

LabellessFace: Fair Metric Learning for Face Recognition without Attribute Labels

Tetsushi Ohki^{1,2}, Yuya Sato¹, Masakatsu Nishigaki¹, Koichi Ito³

¹Shizuoka University, Shizuoka, JP, ²RIKEN AIP, Tokyo, JP, ³Tohoku University, Miyagi, JP

{ohki@, sato@sec, nisigaki@}.inf.shizuoka.ac.jp, ito@aoki.ecei.tohoku.ac.jp

Abstract

Demographic bias is one of the major challenges for face recognition systems. The majority of existing studies on demographic biases are heavily dependent on specific demographic groups or demographic classifier, making it difficult to address performance for unrecognised groups. This paper introduces “LabellessFace”, a novel framework that improves demographic bias in face recognition without requiring demographic group labeling typically required for fairness considerations. We propose a novel fairness enhancement metric called the class favoritism level, which assesses the extent of favoritism towards specific classes across the dataset. Leveraging this metric, we introduce the fair class margin penalty, an extension of existing margin-based metric learning. This method dynamically adjusts learning parameters based on class favoritism levels, promoting fairness across all attributes. By treating each class as an individual in facial recognition systems, we facilitate learning that minimizes biases in authentication accuracy among individuals. Comprehensive experiments have demonstrated that our proposed method is effective for enhancing fairness while maintaining authentication accuracy.

1. Introduction

Face recognition is one of the modalities in biometric authentication systems that have seen rapid social adoption in recent years due to its convenience. As discussions about the responsibility of machine learning systems progress, it has often been pointed out that facial recognition systems also often show inconsistent performance in distinguishing between demographic attributes such as race and gender [2, 10].

Approaches to mitigate bias in inter-attribute discrimination performance can be categorized into two stages: the dataset construction and the model construction. In the dataset construction stage, efforts are made to create

datasets with balanced racial proportions [20, 21], sample or augment data to minimize disparities in recognition accuracy between attributes [13, 16], and propose methods for data augmentation [11]. In the model construction stage, strategies involve mitigating bias in model performance through score normalization between attributes [18] and dynamically adjusting hyperparameters based on attributes [20, 22, 19], under the assumption of the existence of racially biased datasets.

Most previous approaches require sensitive attribute labels (e.g., race and gender) for training the network, which limits scalability to large-scale datasets and cannot guarantee accuracy for unknown attributes. This dependence on human-annotated labels poses challenges in terms of time, cost, and potential biases, especially for emerging attributes.

This paper proposes *LabellessFace*, a novel framework that improves demographic bias in face recognition without requiring demographic group labeling. Our approach aims to maintain authentication accuracy while enhancing fairness. We introduce two key concepts: the *class favoritism level*, quantifying the degree of favoritism towards specific classes across the dataset, and the *fair class margin penalty*, extending existing metric learning methods based on class favoritism level. *LabellessFace* equalizes authentication accuracy across individuals without assuming specific sensitive attributes, achieving fairness even for unknown attributes. We conducted comprehensive experiments using common facial benchmarks, demonstrating that our method successfully improves fairness while maintaining authentication accuracy comparable to existing approaches. The results show the effectiveness of *LabellessFace* in achieving fairness across both known and unknown demographic attributes.

Our contributions are summarized as follows:

- We propose the concept of class favoritism levels, which quantifies the degree of favoritism towards specific class across the entire dataset.
- We propose the fair class margin penalty, which ex-

tends existing metric learning methods based on class favoritism levels. This realizes the LabellessFace framework that improves fairness without the need for labelling based on assumed target attributes.

- Comprehensive experiments have demonstrated that our proposed method is effective for enhancing fairness while maintaining authentication accuracy.

2. Related Work

2.1. Fairness of Facial Recognition

Facial recognition systems, increasingly integrated with surveillance cameras, are being deployed in critical scenarios such as criminal investigations, where the importance of racial fairness has been emphasised and numerous studies have been conducted. Buolamwini et al. [2] reported that facial recognition systems from companies such as Microsoft, IBM and Face++ had a misidentification rate of less than 1% for white males, while the rate for black females was around 35%. The US National Institute of Standards and Technology (NIST) conducted a fairness study of 189 facial recognition software systems. They found disparities in false positive rates between whites and blacks, with differences ranging from ten to a hundred times in 1:1 authentication scenarios. In addition, in 1:N authentication scenarios, black women had higher false detection rates [1]. Garvie et al. [7] pointed out that the bias in the racial proportions of the training data significantly affects the racial bias. They highlighted the inadvertent introduction of human discrimination due to a higher proportion of Caucasians in the data, the effect of skin colour on contrast, and the potential effect of female make-up on authentication accuracy.

