OneFace: One Threshold for All

Jiaheng Liu[1], Zhipeng Yu*[2], Haoyu Qin[3], Yichao Wu[3],
Ding Liang[3], Gangming Zhao[4], and Ke Xu[1]

[1] State Key Lab of Software Development Environment, Beihang University
[2] University of Chinese Academy of Sciences, Beijing, China
[3] SenseTime Group Limited
[4] The University of Hong Kong
liujiaheng@buaa.edu.cn, yuzhipeng21@mails.ucas.ac.cn
{qinhaoyu1,wuyichao,liangding}@sensetime.com

Abstract. Face recognition (FR) has witnessed remarkable progress with the surge of deep learning. Current FR evaluation protocols usually adopt different thresholds to calculate the True Accept Rate (TAR) under a pre-defined False Accept Rate (FAR) for different datasets. However, in practice, when the FR model is deployed on industry systems (e.g., hardware devices), only one fixed threshold is adopted for all scenarios to distinguish whether a face image pair belongs to the same identity. Therefore, current evaluation protocols using different thresholds for different datasets are not fully compatible with the practical evaluation scenarios with one fixed threshold, and it is critical to measure the performance of FR models by using one threshold for all datasets. In this paper, we rethink the limitations of existing evaluation protocols for FR and propose to evaluate the performance of FR models from a new perspective. Specifically, in our OneFace, we first propose the One-Threshold-for-All (OTA) evaluation protocol for FR, which utilizes one fixed threshold called as Calibration Threshold to measure the performance on different datasets. Then, to improve the performance of FR models under the OTA protocol, we propose the Threshold Consistency Penalty (TCP) to improve the consistency of the thresholds among multiple domains, which includes Implicit Domain Division (IDD) as well as Calibration and Domain Thresholds Estimation (CDTE). Extensive experimental results demonstrate the effectiveness of our method for FR.

Keywords: Face recognition, Loss function, Fairness

1 Introduction

Face recognition (FR) based on deep learning has been well investigated for many years [32,33,40,5,4,6,39]. Most of the progress depends on large-scale training data [10,46,16], deep neural network architectures [36,12,13], and effective loss function designs [29,5,39,37,49,30,3,6]. Recently, with the increasing deployment

*Corresponding author.

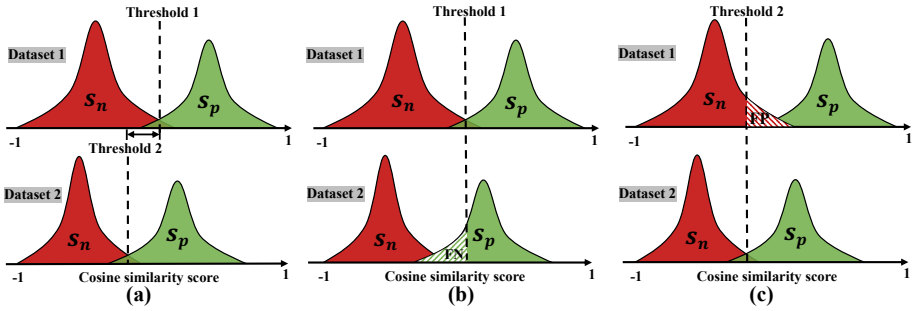

Fig. 1. Similarity distributions of two datasets. The green and red histograms mean the distributions of positive and negative pairs, respectively, where s_p and s_n denote similarities of positive and negative pairs, respectively. (a). Based on existing evaluation protocol, threshold 1 differs from threshold 2 a lot under the same FAR. (b). When we use threshold 1 of dataset 1 to evaluate the dataset 2, a large number of false negative (FN) samples from dataset 2 are produced. (c). When we use threshold 2 to evaluate the dataset 1, many false positive (FP) samples from dataset 1 are generated.

of FR systems, fairness in FR has attracted broad interest from research communities. For example, as reported in the 2019 NIST Face Recognition Vendor Test [9], all participating FR algorithms exhibit different levels of biased performances across various demographic groups (e.g., race, gender, age). However, existing evaluation metrics cannot measure the degree of fairness on threshold across multiple datasets for FR well. Specifically, current practical FR systems usually calculate the True Accept Rate (TAR) under a pre-defined False Accept Rate (FAR) (e.g., 1e-4). As shown in Fig. 1(a), we visualize the distributions of the similarity scores from two datasets, and observe that dataset 1 and dataset 2 have different thresholds under the same FAR. which means that current evaluation protocols adopt different thresholds for performance evaluation for different datasets. We call such phenomenon as Threshold Imbalance on different datasets. Besides, FR models seem to perform well on these two datasets under current evaluation protocols in Fig. 1(a), but in Fig. 1(b) and Fig. 1(c), the performance results of FR models are very sensitive to the changes of the thresholds. Moreover, when the FR models are deployed for industry, only one fixed threshold is adopted for all scenarios, which indicates that the current evaluation protocol with different thresholds for different datasets is not fully compatible with the practical FR. In addition, most existing FR methods are mainly evaluated under the current evaluation protocol in Fig. 1(a), which have not considered the problem of threshold imbalance.

Motivated by the above analysis, in our OneFace, we propose a new One-Threshold-for-All (OTA) evaluation protocol to better exploit the overall performance and fairness on threshold of FR models on multiple datasets, and then introduce an effective Threshold Consistency Penalty (TCP) scheme to tackle the threshold imbalance problem in the training process.

