FAIRCAL: FAIRNESS CALIBRATION FOR FACE VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite being widely used, face recognition models suffer from bias: the probability of a false positive (incorrect face match) strongly depends on sensitive attributes such as the ethnicity of the face. As a result, these models can disproportionately and negatively impact minority groups, particularly when used by law enforcement. The majority of bias reduction methods have several drawbacks: they use an end-to-end retraining approach, may not be feasible due to privacy issues, and often reduce accuracy. An alternative approach is post-processing methods that build fairer decision classifiers using the features of pre-trained models, thus avoiding the cost of retraining. However, they still have drawbacks: they reduce accuracy (AGENDA, FTC), or require retuning for different false positive rates (FSN). In this work, we introduce the Fairness Calibration (FairCal) method, a post-training approach that simultaneously: (i) increases model **accuracy** (improving the state-of-the-art), (ii) produces **fairly-calibrated** probabilities, (iii) significantly reduces the gap in the **false positive rates**, (iv) does not require knowledge of the **sensitive attribute**, and (v) does not require **retraining**, training an additional model, or retuning. We apply it to the task of Face Verification, and obtain state-of-the-art results with all the above advantages.

1 INTRODUCTION

Face recognition (FR) systems are being increasingly deployed worldwide in a variety of contexts, from policing and border control to providing security for everyday consumer electronics. According to Garvie et al. (2016), face images of around half of all American adults are searchable in police databases. FR systems have achieved impressive results in maximizing overall **accuracy** (Jain et al., 2016). However, they have also been shown to exhibit significant bias against certain demographic subgroups (Buolamwini & Gebru, 2018; Orcutt, 2016; Alvi et al., 2018), defined by a **sensitive attribute** such as ethnicity, gender, or age. Many FR systems have much higher false positive rates (FPRs) for non-white faces than white faces (Grother et al., 2019). Therefore, when FR is employed by law enforcement, non-white individuals may be more likely to be falsely detained (Allyn, 2020). Thus, it is of utmost importance to devise an easily-implementable solution to mitigate FR bias.

Most efforts to mitigate FR bias have been directed towards learning less biased representations (Liang et al., 2019; Wang et al., 2019b; Kortylewski et al., 2019; Yin et al., 2019; Gong et al., 2020; Wang & Deng, 2020; Huang et al., 2020). These approaches have enjoyed varying degrees of success: though the bias is reduced, it may still endure (Wang et al., 2019a), and this ends up reducing the **accuracy** of the system (Gong et al., 2020). Moreover, they often require **retraining** the models from scratch, which is computationally expensive. They also often require the **sensitive attribute** of the face (such as ethnicity) (Gong et al., 2020; Dhar et al., 2020; Terhörst et al., 2020a), which may not be feasible to obtain. Hence, an approach that improves **fairness** in existing models without reducing **accuracy**, or knowing the **sensitive attribute**, or require **retraining**, is missing.

In this work, we focus on the face verification problem, a major subproblem in FR: determine whether two face images depict the same person. Current state-of-the-art (SOTA) classifiers are based on deep neural networks that embed images into a low-dimensional space (Taigman et al., 2014). The "Baseline" method is one that deems a pair of images is a match if the cosine similarity between their embeddings exceeds a certain threshold. We illustrate and compare the bias of current models in Figure 1, and compare their characteristics in Table 1.

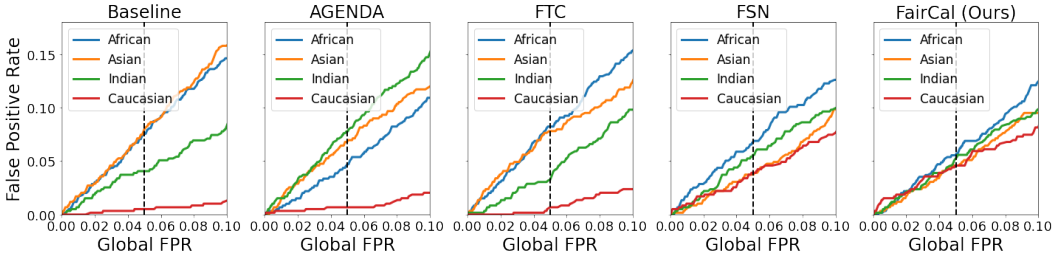


Figure 1: (Lines closer together is better for fairness, best viewed in colour) Illustration of improved fairness / reduction in bias, as measured by the FPRs evaluated on intra-ethnicity pairs on the RFW dataset with the FaceNet (Webface) feature model. The significant subgroup bias in the baseline method is reduced with post-processing methods AGENDA (Dhar et al., 2020), FTC (Terhörst et al., 2020a), FSN (Terhörst et al., 2020b), and FairCal (ours). Our FairCal performs best.

The fairness of face verification classifiers is typically assessed by three quantities, explained below: (i) **fairness calibration** across subgroups, (ii) equal false positive rates (FPRs) across subgroups i.e. **predictive equality**, (iii) equal false negative rates (FNRs) across subgroups, i.e. equal opportunity.

A classifier is said to be globally calibrated if its predicted probability p_{pred} of face matches in a dataset is equal to the fraction p_{true} of true matches in the dataset (Guo et al., 2017). This requires the classifier to additionally output an estimate of the true probability of a match, i.e. confidence. However, we suggest it is even more crucial for models to be **fairly calibrated**, i.e., for calibration to hold conditionally on each subgroup (see Definition 2). Otherwise, if the same predicted probability is known to carry different meanings for different demographic groups, users including law enforcement officers and judges may be motivated to take **sensitive attributes** into account when making critical decisions about arrests or sentencing (Pleiss et al., 2017). Moreover, previous works such as Canetti et al. (2019) and Pleiss et al. (2017) presume the use of a **fairly-calibrated** model for their success.

While global calibration can be achieved with off-the-shelf post-hoc calibration methods (Platt, 1999; Zadrozny & Elkan, 2001; Niculescu-Mizil & Caruana, 2005; Guo et al., 2017), they do not achieve **fairness-calibration**. This is illustrated in Figure 2. However, achieving all three fairness definitions mentioned above: (i) **fairness-calibration**, (ii) **equal FPRs**, (iii) equal FNRs, is impossible in practice as several authors have shown (Hardt et al., 2016; Chouldechova, 2017; Kleinberg et al., 2017; Pleiss et al., 2017). We elaborate further on this in section 3, the key takeaway is that we may aim to satisfy at most two of the three fairness notions. In the particular context of policing, **equal FPRs** i.e. **predictive equality** is considered more important than equal FNRs, as false positive errors (false arrests) risk causing significant harm, especially to members of subgroups already at disproportionate risk for police scrutiny or violence. As a result, our goals are to achieve **fairness-calibration** and **predictive equality** across subgroups, while maintaining **accuracy**, **without retraining**.

We further assume that we do not have access to the **sensitive attributes**, as this may not be feasible in practice due to (a) privacy concerns, (b) challenges in defining the sensitive attribute (e.g. ethnicity cannot be neatly divided into discrete categories), and (c) laborious and expensive to collect.

In this paper, we succeed in achieving all of the objectives mentioned above. The main contribution is the introduction of a post-training **Fairness Calibration (FairCal)** method:

1. **Accuracy**: FairCal achieves state-of-the-art accuracy (Table 2) for face verification on two large datasets.
2. **Fairness-calibration**: Our face verification classifier outputs state-of-the-art fairness-calibrated probabilities without knowledge of the sensitive attribute (Table 3).
3. **Predictive equality:** Our method reduces the gap in FPRs across sensitive attributes (Table 4). For example, in Figure 2, 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).
4. **No sensitive attribute required**: Our approach does not require the sensitive attribute, neither at training nor at test time. In fact, it outperforms models that use this knowledge.
5. **No retraining**: Our method has no need for retraining or tuning the feature representation model, or training an additional model, since it performs statistical calibration *post-hoc*.

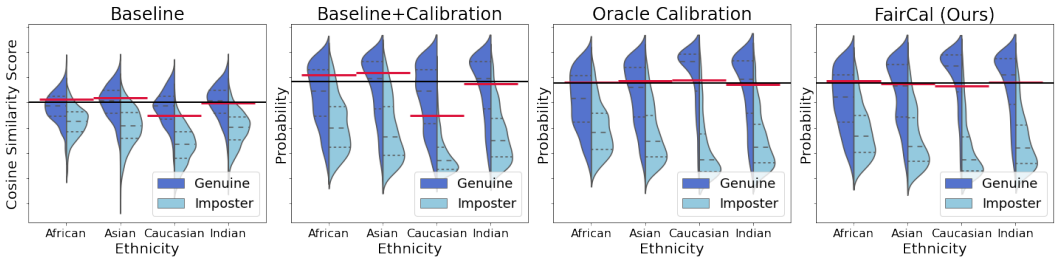


Figure 2: Illustration of bias reduction. False Positives correspond to the imposter pairs above a decision threshold value (y-axis). *Black horizontal line*: threshold which achieves global FPR of 5%; *Red lines*: thresholds which achieve subgroup FPRs of 5%. The deviation of the subgroup FPRs from the global FPR is a measure of bias (smaller is better, red lines closer to black horizontal line is better). The baseline method is biased. Calibration (based on cosine similarity alone) does not reduce the bias of the method. The Oracle method reduces bias by using subgroup membership labels for calibration. The FairCal method (ours) reduces bias by using feature vectors for calibration.

2 Related Work

Fairness, or Bias mitigation: Work on bias mitigation for deep FR models can be divided into two main camps: (i) methods that learn less biased representations, and (ii) post-processing approaches that attempt to remove bias from a pre-trained feature representation model by building fairer decision systems (Srinivas et al., 2019; Dhar et al., 2020; Terhörst et al., 2020a;b). Several approaches have been pursued in (i), including domain adaptation (Wang et al., 2019a), margin-based losses (Khan et al., 2019; Huang et al., 2020), data augmentation (Kortylewski et al., 2019), feature augmentation (Yin et al., 2019; Wang et al., 2019b), reinforcement learning (Wang & Deng, 2020), and adversarial learning (Liu et al., 2018; Gong et al., 2020). While these methods mitigate bias, this is often achieved at the expense of recognition **accuracy** and/or at **high computational cost**. The work we present in this paper fits into (ii), and thus we focus on reviewing those approaches. For an in-depth literature review of bias mitigation for FR models, see Drozdowski et al. (2020). We compare all these methods in Table 1.

Srinivas et al. (2019) proposed an ensemble approach, exploring different strategies for fusing the scores of multiple models. This work does not directly measure bias reduction, and presents only the results of applying the method to a protected subgroup. Terhörst et al. (2020b) show that while this method is effective in increasing overall **accuracy**, it fails to reliably mitigate bias.

Dhar et al. (2020) proposed the Adversarial Gender De-biasing algorithm (AGENDA) to **train** a shallow network that removes the gender information of the embeddings of a pre-trained network. Consequently, this also leads to reduced gender bias in face verification, but at the cost of **accuracy**. Also, it requires the **sensitive attributes** during training.

Terhörst et al. (2020a) proposed the Fair Template Comparison (FTC) method, which replaces the computation of the cosine similarity score by an **additional** shallow neural network trained using cross-entropy loss, with a fairness penalization term and an l_2 penalty term to prevent overfitting. While this method does indeed reduce a model's bias, it comes at the expense of an overall decrease in **accuracy**. Moreover, it requires **training** the shallow neural network and tuning the loss weights. Most importantly, it requires the **sensitive attributes** during training.