2.2. Dataset-Based Approach

In the process of machine learning, the collection of datasets involves human intervention, which can unconsciously introduce bias. Therefore, in facial recognition systems, the issue of racial bias has been mainly related to the racial proportions in the training datasets. Wang et al. [20] created the BUPT-Balancedface dataset with balanced racial proportions and demonstrated that training with it could reduce racial bias compared to training with traditionally biased datasets in terms of racial proportions. Faisal et al. [13] proposed a method of resampling in which data objects are repeatedly replaced to minimise differences in recognition accuracy between races, thereby removing data that could cause racial bias from the dataset. Qraitem et al. [16] proposed a method to improve fairness by creating multiple subsets that mimic the bias in the attributes of the dataset. While techniques such as Data Augmentation are often employed in the collection of training data,

Niharika et al. [11] have pointed out the performance limitations of reducing racial bias through Data Augmentation using GANs. As an approach to optimizing a score threshold for a dataset, Pereira et al. [3] introduced the Fairness Discrepancy Rate (FDR) to assess demographic differences by assuming a single decision threshold in biometric verification systems.

2.3. Model-Based Approach

Dataset approach is an efficient method for improving fairness among races, yet constructing large-scale datasets with fairness considerations is not straightforward. Additionally, it has been pointed out that racial boundaries are ambiguous [6] and that the impact of fairness can also arise from the interrelationship between racial and environmental factors [12, 17]. Many researches has also been conducted to address racial bias by innovating the structure of models to allow fair learning even with existing datasets that are biased towards certain races. Most of the state-of-the-art methods and our proposed LabellessFace lies in this category. Philipp et al. [18] used an approach that normalizes scores between races, using the race of the authentication subject as prior information. Dooley et al. [6] proposed an approach to search for models that represent the Pareto optimal solution in terms of fairness and accuracy among multiple models. Xu et al. [22] proposed an approach where they dynamically balances the false positive rate (FPR) between each training sample without the need for demographic labels. Wang et al. [19] proposed MixFair Adapter to estimate and reduce the identity bias, the performance inconsistency between different identities, by reducing the feature discriminability differences.

Our proposed LabellessFace is inspired by the techniques of Xu et al. [22] and Wang et al. [19], particularly because it does not require demographic labels. Unlike their approaches, we consider the degree of favoritism towards specific identity across the entire dataset.

3. Proposed Method

In this section, we introduce each component of the *LabellessFace* framework using the *fair class margin penalty*. The overview of the proposed method is shown in Figure 1. In addition to existing softmax-based metric learning (section 3.1), our method dynamically sets different margins for each class based on *class favoritism level* while progressing the training through the fair class margin penalty process (section 3.2), and updates the class favoritism level at the end of each epoch (section 3.3). Here, the class favoritism level is determined based on how much the recognition accuracy for each individual deviates from the overall average using the training samples.

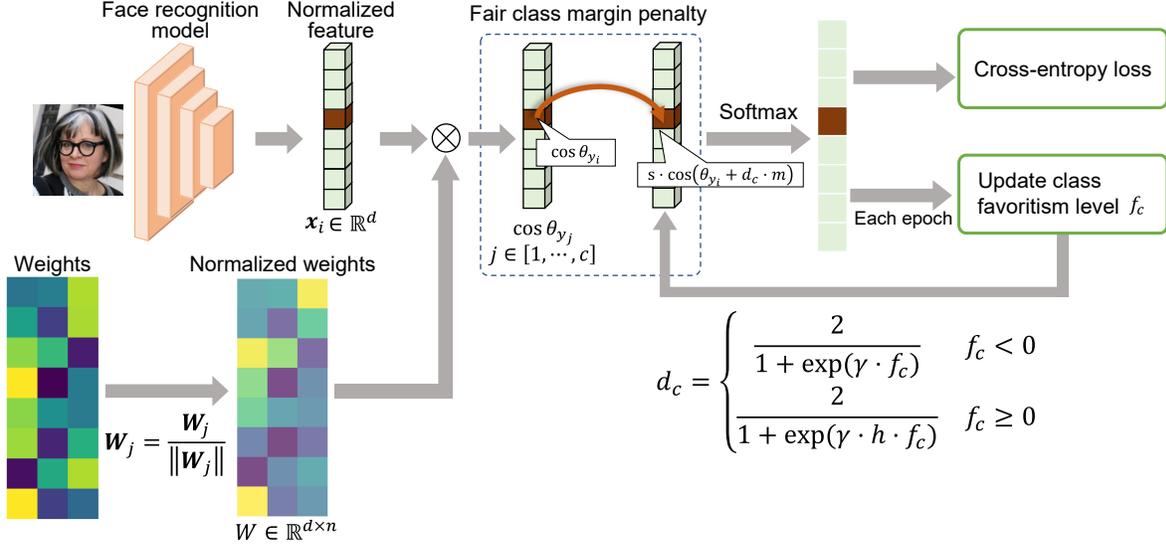


Figure 1. Overview of the LabellessFace framework.

