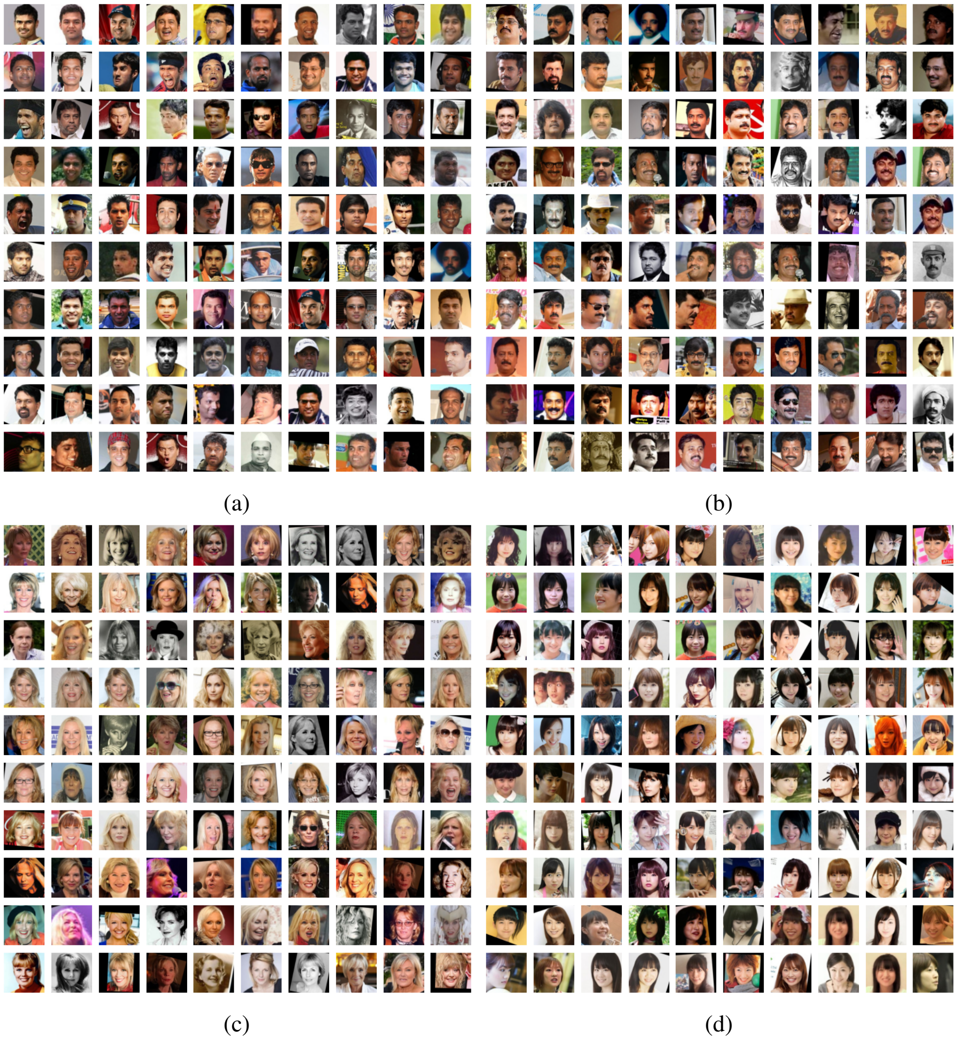
# Supplemental Material

## 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.

The ECE and KS can be seen as less informative, since the Brier score combines both calibration and sharpness (the spread of the model confidences).

## 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 $S^{\text{cal}}$ into $m$ bins $B_i$, where $i = 1, \ldots, m$. 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.

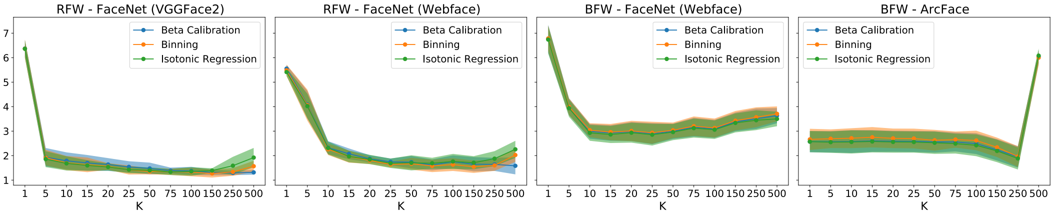


Figure 4: Comparison of **fairness-calibration** as measured by the subgroup mean of the KS across the sensitive subgroups for different values of $K$ and different choices of post-hoc calibration methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