In OTA evaluation protocol, we directly measure the performance of the different datasets using only one fixed threshold, which is more consistent with the practical FR and can be easily combined with existing evaluation protocols Specifically, we call the fixed threshold for deployment as calibration threshold t_c. Given G datasets and the overall FAR (e.g., 1e-4), we introduce two types of calibration threshold estimation methods. The first type is to combine all datasets into one dataset directly and estimate t_c under the overall FAR based on the negative pairs constructed from the dataset, which is straight-forward and feasible. However, as the computation cost is proportional to the number of negative pairs, it will be unaffordable when the number of negative pairs of the whole dataset is large. Thus, we also propose another calibration threshold estimation method by using an extra dataset called as calibration dataset, where the number of negative pairs in the calibration dataset is relatively acceptable. Besides, the calibration dataset is supposed to cover images from as many domains as possible. Once the calibration dataset is prepared, we can directly adopt the threshold of this dataset under the overall FAR as t_c. After obtaining the calibration threshold, under the OTA evaluation protocol, we calculate the TAR and FAR results for these G datasets based on t_c, respectively. Meanwhile, we propose a new fairness metric γ to denote the degree of the threshold imbalance across G datasets. Specifically, we calculate the thresholds $\{t_d^g\}_{g=1}^G$ within each dataset, where t_d^g denotes the domain threshold for the g-th dataset under the overall FAR. Then, we can obtain the γ based on t_c and $\{t_d^g\}_{g=1}^G$.

To improve the performance under the OTA evaluation protocol, our TCP aims to mitigate the threshold imbalance among different domains in the training process. Specifically, first, as domain labels of the training dataset are usually not available and it is necessary to obtain the domain label for computing the domain threshold, we propose to adopt an Implicit Domain Division (IDD) module to assign the domain labels for samples implicitly following the GroupFace [18], which aims to divide the samples of each mini-batch into M domains. M is a pre-defined hyperparameter on the number of implicit domains. Then, we need to construct negative pairs to calculate the calibration and domain thresholds. Existing work [45] uses the weights of the last FC layer and the features of the current batch to construct the negative pairs. However, as discussed in VPL [6], the weights of the last FC layer update very slowly and the similarity distributions of the sample-to-prototype comparisons used in [45] are also different from distributions of the sample-to-sample comparisons used in evaluation process. Thus, this method [45] may lead to inaccurate threshold estimation for FR. To generate accurate thresholds, inspired by MoCo [11], we propose to build a feature queue to maintain the features of the previous iterations. Then, in training, we can construct the negative pairs based on the features of the current batch and the features of the feature queue. After that, we calculate the similarities of these negative pairs to obtain the calibration threshold t_c. Besides, we generate the domain threshold t_d^m for the m-th domain using the features from the m-th domain of the current batch and the features of the feature queue, where the domain labels in each mini-batch are predicted by IDD module. Finally, we adopt

the ratio of t_d^m and t_c as the loss weight for the samples of m-th domain to re-weighting the samples with high domain thresholds, which makes the thresholds across domains more consistent and reduce the degree of threshold imbalance.

The contributions of our proposed OneFace are summarized as follows:

- We first investigate the limitations of the existing evaluation protocols and propose a new One-Threshold-for-All (OTA) evaluation protocol to measure the performance of different datasets under one fixed calibration threshold, which is more consistent with the industry FR scenarios.
- Our proposed Threshold Consistency Penalty (TCP) scheme can improve the fairness on threshold across the domains by penalizing the domains with high thresholds, where we introduce the Implicit Domain Division (IDD) as well as the Calibration and Domain Thresholds Estimation (CDTE).
- Extensive experiments on multiple face recognition benchmarks demonstrate the effectiveness of our proposed method.

2 Related Work

Face Recognition. Face recognition (FR) is a key technique for biometric authentication in many applications (e.g., electronic payment, video surveillance). The success of deep FR can be credited to the following three important reasons: large-scale datasets [50,2,46,16], powerful deep neural networks [36,32,31,35,25] and effective loss functions [19,38,48,26,34,37,5,4,28,6,23,22,24,20,17,21,1]. The mainstream of recent studies is to introduce a new objective function to maximize inter-class discriminability and intra-class compactness. For example, Triplet loss [30] enlarges the distances of negative pairs and reduce the distances of positive pairs in the Euclidean space. Recently, the angular constraint is introduced into the cross-entropy loss function to improve the discriminative ability of the learned representation [27,26]. For example, CosFace [39] and ArcFace [5] utilize a margin item for better discriminative capability of the feature representation. Besides, some mining-based loss functions (e.g., CurricularFace [14] and MV-Arc-Softmax [43]) further consider the difficulty degree of samples to emphasize the mis-classified samples and achieve better results. In addition, GroupFace[18] aggregates implicit group-aware representations to improve the discriminative ability of feature representations by using self-distributed labeling trick. Moreover, VPL [6] additionally introduces the sample-to-sample comparisons into the training process for reducing the gap between the training and evaluation processes for FR. Overall, existing methods mainly aim to improve the generalization and discriminative abilities of the learned feature representation, but they have not considered the limitations of existing evaluation protocols, where different datasets use different thresholds. In contrast, OneFace investigates the gap between existing evaluation protocols and the practical deployment scenarios from a new perspective, and we propose the OTA evaluation protocol to evaluate the performance of different datasets under the same calibration threshold.