3.1. Softmax-Based Metric Learning

The original softmax loss function is formulated as follows:

$$\mathcal{L} = -\log \frac{e^{\mathbf{W}_{y_i}^T \mathbf{x}_i + b_{y_i}}}{\sum_{j=1}^{|C|} e^{\mathbf{W}_j^T \mathbf{x}_i + b_j}}, \quad (1)$$

where $\mathbf{x}_i \in \mathbb{R}^d$ is a d -dimensional feature vector that serves as the input to the fully-connected layer corresponding to class y_i , and $\mathbf{W}_j \in \mathbb{R}^d$ represents the j -th column of the weight matrix $W \in \mathbb{R}^{d \times |C|}$. Additionally, b_j represents the bias term. Under the conditions where the bias term $b_j = 0$ and L2 regularization constraints $\|\mathbf{W}_j\| = 1$ and $\|\mathbf{x}_i\| = 1$ are applied, equation (1) expresses the angle between \mathbf{W}_{y_i} and \mathbf{x}_i as θ_{y_i} , and can be represented by

$$\mathcal{L} = -\log \frac{e^{\cos \cdot \theta_{y_i}}}{e^{\cos \cdot \theta_{y_i}} + \sum_{j=1, j \neq y_i}^{|C|} e^{\cos \cdot \theta_j}}. \quad (2)$$

In metric learning based on the softmax loss, the Angular Margin Penalty is employed to reduce intra-class variance for the correct class y_i , and a scale parameter s is used to scale $\cos \theta_i$. For example, in ArcFace [4], the equation is defined by adding the margin parameter m to equation (2) as follows:

$$\mathcal{L} = -\log \frac{e^{s \cdot (\cos \theta_{y_i} + m)}}{e^{s \cdot (\cos \theta_{y_i} + m)} + \sum_{j=1, j \neq y_i}^{|C|} e^{s \cdot (\cos \theta_j)}}. \quad (3)$$

In the proposed method, a coefficient is added to the equation (3) that allows the optimal margin to vary dynamically for each class.

3.2. Fair Class Margin Penalty

In this proposal, to minimize the bias in individual authentication accuracy, a coefficient d_c (hereafter referred to as the margin coefficient) is added to equation (3) so that the optimal margin for each class dynamically changes during training. Consequently, equation (3) is modified as

$$\mathcal{L} = -\log \frac{e^{s(\cos \theta_{y_i} + d_c \cdot m)}}{e^{s(\cos \theta_{y_i} + d_c \cdot m)} + \sum_{j=1, j \neq y_i}^{|C|} e^{s \cdot (\cos \theta_j)}}. \quad (4)$$

In this case, d_c takes different values for each class and is determined at the end of each epoch based on the class favoritism level f_c , which indicates the extent to which each class $c \in \mathcal{C}$ is favored among all classes. The margin coefficient d_c is defined by

$$d_c = \begin{cases} \frac{2}{1 + \exp(\gamma \cdot f_c)} & (f_c < 0) \\ \frac{2}{1 + \exp(\gamma \cdot h \cdot f_c)} & (f_c \geq 0) \end{cases}. \quad (5)$$

The margin coefficient d_c is designed to increase the margin's impact on classes that are less favored, thereby enlarging the facial feature space, while reducing the margin's impact on classes that are more favored, thus narrowing the facial feature space to enhance fairness.

Hyperparameters. Figure 3 shows the relationship between d_c and the class favoritism level f_c . In equation (5), the coefficient γ is a real number within the range $[0, \infty)$ and serves as a hyperparameter that determines the gradient of the margin coefficient d_c . When the gradient coefficient $\gamma = 0$, the margin coefficient d_c becomes 1, irrespective of the value of class favoritism level f_c , making it equivalent to ArcFace [4]. Additionally, the coefficient h , taking a real