Finally, Terhörst et al. (2020b) proposed the Fair Score Normalization (FSN) method, which normalizes the scores by requiring the model's FPRs across unsupervised clusters to be the same at a predefined global FPR. This method is not designed to be **fairly-calibrated**. Moreover, if a different global FPR is desired, the method needs to be recomputed. In contrast, instead of fixing one FPR threshold, our method converts the cosine similarity scores into calibrated probabilities, which can then be used for any choice of fair FPRs. In addition, our approach can be extended to group fairness for K-classification problems, which is not possible with FSN.

Calibration: Calibration is closely related to uncertainty estimation for deep networks (Guo et al., 2017). Several post-hoc calibration methods have been proposed such as Platt's scaling or temperature scaling (Platt, 1999; Guo et al., 2017), histogram binning (Zadrozny & Elkan, 2001), isotonic

Table 1: Comparison of desirable features of the different fairness methods for face verification.

Method (requires re-training)	Improves accuracy	Fairly calibrated	Predictive equality	Does not require sensitive attribute during training	Does not require sensitive attribute at test time	Does not require re-training
D^2AE (Liu et al., 2018)	✔	✘	✘	✔	✔	✘
FTL (Yin et al., 2019)	✘	✘	✘	✔	✔	✘
LMFA+TDN (Wang et al., 2019b)	✘	✘	✘	✘	✔	✘
SYN (Kortylewski et al., 2019)	✔	✘	✘	✔	✔	✘
UMML (Khan et al., 2019)	✔	✘	✘	✔	✔	✘
CLMLE (Huang et al., 2020)	✔	✘	✘	✔	✔	✘
DebFace-ID (Gong et al., 2020)	✘	✘	✔	✘	✔	✘
RL-RBN (Wang & Deng, 2020)	✔	✘	✔	✘	✔	✘
Method (post-training)	**Improves accuracy**	**Fairly calibrated**	**Predictive equality**	**Does not require sensitive attribute during training**	**Does not require sensitive attribute at test time**	**Does not require additional training**
AGENDA (Dhar et al., 2020)	✘	✘	✔	✘	✔	✘
FTC (Terhörst et al., 2020a)	✘	✘	✔	✘	✔	✘
FSN (Terhörst et al., 2020b)	✔	✘	✔	✔	✔	✔
Oracle (Ours)	✔	✔	✔	✘	✘	✔
FairCal (Ours)	✔	✔	✔	✔	✔	✔

regression (Niculescu-Mizil & Caruana, 2005), spline calibration (Gupta et al., 2021), and beta calibration (Kull et al., 2017), among others. All of these methods involve computing a calibration function. As such, any calibration method can be readily applied, leading to models that are calibrated but not **fairly-calibrated**. An algorithm to achieve the latter for a binary classifier has been proposed in Hebert-Johnson et al. (2018), but the work remains theoretical and no practical implementation is known. **Fairly calibrated** models are the missing ingredient in the works of Canetti et al. (2019) and Pleiss et al. (2017), since they work with the presumption that a **fairly calibrated** model exists.

Closely related to face verification is the problem of face identification: to determine the identity of a person from a set of known people. However, previous works (Eickeler et al., 2000; Kral & Lenc, 2015; Eliades et al., 2019; Xie et al., 2020) did not address fairness and bias.

3 Face Verification and Fairness

In order to discuss bias mitigation strategies and their effectiveness, one must first agree on what constitutes a fair algorithm. We start by rigorously defining the face verification problem as a binary classification problem. Then we present the notion of a probabilistic classifier, which outputs calibrated probabilities that the classification is correct. Finally, since several different definitions of fairness have been proposed (see Verma & Rubin (2018); Garg et al. (2020) for a comparative analysis), we review the ones pertinent to our work.

3.1 Binary classifiers and score outputs

Let f denote a trained neural network that encodes an image $\boldsymbol{x}$ into an embedding $f(\boldsymbol{x}) = \boldsymbol{z} \in \mathcal{Z} = \mathbb{R}^d$. Pairs of images of faces are drawn from a global pair distribution $(\boldsymbol{x}_1, \boldsymbol{x}_2) \sim \mathcal{P}$. Given such a pair, let $Y(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1$ if the identities of the two images are the same and $Y = 0$ otherwise. The face verification problem consists of learning a binary classifier for Y.

The baseline classifier for the face verification problem is based on the cosine similarity between the feature embeddings of the two images, $s(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{f(\boldsymbol{x}_1)^T f(\boldsymbol{x}_2)}{\|f(\boldsymbol{x}_1)\| \|f(\boldsymbol{x}_2)\|}$. The cosine similarity score is used to define a binary classifier $\widehat{Y} : \mathcal{X} \times \mathcal{X} \to \{0, 1\}$ by thresholding with a choice of $s_{\text{thr}} \in [-1, 1]$, which is determined by a target FPR.

$$\widehat{Y}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \begin{cases} 1 & \text{if } s(\boldsymbol{x}_1, \boldsymbol{x}_2) \geq s_{\text{thr}} \\ 0 & \text{otherwise} \end{cases}$$

3.2 Calibration

A calibrated probabilistic model provides a meaningful output: its confidence reflects the probability of a genuine match, in the sense of the following definition.

Definition 1. *The probabilistic model $\widehat{C} : \mathcal{X} \times \mathcal{X} \to [0, 1]$ is said to be **calibrated** if the true probability of a match is equal to the model's confidence output c,*

$$\mathbb{P}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathcal{P}}(Y = 1 \mid \widehat{C} = c) = c.$$

Any score-based classifier can be converted to a calibrated probabilistic model using standard *post-hoc* calibration methods (Zadrozny & Elkan, 2001; Niculescu-Mizil & Caruana, 2005; Platt, 1999; Kull et al., 2017) (see Appendix F for a brief description of ones used in this work). This probabilistic model can be converted to a binary classifier by thresholding the output probability.

Applied to face verification, a calibrated probabilistic model outputs the likelihood/confidence that a pair of images $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ are a match. Calibration is achieved by using a calibration set S^{cal} to learn a calibration map μ from the cosine similarity scores to probabilities. Since calibration defines a monotonically increasing calibration map μ, a binary classifier with a score threshold s_{thr} can be obtained by thresholding the probabilistic classifier at $\mu(s_{\text{thr}})$.

3.3 Fairness

In the context of face verification, fairness implies that the classifier has similar performance on different population subgroups, i.e. being calibrated conditional on subgroup membership. Let $g(\boldsymbol{x}) \in G = \{g_i\}$ denote the **sensitive/protected attribute** of $\boldsymbol{x}$, such as ethnicity, gender, or age. Each **sensitive attribute** $g_i \in G$ induces a population subgroup, whose distribution on the intra-subgroup pairs we denote by $\mathcal{G}_i$.

Definition 2. *The probabilistic model $\widehat{C}$ is **fairly-calibrated**[1] for subgroups g_1 and g_2 if the classifier is calibrated when conditioned on each subgroup.*

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

Intuitively, a model is considered biased if its **accuracy** alters when tested on different subgroups. In the case of face recognition applications, **predictive equality**, meaning equal FPRs across subgroups, is often of primary importance.

Definition 3. *A binary classifier $\widehat{Y}$ exhibits **predictive equality** for subgroups g_1 and g_2 if the classifier has equal FPRs for each subgroup,*

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

In certain applications of FR (such as office security, where high FNRs would cause disruption), equal opportunity, meaning equal FNRs across subgroups, is also important.

Definition 4. *A binary classifier $\widehat{Y}$ exhibits **equal opportunity** for subgroups g_1 and g_2 if the classifier has equal FNRs for each subgroup,*

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

Prior works (Hardt et al., 2016; Chouldechova, 2017; Kleinberg et al., 2017; Pleiss et al., 2017) have shown that it is impossible in practice to satisfy all three fairness properties at the same time: (i) **fairness calibration** (definition 2), (ii) **predictive equality** (definition 3), (iii) equal opportunity (definition 4). At most, two of these three desirable properties can be achieved simultaneously in practice. In the context of our application, **predictive equality** is more critical than equal opportunity. Hence we choose to omit equal opportunity as our goal.

4 Fairness Calibration (FairCal)

In this section we describe our FairCal method, which constitutes the main contribution of this work. Simply calibrating a model based on cosine similarity scores will not improve fairness: if the baseline score-based classifier is biased, the resulting probabilistic classifier remains equally biased. This is illustrated in Figure 2, where for both Baseline and Baseline Calibration, any choice of global threshold will lead to different FPRs across **sensitive** subgroups; consequently, these models will fail to achieve **predictive equality**.

Our proposed solution is to introduce *conditional* calibration methods, which involve partitioning the pairs into sets, and calibrating each set. Given an image pair, we first identify its set membership, and then apply the corresponding calibration map. In our FairCal method, the sets are formed by clustering the images' feature vectors in an unsupervised fashion. In our Oracle method, the sets are defined by the **sensitive attributes** themselves. Both methods are designed to achieve **fairness-calibration**.

[1]This notion is also referred to as well-calibration (Verma & Rubin, 2018) and calibration within subgroups (Kleinberg et al., 2017; Berk et al., 2021)

4.1 FAIRNESS CALIBRATION (FAIRCAL, OUR METHOD)

(i) Apply the K-means algorithm to the image features, $\mathcal{Z}^{\text{cal}}$, partitioning the embedding space $\mathcal{Z}$ into K clusters $\mathcal{Z}_1, \ldots, \mathcal{Z}_k$. These form the K calibration sets:

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

(ii) For each calibration set S_k^{cal}, use a post-hoc calibration method to compute the calibration map μ_k that maps scores $s(\boldsymbol{x}_1, \boldsymbol{x}_2)$ to cluster-conditional probabilities $\mu_k(s(\boldsymbol{x}_1, \boldsymbol{x}_2))$. We use beta calibration (Kull et al., 2017), the recommended method when dealing with bounded scores.

(iii) For an image pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$, find the clusters they belong to. If both images fall into the same cluster k, define the pair's calibrated score as the cluster's calibrated score:

$$c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mu_k(s(\boldsymbol{x}_1, \boldsymbol{x}_2)) \tag{2}$$

Else, if the images are in different clusters, k_1 and k_2, respectively, define the pair's calibrated score as the weighted average of the calibrated scores in each cluster:

$$c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \theta\ \mu_{k_1}(s(\boldsymbol{x}_1, \boldsymbol{x}_2))\ +\ (1-\theta)\ \mu_{k_2}(s(\boldsymbol{x}_1, \boldsymbol{x}_2)), \tag{3}$$

where the weight θ is the relative population fraction of the two clusters,

$$\theta = \left|S_{k^1}^{\text{cal}}\right| / (\left|S_{k^1}^{\text{cal}}\right| + \left|S_{k^2}^{\text{cal}}\right|) \tag{4}$$

Since FairCal uses unsupervised clusters, it does not require knowledge of the **sensitive attributes**. Moreover, FairCal is a post-training statistical method, hence it does not require any **retraining**.