Fairness. Recently, more and more attention has been attracted to the fairness for FR models. A straightforward way to tackle the fairness issue is to

build large-scale training datasets (e.g., MS-Celeb-1M [10], Glint360k [2] and WebFace260M [50]). Unfortunately, the time-consuming collected datasets often include unbalanced distributions of different attributes (e.g., race, gender, age), which also introduce inherent bias on different attributes. Recently, Wang et al. [41] introduce the BUPT-balanced as a balanced dataset on race, and BUPT-Globalface to reveal the real distribution of the world's population for fairness study. However, it is still difficult to collect the FR datasets with balanced distribution on different attributes and it is also unclear if the FR models trained on attribute-balanced datasets can eliminate the fairness bias completely. Therefore, some works have proposed to design effective algorithms instead of collecting datasets. For example, Wang et al. [42] propose a deep information maximization adaptation network to transfer the knowledge from Caucasians to other races, and propose another reinforcement learning-based method [41] to learn the optimal margins for different racial groups. Gong et al. [7,8] further utilize a debiasing adversarial network with four specific classifiers, where one classifier is designed for identification and the others are designed for demographic attributes. Xu et al. [45] propose to promote the consistency of instance FPRs to improve fairness across different races. In general, existing methods usually consider improving the accuracy and fairness on all races, but different thresholds are still used for different domains (e.g., races). In contrast to these methods, our proposed OneFace focuses on improving fairness of different domains with the same calibration threshold, which is more compatible with the real-world scenarios and provides new insight for FR.

3 Preliminary

In this section, we take the widely used 1:1 face verification evaluation protocol as an example to show the evaluation process of FR. False Accept Rate (FAR) and True Accept Rate (TAR) are used in face verification. Given N_p positive pairs, the TAR α is computed as follows:

$$\alpha = \frac{1}{N_p} \sum_{j=1}^{N_p} \mathbb{1}(s_p^j > t), \tag{1}$$

where t is the chosen similarity score threshold and s_p^j is the similarity score of the j-th positive pair. $\mathbb{1}(x)$ is the indicator function, which returns 1 when x is true and returns 0 when x is false. Similarly, given N_n negative pairs, the FAR β is defined as follows:

$$\beta = \frac{1}{N_n} \sum_{i=1}^{N_n} \mathbb{1}(s_n^i > t), \tag{2}$$

where s_n^i is the similarity score of the i-th negative pair.

In the testing process, we fix a FAR (e.g., 1e-4) and calculate the corresponding TAR to represent the performance of the FR models. For each dataset, the threshold t under the specific FAR β in Eq. 2 can be generated by the quantile

of the similarity scores of all negative pairs. Then, based on the similarities of all positive pairs and the threshold t, we calculate the TAR α.

4 OneFace

In this section, we describe our OneFace framework, which includes the newly proposed One-Threshold-for-All (OTA) evaluation protocol and Threshold Consistency Penalty (TCP) scheme.

4.1 One-Threshold-for-All Evaluation Protocol

We first discuss the necessity of One-Threshold-for-All (OTA) evaluation protocol, and then introduce the details of the OTA evaluation protocol for FR, where we first estimate the calibration threshold under the overall FAR (e.g., 1e-4) and then calculate the performance results for different datasets.

Necessity of the OTA evaluation protocol. The above-mentioned evaluation protocol (TAR@FAR) in Sec. 3 has been adopted in many works. Nevertheless, we argue that the current evaluation protocols are not compatible with the practical FR, as different testing datasets use different thresholds even if the model and the pre-defined FAR are the same. Moreover, we can observe the Threshold Imbalance phenomenon, where these thresholds vary a lot when these datasets are from different domains. As shown in Table 1, we report the TAR and threshold results of different races from the RFW [42] dataset under the FAR of 1e-4 and 1e-5, where the thresholds are estimated within each race under the same FAR. It can be easily observed that current evaluation protocol adopts different thresholds for different races and the thresholds differ a lot in some races. For example, the threshold of African is much higher than the threshold of Caucasian under the same FAR. In contrast, when FR model is deployed in practice, only one fixed threshold score (i.e., calibration threshold) is applied for all scenarios, which means that the current evaluation protocol is not fully consistent with the practical FR applications. Therefore, it is critical to evaluate the performance of all domains using one fixed threshold for FR.

Here, we describe the evaluation process of OTA for G testing datasets. Given the overall FAR (e.g., 1e-4), we first generate the fixed calibration threshold. Then, based on the calibration threshold, we calculate the TAR and FAR results for different datasets. Finally, we also define the fairness metric to represent the degree of the threshold imbalance for these datasets.

Table 1. The threshold and TAR results based on classical 1:1 verification evaluation protocol when FAR is 1e-4 and 1e-5 on different races from the RFW dataset [42].

Results	FAR=1e-4				FAR=1e-5			
	African	Asian	Caucasian	Indian	African	Asian	Caucasian	Indian
Threshold	0.455	0.403	0.384	0.419	0.511	0.465	0.443	0.478
TAR	92.29	92.61	95.58	94.47	86.25	85.90	91.05	89.54