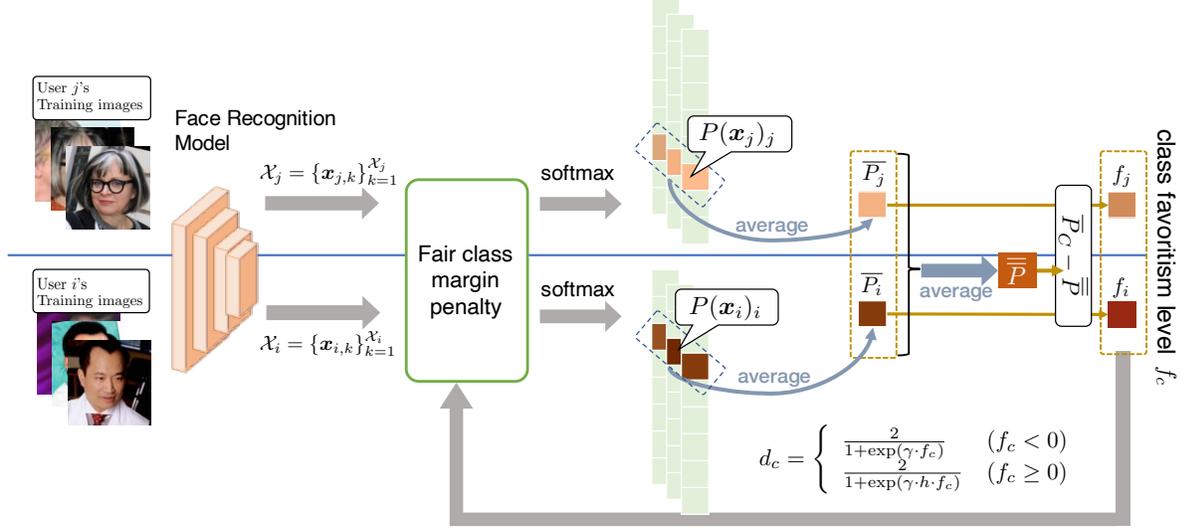


Figure 2. Overview of the Class Favoritism Level calculation.

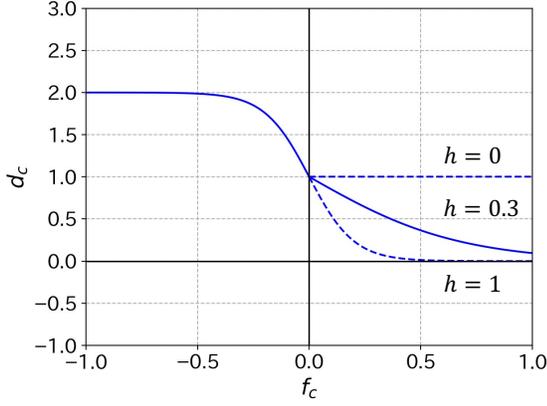


Figure 3. The gradient change of the margin coefficient d_c with respect to the value of the harmony coefficient.

number in the range $[0, 1]$, determines how much importance is placed on improving fairness among class sets. A lower value of h indicates a reluctance to allow a decrease in accuracy for favored classes, prioritizing authentication accuracy over fairness. Conversely, a higher value of h allows for a decrease in accuracy of favored classes, emphasizing fairness over authentication accuracy. In this paper, γ is referred to as the *gradient coefficient* and h as the *harmony coefficient*.

3.3. Class Favoritism Level Calculation

The class favoritism level f_c for class c is calculated at the end of each epoch using the training data and is reflected in the calculation of the margin coefficient d_c for the subsequent epoch.

The confidence derived from the softmax output in-

creases for the class to which a sample belongs as training progresses, maximizing the confidence for the sample's class while reducing the confidence for other classes. However, this tendency varies depending on the class to which the sample belongs. At this point, classes identified with relatively high confidence within the class set can be considered as favoured, whereas classes identified with low confidence are perceived as neglected. The class favoritism level f_c interprets and quantifies the difference in confidence levels among classes as a measure of fairness. An overview of the method for calculating the class favoritism level f_c is shown in Figure 2.

Let the features extracted from the training data of class c be denoted as $\mathcal{X}_c = \{\mathbf{x}_{c,k}\} (k = 1, \dots, |\mathcal{X}_c|)$, where $|\mathcal{X}_c|$ represents the total number of training data for class c . Inputting $\mathbf{x}_{c,k}$ into the face recognition model yields the softmax output $P(\mathbf{x}_{c,k})$. Here, denoting the confidence component corresponding to class c in $P(\mathbf{x}_{c,k})$ by $P(\mathbf{x}_{c,k})_c$, then the average confidence for $\mathbf{x}_{c,k}$ can be expressed as \bar{P}_c , which is given by

$$\bar{P}_c = \text{Mean}(P(\mathbf{x}_{c,k})_c) = \frac{1}{|\mathcal{X}_c|} \sum_{k=1}^{|\mathcal{X}_c|} P(\mathbf{x}_{c,k})_c. \quad (6)$$

After calculating \bar{P}_c for all classes $c \in \mathcal{C}$, the average of these values, \bar{P} , is derived as follows:

$$\bar{P} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \bar{P}_c. \quad (7)$$

The class favoritism level f_c for class c can be derived as the relative favoritism of class c within the class set \mathcal{C} . Therefore, the class favoritism level f_c is defined by the difference

Table 1. Experimental environments.