4.2 COMPARISON WITH FSN (TERHÖRST ET AL., 2020B)

Building the unsupervised clusters allows us to not rely on knowing the **sensitive attributes**. In contrast, the FSN method normalizes the scores by shifting them such that thresholding at a predefined global FPR leads to that same FPR on each calibration set. This is limiting in two crucial ways: 1) FSN only applies to the binary classification problem (e.g. face verification); 2) the normalizing shift in FSN depends on the global FPR chosen a priori. Our FairCal method has neither of these limitations. By converting the scores to calibrated probabilities, we can extended it to the multi-class setting, and we do not need to choose a global FPR a priori.

A simple analogy to explain the difference is : consider the problem of fairly assessing two classrooms of students with different distribution of grades. Fair calibration means the same grade should mean the same for both classrooms. Equal fail rate implies the percentage of students who fail is the same in both classrooms. Global (unfair) methods would choose to pass or fail students from both classrooms using the same threshold. Two possible fair approaches are: (A) Estimate a different threshold for each classroom based on a fixed fail rate for both. (B) Calibrate the grades so the distributions of the two classrooms match, i.e. the same grade means the same in both classrooms, and then choose a fair threshold. Method B automatically ensures equal fail rate for both classes. Method A is what FSN does, Method B is what our FairCal method achieves.

4.3 ORACLE CALIBRATION (SUPERVISED FAIRCAL, ALSO OURS)

We include a second calibration method, Oracle, which proceeds like the FairCal defined above, but creates the calibration sets based on the **sensitive attribute** instead in an unsupervised fashion. If the images belong to different subgroups i.e. $g(\boldsymbol{x}_1) \neq g(\boldsymbol{x}_2)$, then the classifier correctly outputs zero. This method is not feasible in practice, since the **sensitive attribute** may not be available, or because using the sensitive attribute may be not permitted for reasons of discrimination or privacy. However, the Oracle method represents an ideal baseline for **fairness-calibration**.

5 EXPERIMENTAL DETAILS

Models: We used three distinct pretrained models: two Inception Resnet models (Szegedy et al., 2017) obtained from Esler (2021) (MIT License), one trained on the VGGFace2 dataset (Cao et al., 2018) and another on the CASIA-Webface dataset (Yi et al., 2014), and an ArcFace model obtained

from Sharma (2021) (Apache 2.0 license) and trained on the refined version of MS-Celeb-1M (Guo et al., 2016). We will refer to the models as Facenet (VGGFace2), Facenet (Webface), and ArcFace, respectively. As is standard for face recognition models, we pre-processed the images by cropping them using a Multi-Task Convolution Neural Network (MTCNN) algorithm (Zhang et al., 2016). If the algorithm failed to identify a face, the pair was removed from the analysis.

Datasets: We present experiments on two different datasets: Racial Faces in the Wild (RFW) (Wang et al., 2019a) and Balanced Faces in the Wild (BFW) (Robinson et al., 2020), both of which are available under licenses for non-commercial research purposes only. Both datasets already include predefined pairs separated into five folds. The results we present are the product of leave-one-out cross-validation. The RFW dataset contains a 1:1 ratio of genuine/imposter pairs and 23,541 pairs in total (after applying the MTCNN). The dataset's images are labeled by ethnicity (African, Asian, Caucasian, or Indian), with all pairs consisting of same-ethnicity images. The BFW dataset, which possesses a 1:3 ratio of genuine/imposter pairs, is comprised of 890,347 pairs (after applying the MTCNN). Its images are labeled by ethnicity (African, Asian, Caucasian, or Indian) and gender (Female or Male), and it includes mixed-gender and mixed-ethnicity pairs. The RFW and BFW datasets are made up of images taken from MS-Celeb-1M (Guo et al., 2016) and VGGFace (Cao et al., 2018), respectively. Thus, to expose the models to new images, only the two FaceNet models can be evaluated on the RFW dataset, while only the FaceNet (Webface) and ArcFace models can be evaluated on the BFW dataset.

Methods: For both the FSN and FairCal method we used $K = 100$ clusters for the K-means algorithm, as recommended by Terhörst et al. (2020b). For FairCal, we employed the recently proposed beta calibration method (Kull et al., 2017). Our FairCal method is robust to the choice of number of clusters and post-hoc calibration method (see Appendix G for more details). We discuss the parameters used in training AGENDA (Dhar et al., 2020) and FTC (Terhörst et al., 2020a) in Appendix B and Appendix C, respectively.

Notice that the prior relevant methods (Baseline, AGENDA, FTC, and FSN) output scores that, even when rescaled to [0,1], do not result in calibrated probabilities and hence, by design, *cannot* be **fairly-calibrated**. Therefore, in order to fully demonstrate that our method is superior to those approaches, when measuring **fairness-calibration** we apply beta calibration to their final score outputs as well, which is the same post-hoc calibration method used in our FairCal method.

6 Results

In this section we report the performance of our models with respect to the three metrics: (i) **accuracy**, (ii) **fairness calibration**, and (iii) **predictive equality**. Our results show that among post hoc calibration methods,

1. FairCal is best at global **accuracy**, see Table 2
2. FairCal is best on **fairness calibration**, see Table 3.
3. FairCal is best on **predictive equality**, i.e., equal FPRs, see Table 4.
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.

We discuss these results in further detail below. We provide additional detailed discussion and results, including on equal opportunity and the standard deviations that result from the 5-fold cross-validation study, in the Appendix. Overall, our method that satisfies both fairness definitions without decreasing baseline model **accuracy**. In contrast, while the FTC method obtains slightly better **predictive equality** results than our method in one situation, this is achieved only at the expense of a significant decrease in **accuracy**.

It is important to emphasize that **accuracy** and **predictive equality** metrics are determined at a specified global FPR. This entails determining for each prior method a threshold that achieves the desired global FPR. However, this itself promotes our method's advantages: (a) previous methods such as FSN rely on a predefined FPR, while ours does not; (b) while other methods need recomputation for different thresholds, our method simultaneously removes the bias at different global FPRs.

Table 2: Global **accuracy** (higher is better) measured by AUROC, and TPR at two FPR thresholds.

	RFW						BFW					
	FaceNet (VGGFace2)			FaceNet (Webface)			FaceNet (Webface)			ArcFace		
(↑)	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR
Baseline	88.26	18.42	34.88	83.95	11.18	26.04	96.06	33.61	58.87	97.41	86.27	90.11
AGENDA	76.83	8.32	18.01	74.51	6.38	14.98	82.42	15.95	32.51	95.09	69.61	79.67
FTC	86.46	6.86	23.66	81.61	4.65	18.40	93.30	13.60	43.09	96.41	82.09	88.24
FSN	90.05	23.01	40.21	85.84	17.33	32.80	96.77	**47.11**	68.92	97.35	86.19	90.06
FairCal (Ours)	**90.58**	**23.55**	**41.88**	**86.71**	**20.64**	**33.13**	**96.9**	46.74	**69.21**	**97.44**	**86.28**	**90.14**
Oracle (Ours)	*89.74*	*21.4*	*41.83*	*85.23*	*16.71*	*31.6*	*97.28*	*45.13*	*67.56*	*98.91*	*86.41*	90.40

Table 3: **Fairness calibration** measured by the mean KS across the sensitive subgroups. **Bias:** measured by the deviations of KS across subgroups: Average Absolute Deviation (AAD), Maximum Absolute Deviation (MAD), and Standard Deviation (STD). (lower is better in all cases)

	RFW								BFW							
	FaceNet (VGGFace2)				FaceNet (Webface)				FaceNet (Webface)				ArcFace			
(↓)	Mean	AAD	MAD	STD	Mean	AAD	MAD	STD	Mean	AAD	MAD	STD	Mean	AAD	MAD	STD
Baseline	6.37	2.89	5.73	3.77	5.55	2.48	4.97	2.91	6.77	3.63	5.96	4.03	2.57	1.39	2.94	1.63
AGENDA	7.71	3.11	6.09	3.86	5.71	2.37	4.28	2.85	13.21	6.37	12.91	7.55	5.14	2.48	5.92	3.04
FTC	5.69	2.32	4.51	2.95	4.73	1.93	3.86	2.28	6.64	2.80	5.61	3.27	2.95	1.48	3.03	1.74
FSN	1.43	0.35	0.57	0.40	2.49	0.84	1.19	0.91	**2.76**	1.38	2.67	1.60	2.65	1.45	3.23	1.71
FairCal (Ours)	**1.37**	**0.28**	**0.50**	**0.34**	**1.75**	**0.41**	**0.64**	**0.45**	3.09	**1.34**	**2.48**	**1.55**	**2.49**	**1.30**	**2.68**	**1.52**
Oracle (Ours)	*1.18*	*0.28*	*0.53*	*0.33*	*1.35*	*0.38*	*0.66*	*0.43*	*2.23*	*1.15*	*2.63*	*1.40*	*1.41*	*0.59*	*1.30*	0.69

Table 4: **Predictive equality**: For two choices of global FPR (two blocks of rows): 0.1% and 1%, we compare the deviations in subgroup FPRs in terms of: Average Absolute Deviation (AAD), Maximum Absolute Deviation (MAD), and Standard Deviation (STD). (lower is better in all cases)

		RFW						BFW					
		FaceNet (VGGFace2)			FaceNet (Webface)			FaceNet (Webface)			ArcFace		
	(↓)	AAD	MAD	STD	AAD	MAD	STD	AAD	MAD	STD	AAD	MAD	STD
0.1% FPR	Baseline	0.10	0.15	0.10	0.14	0.26	0.16	0.29	1.00	0.40	0.12	0.30	0.15
	AGENDA	0.11	0.20	0.13	0.12	0.23	0.14	0.14	0.40	0.18	**0.09**	0.23	**0.11**
	FTC	0.10	0.15	0.11	0.12	0.23	0.14	0.24	0.74	0.32	**0.09**	**0.20**	**0.11**
	FSN	0.10	0.18	0.11	0.11	0.23	0.13	**0.09**	**0.20**	**0.11**	0.11	0.28	0.14
	FairCal (Ours)	**0.09**	**0.14**	**0.10**	**0.09**	**0.16**	**0.10**	**0.09**	**0.20**	**0.11**	0.11	0.31	0.15
	Oracle (Ours)	*0.11*	*0.19*	*0.12*	*0.11*	*0.20*	*0.13*	*0.12*	*0.25*	*0.15*	*0.12*	*0.27*	0.14
1% FPR	Baseline	0.68	1.02	0.74	0.67	1.23	0.79	2.42	7.48	3.22	0.72	1.51	0.85
	AGENDA	0.73	1.14	0.81	0.73	1.08	0.78	1.21	3.09	1.51	0.65	1.78	0.84
	FTC	0.60	0.91	0.66	0.54	1.05	0.66	1.94	5.74	2.57	**0.54**	**1.04**	**0.61**
	FSN	0.37	0.68	0.46	0.35	0.61	0.40	0.87	2.19	1.05	0.55	1.27	0.68
	FairCal (Ours)	**0.28**	**0.46**	**0.32**	**0.29**	**0.57**	**0.35**	**0.80**	**1.79**	**0.95**	0.63	1.46	0.78
	Oracle (Ours)	*0.40*	*0.69*	*0.45*	*0.41*	*0.74*	*0.48*	*0.77*	*1.71*	*0.91*	*0.83*	*2.08*	1.07

6.1 GLOBAL ACCURACY

The receiver operating characteristic (ROC) curve plots the true positive rate (TPR, equal to $1-$FNR) against the FPR, obtained by thresholding the model at different values. The area under the ROC curve (AUROC) is thus a holistic metric that summarizes the **accuracy** of the classifiers (Davis & Goadrich, 2006). A higher AUROC is better, an uninformative classifier has an AUROC of 50%.