Calibration Threshold Estimation. In OTA, given G testing datasets, we describe two types of calibration threshold estimation methods. In the first type, we directly combine these G datasets into one whole dataset and extract the features of all negative pairs to calculate the similarities of these pairs. Then, we obtain the threshold score under the overall FAR (e.g., 1e-4) as the calibration threshold t_c. The first type is to estimate the threshold under the overall FAR based on the negative pairs constructed from all datasets, which is feasible and effective when the number of negative pairs is relatively acceptable. However, when the number of negative pairs increases, large computation costs for threshold estimation are needed. Therefore, we propose another calibration threshold estimation method by adopting an extra dataset (e.g., FairFace dataset [15]), which is called as calibration dataset. Specifically, we suppose that the calibration dataset covers images from as many domains as possible, which aims to make the calibration threshold t_c general and accurate. Besides, the size of the calibration dataset should be acceptable, which leads to affordable computation costs. Similarly, we calculate the similarities of all negative pairs from the calibration dataset, and obtain the calibration threshold t_c under the overall FAR.
Performance under the Calibration Threshold. After obtaining the calibration threshold t_c, for g-th dataset, where $g \in \{1, ..., G\}$, we can easily produce the TAR α_g and FAR β_g within g-th dataset under the threshold t_c. Then, we directly adopt the mean $\mu_\alpha = \frac{1}{G}\sum_{g=1}^{G}\alpha_g$ and the variance $\sigma_\alpha^2 = \frac{1}{G}\sum_{g=1}^{G}(\alpha_g - \mu_\alpha)^2$ of all datasets to represent the overall TAR performance for these datasets. As the FAR value is usually small (e.g., 1e-4) and the magnitude of the FARs under the same threshold among different domains varies greatly, we adopt the $\log_{10}$ operation to maintain the monotonicity and simplify the calculation. Additionally, without this operation, statistics values are dominated by the large FAR value. For example, the mean of {1e-3, 1e-4, 1e-5} is dominated by 1e-3, where 1e-5 is ignored. Thus, for the results of FAR, we utilize the mean $\mu_\beta = \frac{1}{G}\sum_{g=1}^{G} -\log_{10}\beta_g$ and the variance $\sigma_\beta^2 = \frac{1}{G}\sum_{g=1}^{G}(-\log_{10}\beta_g - \mu_\beta)^2$ of all datasets. Furthermore, we propose a fairness metric denoted as γ to represent the degree of the threshold imbalance across the G datasets, Specifically, we also compute the threshold t_d^g within g-th dataset under the overall FAR, where we call t_d^g as the domain threshold for the g-th dataset. Meanwhile, we define the deviations between G domain thresholds (i.e., $\{t_d^g\}_{g=1}^{G}$) and the calibration threshold t_c as the fairness metric γ, which is illustrated as follows:

$$\gamma = \sqrt{\frac{1}{G}\sum_{g=1}^{G}(t_d^g - t_c)^2}. \tag{3}$$

In Eq. 3, γ is larger when the degree of threshold imbalance is more serious.

4.2 Threshold Consistency Penalty

To improve the performance under the OTA evaluation protocol, we propose the Threshold Consistency Penalty (TCP) scheme to mitigate the threshold

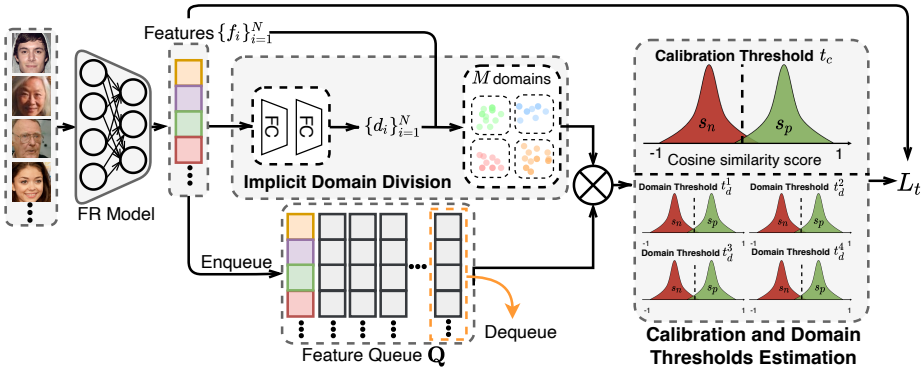

Fig. 2. The framework of our Threshold Consistency Penalty (TCP) scheme, which includes the Implicit Domain Division (IDE) as well as the Calibration and Domain Thresholds Estimation (CDTE). In each iteration, we first use the FR model to extract the features $\{f_i\}_{i=1}^N$ of each mini-batch and update the feature queue $\mathbf{Q}$, where N is the number of samples in each mini-batch. Then, we use the IDD to divide the $\{f_i\}_{i=1}^N$ into M domains implicitly, where the domain loss L_d is used. After that, we calculate the calibration threshold t_c and domain thresholds $\{t_d^m\}_{m=1}^M$. Finally, based on the t_c, $\{t_d^m\}_{m=1}^M$, $\{f_i\}_{i=1}^N$ and ground-truth identity labels, we calculate the TCP loss L_t.

imbalance among different domains in training as shown in Fig. 2, where we define a domain as a set of samples that share any common visual-or-non-visual properties for FR.

Specifically, TCP includes Implicit Domain Division (IDD) as well as Calibration and Domain Thresholds Estimation (CDTE). In IDD, we propose to divide the images of each mini-batch into several domains without additional annotations. In CDTE, we first build sufficient negative pairs using the features of the current batch and the features in our proposed feature queue. Then, we compute the similarities of these negative pairs and estimate the calibration and domain thresholds. Finally, based on the calibration and domain thresholds, we adaptively adjust the loss weights of samples from each domain.