Language	Python3.8.10
GPU	NVIDIA RTX 6000 Ada \times 4
CPU	Intel(R) Xeon(R) Gold 5418Y
Memory	256GB
Kernel	6.2.0-32-generic
Distribution	Ubuntu 22.04.3 LT

between \overline{P}_c for each class and the average \overline{P} as follows:

$$f_c = \overline{P}_c - \overline{P}, \quad (8)$$

where the range of \overline{P}_c and \overline{P} is $[0,1]$, and the range of f_c is $[-1,1]$. For instance, a class with a negative class favoritism level f_c indicates that samples belonging to that class are identified with relatively low confidence, thus they are more prone to misclassification compared to other classes, and such classes can be considered as neglected. Therefore, the dynamic margin coefficient d_c monitors whether each class is favored or neglected by referencing the class favoritism level f_c regularly throughout the training process, and adjusts the optimal margin for each class at every epoch accordingly.

4. Experiments

In our experiment, we evaluate the effectiveness of a our proposed Labelless Face framework by addressing the following two viewpoints: “(1) Can our proposed method improve the fairness of certain sensitive attributes (e.g. race or gender) while maintaining accuracy? (Section 4.3.1)” and “(2) Can our proposed method improve fairness independent of annotated labels? (Section 4.3.2)”. We describe the experimental protocols used to address each research question, present the results, and discuss their implications.

4.1. Protocol

Training Dataset. For model training, the BUPT-Balancedface dataset [20], which has an equal proportion of races, was used. BUPT-Balancedface contains 7,000 classes of four races: African, Asian, Caucasian, and Indian, with racial labeling provided for each data point. The dataset, comprising 28,000 classes of facial images from these four races, was split into a training and validation ratio of 9:1. The validation data is used for early stopping decisions and for calculating the class favoritism levels f_c .

Evaluation Datasets. In the evaluation, the Labeled Faces in the Wild (LFW) [9] and Racial Faces in the Wild (RFW) [21] datasets were used. LFW is a dataset with approximately 3,000 pairs each of genuine and imposter facial image pairs prepared in advance, and each class is labeled with 74 attributes including race, age, hair color, and the presence of glasses. This dataset is used for evaluating the

discriminative performance of the model and assessing fairness across various attribute domains. RFW contains about 3,000 classes for each of the four races: African, Asian, Caucasian, and Indian, with racial labeling provided for each data point. Approximately 3,000 pairs each of genuine and imposter facial image pairs were created using RFW, and it was used to evaluate fairness in 1:1 verification.

Implementation Detail. We utilized ResNet34 [8] as a face recognition model architecture, with the layer prior to the final layer connected to a metric learning layer, trained as a classifier for 28,000 classes. The parameters of ResNet34 were initialized randomly. The batch size during training was set to 256, and the learning rate was linearly adjusted from $1e-1$ to $1e-4$. The weight decay was set at $5e-5$ and momentum at 0.9. The optimization algorithm used was SGD, and training was conducted over 30 epochs. For metric learning, the scale s was set to 64, and the initial margin m was set to 0.3. The hyperparameters for the proposed method were set to the gradient coefficient $\gamma = 10$ and the harmony coefficient $h = 1$ for the experiments.

4.2. Metrics

We quantitatively evaluated the discriminative performance of the proposed method by the Equal Error Rate (EER) and the Area Under the Curve (AUC). The fairness of the model is evaluated by the standard deviation of EER (STD), the Gini Index (Gini), and the Skewed Error Ratio (SER) across different classes. The Gini coefficient [5] is a fairness metric that represents the disparity in the cumulative distribution ratios of EER among different classes, and is defined by

$$\text{Gini} = \frac{\sum_{i=1}^{|\mathcal{C}|} \sum_{j=1}^{|\mathcal{C}|} |EER_i - EER_j|}{2|\mathcal{C}|^2 \overline{EER}}, \quad (9)$$

where $|\mathcal{C}|$ represents the total number of classes, and EER_i , EER_j refer to EER for classes i and j , respectively, \overline{EER} represents the average of EERs. SER is a fairness metric that represents the ratio between the highest and lowest error rates among attributes, and is defined by

$$\text{SER} = \frac{\max_{c \in \mathcal{C}} (EER_c)}{\min_{c \in \mathcal{C}} (EER_c)}. \quad (10)$$

4.3. Results

In our experiments, we compared the proposed method with the following approaches: (1) ArcFace [4]: A basic method for learning facial representations, (2) MagFace [15]: A method for learning facial representations considering sample quality, (3) CIFP [22]: A method to minimize the disparity in false positive rates across different races, (4) MixFairFace [19]: A method to equalize feature distances among individuals, and (5) Proposed: A method to equalize confidence level among individuals.

