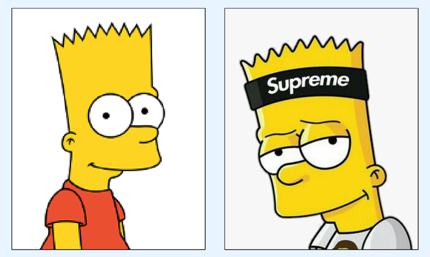
FairCal: Fairness Calibration for Face Verification

Tiago Salvador [1,3] Stephanie Cairns [1,3] Vikram Voleti [2,3] Noah Marshall [1,3] Adam Oberman [1,3]

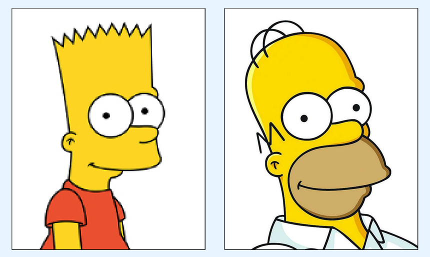
[1]McGill University [2]Université de Montréal [3]Mila

Face Verification Problem

Given two images, decide if it is a genuine/imposter pair.



genuine pair



imposter pair

Fairness in Face Verification

Chouldechova [1] showed that maximum two of the following three conditions can be satisfied:

1. **Fairness Calibration** i.e. calibrated fairly for different subgroups:

$$\mathbb{P}_{x_1,x_2\sim\mathcal{G}_1}(Y=1 \mid \widehat{C}=c) = \mathbb{P}_{x_1,x_2\sim\mathcal{G}_2}(Y=1 \mid \widehat{C}=c) = c$$

2. **Predictive Equality** i.e. equal False Positive Rates (FPRs) across different subgroups:

$$\mathbb{P}_{(x_1,x_2)\sim\mathcal{G}_1}(\widehat{Y}=1 \mid Y=0) = \mathbb{P}_{(x_1,x_2)\sim\mathcal{G}_2}(\widehat{Y}=1 \mid Y=0)$$

3. Equal Opportunity i.e. equal False Negative Rates across different subgroups:

$$\mathbb{P}_{(x_1,x_2)\sim\mathcal{G}_1}(\widehat{Y}=0 \mid Y=1) = \mathbb{P}_{(x_1,x_2)\sim\mathcal{G}_2}(\widehat{Y}=0 \mid Y=1)$$

We satisfy 1. Fairness Calibration and 2. Predictive Equality.

Bias in Face Verification

1. No prior method has targeted **Fairness Calibration**.
2. **Predictive equality** is measured by comparing the FPR on each subgroup at one global FPR:

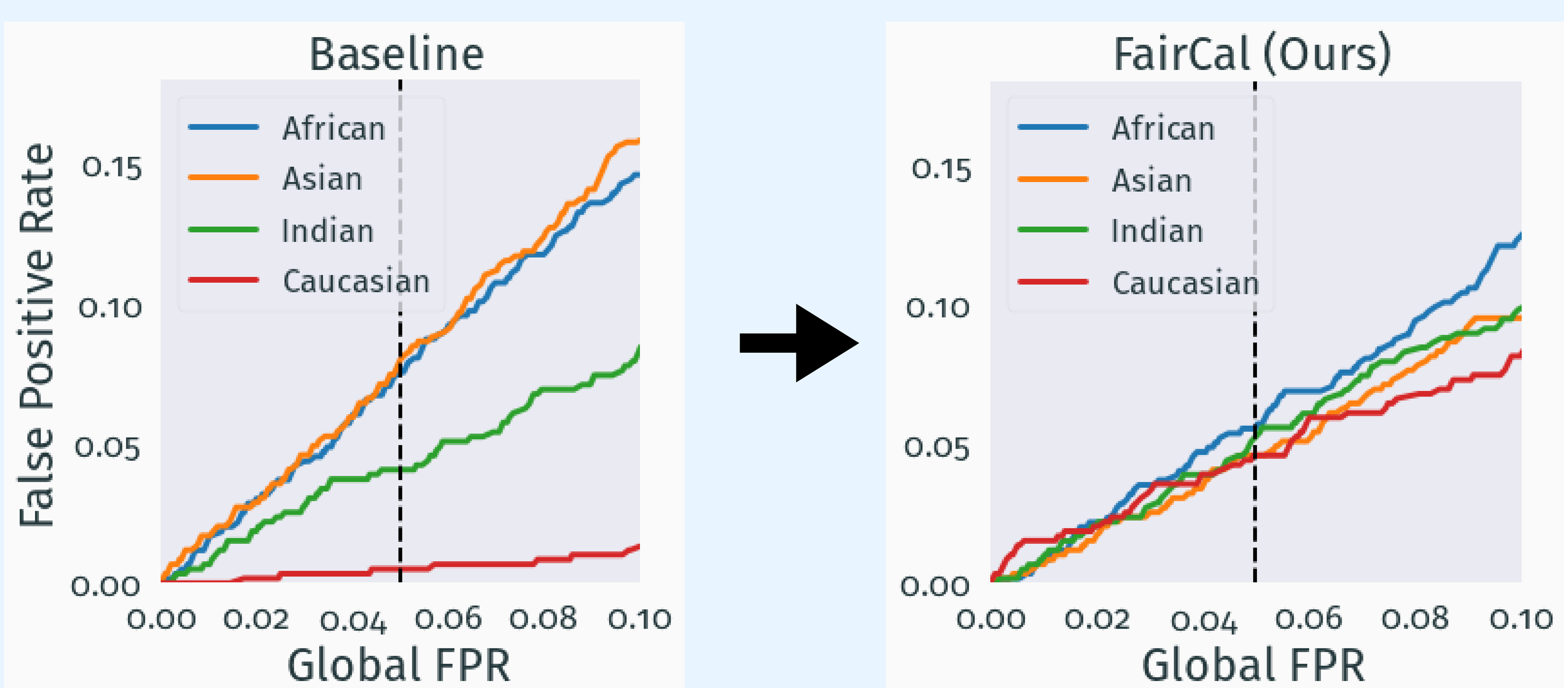


Figure 1. Predictive equality for the FaceNet (Webface) model on the RFW dataset. Lines closer together is better for fairness. At a Global FPR of 5% using the baseline method Black people are 15X more likely to false match than white people. Our method reduces this to 1.2X (while SOTA for post-hoc methods is 1.7X).

Baseline Approach

$f :=$ a trained neural network that encodes an image $\boldsymbol{x}$ into an embedding $\boldsymbol{z} = f(\boldsymbol{x})$.

1. Given an image pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$: compute the feature embedding pair $(\boldsymbol{z}_1, \boldsymbol{z}_2)$
2. Compute the cosine similarity score $s(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\boldsymbol{z}_1^T \boldsymbol{z}_2}{\|\boldsymbol{z}_1\|\|\boldsymbol{z}_2\|}$
3. Given a predefined threshold s_{thr} : $s(\boldsymbol{x}_1, \boldsymbol{x}_2) > s_{\text{thr}} \implies$ genuine pair!

We remove bias by calibrating pseudo-subgroups from unsupervised clustering. We improve Fairness Calibration, Predictive Equality, and accuracy, without knowing the sensitive attribute (group identity such as race, ethnicity, etc.), **without any additional training.**

Goals and Related Work

Work on bias mitigation for deep Face Verification models can be divided into two main camps:

(i) methods that let a model learn less-biased representations during training, and
(ii) post-processing approaches that attempt to remove bias *after* a model is trained.

Our work focus on (ii) post-hoc methods:

Post-Hoc Methods	**Fair Calibration**	Predictive Equality	Improves accuracy	Does not require sensitive attribute during training	Does not require sensitive attribute at test time	Does not require additional training
AGENDA [3]	✘	✔	✘	✘	✔	✘
PASS [2]	✘	✔	✘	✘	✔	✘
FTC [7]	✘	✔	✘	✘	✔	✘
GST [5]	✘	✔	✔	✘	✘	✔
FSN [8]	✘	✔	✔	✔	✔	✔
Oracle (Ours) [6]	✔	✔	✔	✘	✘	✔
FairCal (Ours) [6]	✔	✔	✔	✔	✔	✔

FairCal: Calibration stage

Let $\mathcal{Z}^{\text{cal}}$ denote the feature embeddings of a set of face images.

1. Apply K-means algorithm to $\mathcal{Z}^{\text{cal}}$, partitioning the embedding space into K clusters $\mathcal{Z}_1, \dots, \mathcal{Z}_K$
2. Form the K calibration sets of cosine similarity scores:

$$S_k^{\text{cal}} = \{s(\boldsymbol{x}_1, \boldsymbol{x}_2) : f(\boldsymbol{x}_1) \in \mathcal{Z}_k \text{ or } f(\boldsymbol{x}_2) \in \mathcal{Z}_k\}, \quad k = 1, \dots, K$$

3. For $k = 1, \dots, K$ estimate the calibration map μ_k that calibrates the scores:

$$\mu_k(s(\boldsymbol{x}_1, \boldsymbol{x}_2)) = \mathbb{P}[Y=1 \mid S=s, f(\boldsymbol{x}_1) \in \mathcal{Z}_k \text{ or } f(\boldsymbol{x}_2) \in \mathcal{Z}_k]$$

For **FairCal** we chose Beta Calibration [4], but experiments show similar performance with other calibration methods.