Implicit Domain Division. As the domain labels are usually unavailable in the training dataset, our IDD is trained in a self-supervised manner, which predicts the domain label for each sample without any explicit ground-truth information. Specifically, inspired by GroupFace [18], our IDD is implemented by two fully-connected layers and a softmax layer, which takes the feature representation f_i of the i-th sample as input and predicts the domain probabilities as follows:

$$\{p_i^m\}_{m=1}^M = \mathcal{H}(f_i). \tag{4}$$

$\mathcal{H}$ denotes the neural network of IDD. M is a pre-defined hyperparameter on the number of implicit domains, which is not related to the number of evaluation datasets (i.e., G in OTA). p_i^m is the domain probability for m-th domain. Additionally, $\mathcal{H}$ is trained by the self-distributed labeling strategy. Specifically,

$\{p_i^m\}_{m=1}^M$ is the initial predictive domain probabilities for i-th sample. Following [18], to generate uniformly-distributed domain labels, we use modified probability regulated by a prior probability, where an expectation-normalized strategy is used. The updated domain probability $\tilde{p}_i^m$ for m-th domain is as follows:

$$\tilde{p}_i^m = \frac{1}{M}(p_i^m - \frac{1}{T}\sum_{i=1}^{T} p_i^m) + \frac{1}{M}, \tag{5}$$

where T is the number of samples to calculate the expectation value. We directly set T as the number of samples in each mini-batch. Thus, the expectation of the expectation-normalized probability $\frac{1}{T}\sum_{i=1}^{T}\tilde{p}_i^m = \frac{1}{M}$. The domain label $d_i \in \{1, ..., M\}$ for i-th sample is obtained as $d_i = \arg\max_m \tilde{p}_i^m$. Meanwhile, to reduce the divergence between the prediction probabilities and the generated domain label, a domain loss L_d based on cross-entropy loss is defined as follows:

$$L_d = -\frac{1}{N}\sum_{i=1}^{N}\log(\frac{e^{p_i^{d_i}}}{\sum_{m=1}^{M} e^{p_i^m}}), \tag{6}$$

where N is the number of samples in each iteration.

Calibration and Domain Thresholds Estimation. To estimate the calibration and domain thresholds in the training process, we first need to construct sufficient negative pairs with high qualities. Inspired by MoCo [11], for unsupervised learning, which adopts a memory bank from the previous mini-batches to obtain sufficient negative samples, we propose to build a feature queue $\mathbf{Q} \in \mathbb{R}^{K\times N\times d}$ to construct sufficient negative pairs, as shown in Fig. 2, where K is the number of iterations, and d represents the dimension of the feature representation extracted by the neural network for each face image. Meanwhile, as discussed in VPL [6], features drift slowly for FR models, which represents that features extracted previously can be considered as an approximation of the output of the current network within a certain number of training steps. Thus, we could set K as a relatively large value ($K = 1000$ in our work) to generate sufficient negative pairs with high qualities. Furthermore, we establish an auxiliary label queue, $\mathbf{Q'} \in \mathbb{R}^{K\times N}$ to store the identity labels for the features in $\mathbf{Q}$. In each iteration, we first extract the features $\{f_i\}_{i=1}^N$ of the current batch, where y_i is the corresponding label of f_i. Then, the features $\{f_i\}_{i=1}^N$ and the labels $\{y_i\}_{i=1}^N$ are enqueued into feature queue $\mathbf{Q}$ and label queue $\mathbf{Q'}$, respectively. After that, the features and labels from the oldest batch in $\mathbf{Q}$ and $\mathbf{Q'}$ are also dequeued. Finally, we can construct the negative pairs based on $\{f_i\}_{i=1}^N$ and $\mathbf{Q}$.

For Calibration Threshold Estimation, we calculate the similarities of all negative pairs, and generate the calibration threshold t_c in training under the overall FAR (e.g., 1e-4). For Domain Threshold Estimation, we first generate the domain labels for the samples of the current mini-batch based on IDD. Then, to accurately estimate the threshold distribution for each domain m, only the features of samples with the same domain label (i.e., m) are selected to construct domain-specific negative pairs with the features from $\mathbf{Q}$. By calculating

the similarities of such domain-specific negative pairs, we can obtain the domain threshold t_d^m under the same FAR value (e.g., 1e-4) for the m-th domain. Finally, the calibration threshold t_c and domain thresholds $\{t_d^m\}_{m=1}^M$ are obtained.

Loss formulation. After generating the calibration and domain thresholds, we define the TCP loss L_t from the domain level as follows:

$$L_t = \frac{1}{N} \sum_{m=1}^{M} \sum_{i \in \mathbf{T}_m} (\frac{t_d^m}{t_c} \cdot L_i). \tag{7}$$

$\mathbf{T}_m$ is an index set, which contains the indices of samples with domain label m in each mini-batch. N is the number of samples in each mini-match, and L_i denotes the classification loss for i-th sample. In our work, we utilize the widely-used ArcFace [5] loss as L_i. To this end, our TCP loss will enforce the neural network to pay more attention for these samples from domains with $t_d^m > t_c$ as there are more false positive pairs with higher scores than t_c, and we can automatically down-weight the contribution of these samples from domains with $t_d^m < t_c$ during training. In other words, our TCP loss aims to reduce the degree of threshold imbalance across multiple domains by dynamically adjusting the loss weights for M domains. It should be mentioned that L_i can be replaced with many existing loss functions [39,14]. Finally, the overall loss function of our proposed method is defined as follows:

$$L = L_t + \lambda L_d, \tag{8}$$

where λ is the loss weight for the domain loss L_d of in our IDD.

5 Experiments

In this section, we first report the results of different methods on multiple cross-domain settings under our proposed OTA evaluation protocol, where one fixed calibration threshold for different datasets. Then, we perform detailed analysis and discussion to further show the effectiveness of our method.

5.1 Implementation Details

Dataset. Our experimental settings include two settings (i.e., cross-race and cross-gender settings) as follows. For cross-race setting, we follow [45] to employ the BUPT-Balancedface [41] as the training dataset, and use the RFW dataset [42] as the testing dataset with four race groups (i.e., African, Asian, Caucasian, and Indian), where we directly use the whole RFW dataset to estimate the calibration threshold under the overall FAR for OTA evaluation protocol. For the cross-gender setting, we follow many existing works [5,14] to use the refined version of MS-1M [10] dataset as the training dataset. For the testing dataset of cross-gender setting, we split the IJB-B dataset [44] into two datasets (i.e., IJB-B(F), IJB-B(M)) manually based on the gender attribute (i.e., female or male), where we also use the whole IJB-B dataset to estimate the calibration threshold under the overall FAR for OTA evaluation protocol.