Table 2. The performance and fairness evaluation results trained on BUPT-Balancedface dataset and evaluated on RFW dataset: Racial attributes (Asian, African, Caucasian, Indian, referred to as As, Af, Ca, In, respectively) were selected as the subjects for evaluation. STD, Gini, SER were assessed when users were divided according to these attributes, respectively.

	EER-Af(↓)	EER-As(↓)	EER-Ca(↓)	EER-In(↓)	STD(↓)	Gini(↓)	SER(↓)
ArcFace	0.1847	0.1975	0.1145	0.1621	0.03163	0.01031	0.1725
MagFace	0.2034	0.1905	0.0989	0.1540	0.04054	0.1353	2.056
CIFP	0.1683	0.1730	0.0970	0.1293	0.03097	0.1175	1.78
MixFairFace	0.4661	0.2869	0.2928	0.3155	0.07349	0.1032	1.627
Proposed	0.1810	0.1871	0.1163	0.1625	0.02775	0.08922	1.609

Table 3. The performance and fairness evaluation results trained on BUPT-Balancedface dataset and evaluated on LFW dataset: For the LFW dataset, 26 attributes with more than 100 samples each were selected as the subjects for evaluation. STD, Gini, SER were assessed when users were divided according to these attributes, respectively.

	EER(↓)	AUC(↑)	STD(↓)	Gini(↓)	SER(↓)
ArcFace	0.09300	0.9665	0.01170	0.08292	2.766
MagFace	0.09867	0.9590	0.01127	0.08279	2.766
CIFP	0.09100	0.9614	0.01157	0.08845	3.038
Proposed	0.09100	0.9681	0.01019	0.07398	2.525

4.3.1 Fairness-Accuracy Trade-off

Our study investigated whether the proposed method could improve fairness with respect to sensitive attributes (e.g., race or gender) while maintaining accuracy. Our performance and fairness evaluation results trained on BUPT-Balancedface dataset and evaluated on the RFW datasets are shown in Table 2. EER for each race is denoted as EER-Af (African), EER-As (Asian), EER-Ca (Caucasian), and EER-In (Indian). As shown in Table 2, CIFP achieved the highest performance in EER across all racials on RFW, which is believed to be due to CIFP utilizing a training algorithm that takes into account pre-labeled racial information. In contrast, the Proposed method exhibits a lower (the best) STD/Gini/SER values compared to other methods. The differences of EERs between Proposed and ArcFace are small, indicating that Proposed improves fairness while maintaining authentication accuracy. As for MixFairFace, despite conducting replication experiments using the implementation and parameters published by the authors of original paper¹, we could not achieve high performance.

4.3.2 Label-Independent Fairness Improvement

Next, our study explored whether our proposed method could improve fairness independent of annotated labels. Fairness evaluation results trained on BUPT-Balancedface dataset and evaluated on the LFW dataset are shown in Table 3. Additionally, a comparison of fairness across the 26 attribute domains in LFW is shown as a heatmap in Figure 4. For the evaluation with the LFW dataset, 26 attributes

with more than 100 samples each were selected for analysis. STD, Gini, SER were assessed when users were divided according to these attributes, respectively. The 74 attributes of LFW are labeled with continuous values, where higher values indicate a stronger presence of the attribute [14]. Hence, continuous values were MinMax scaled to the range $[-1, 1]$, and values above 0.5 were considered to indicate the presence of the attribute in the images. MixFairFace, which did not perform well in section 4.3.1, was excluded here.

As shown in Table 3, CIFP has shown poorer performance on all fairness metrics compared to other methods, suggesting that its fairness for attributes not considered in training has deteriorated because it focuses only on racial attributes. In contrast, the Proposed method performs best on all performance and fairness metrics. These results suggest that the label-free fairness training at an individual level proposed by this method can achieve a high trade-off between accuracy and fairness even for unknown attributes.

5. Discussions

Computational cost. The calculation of the Class Favoritism Level, as shown in Figure 2, can keep the required memory capacity low by sequentially calculating \bar{P}_j during training. However, since the computation increases in proportion to the number of training data, this aspect needs to be considered.