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 |

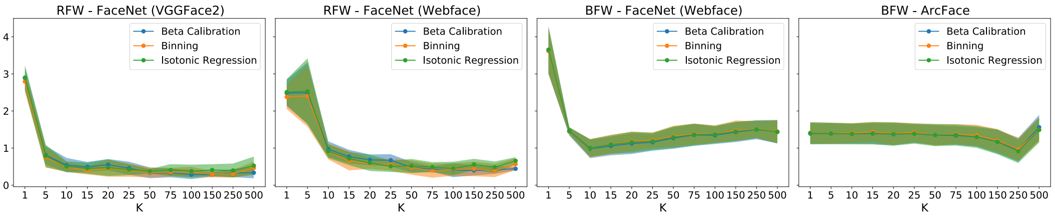


Figure 5: Bias in **fairness-calibration** as measured by the AAD (Average Absolute Deviation) in the KS across the sensitive subgroups for different values of $K$ and different choices of post-hoc calibration methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

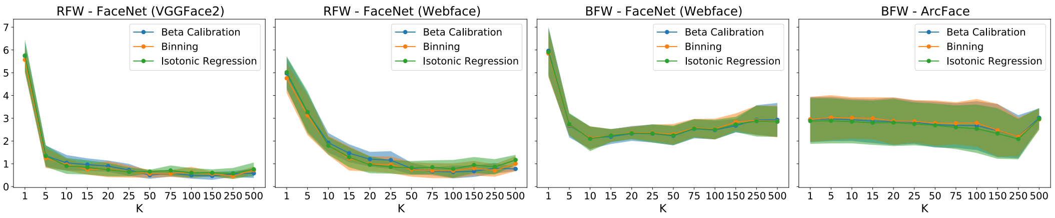


Figure 6: Bias in **fairness-calibration** as measured by the MAD (Maximum Absolute Deviation) in the KS across the sensitive subgroups for different values of $K$ and different choices of post-hoc calibration methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

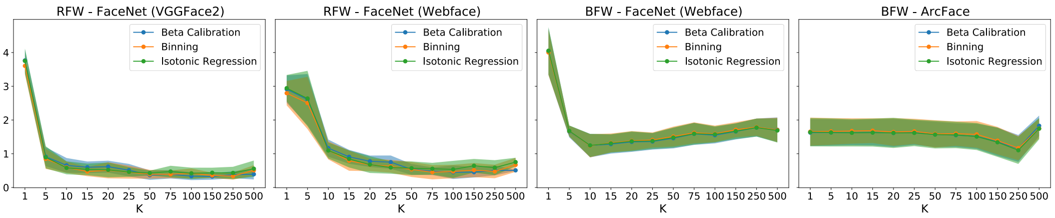


Figure 7: Bias in **fairness-calibration** as measured by the STD (Standard Deviation) in the KS across the sensitive subgroups for different values of $K$ and different choices of post-hoc calibration methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

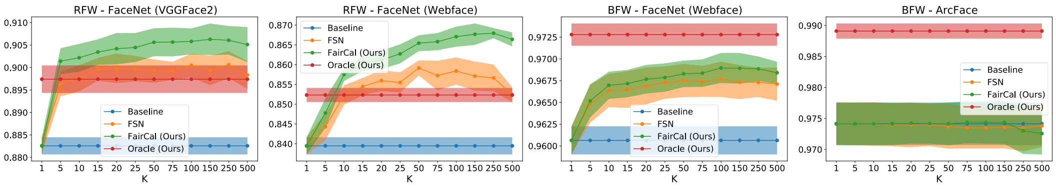


Figure 8: Global **accuracy** measured by the AUROC for different values of $K$ for Baseline, FSN Terhörst et al. (2020b), FairCal, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

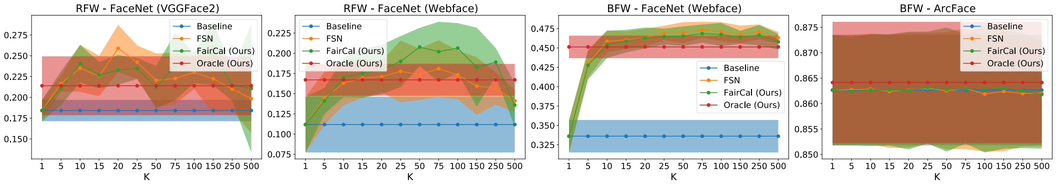


Figure 9: Global **accuracy** measure by the TPR at different a global 0.1% FPR for different values of $K$ for Baseline, FSN Terhörst et al. (2020b), FairCal, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

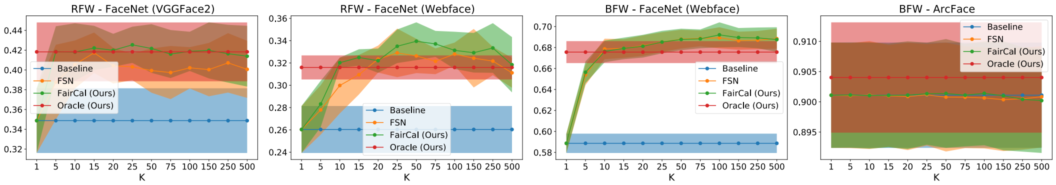


Figure 10: Global **accuracy** measure by the TPR at different a global 1% FPR for different values of $K$ for Baseline, FSN Terhörst et al. (2020b), FairCal, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the datasets.