Experimental setting. For the pre-processing of the training data and testing data, we follow [5,6,14] to generate the normalized face crops (112×112) with five landmarks detected by MTCNN [47]. For the backbone network for cross-race and cross-dataset settings, we follow the state-of-the-art method [45] to use the ResNet34 [12] to produce 512-dim feature representation. For the backbone network for cross-gender setting, we use the ResNet100 [12] for all methods. For the training process on BUPT-Balancedface [41], the initial learning rate is 0.1 and divided by 10 at the 55k, 88k, 99k iterations, where the total iteration is set as 110k. For the training process on the refined MS-1M [10], the initial learning rate is 0.1 and divided by 10 at the 110k, 190k, 220k iterations, where the total iteration is set as 240k. The batchsize is set as 512 for all experiments. For the feature queue $\mathbf{Q}$, d is set as 512, the number of iterations (i.e., K) in the feature queue and label queue is set as 1000. In training, the number of implicit domains (i.e., M) in IDD is set as 8. The loss weight (i.e., λ) of the domain loss L_d is set as 0.05. Under the OTA evaluation protocol, G is set as 4, 2 for cross-race and cross-gender settings, respectively, and we report the results of our method and recent widely-used loss functions [5,39,14,18].

5.2 Experimental results under the OTA evaluation protocol

Results on cross-race setting. As shown in Table 2, for cross-race setting, we report the results of different methods among four demographic groups in the RFW dataset [42] under the OTA evaluation protocol, where we use the same calibration threshold for these groups. In Table 2, for the TAR results, when compared with methods, we observe that our method achieves higher TAR with lower variance. For the $-\log_{10}$FAR, our method also achieves lower variance, which shows the effectiveness of our method.

Results on cross-gender setting. As shown in Table 3, we report the results of different methods under the OTA evaluation protocol for cross-gender setting. Specifically, we divide the original IJB-B dataset [44] into two datasets (i.e., IJB-B(F) and IJB-B(M)) based on the gender attribute, where IJB-B(F) and IJB-B(M) represents the female and male datasets, respectively. Then, we use the whole IJB-B dataset to estimate the calibration threshold under the overall FAR of 1e-4. In Table 3, we have the following observations: (1) The performance of IJB-B(F) is lower than IJB-B(M) a lot, which indicates the gender attribute influences the FR performance greatly. (2) Our method achieves better average performance results and lower variances when compared with other baseline methods on both TAR and FAR metrics.

5.3 Analysis

Analysis on the classical evaluation results on the RFW dataset. As shown in Table 4, we also report the results of different methods based on the classical 1:1 verification results on the RFW dataset, where different races use different thresholds. In Table 4, we observe that our method also achieves better results when compared with other methods, which further shows the effectiveness

Table 2. The performance of different methods under the OTA evaluation protocol when overall FAR is 1e-4 for cross-race setting.

Method	TAR						$-\log_{10}$FAR					
	African	Asian	Caucasian	Indian	Avg.↑	Std.↓	African	Asian	Caucasian	Indian	Avg.↑	Std.↓
ArcFace [5]	96.71	93.80	95.33	96.22	95.51	1.107	3.030	3.791	4.052	3.519	3.598	0.378
CosFace [39]	96.44	91.83	94.36	94.99	94.41	1.667	3.024	3.810	4.053	3.477	3.591	0.386
CurricularFace [14]	96.79	93.80	95.41	96.14	95.54	1.114	2.999	3.811	4.045	3.554	**3.602**	0.389
GroupFace [18]	96.76	94.00	95.62	96.27	95.66	1.041	3.010	3.778	4.028	3.552	3.592	0.376
Ours	96.96	95.02	95.92	96.43	**96.08**	**0.715**	3.057	3.604	3.980	3.641	3.571	**0.331**

Table 3. The performance of different methods under the OTA evaluation protocol when overall FAR is 1e-4 for cross-gender setting.

Method	TAR				$-\log_{10}$FAR			
	IJB-B(F)	IJB-B(M)	Avg.↑	Std.↓	IJB-B(F)	IJB-B(M)	Avg.↑	Std.↓
ArcFace [5]	91.85	96.96	94.41	2.555	3.535	4.079	3.807	0.272
CosFace [39]	91.46	96.69	94.07	2.615	3.550	4.037	3.794	0.244
CurricularFace [14]	91.71	96.98	94.35	2.635	3.559	4.037	3.798	0.239
GroupFace [18]	91.93	97.09	94.51	2.580	3.548	4.037	3.793	0.245
Ours	92.29	97.26	**95.03**	**2.335**	3.591	4.062	**3.827**	**0.236**

of our method. Moreover, as shown in Table 5, we also report the verification accuracy results of different methods on the RFW dataset, where different races use different thresholds. Note that the results of other methods are directly quoted from [45]. In Table 5, we observe that our method also achieves better results on most cases when compared with other methods, which further shows the effectiveness of our method.

Analysis on the fairness metric. As shown in Table 6, we provide the fairness results of different methods under the OTA evaluation protocol when overall FAR is 1e-4 for cross-race setting, and we observe the fairness metric in Eq. 3 of our method is also lower than other methods, which shows that our method can mitigate the threshold imbalance greatly for cross-race setting.