Selection of Hyperparameters. In this method, the parameters γ and h are employed to balance the trade-off between fairness and accuracy. A grid search was performed for h and γ , revealing that the best performance was achieved with $h = 1$ and $\gamma = 10$. Higher values of h suppress the learning of favored attributes while encouraging the learn-

¹<https://github.com/fuenwang/MixFairFace>

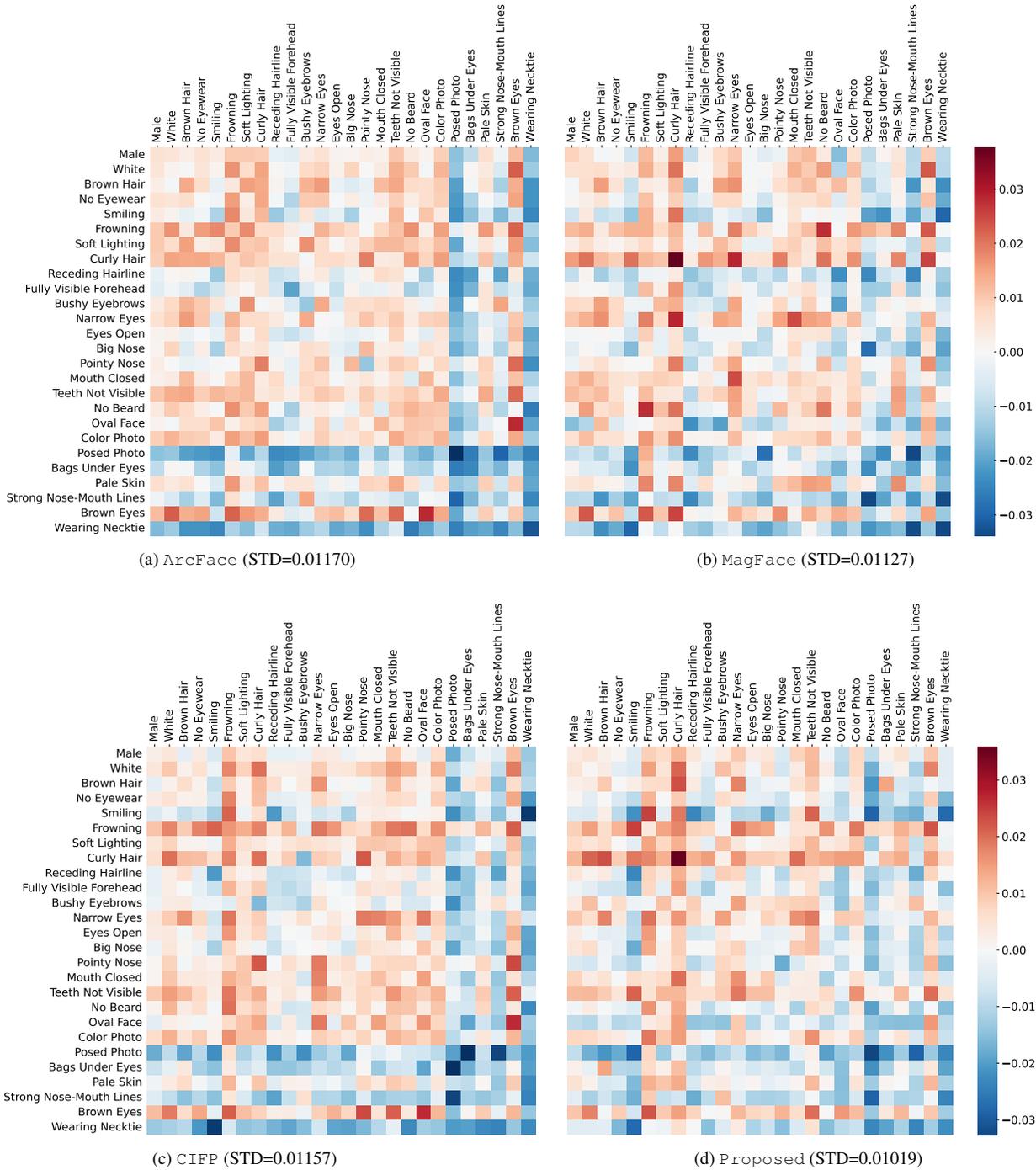


Figure 4. The fairness heatmap of each model across 26 attributes on LFW: Each cell indicates the deviation of EER, with blue indicating lower EER than the average and red indicating higher EER than the average. For reference, we include the values of the standard deviation (STD) from Table 3 in parentheses.

ing of neglected attributes. γ determines the intensity of this effect. If significant latent attribute biases are expected in the dataset, it is suggested that using larger values for h or γ could lead to greater fairness improvements. However, it should be noted that excessively large values may cause

instability in the learning process. This paper does not propose an optimal method for determining these parameters, leaving it as a topic for future research.