Our FairCal method achieves SOTA results, as shown in Table 2. We report the AUROC for the different pre-trained models and datasets, as well as the TPRs at 0.1% and 1% global FPR thresholds. FairCal achieves the best values of AUROC in all cases, and the highest TPR in seven of the eight cases. In addition, our FairCal method surpasses our Oracle method (which uses **subgroup information**) on the RFW dataset.

This overall **accuracy** improvement of our FairCal method can be explained as follows: the feature vectors contain more information than the score alone, so they can be used to identify pairs where the probability of an error (at a given similarity score) is higher or lower, allowing per-cluster calibration to give better **accuracy**.

In our FairCal method, the subgroups are formed based on the features of the model, and not on manual human-assigned attributes. This allows for potentially more complex and effective subgroups. We confirmed this by plotting the images of each cluster. We found the clusters to indeed have semantic meaning, see Appendix A.

6.2 FAIRNESS CALIBRATION

The prior relevant methods (Baseline, AGENDA, FTC, and FSN) do not output probabilities, hence by design they cannot be **fairly-calibrated**. However, in order to fully demonstrate that our method is superior to those approaches, we apply beta calibration (the same post-hoc calibration method used in our FairCal method) to their score outputs.

Calibration error is a measure of the deviation in a model's output probabilities of matches from the actual results. When considering subgroups, this is measured by two quantities: (i) the calibration error on each subgroup (lower is better, since it means fewer errors), which can be measured using the Brier score (BS) (DeGroot & Fienberg, 1983), Expected Calibration Error (ECE) (Guo et al., 2017), or Kolmogorov-Smirnov (KS) (Gupta et al., 2021) (discussed in Appendix D). We present results for KS, which has been established to be the best measure (Gupta et al., 2021). (ii) The deviation in these quantities across subgroups (lower is better, since it is more fair), measured as: Average Absolute Deviation (AAD), Maximum Absolute Deviation (MAD), and Standard Deviation (STD).

Our FairCal method achieves the best results, as shown in Table 3: the best mean in three of the four cases (smaller errors) and the best deviation in nine of the nine cases (more fair across subgroups). The improvement is most significant on the less **accurate** models. We discuss these results in more detail in the Appendix.

6.3 PREDICTIVE EQUALITY

As **predictive equality** is achieved through equal FPRs between different subgroups, we can measure bias by quantifying the deviation of these FPRs. We report three measures of deviation (AAD, MAD, and STD) at two different global FPRs: 0.1% and at 1.0%.

Our FairCal method achieves the best results, as shown in Table 4: the best or tied measure of **predictive equality** in 18 of the 24 cases. In the cases where FairCal is not best, i.e., when the ArcFace model evaluated on the BFW dataset, FTC provides the best results, but the differences between AGENDA's, FTC's, FSN's, and FairCal's deviation measures are within a fraction of 1%. Moreover, when applied to ArcFace, the FTC method reduces the model's **accuracy**: at FPRs of 0.1% and 1%, the TPRs are, respectively, 4% and 2% lower.

7 CONCLUSION

We introduce FairCal, a post-training calibration method that (i) achieves **SOTA accuracy**, (ii) is **fairly-calibrated**, (iii) achieves **predictive equality (equal FPRs)** on challenging face verification datasets, (iv) **without the knowledge of any sensitive attribute**. It can be readily applied to existing FR systems, as it is a statistical method that (v) **does not require any retraining**. It can aid human-users in using FR systems to make better informed decisions based on interpretable, unbiased results.

Societal Impact: Bolstered by dramatically improving **accuracy**, Facial Recognition (FR) systems have exploded in popularity in recent years and have been applied to innumerable settings under various use cases. However, the severe biases within these systems limit their applicability and open the door to widespread social harm. In this work, we address this issue by proposing a novel post-processing method that significantly reduces the false positive rates across demographic subgroups without requiring knowledge of any sensitive attributes.

Limitations: Only under very strict conditions (e.g. a perfect classifier) can all three fairness notions (**fairness calibration**, **predictive equality**, equal opportunity) be satisfied. While we have shown that our FairCal method can achieve two out of the three definitions—which is the best we can hope for in practice—it requires the use of a calibration dataset. Deploying our method on a substantially different dataset would most likely require further calibration. Calibrating models for never-seen data is currently an open problem.

ETHICS STATEMENT

Our paper is primarily focused on addressing the lack of fairness as well as potential unethical uses of current face recognition systems, specifically in the context of face verification. At many places in the paper, we have referenced several prior works that mention and analyze the current and potential risks posed by such systems. We identified the shortcomings in the previous methods, and methodically established a scientific way of identifying and measuring fairness in the current systems. Our experiments involve large public datasets of faces. Our work explicitly addresses problems with face recognition systems that could misuse the sensitive attributes in these datasets, such as ethnicity, race, gender, etc. The motivation of our method is in trying to mitigate bias and improve fairness in the systems using these datasets.

REPRODUCIBILITY STATEMENT

In order to ensure the reproducibility of our work, we described our proposed methods, FairCal and Oracle, in detail in section 4. The details of our experiments, including the datasets and pre-processing steps, can be found in section 5. Additional details on how we implemented AGENDA and FTC, which require training additional models, are provided in Appendix B and Appendix C, respectively. The embeddings from the pretrained models were obtained on a machine with one GeForce GTX 1080 Ti GPU. All methods were implemented in Python, and the code is provided in the supplemental material.

REFERENCES

Bobby Allyn. 'The Computer Got It Wrong': How Facial Recognition Led To False Arrest Of Black Man. https://www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong-how-facial-recognition-led-to-a-false-arrest-in-michig, June 2020.

Mohsan Alvi, Andrew Zisserman, and Christoffer Nellaaker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50 (1):3–44, 2021. doi: 10.1177/0049124118782533. URL https://doi.org/10.1177/0049124118782533.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR. URL http://proceedings.mlr.press/v81/buolamwini18a.html.

Ran Canetti, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam Smith. From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 309–318, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287561. URL https://doi.org/10.1145/3287560.3287561.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

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

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 233–240, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143874. URL https://doi.org/10.1145/1143844.1143874.

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

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. URL https://arxiv.org/abs/2006.07845.

Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020. doi: 10.1109/TTS.2020.2992344.

S. Eickeler, M. Jabs, and G. Rigoll. Comparison of confidence measures for face recognition. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. IEEE Comput. Soc, 2000. doi: 10.1109/afgr.2000.840644. URL https://doi.org/10.1109/afgr.2000.840644.

Charalambos Eliades, Ladislav Lenc, Pavel Král, and Harris Papadopoulos. Automatic face recognition with well-calibrated confidence measures. *Machine Learning*, 108(3):511–534, 2019. doi: 10.1007/s10994-018-5756-7. URL https://doi.org/10.1007/s10994-018-5756-7.

Tim Esler. Face recognition using pytorch. https://github.com/timesler/facenet-pytorch, 2021.

Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 3662–3666, 2020. doi: 10.1109/BigData50022.2020.9378025.

C. Garvie, Georgetown University. Center on Privacy, Technology, and Georgetown University. Law Center. Center on Privacy & Technology. *The Perpetual Line-up: Unregulated Police Face Recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016. URL https://books.google.de/books?id=uAYjngAACAAJ.

Sixue Gong, Xiaoming Liu, and A Jain. Jointly de-biasing face recognition and demographic attribute estimation. ECCV, 2020.

Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test part 3:. Technical report, National Institute of Standards and Technology, December 2019. URL https://doi.org/10.6028/nist.ir.8280.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 1321–1330, 2017.

Yandong Guo, Lei Zhang, Yuxiao Hu, X. He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016.

Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eQe8DEWNN2W.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29, pp. 3315–3323. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/hebert-johnson18a.html.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2781–2794, 2020. doi: 10.1109/TPAMI.2019.2914680.

Anil K. Jain, Karthik Nandakumar, and Arun Ross. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters*, 79:80–105, 2016. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2015.12.013. URL https://www.sciencedirect.com/science/article/pii/S0167865515004365.

Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany*, 2017. doi: 10.4230/LIPICS.ITCS.2017.43. URL http://drops.dagstuhl.de/opus/volltexte/2017/8156/.

Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2261–2268, 2019. doi: 10.1109/CVPRW.2019.00279.

Pavel Kral and Ladislav Lenc. Confidence measure for experimental automatic face recognition system. In *Lecture Notes in Computer Science*, pp. 362–378. Springer International Publishing, 2015. doi: 10.1007/978-3-319-25210-0_22. URL https://doi.org/10.1007/978-3-319-25210-0_22.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 623–631, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. URL http://proceedings.mlr.press/v54/kull17a.html.

Jian Liang, Yuren Cao, Chenbin Zhang, Shiyu Chang, Kun Bai, and Zenglin Xu. Additive adversarial learning for unbiased authentication. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11420–11429, 2019. doi: 10.1109/CVPR.2019.01169.

Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Exploring disentangled feature representation beyond face identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2080–2089, 2018.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pp. 625–632, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102430. URL https://doi.org/10.1145/1102351.1102430.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2, 2019.

M. Orcutt. Are face recognition systems accurate? depends on your race. In *MIT Technology Review*. 2016.

Kanil Patel, William H. Beluch, Bin Yang, Michael Pfeiffer, and Dan Zhang. Multi-class uncertainty calibration via mutual information maximization-based binning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=AICNpd8ke-m.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pp. 61–74. MIT Press, 1999.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 5680–5689. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf.

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.

Abhinav Sharma. ONNX Model Zoo. https://github.com/onnx/models, 2021.

N. Srinivas, K. Ricanek, D. Michalski, D. S. Bolme, and M. King. Face recognition algorithm bias: Performance differences on images of children and adults. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2269–2277, 2019. doi: 10.1109/CVPRW.2019.00280.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. URL https://ojs.aaai.org/index.php/AAAI/article/view/11231.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708, 2014. doi: 10.1109/CVPR.2014.220.

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)*, pp. 1–6, 2020a. doi: 10.1109/IWBF49977.2020.9107956.

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, 2020b. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2020.11.007. URL http://www.sciencedirect.com/science/article/pii/S0167865520304128.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, pp. 1–7, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357463. doi: 10.1145/3194770.3194776. URL https://doi.org/10.1145/3194770.3194776.

Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in-the-wild: Reducing racial bias by information maximization adaptation network, 2019a.

Pingyu Wang, Fei Su, Zhicheng Zhao, Yandong Guo, Yanyun Zhao, and Bojin Zhuang. Deep class-skewed learning for face recognition. *Neurocomputing*, 363:35–45, 2019b. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.04.085. URL https://www.sciencedirect.com/science/article/pii/S092523121930832X.

Weidi Xie, Jeffrey Byrne, and Andrew Zisserman. Inducing predictive uncertainty estimation for face verification. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020. URL https://www.bmvc2020-conference.com/assets/papers/0149.pdf.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch, 2014.

Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

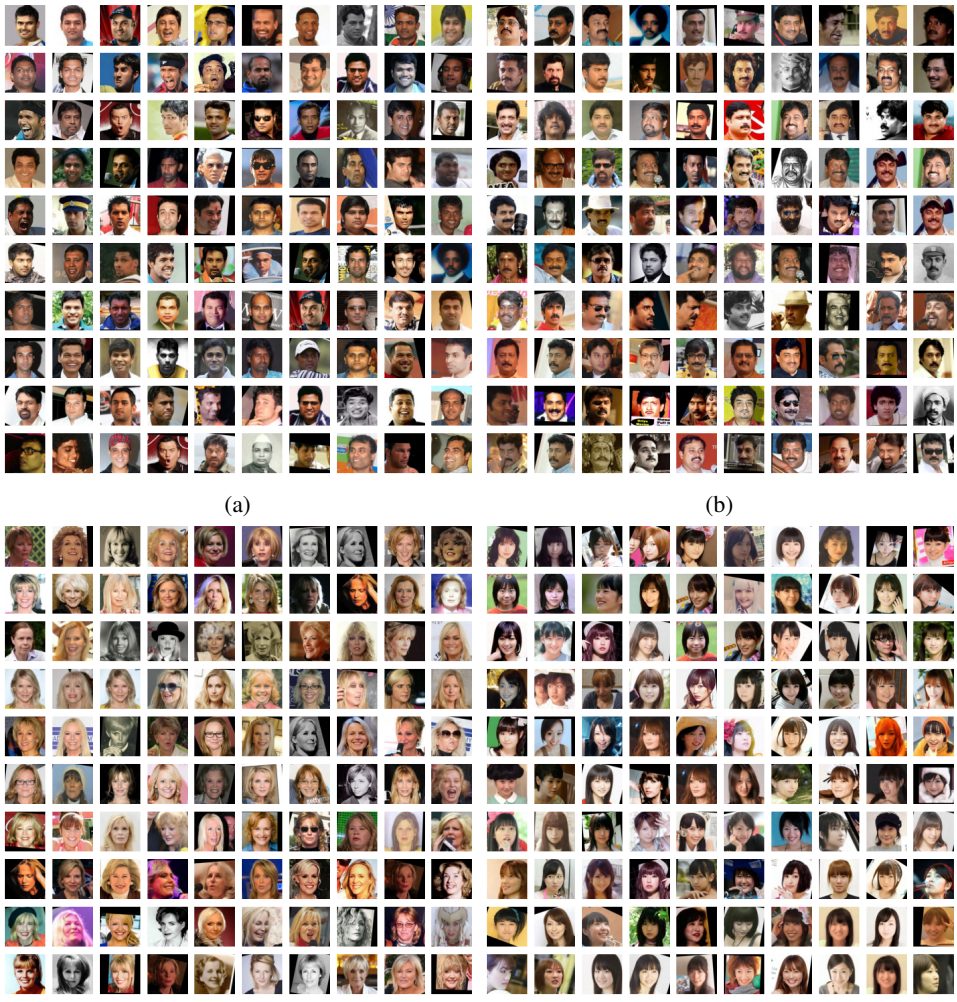
Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 609–616, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 10 2016. ISSN 1558-2361. doi: 10.1109/lsp.2016.2603342. URL http://dx.doi.org/10.1109/LSP.2016.2603342.

Supplemental Material

A EXAMPLES OF THE 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. Some examples are included in Figure 3.



(a) (b) (c) (d)

Figure 3: 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: (a) Indian men with no facial hair, (b) Indian men with moustaches, (c) Caucasian women with blond hair, (d) young Asian women with dark hair.

B AGENDA METHOD

The Adversarial Gender De-biasing algorithm (AGENDA) learns a shallow network that removes the gender information of the embeddings from a pre-trained network producing new embeddings. The algorithm entails training a generator model M, a classifier C and a discriminator E. As proposed, the algorithm only removes the gender information using an adversarial loss that encourages the discriminator to produce equal probabilities for both male and female gender. Since the RFW dataset

contains the ethnicity attribute (as opposed to gender) and the BFW dataset contains both gender and ethnicity, we modify the loss to encourage the discriminator to produce equal probabilities amongst all possible sensitive attributes.

The shallow networks used for M, C and E are the same as the ones specified in Dhar et al. (2020), with the exception that E has as many outputs as possible values of sensitive attributes (4 for RFW and 8 for BFW).

AGENDA requires the use of a validation set to determine if the discriminator should continue to be updated or not. Hence the embeddings used for training into a 80/20 split, with the 20 split used for the validation.

In Stage 1, the generator M and classifier C are trained for 50 epochs. Then the bulk of training consists of 100 episodes: (i) Stage 2 is repeated every 10 episodes and consists in training the discriminator E for 25 epochs; (ii) in Stage 3 both M and C are trained with the adversarial loss for 5 epochs with $\lambda = 10$; (iii) in Stage 4 the discriminator is updated for 5 epochs, unless its accuracy on the validations set is higher than 90%. All training is done with a batch size of 400 and an ADAM optimizer with a learning rate of 10^{-3}. For more details, we refer to the code provided in the supplemental material.

C Fair Template Comparison (FTC) method

The Fair Template Comparison (FTC) method Terhörst et al. (2020a) learns a shallow network with the goal of outputting fairer decisions. We implemented the FTC method as follow. In order to keep the ratios between the dimensions of layers the same as in the original paper Terhörst et al. (2020a), we used a 512-dimensional input layer, followed by two 2048-dimensional intermediate layers. The final layer is a fully connected linear layer with 2-dimensional output with a softmax activation. All intermediate layers are followed by a ReLU activation function and dropout (with $p = 0.3$). The network was trained with a batchsize of $b = 200$ over 50 epochs, using an Adam optimizer with a learning rate of 10^{-3} and weight decay of $10^{-}4$. Two losses, one based on subgroup fairness and the other on both subgroup and individual fairness, were proposed in Terhörst et al. (2020a). Based on the paper's recommendations, we used the individual fairness loss with a trade-off parameter of $\lambda = 0.5$.

D Measuring Calibration Error

There are different metrics available to measure if a probabilistic classifier is calibrated or fairly-calibrated. Calibration error is the error between the true and estimated confidences and is typically measured by the Expected Calibration Error (ECE) Guo et al. (2017):

Despite being the most popular calibration error metric, the ECE has several weaknesses, chief among which is its dependence on the binning scheme Nixon et al. (2019). Recently, Gupta et al. (2021) introduced a simple, bin-free calibration measure. For calibrated scores $P(Y = 1|C = c) = c$ we have, by Bayes' rule:

$$P(Y = 1, C = c) = cP(C = c).$$

Inspired by the Kolmogorov-Smirnov (KS↓) statistic test, Gupta et al. (2021) proposed to measure the calibration error by comparing the cumulative distributions of $P(Y = 1, C = c)$ and $cP(C = c)$, which empirically correspond to computing the sequences

$$h_i = h_{i-1} + \mathbf{1}_{y_i=1}/N \quad \text{and} \quad \tilde{h}_i = \tilde{h}_{i-1} + c_i/N$$

with $h_0 = \tilde{h}_0 = 0$, and N is the total number of samples. Then the KS calibration error metric is given by

$$KS = \max_i \left| h_i - \tilde{h_i} \right|.$$

Another measure is the Brier score (BS↓) (DeGroot & Fienberg, 1983), which estimates the mean squared error between the correctness of prediction and the confidence score:

$$BS(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(c_i, y_i) \in \mathcal{D}} \left(\mathbf{1}_{\hat{y}_i = y_i} - c_i\right)^2 \tag{5}$$

Table 5: KS on all the pairs (Global (Gl)) and on each ethnicity subgroup (African (Af), Asian (As), Caucasian (Ca), Indian (In) using beta calibration on the RFW dataset.

	FaceNet (VGGFace2)					FaceNet (Webface)				
(↓)	Gl	Af	As	Ca	In	Gl	Af	As	Ca	In
Baseline	0.78	6.16	5.74	12.06	1.53	0.69	3.89	4.34	10.52	3.46
AGENDA	1.02	3.66	6.97	13.76	6.46	1.21	1.75	4.94	9.39	6.77
FTC	1.12	5.13	5.41	10.19	2.02	1.25	3.19	3.81	8.59	3.35
FSN	0.77	1.27	1.62	1.49	1.35	0.85	1.94	3.06	1.70	3.27
FairCal (Ours)	0.81	1.11	1.41	1.29	1.70	0.70	1.48	1.62	1.68	2.21
Oracle (Ours)	*0.76*	*0.99*	*1.28*	*1.2*	*1.25*	*0.62*	*1.54*	*1.46*	*1.13*	1.25

Table 6: KS on all the pairs (Global (Gl)) and on each ethnicity and gender subgroup (African Females (AfF), African Males (AfM), Asian Females (AsF), Asian Males (AsM), Caucasian Females (CF), Caucasian Males (CM), Indian Females (IF), Indian Males (IM)) using beta calibration on the BFW dataset.

	FaceNet (Webface)									ArcFace								
(↓)	Gl	AfF	AfM	AsF	AsM	CF	CM	IF	IM	Gl	AfF	AfM	AsF	AsM	CF	CM	IF	IM
Baseline	0.48	5.00	2.17	11.19	2.93	12.06	10.41	5.58	4.80	0.37	1.52	3.17	5.30	4.28	1.31	1.10	2.09	1.81
AGENDA	1.44	13.49	12.66	5.64	8.52	25.18	23.43	5.72	11.06	0.99	5.39	5.44	11.07	7.91	2.66	2.26	3.33	3.09
FTC	0.56	7.33	4.06	5.71	3.68	12.25	10.47	4.13	5.51	0.49	2.02	3.56	5.77	4.62	1.80	1.03	2.65	2.15
FSN	0.39	2.35	3.12	4.16	4.40	1.50	0.99	3.54	2.02	0.38	1.74	3.01	5.70	4.30	1.02	1.15	2.45	1.81
FairCal (Ours)	0.59	3.83	2.55	2.92	3.79	3.70	2.43	3.21	2.32	0.49	1.73	3.12	4.79	3.81	1.05	1.16	2.28	1.97
Oracle (Ours)	*0.43*	*1.67*	*2.3*	*2.83*	*2.49*	*0.67*	*1.24*	*4.6*	*2.02*	*0.32*	*1.26*	*0.99*	*1.93*	*1.72*	*0.86*	*1.15*	*1.64*	1.74

For all the above metrics (ECE, KS, BS), lower is better.

E Fairness calibration and Equal Opportunity (equal FNR)

E.1 Fairness calibration

Since the calibration map produced by beta calibration is monotone, the ordering of the images provided by the scores is the same as the ordering provided by the probabilities; therefore, the accuracy of the methods wheh thresholding remains unchanged. The calibration error (CE) measured with an adaptation of the Kolmogorov-Smirnov (KS) test (described in the Appendix) is computed for each subgroup of interest. Notice that for the BFW dataset we consider the eight subgroups that result from the intersection of the ethnicity and gender subgroups.

We first observe that all methods are equally globally calibrated (i.e., the calibration error is low) after the post-hoc calibration method is applied, except for the FTC on the RFW dataset (see the Global column in Table 5 and Table 6).