Analysis on the computation costs. No extra costs (e.g., GPU memory usage, time) are required at inference. Besides, in training, when compared with ArcFace baseline method, the training time and GPU memory usage of our method are 1.183 times and 1.005 times, respectively, which is acceptable.

Analysis on the effectiveness of TCP. In Fig. 3(a) and Fig. 3(b), we visualize the distributions of similarity scores on the African and Caucasian from the RFW dataset of different methods (i.e., ArcFace and Ours) and the red vertical line denotes the domain threshold within each group under the FAR of 1e-4. When compared with the ArcFace, the difference of the domain thresholds between African and Caucasian is smaller in our method, which demonstrates that our method can mitigate the threshold imbalance among different domains.

Table 4. 1:1 verification TAR results on the RFW dataset.

Method	TAR@FAR=1e-4			
	African	Asian	Caucasian	Indian
ArcFace [5]	92.29	92.61	95.58	94.47
CosFace [39]	91.24	90.66	94.59	92.57
CurricularFace [14]	92.41	92.88	95.56	94.73
GroupFace [18]	92.48	92.82	95.73	94.69
Xu et al. [45]	93.31	93.05	95.71	92.89
Ours	**93.43**	**93.39**	**95.93**	**95.38**

Table 5. Verification accuracy (%) on the RFW dataset.

Methods	African	Asian	Caucasian	Indian	Avg.↑	Std.↓
ArcFace [5] (R34)	93.98	93.72	96.18	94.67	94.64	1.11
CosFace [39] (R34)	92.93	92.98	95.12	93.93	93.74	1.03
RL-RBN [41] (R34)	95.00	94.82	96.27	94.68	95.19	0.93
Xu et al. [45] (R34)	95.95	95.17	96.78	96.38	96.07	**0.69**
ArcFace [5] (R100)	96.43	94.98	97.37	96.17	96.24	0.98
Xu et al. [45] (R100)	97.03	95.65	97.60	**96.82**	96.78	0.82
Ours (R100)	**97.33**	**95.95**	**98.10**	96.55	**97.01**	0.79

5.4 Discussion

Discussion on the OTA evaluation protocol. In our work, for FR, we propose the OTA protocol to measure the fairness problem by evaluating the results of different datasets under one fixed calibration threshold, and we are not to search for one threshold given fixed mixture distributions. Specifically, when an FR model is deployed on FR systems (e.g., hardware devices), only one fixed threshold is used and it is infeasible to select a threshold for each face image. The reasons are as follows: (1) FR systems usually focus on unconstrained (in the wild) scenarios (e.g., environment), which indicates that it is difficult to define or distinguish the specific scenarios for different face images (e.g., probe/gallery). In other words, FR models are supposed to work well for different scenarios (e.g. airport/train station, sunny day/rainy day). (2) Even if the scenario is constrained, it is still difficult to define the number of domains (e.g., facial appearance), as there are many different aspects to describe the property of each domain. For example, age (old, youth, child), gender (male, female), glasses (w, w/o) and many other implicit domains that cannot be observed. (3) If we select a threshold for each image, extra costs (e.g., domain prediction model) are needed, and accumulation errors will be brought by domain prediction and face verification tasks. (4) Setting different thresholds for some domains (e.g., gender, race) may also bring ethical risks. Overall, when compared with existing evaluation protocol, our OTA protocol is more consistent with real-world scenarios, and our proposed TCP method aims to align the similarity distributions of different domains and not to search for optimal thresholds.

Table 6. The fairness of different methods under the OTA evaluation protocol when overall FAR is 1e-4 for cross-race setting.

Models	ArcFace [5]	CosFace [39]	CurricularFace [14]	GroupFace [18]	Ours
Fairness (γ)	0.038	0.043	0.037	0.038	**0.030**

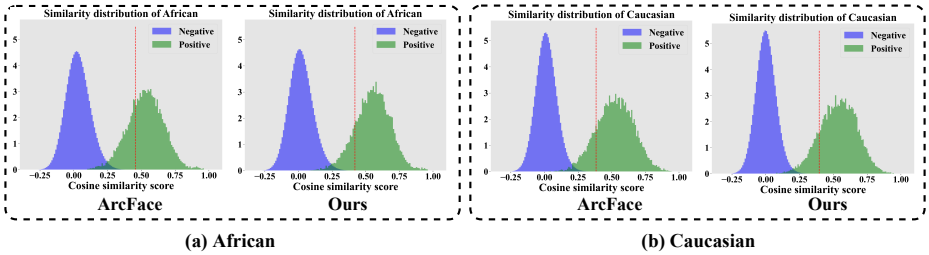

Fig. 3. (a) The similarity distributions of African. (b) The similarity distributions of Caucasian. The red vertical lines in (a) and (b) denote the thresholds under the FAR of 1e-4 within each race dataset.

Discussion on the calibration threshold. In the industry scenarios, we cannot obtain the similarity distributions of all datasets. Thus, we use the calibration threshold generated by the similarity distribution from the well-constructed calibration dataset to distinguish whether a face image pair belongs to the same identity. Specifically, to improve the robustness of the calibration threshold for practical FR, when the model is deployed, we can build the calibration dataset to generate the fixed calibration threshold, and the calibration threshold will be more suitable when the distribution of the calibration dataset is closer to the distribution of the real-world scenarios.

6 Conclusion

In our OneFace, we first investigate the limitations of the existing evaluation protocols for FR and propose the One-Threshold-for-All (OTA) evaluation protocol, which is more consistent with the deployment phase. Besides, we also propose the Threshold Consistency Penalty (TCP) scheme to improve the performance of FR models under the OTA protocol. Extensive experiments on multiple FR benchmark datasets demonstrate the effectiveness of our proposed method. Moreover, we hope our method can motivate other researchers to investigate the fairness problem on practical FR systems (e.g., more reliable fairness metric), and explore more research areas on the fairness in the future work.