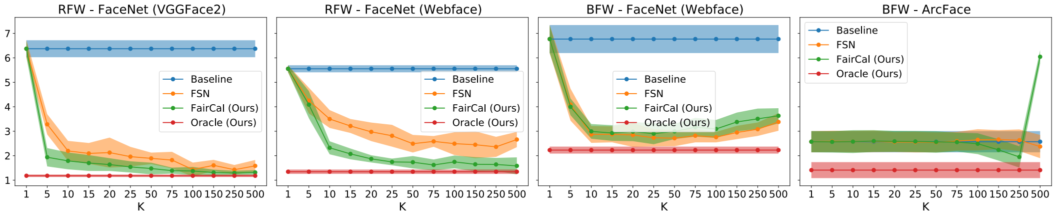


Figure 11: Comparison of **fairness-calibration** as measured by the subgroup mean of the KS across the sensitive subgroups for different values of $K$ for Baseline, FSN Terhörst et al. (2020b), FairCal, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds in the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

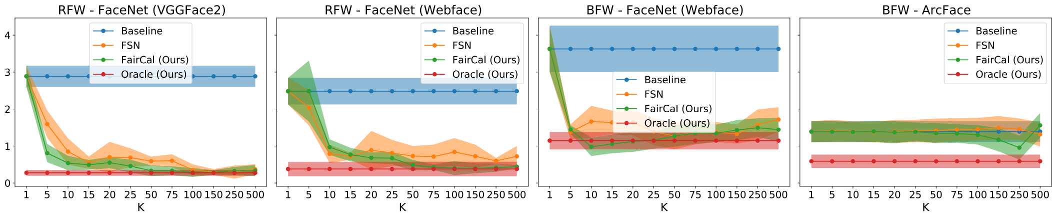


Figure 12: Bias in **fairness-calibration** as measured by the AAD (Average Absolute Deviation) in the KS across the sensitive subgroups for different values of $K$ for Baseline, FSN Terhörst et al. (2020b), FairCal, and Oracle methods. Shaded regions refer to the standard error across the 5 different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

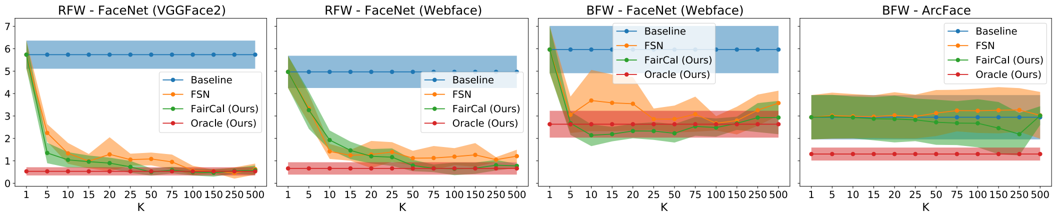


Figure 13: Bias in **fairness-calibration** as measured by the MAD (Maximum Absolute Deviation) in the KS across the sensitive subgroups for different values of $K$ for Baseline, FSN Terhörst et al. (2020b), FairCal, and Oracle methods. Shaded regions refer to the standard error across the 5 different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

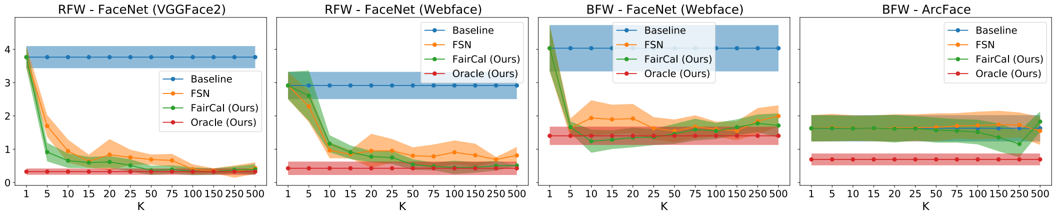


Figure 14: Bias in **fairness-calibration** as measured by the STD (Standard Deviation) in the KS across the sensitive subgroups for different values of $K$ for Baseline, FSN Terhörst et al. (2020b), FairCal, and Oracle methods. Shaded regions refer to the standard error across the 5 different folds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