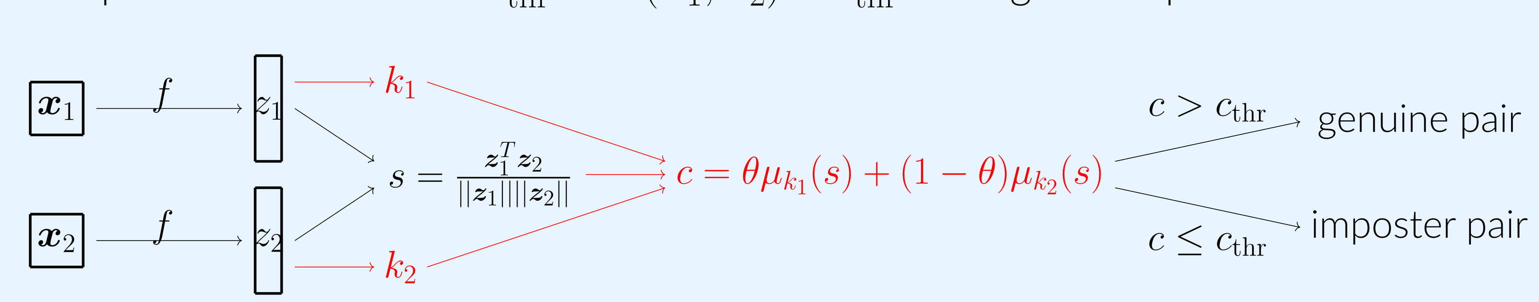
FairCal: Test stage

1. Given an image pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$, compute $(\boldsymbol{z}_1, \boldsymbol{z}_2)$, and the cluster of each image feature: k_1 and k_2
2. The model's confidence c in it being a genuine pair is:

$$c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \theta\ \mu_{k_1}(s(\boldsymbol{x}_1, \boldsymbol{x}_2)) + (1-\theta)\ \mu_{k_2}(s(\boldsymbol{x}_1, \boldsymbol{x}_2))$$

where $\theta = \frac{|S_{k_1}^{\text{cal}}|}{|S_{k_1}^{\text{cal}}| + |S_{k_2}^{\text{cal}}|}$ is the relative population fraction of the two clusters.

3. Given a predefined threshold c_{thr} : $c(\boldsymbol{x}_1, \boldsymbol{x}_2) > c_{\text{thr}} \implies$ genuine pair!



Results

Our results show that among post hoc calibration methods,

1. **FairCal** has the **best Fairness Calibration**.
2. **FairCal** has the **best Predictive Equality**, i.e., equal FPRs.
3. **FairCal** has the **best global accuracy**,
4. **FairCal does not require the sensitive attribute**, and outperforms methods that use this knowledge, including a variant of FairCal that uses the sensitive attribute (Oracle).
5. **FairCal does not require retraining** of the classifier, or any additional training.

Unsupervised Clusters

In order to not rely on the sensitive attribute like the Oracle method, our **FairCal** method uses unsupervised clusters computed with the K-means algorithm based on the feature embeddings of the images. We found them to have semantic meaning.

Caucasian Blonde Women   Indian Men with Moustache

Figure 2. Examples of clusters obtained with the K-means algorithm ($K = 100$) on the RFW dataset based on the feature embeddings computed with the FaceNet model.

References

[1] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[2] Prithviraj Dhar, Joshua Gleason, Aniket Roy, Carlos D Castillo, and Rama Chellappa. Pass: Protected attribute suppression system for mitigating bias in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15087–15096, 2021.

[3] Prithviraj Dhar, Joshua Gleason, Hossein Souri, Carlos Domingo Castillo, and Rama Chellappa. An adversarial learning algorithm for mitigating gender bias in face recognition. *CoRR*, abs/2006.07845, 2020.

[4] Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, pages 623–631. PMLR, 2017.

[5] Joseph P. Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: Too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 6 2020.

[6] Tiago Salvador, Stephanie Cairns, Vikram Voleti, Noah Marshall, and Adam M Oberman. Faircal: Fairness calibration for face verification. In *International Conference on Learning Representations*, 2022.

[7] P. Terhörst, M. L. Tran, N. Damer, F. Kirchbuchner, and A. Kuijper. Comparison-level mitigation of ethnic bias in face recognition. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2020.

[8] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognition Letters*, 140:332 – 338, 2020.

tiago.saldanhasalvador@mcgill.ca