7 Acknowledgments

This research was supported by National Natural Science Foundation of China under Grant 61932002.

References

1. An, X., Deng, J., Guo, J., Feng, Z., Zhu, X., Yang, J., Liu, T.: Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4042–4051 (2022)
2. An, X., Zhu, X., Gao, Y., Xiao, Y., Zhao, Y., Feng, Z., Wu, L., Qin, B., Zhang, M., Zhang, D., Fu, Y.: Partial fc: Training 10 million identities on a single machine. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 1445–1449 (October 2021)
3. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 539–546. IEEE (2005)
4. Deng, J., Guo, J., Liu, T., Gong, M., Zafeiriou, S.: Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In: Proceedings of the IEEE Conference on European Conference on Computer Vision (2020)
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
6. Deng, J., Guo, J., Yang, J., Lattas, A., Zafeiriou, S.: Variational prototype learning for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11906–11915 (June 2021)
7. Gong, S., Liu, X., Jain, A.K.: Jointly de-biasing face recognition and demographic attribute estimation. In: European conference on computer vision. pp. 330–347. Springer (2020)
8. Gong, S., Liu, X., Jain, A.K.: Mitigating face recognition bias via group adaptive classifier. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3414–3424 (2021)
9. Grother, P.J., Ngan, M.L., Hanaoka, K.K., et al.: Ongoing face recognition vendor test (frvt) part 3: Demographic effects. In: NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD (2019)
10. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European Conference on Computer Vision. pp. 87–102. Springer (2016)
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
14. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., Huang, F.: Curricularface: adaptive curriculum learning loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5901–5910 (2020)
15. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1548–1558 (2021)

16. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4873–4882 (2016)
17. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18750–18759 (2022)
18. Kim, Y., Park, W., Roh, M.C., Shin, J.: Groupface: Learning latent groups and constructing group-based representations for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5621–5630 (2020)
19. Kim, Y., Park, W., Shin, J.: Broadface: Looking at tens of thousands of people at once for face recognition. ECCV (2020)
20. Li, Z., Wu, Y., Chen, K., Wu, Y., Zhou, S., Liu, J., Yan, J.: Learning to auto weight: Entirely data-driven and highly efficient weighting framework. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 4788–4795 (2020)
21. Liu, C., Yu, X., Tsai, Y.H., Faraki, M., Moslemi, R., Chandraker, M., Fu, Y.: Learning to learn across diverse data biases in deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4072–4082 (2022)
22. Liu, J., Qin, H., Wu, Y., Guo, J., Liang, D., Xu, K.: Coupleface: Relation matters for face recognition distillation. In: Proceedings of the European Conference on Computer Vision (2022)
23. Liu, J., Qin, H., Wu, Y., Liang, D.: Anchorface: Boosting tar@ far for practical face recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (2022)
24. Liu, J., Wu, Y., Wu, Y., Li, C., Hu, X., Liang, D., Wang, M.: Dam: Discrepancy alignment metric for face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3814–3823 (2021)
25. Liu, J., Zhou, S., Wu, Y., Chen, K., Ouyang, W., Xu, D.: Block proposal neural architecture search. IEEE Transactions on Image Processing **30**, 15–25 (2020)
26. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
27. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: ICML. vol. 2, p. 7 (2016)
28. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: A universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14225–14234 (2021)
29. Ranjan, R., Castillo, C.D., Chellappa, R.: L2-constrained softmax loss for discriminative face verification. arXiv preprint arXiv:1703.09507 (2017)
30. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. pp. 815–823 (2015)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
32. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems. pp. 1988–1996 (2014)
33. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2892–2900 (2015)

34. Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle loss: A unified perspective of pair similarity optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6398–6407 (2020)
35. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
36. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1701–1708 (2014)
37. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Processing Letters **25**(7), 926–930 (2018)
38. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: L2 hypersphere embedding for face verification. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1041–1049 (2017)
39. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018)
40. Wang, M., Deng, W.: Deep face recognition: A survey. arxiv 2018. arXiv preprint arXiv:1804.06655 (2018)
41. Wang, M., Deng, W.: Mitigating bias in face recognition using skewness-aware reinforcement learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9322–9331 (2020)
42. Wang, M., Deng, W., Hu, J., Tao, X., Huang, Y.: Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In: Proceedings of the ieee/cvf international conference on computer vision. pp. 692–702 (2019)
43. Wang, X., Zhang, S., Wang, S., Fu, T., Shi, H., Mei, T.: Mis-classified vector guided softmax loss for face recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12241–12248 (2020)
44. Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A.K., Duncan, J.A., Allen, K., et al.: Iarpa janus benchmark-b face dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 90–98 (2017)
45. Xu, X., Huang, Y., Shen, P., Li, S., Li, J., Huang, F., Li, Y., Cui, Z.: Consistent instance false positive improves fairness in face recognition. In: CVPR. pp. 578–586 (2021)
46. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
47. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters **23**(10), 1499–1503 (2016)
48. Zhang, X., Fang, Z., Wen, Y., Li, Z., Qiao, Y.: Range loss for deep face recognition with long-tailed training data. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5409–5418 (2017)
49. Zhang, X., Zhao, R., Qiao, Y., Wang, X., Li, H.: Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10823–10832 (2019)
50. Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., Zhu, J., Yang, T., Lu, J., Du, D., Zhou, J.: Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In: CVPR. pp. 10492–10502 (June 2021)