By inspecting Table 5 and Table 6, we notice that, after calibration, the Baseline method results in models that are not fairly-calibrated, though perhaps not in the way one would expect. Typically, bias is directed against minority groups, but in this case, it is the Caucasian subgroups that have the higher CEs. This is a consequence of the models' above average accuracy on this subgroup, which is underestimated and therefore not captured by the calibration procedure. It is important to point out that this is not a failure of the calibration procedure, since the global CE (i.e., the CE measured on all pairs) is low, as discussed above.

E.2 Equal Opportunity

While equal opportunity (equal FNRs between subgroups) is not prioritized for the FR systems when used by law enforcement, it may be prioritized in different contexts such as office building security. Empirically, our method also mitigates the equal opportunity bias at low global FNRs, as can be seen in Table 7.

Table 7: **Equal opportunity**: Each block of rows represents a choice of global FNR: 0.1% and1%. For a fixed a global FNR, compare the deviations in subgroup FNRs in terms of three deviation measures: Average Absolute Deviation (AAD), Maximum Absolute Deviation (MAD), and Standard Deviation (STD) (lower is better).

		RFW						BFW					
		FaceNet (VGGFace2)			FaceNet (Webface)			FaceNet (Webface)			ArcFace		
	(↓)	AAD	MAD	STD	AAD	MAD	STD	AAD	MAD	STD	AAD	MAD	STD
0.1% FNR	Baseline	**0.09**	0.13	0.10	0.10	0.16	0.11	0.09	0.23	0.11	0.11	0.31	0.14
	AGENDA	0.11	0.22	0.13	0.10	**0.14**	0.11	0.14	0.34	0.16	0.09	0.24	0.12
	FTC	**0.09**	**0.11**	**0.09**	**0.08**	**0.14**	**0.1**	**0.04**	**0.09**	**0.05**	**0.06**	**0.14**	**0.07**
	FSN	**0.09**	0.13	**0.09**	0.09	**0.14**	0.10	0.07	0.22	0.10	0.12	0.33	0.15
	FairCal (Ours)	0.10	0.14	0.10	0.11	0.17	0.12	0.10	0.27	0.13	0.09	0.17	0.10
	Oracle (Ours)	*0.11*	*0.18*	*0.12*	*0.12*	*0.21*	*0.13*	*0.09*	*0.24*	*0.11*	*0.11*	*0.32*	0.14
1% FNR	Baseline	0.60	0.96	0.67	0.45	0.81	0.53	0.39	0.84	0.47	0.75	1.85	0.93
	AGENDA	0.99	1.97	1.16	0.67	1.33	0.81	0.90	2.39	1.15	0.72	1.54	0.84
	FTC	0.48	0.83	0.56	**0.32**	**0.58**	**0.38**	**0.30**	**0.62**	**0.34**	**0.49**	**1.12**	**0.60**
	FSN	**0.28**	**0.47**	**0.32**	0.40	0.78	0.48	0.41	0.92	0.49	0.77	1.91	0.96
	FairCal (Ours)	0.30	0.51	0.34	0.39	0.72	0.48	0.32	0.74	0.40	0.65	1.48	0.80
	Oracle (Ours)	*0.38*	*0.61*	*0.42*	*0.56*	*1.06*	*0.67*	*0.37*	*0.77*	*0.44*	*0.50*	*1.11*	0.60

F STANDARD POST-HOC CALIBRATION METHODS

For completeness, we provide a brief description of the post-hoc calibration methods used in this work. Beta calibration Kull et al. (2017) was used to obtain our main results, but we show below that choosing another method (histogram binning Zadrozny & Elkan (2001), isotonic regression (Zadrozny & Elkan, 2001; Niculescu-Mizil & Caruana, 2005)) does not impact the performance of our FairCal method.

F.1 HISTOGRAM BINNING

In histogram binning Zadrozny & Elkan (2001), we partition S^{cal} into m bins B_i, where $i = 1, \dots, m$. Then, given a pair of images $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ with score $s(\boldsymbol{x}_1, \boldsymbol{x}_2) \in B_i$, we define

$$c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{|B_i|} \sum_{\substack{s(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2) \in B_i \\ (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{P}^{\text{cal}}}} \mathbf{1}_{I(\hat{\boldsymbol{x}}_1) = I(\hat{\boldsymbol{x}}_2)} \tag{6}$$

In other words, we simply count the number of scores in each bin that correspond to genuine pairs of images, i.e., images that belong to the same person. By construction, a confidence score c (Equation 6) satisfies the binned version of the standard calibration (Definition 1). As for the bins, they can be chosen so as to have equal mass or to be equally spaced, or else by maximizing mutual information, as recently proposed in Patel et al. (2021). In this work, we created bins with equal mass.

Despite being an extremely computationally efficient method and providing good calibration, histogram binning is not guaranteed to preserve the monotonicity between scores and confidences, which is typically a desired property. Monotonicity ensures that the accuracy of the classifier is the same when thresholding either the scores or the calibrated confidences.

F.2 ISOTONIC REGRESSION

Isotonic Regression (Zadrozny & Elkan, 2001; Niculescu-Mizil & Caruana, 2005) learns a monotonic function $\mu : \mathbb{R} \to \mathbb{R}$ by solving

$$\arg\min_{\mu} \frac{1}{|\mathcal{P}^{\text{cal}}|} \sum_{(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{P}^{\text{cal}}} \left(\mu(s(\boldsymbol{x}_1, \boldsymbol{x}_2)) - \mathbf{1}_{I(\hat{\boldsymbol{x}}_1) = I(\hat{\boldsymbol{x}}_2)} \right)^2$$

The confidence score is then given by $c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mu(s(\boldsymbol{x}_1, \boldsymbol{x}_2))$.

F.3 BETA CALIBRATION

Beta calibration Patel et al. (2021) is a parametric calibration method, which learns a calibration map $\mu : \mathbb{R} \to \mathbb{R}$ of the form

$$c_\theta(s) = \mu(s; \theta_1, \theta_2, \theta_3) = \frac{1}{1 + 1/\left(e^{\theta_3} \frac{s^{\theta_1}}{(1-s)^{\theta_2}}\right)}$$

where the parameters $\theta_1, \theta_2, \theta_3 \in \mathbb{R}$ are chosen by minimizing the log-loss function

$$LL(c, y) = y(-\log(c)) + (1 - y)(-\log(1 - c))$$

where $c = \mu(s(\boldsymbol{x}_1, \boldsymbol{x}_2))$. By restricting, a and b to be positive, the calibration map is monotone.

G ROBUSTNESS OF FAIRCAL RESULTS TO PARAMETERS

In this section we show that the results presented in the main paper still hold if we vary model hyperparameters, such as the number K of clusters used in FairCal, and the calibration method.

Choice of post-hoc calibration: The implementation of the FairCal method requires choosing a post-hoc calibration method and the number of clusters K in the K-means algorithm. Our method is robust to the choice of both with respect to fairness-calibration (the metric of interest when it comes to calibration) and its bias as depicted in Figure 4, Figure 5, Figure 6 and Figure 7. We compare with binning and isotonic regression. For the former, we chose 10 and 25 bins for the RFW and BFW datasets, respectively, given the different number of pairs in each dataset.

Comparison to FSN (Terhörst et al., 2020b): The improved performance of FairCal over FSN is consistent across different choices of K. Fixing the choice of the post-hoc calibration method as beta calibration as in the results in the paper, we compare the two, together with Baseline and Oracle for additional baselines. Results are displayed in Figure 8, Figure 9, Figure 10, Figure 11, Figure 12, Figure 13 and Figure 14

H RESULTS PRESENTED WITH STANDARD DEVIATIONS

Recall that the results presented in the main text were computed by taking the mean of a 5-fold leave-one-out cross-validation. Below, we report the corresponding standard deviations of the five folds. The standard deviations for the results on **accuracy** reported in Table 2 can be found in Table 8, Table 9, Table 10. For fairness-calibration in Table 3, they can be found in Table 11, Table 12,Table 13, Table 14. Finally, for **predictive equality** and equal opportunity in Table 4 and Table 7, they can be found in Table 15, Table 16, Table 17 and Table 18, Table 19, Table 20.

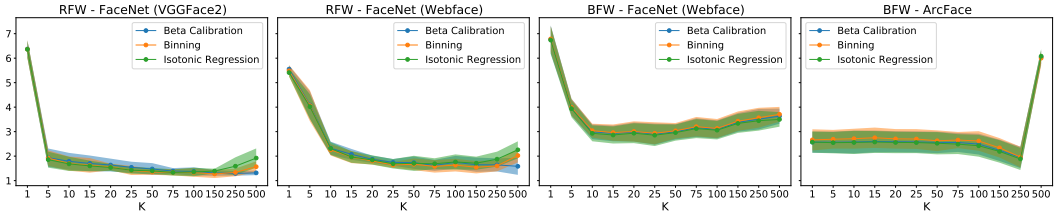


Figure 4: Comparison of **fairness-calibration** as measured by the subgroup mean of the KS across the sensitive subgroups for different values of K and different choices of post-hoc calibration methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

Table 8: Global **accuracy** measured by the AUROC.

	RFW		BFW	
(↑)	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Baseline	88.26± 0.19	83.95± 0.22	96.06± 0.16	97.41± 0.34
AGENDA	76.83± 0.57	74.51± 0.94	82.42± 0.45	95.09± 0.55
FTC	86.46± 0.17	81.61± 0.57	93.30± 0.70	96.41± 0.53
FSN	90.05± 0.26	85.84± 0.34	96.77± 0.20	97.35± 0.33
FairCal (Ours)	**90.58± 0.29**	**86.71± 0.25**	**96.90± 0.17**	**97.44± 0.34**
Oracle (Ours)	*89.74± 0.31*	*85.23± 0.18*	*97.28± 0.13*	98.91± 0.12

Table 9: Global **accuracy** measured by the TPR at 0.1% FPR threshold.

	RFW		BFW	
(↑)	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Baseline	18.42± 1.28	11.18± 3.45	33.61± 2.10	86.27± 1.09
AGENDA	8.32± 1.86	6.38± 0.78	15.95± 1.53	69.61± 2.40
FTC	6.86± 5.24	4.65± 2.10	13.60± 4.92	82.09± 1.11
FSN	23.01± 2.00	17.33± 3.01	**47.11± 1.23**	86.19± 1.13
FairCal (Ours)	**23.55± 1.82**	**20.64± 3.09**	46.74± 1.49	**86.28± 1.24**
Oracle (Ours)	*21.40± 3.54*	*16.71± 1.98*	*45.13± 1.45*	86.41± 1.19

Table 10: Global **accuracy** measured by the TPR at 1% FPR threshold.

	RFW		BFW	
(↑)	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Baseline	34.88± 3.27	26.04± 2.11	58.87± 0.92	90.11± 0.87
AGENDA	18.01± 1.44	14.98± 1.11	32.51± 1.24	79.67± 2.06
FTC	23.66± 6.58	18.40± 4.02	43.09± 5.70	88.24± 0.63
FSN	40.21± 2.09	32.80± 1.03	68.92± 1.01	90.06± 0.84
FairCal (Ours)	**41.88± 1.99**	**33.13± 1.67**	**69.21± 1.19**	**90.14± 0.86**
Oracle (Ours)	*41.83± 2.98*	*31.60± 1.08*	*67.56± 1.05*	90.40± 0.91

Table 11: **Fairness-calibration** as measured by the mean KS across sensitive subgroups.