6. Conclusion

In this paper, we proposed a new framework, Labelless-Face, aimed at reducing authentication bias in facial recognition. This method was designed to realize a label-free training method that does not require pre-labeling for demographic groups. To achieve this goal, we introduced the concept of a fair class margin penalty based on class favoritism levels, utilizing individual class units to avoid the need for demographic group labeling for fairness considerations. Extensive experiments using common facial benchmarks demonstrated the effectiveness of our proposed method compared to other baselines, particularly in achieving fairness across a broad range of attributes without the need for consideration during training. Future work can explore further research extensions in various aspects, such as determining more optimal margin coefficients and optimizing hyperparameters.

Acknowledgement

This work was supported in part by JSPS KAKENHI JP 23K28085, and JST Moonshot R&D Grant Number JP-MJMS2215.

References

- [1] NIST study evaluates effects of race, age, sex on face recognition software. <https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software>, 2019. Accessed:2021/1/3. 2
- [2] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proc. 1st Conf. Fairness, Accountability and Transparency*, volume 81, pages 77–91. PMLR, 2018. 1, 2
- [3] T. de Freitas Pereira and S. Marcel. Fairness in biometrics: A figure of merit to assess biometric verification systems. *IEEE Trans. Biometrics, Behavior, and Identity Science*, 4(1):19–29, Jan. 2022. 2
- [4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 4685–4694. IEEE, June 2019. 3, 5
- [5] P. M. Dixon, J. Weiner, T. Mitchell-Olds, and R. Woodley. Bootstrapping the Gini coefficient of inequality. *Ecology*, 68(5):1548–1551, 1987. 5
- [6] S. Dooley, R. Sukthankar, J. Dickerson, C. White, F. Hutter, and M. Goldblum. Rethinking bias mitigation: Fairer architectures make for fairer face recognition. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [7] C. Garvie, A. Bedoya, and J. Frankle. Unregulated police face recognition in America. <https://www.perpetuallineup.org/findings/racial-bias>, 2016. Accessed: 2024/4/30. 2
- [8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 770–778, June 2016. 5
- [9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 5
- [10] A. K. Jain, D. Deb, and J. J. Engelsma. Biometrics: Trust, but verify. *IEEE Trans. Biometrics, Behavior, and Identity Science*, 4(3):303–323, July 2022. 1
- [11] N. Jain, A. Olmo, S. Sengupta, L. Manikonda, and S. Kambhampati. Imperfect ImGANation: Implications of gans Exacerbating Biases on Facial Data Augmentation and Snapchat Selfie Lenses. *arXiv:2001.09528*, 2020. 1, 2
- [12] A. R. Joshi, X. Suau, N. Sivakumar, L. Zappella, and N. Apostoloff. Fair SA: Sensitivity analysis for fairness in face recognition. In *NeurIPS*, 2021. 2
- [13] F. Kamiran and T. Calders. Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, pages 1–6, 2010. 1, 2
- [14] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proc. Int’l Conf. Computer Vision*, pages 365–372. IEEE, 2009. 6
- [15] Q. Meng, S. Zhao, Z. Huang, and F. Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021. 5
- [16] M. Qraitem, K. Saenko, and B. A. Plummer. Bias mimicking: A simple sampling approach for bias mitigation. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 20311–20320, June 2023. 1, 2
- [17] Y. Sato, S. Maeda, M. Akasaka, M. Nishigaki, and T. Ohki. A fair model is not fair in a biased environment. In *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1577–1582. IEEE, Nov. 2022. 2
- [18] P. Terhorst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. Post-Comparison Mitigation of Demographic Bias in Face Recognition Using Fair Score Normalization. *Pattern Recognit. Lett.*, 140:332–338, 2020. 1, 2
- [19] F.-E. Wang, C.-Y. Wang, M. Sun, and S.-H. Lai. Mixfairface: Towards ultimate fairness via mixfair adapter in face recognition. In *Proc. AAAI Conf. Artificial Intelligence*, volume 37, pages 14531–14538, 2023. 1, 2, 5
- [20] M. Wang and W. Deng. Mitigating Bias in Face Recognition Using Skewness-Aware Reinforcement Learning. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 9322–9331, June 2020. 1, 2, 5
- [21] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proc. Int’l Conf. Computer Vision*, pages 692–702, 2019. 1, 5
- [22] X. Xu, Y. Huang, P. Shen, S. Li, J. Li, F. Huang, Y. Li, and Z. Cui. Consistent instance false positive improves fairness in face recognition. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pages 578–586, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society. 1, 2, 5