian	66	322	388
	Other	82	93	175
	Unknown	10	102	112
Total		136604	265451	402055

Table 3: **RFW Demographics.** Racial Faces in-the-Wild (RFW) dataset image statistics breakdown by ethnicity.

Group	Num of Identities	Num of Images
Caucasian	2959	10196
Indian	2984	10308
Asian	2492	9688
African	2995	10415

Table 4: **MORPH 40K Protocol Details.** MORPH 40K Protocol statistics breakdown by ethnicity, gender, and whether the pair is genuine match.

Ethnicity	Gender	Num of Match Pairs	Num of Non-Match Pairs
african	female	25840	40000
	male	40000	40000
asian	female	777	24899
	male	1125	40000
european	female	40000	40000
	male	40000	40000
hispanic	female	2715	40000
	male	9136	40000

G.2 BUPT-BalancedFace dataset

In order to train face recognition models for our experiments, we employed the BUPT-BalancedFace dataset [23] as our training set. This dataset provides images that are balanced across four ethnicities, namely African, Asian, Caucasian and Indian, each with 7,000 identities. The images are sourced from MS-Celeb-1M dataset [12] and downloaded from Google FreeBase images. The ethnicity labels are obtained with the help of the 'Nationality' attribute of the FreeBase images and Face++ API prediction. Note that since this dataset does not provide gender annotations, we use Insightface*, an open-source face analysis repository [10, 1, 5, 7, 11, 8, 6], to assign gender labels to the images in the BUPT-BalancedFace dataset.

The demographics of the training set BUPT-BalancedFace dataset [23] are shown in Table 5.

Table 5: **BUPT BalancedFace Demographics.** BUPT BalancedFace dataset image statistics breakdown by ethnicity.

BUPT-BalancedFace	Num of Identities	Num of Images
African	7000	324376
Asian	7000	325475
Caucasian	7000	326484
Indian	7000	275095

* Insightface is an open-source face analysis software. Repository can be accessed at https://github.com/deepinsight/insightface

H Bayesian Calibration

As introduced in main paper Sec.5, we extend experiments on the AA model to the recent Bayesian calibration (BC) [16] method on the same Morph 40K subset face verification benchmark protocol (detailed in Table 4) at the same threshold. The Bayesian calibration method was implemented using the open-source code from the paper *.

To adapt the method from binary classification to distance based face verification, we made several changes to accommodate to the implementation and the given priors on hyperparameters. Specifically, input to binary classification calibration is classification probability while our input is face verification distance; classification accuracy in our case can still be classification accuracy, but for the binary indication of if a face pair is a genuine match from the same identity.

We empirically found that directly using the pairwise Euclidean distance for accuracy calibration yields nearly random accuracies (close to 50%) and very high error compared to the ground truth of using all id labels, such that the calibration is not useful. We speculate this is due to the priors been given on binary classification probabilities in the numerical range [0, 1]. Therefore, we explored several ways of mapping the Euclidean distance to the range and found the intuitive mapping to cosine similarity (clipped to fit the range [0, 1]) to give the best results (see Table 6). As for the threshold at which we decide the classification boundary and calculate accuracy, the default setting for binary classification is 0.5. However, we adapted the threshold selected on FNMR (False Non-Match Rate) at FMR=0.001 over the entire Morph dataset. This is consistent with the SPE threshold in the main text Figure 4(b) plot, and the generated results have been shown in the main text Figure 6 for comparison. We conclude that although the the Bayesian calibration method has fairly good estimation of the pair classification accuracy, its estimation of FNMR is error-prone, biased towards over-estimation, and generate larger error than our method (main text Figure 6).

Though the BC method works very well in the experiments reported in the original paper and our accuracy estimation of binary classification, our experiments on FNMR@FMR consider models with very different operating characteristics and assess error rates that are very small compared to those in the original work. We also did not consider fine-tuning the calibration approach to better tailor it to the face recognition context. So while our results demonstrate that an off-the-shelf application of BC does not produce accurate estimates of face recognition system performance, we believe that tailoring BC to this setting is a promising direction for future work.

* https://github.com/disiji/bayesian-fairness-assess

Table 6: **Bayesian Calibration Results for Accuracy.** The estimated accuracy are compared with the ground truth of the accuracy calculated when using full identity labels. 90% confidence intervals, cut points being 5% and 95%, are provided in the brackets. It can been seem that the Bayesian calibration method has fairly good estimation of the pair classification accuracy.

	Ground Truth	BC Estimation [90% CI]
Caucasian Male	0.97	0.97 [0.95, 0.98]
African Male	0.99	0.97 [0.95, 0.98]
Asian Male	1.00	0.99 [0.97, 1.00]
Hispanic Male	1.00	0.99 [0.97, 0.99]
Caucasian Female	0.98	0.97 [0.94, 0.98]
African Female	0.99	0.96 [0.94, 0.98]
Asian Female	1.00	0.99 [0.96, 1.00]
Hispanic Female	1.00	0.98 [0.96, 1.00]

Bibliography

[1] 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) 24

[2] Brooks, S.P., Gelman, A.: General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics **7**(4), 434–455 (1998) 20

[3] Deepglint: http://trillionpairs.deepglint.com/overview, http://trillionpairs.deepglint.com/overview 21

[4] 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) 21

[5] 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) 24

[6] Deng, J., Guo, J., Niannan, X., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019) 24

[7] Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: CVPR (2020) 24

[8] 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) 24

[9] Efron, B.: Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. Journal of the American Statistical Association **99**(465), 96–104 (2004) 20

[10] Guo, J., Deng, J., Lattas, A., Zafeiriou, S.: Sample and computation redistribution for efficient face detection. arXiv preprint arXiv:2105.04714 (2021) 24

[11] Guo, J., Deng, J., Xue, N., Zafeiriou, S.: Stacked dense u-nets with dual transformers for robust face alignment. In: BMVC (2018) 24

[12] Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European conference on computer vision. pp. 87–102. Springer (2016) 24

[13] Hall, P., York, M.: On the calibration of silverman's test for multimodality. Statistica Sinica pp. 515–536 (2001) 17, 18

[14] 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) 21

[15] 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) 21

[16] 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) 1, 25
[17] Ranjan, R., Castillo, C.D., Chellappa, R.: L2-constrained softmax loss for discriminative face verification. arXiv preprint arXiv:1703.09507 (2017) 21
[18] 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) 22
[19] Storey, J.D.: The positive false discovery rate: a bayesian interpretation and the q-value. The Annals of Statistics **31**(6), 2013–2035 (2003) 17, 18
[20] 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) 21
[21] 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) 21
[22] 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) 22
[23] Wang, M., Zhang, Y., Deng, W.: Meta balanced network for fair face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 21, 24