	RFW		BFW	
(↓)	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Baseline	6.37± 0.35	5.55± 0.14	6.77± 0.57	2.57± 0.43
AGENDA	7.71± 0.27	5.71± 0.28	13.2± 1.04	5.14± 0.40
FTC	5.69± 0.14	4.73± 0.53	6.64± 0.41	2.95± 0.45
FSN	1.43± 0.28	2.49± 0.46	**2.76± 0.21**	2.65± 0.43
FairCal (Ours)	**1.37± 0.17**	**1.75± 0.26**	3.09± 0.37	**2.49± 0.43**
Oracle (Ours)	*1.18± 0.05*	*1.35± 0.09*	*2.23± 0.14*	1.41± 0.33

Table 12: Bias in **fairness-calibration** as measured by the deviations of KS across subgroups in terms of AAD (Average Absolute Deviation).

	RFW		BFW	
(↓)	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Baseline	2.89± 0.29	2.48± 0.36	3.63± 0.63	1.39± 0.28
AGENDA	3.11± 0.25	2.37± 0.33	6.37± 0.62	2.48± 0.50
FTC	2.32± 0.28	1.93± 0.35	2.80± 0.55	1.48± 0.31
FSN	0.35± 0.15	0.84± 0.38	1.38± 0.27	1.45± 0.31
FairCal (Ours)	**0.28± 0.12**	**0.41± 0.19**	**1.34± 0.24**	**1.30± 0.26**
Oracle (Ours)	*0.28± 0.08*	*0.38± 0.20*	*1.15± 0.24*	0.59± 0.18

Table 13: Bias in **fairness-calibration** as measured by the deviations of KS across subgroups in terms of MAD (Maximum Absolute Deviation).

	RFW		BFW	
(↓)	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Baseline	5.73± 0.63	4.97± 0.72	5.96± 1.05	2.94± 0.99
AGENDA	6.09± 0.65	4.28± 0.38	12.9± 0.47	5.92± 1.86
FTC	4.51± 0.64	3.86± 0.70	5.61± 0.66	3.03± 0.88
FSN	0.57± 0.21	1.19± 0.38	2.67± 0.32	3.23± 0.99
FairCal (Ours)	**0.50± 0.15**	**0.64± 0.28**	**2.48± 0.41**	**2.68± 1.07**
Oracle (Ours)	*0.53± 0.18*	*0.66± 0.28*	*2.63± 0.60*	1.30± 0.29

Table 14: Bias in **fairness-calibration** as measured by the deviations of KS across subgroups in terms of STD (Standard Deviation).

	RFW		BFW	
(↓)	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
Baseline	3.77± 0.33	2.91± 0.41	4.03± 0.70	1.63± 0.40
AGENDA	3.86± 0.24	2.85± 0.33	7.55± 0.60	3.04± 0.65
FTC	2.95± 0.32	2.28± 0.43	3.27± 0.46	1.74± 0.42
FSN	0.40± 0.15	0.91± 0.36	1.60± 0.23	1.71± 0.41
FairCal (Ours)	**0.34± 0.12**	**0.45± 0.20**	**1.55± 0.24**	**1.52± 0.37**
Oracle (Ours)	*0.33± 0.10*	*0.43± 0.20*	*1.40± 0.27*	0.69± 0.18

Table 15: **Predictive equality:** Each block of rows represents a choice of global FPR: 0.1% and 1%. For a fixed a global FPR, compare the deviations in subgroup FPRs in terms of AAD (Average Absolute Deviation). We report the average and standard deviation error across the 5 folds.

		RFW		BFW	
	(↓)	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
0.1% FPR	Baseline	0.10± 0.02	0.14± 0.03	0.29± 0.04	0.12± 0.03
	AGENDA	0.11± 0.04	0.12± 0.03	0.14± 0.04	**0.09± 0.03**
	FTC	0.10± 0.02	0.12± 0.04	0.24± 0.02	0.09± 0.02
	FSN	0.10± 0.05	0.11± 0.04	**0.09± 0.03**	0.11± 0.02
	FairCal (Ours)	**0.09± 0.03**	**0.09± 0.03**	0.09± 0.02	0.11± 0.03
	Oracle (Ours)	*0.11± 0.05*	*0.11± 0.03*	*0.12± 0.03*	0.12± 0.04
1% FPR	Baseline	0.68± 0.06	0.67± 0.15	2.42± 0.14	0.72± 0.19
	AGENDA	0.73± 0.11	0.73± 0.08	1.21± 0.27	0.65± 0.13
	FTC	0.60± 0.11	0.54± 0.12	1.94± 0.22	**0.54± 0.09**
	FSN	0.37± 0.12	0.35± 0.16	0.87± 0.11	0.55± 0.11
	FairCal (Ours)	**0.28± 0.11**	**0.29± 0.10**	**0.80± 0.10**	0.63± 0.15
	Oracle (Ours)	*0.40± 0.09*	*0.41± 0.10*	*0.77± 0.17*	0.83± 0.15

Table 16: **Predictive equality:** Each block of rows represents a choice of global FPR: 0.1% and 1%. For a fixed a global FPR, compare the deviations in subgroup FPRs in terms of MAD (Maximum Absolute Deviation). We report the average and standard deviation error across the 5 folds.

		RFW		BFW	
	(↓)	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
0.1% FPR	Baseline	0.15± 0.05	0.26± 0.09	1.00± 0.28	0.30± 0.08
	AGENDA	0.20± 0.10	0.23± 0.07	0.40± 0.16	0.23± 0.10
	FTC	0.15± 0.03	0.23± 0.08	0.74± 0.22	**0.20± 0.03**
	FSN	0.18± 0.10	0.23± 0.07	0.20± 0.06	0.28± 0.08
	FairCal (Ours)	**0.14± 0.04**	**0.16± 0.06**	**0.20± 0.04**	0.31± 0.10
	Oracle (Ours)	*0.19± 0.10*	*0.20± 0.07*	*0.25± 0.06*	0.27± 0.09
1% FPR	Baseline	1.02± 0.01	1.23± 0.30	7.48± 1.75	1.51± 0.44
	AGENDA	1.14± 0.22	1.08± 0.10	3.09± 1.06	1.78± 0.76
	FTC	0.91± 0.08	1.05± 0.17	5.74± 1.73	**1.04± 0.15**
	FSN	0.68± 0.23	0.61± 0.25	2.19± 0.58	1.27± 0.35
	FairCal (Ours)	**0.46± 0.16**	**0.57± 0.23**	**1.79± 0.54**	1.46± 0.29
	Oracle (Ours)	*0.69± 0.19*	*0.74± 0.23*	*1.71± 0.59*	2.08± 0.57

Table 17: **Predictive equality:** Each block of rows represents a choice of global FPR: 0.1% and 1%. For a fixed a global FPR, compare the deviations in subgroup FPRs in terms of STD (Standard Deviation). We report the average and standard deviation error across the 5 folds.

		RFW		BFW	
	(↓)	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
0.1% FPR	Baseline	0.10± 0.03	0.16± 0.04	0.40± 0.09	0.15± 0.04
	AGENDA	0.13± 0.05	0.14± 0.04	0.18± 0.05	**0.11± 0.04**
	FTC	0.11± 0.02	0.14± 0.05	0.32± 0.05	0.11± 0.02
	FSN	0.11± 0.06	0.13± 0.04	**0.11± 0.03**	0.14± 0.03
	FairCal (Ours)	**0.10± 0.03**	**0.10± 0.03**	0.11± 0.03	0.15± 0.03
	Oracle (Ours)	*0.12± 0.05*	*0.13± 0.03*	*0.15± 0.03*	0.14± 0.04
1% FPR	Baseline	0.74± 0.04	0.79± 0.18	3.22± 0.44	0.85± 0.20
	AGENDA	0.81± 0.11	0.78± 0.06	1.51± 0.33	0.84± 0.23
	FTC	0.66± 0.09	0.66± 0.12	2.57± 0.45	**0.61± 0.08**
	FSN	0.46± 0.14	0.40± 0.17	1.05± 0.18	0.68± 0.14
	FairCal (Ours)	**0.32± 0.12**	**0.35± 0.13**	**0.95± 0.16**	0.78± 0.15
	Oracle (Ours)	*0.45± 0.11*	*0.48± 0.12*	*0.91± 0.22*	1.07± 0.18

Table 18: **Equal opportunity:** Each block of rows represents a choice of global FNR: 0.1% and 1%. For a fixed a global FNR, compare the deviations in subgroup FNRs in terms of AAD (Average Absolute Deviation). We report the average and standard deviation error across the 5 folds.

		RFW		BFW	
	(↓)	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
0.1% FPR	Baseline	0.09± 0.01	0.10± 0.02	0.09± 0.03	0.11± 0.02
	AGENDA	0.11± 0.04	0.10± 0.02	0.14± 0.01	0.09± 0.02
	FTC	0.09± 0.01	**0.08± 0.03**	**0.04± 0.02**	**0.06± 0.01**
	FSN	**0.09± 0.02**	0.09± 0.02	0.07± 0.02	0.12± 0.01
	FairCal (Ours)	0.10± 0.02	0.11± 0.02	0.10± 0.02	0.09± 0.02
	Oracle (Ours)	*0.11± 0.02*	*0.12± 0.02*	*0.09± 0.02*	0.11± 0.02
1% FPR	Baseline	0.60± 0.17	0.45± 0.09	0.39± 0.05	0.75± 0.16
	AGENDA	0.99± 0.24	0.67± 0.17	0.90± 0.09	0.72± 0.19
	FTC	0.48± 0.06	**0.32± 0.12**	**0.30± 0.07**	**0.49± 0.14**
	FSN	**0.28± 0.06**	0.40± 0.19	0.41± 0.10	0.77± 0.17
	FairCal (Ours)	0.30± 0.14	0.39± 0.12	0.32± 0.10	0.65± 0.11
	Oracle (Ours)	*0.38± 0.15*	*0.56± 0.11*	*0.37± 0.09*	0.50± 0.10

Table 19: **Equal opportunity:** Each block of rows represents a choice of global FNR: 0.1% and 1%. For a fixed a global FNR, compare the deviations in subgroup FNRs in terms of MAD (Maximum Absolute Deviation). We report the average and standard deviation error across the 5 folds.

		RFW		BFW	
	(↓)	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
0.1% FPR	Baseline	0.13± 0.02	0.16± 0.08	0.23± 0.11	0.31± 0.07
	AGENDA	0.22± 0.08	0.14± 0.08	0.34± 0.05	0.24± 0.09
	FTC	**0.11± 0.02**	0.14± 0.07	**0.09± 0.03**	**0.14± 0.04**
	FSN	0.13± 0.06	**0.14± 0.06**	0.22± 0.11	0.33± 0.06
	FairCal (Ours)	0.14± 0.06	0.17± 0.09	0.27± 0.09	0.17± 0.05
	Oracle (Ours)	*0.18± 0.07*	*0.21± 0.08*	*0.24± 0.08*	0.32± 0.15
1% FPR	Baseline	0.96± 0.21	0.81± 0.14	0.84± 0.14	1.85± 0.66
	AGENDA	1.97± 0.48	1.33± 0.35	2.39± 0.74	1.54± 0.42
	FTC	0.83± 0.21	**0.58± 0.24**	**0.62± 0.11**	**1.12± 0.29**
	FSN	**0.47± 0.15**	0.78± 0.38	0.92± 0.28	1.91± 0.67
	FairCal (Ours)	0.51± 0.25	0.72± 0.20	0.74± 0.18	1.48± 0.36
	Oracle (Ours)	*0.61± 0.15*	*1.06± 0.18*	*0.77± 0.19*	1.11± 0.27

Table 20: **Equal opportunity:** Each block of rows represents a choice of global FNR: 0.1% and 1%. For a fixed a global FNR, compare the deviations in subgroup FNRs in terms of STD (Standard Deviation). We report the average and standard deviation error across the 5 folds.

		RFW		BFW	
	(↓)	FaceNet (VGGFace2)	FaceNet (Webface)	FaceNet (Webface)	ArcFace
0.1% FPR	Baseline	0.10± 0.01	0.11± 0.03	0.11± 0.04	0.14± 0.02
	AGENDA	0.13± 0.04	0.11± 0.03	0.16± 0.02	0.12± 0.03
	FTC	**0.09± 0.01**	**0.10± 0.03**	**0.05± 0.02**	**0.07± 0.02**
	FSN	0.09± 0.03	0.10± 0.03	0.10± 0.04	0.15± 0.02
	FairCal (Ours)	0.10± 0.02	0.12± 0.03	0.13± 0.03	0.10± 0.02
	Oracle (Ours)	*0.12± 0.03*	*0.13± 0.03*	*0.11± 0.03*	0.14± 0.04
1% FPR	Baseline	0.67± 0.15	0.53± 0.09	0.47± 0.06	0.93± 0.23
	AGENDA	1.16± 0.27	0.81± 0.19	1.15± 0.17	0.84± 0.22
	FTC	0.56± 0.10	**0.38± 0.15**	**0.34± 0.06**	**0.60± 0.16**
	FSN	**0.32± 0.09**	0.48± 0.23	0.49± 0.12	0.96± 0.23
	FairCal (Ours)	0.34± 0.17	0.48± 0.14	0.40± 0.10	0.80± 0.13
	Oracle (Ours)	*0.42± 0.14*	*0.67± 0.11*	*0.44± 0.10*	0.60± 0.12

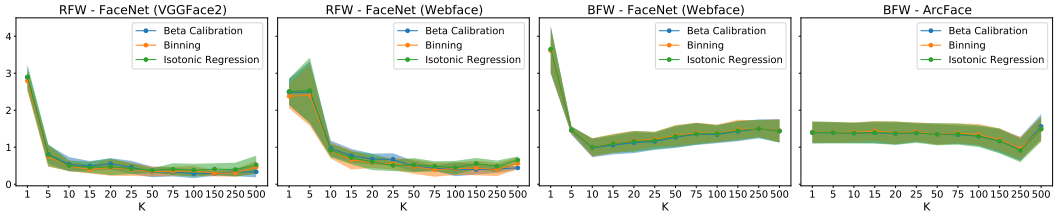


Figure 5: Bias in **fairness-calibration** as measured by the AAD (Average Absolute Deviation) in the KS across the sensitive subgroups for different values of K and different choices of post-hoc calibration methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

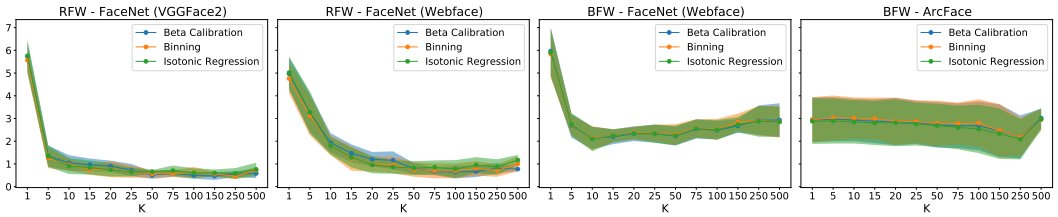


Figure 6: Bias in **fairness-calibration** as measured by the MAD (Maximum Absolute Deviation) in the KS across the sensitive subgroups for different values of K and different choices of post-hoc calibration methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

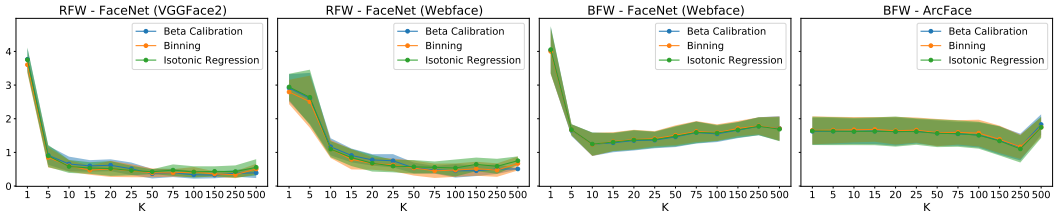


Figure 7: Bias in **fairness-calibration** as measured by the STD (Standard Deviation) in the KS across the sensitive subgroups for different values of K and different choices of post-hoc calibration methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

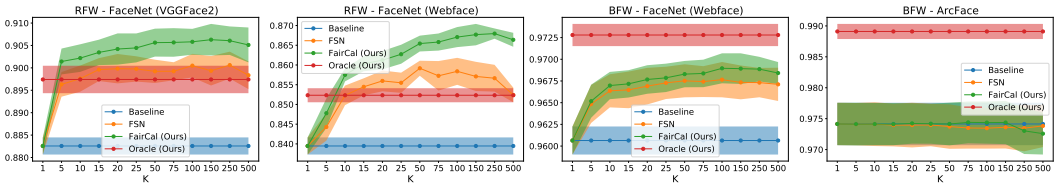


Figure 8: Global **accuracy** measured by the AUROC for different values of K for Baseline, FSN Terhörst et al. (2020b), FairCal, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

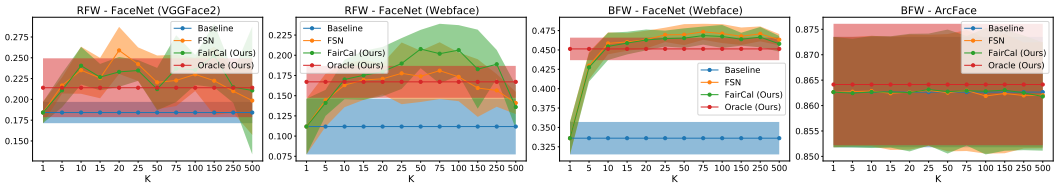


Figure 9: Global **accuracy** measure by the TPR at different a global 0.1% FPR for different values of K for Baseline, FSN Terhörst et al. (2020b), FairCal, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

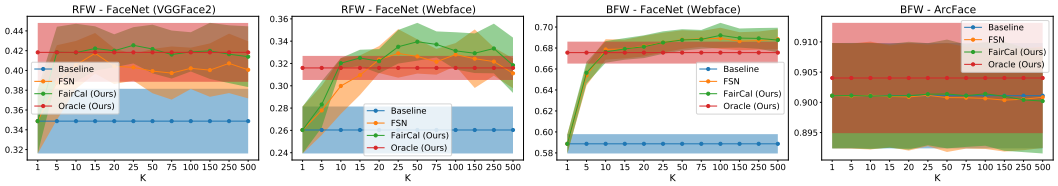


Figure 10: Global **accuracy** measure by the TPR at different a global 1% FPR for different values of K for Baseline, FSN Terhörst et al. (2020b), FairCal, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

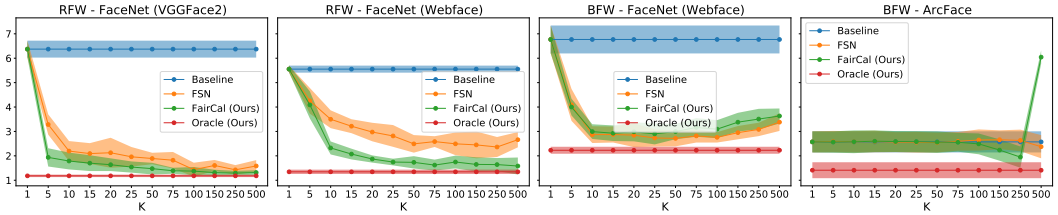


Figure 11: Comparison of **fairness-calibration** as measured by the subgroup mean of the KS across the sensitive subgroups for different values of K for Baseline, FSN Terhörst et al. (2020b), FairCal, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

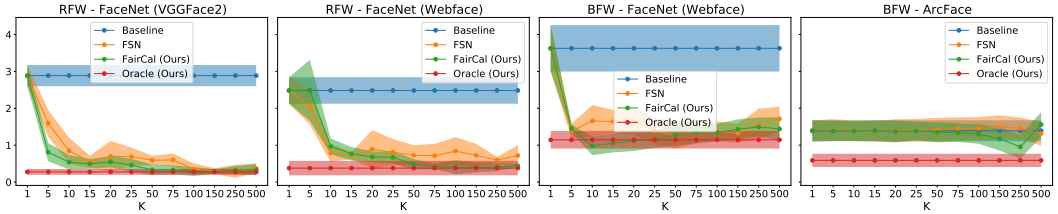


Figure 12: Bias in **fairness-calibration** as measured by the AAD (Average Absolute Deviation) in the KS across the sensitive subgroups for different values of K for Baseline, FSN Terhörst et al. (2020b), FairCal, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

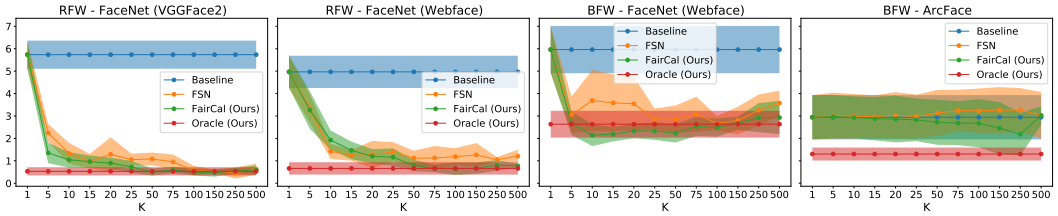


Figure 13: Bias in **fairness-calibration** as measured by the MAD (Maximum Absolute Deviation) in the KS across the sensitive subgroups for different values of K for Baseline, FSN Terhörst et al. (2020b), FairCal, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

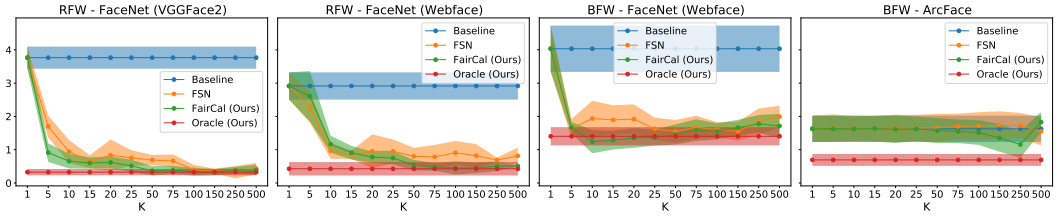


Figure 14: Bias in **fairness-calibration** as measured by the STD (Standard Deviation) in the KS across the sensitive subgroups for different values of K for Baseline, FSN Terhörst et al. (2020b), FairCal, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.