# Emphasizing Closeness and Diversity Simultaneously for Deep Face Representation

Chaoyu Zhao[1], Jianjun Qian[1](✉), Shumin Zhu[2], Jin Xie[1], and Jian Yang[1]

[1] PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{cyzhao,csjqian}@njust.edu.cn
[2] AiDLab, Laboratory for Artificial Intelligence in Design, School of Fashion and Textiles, The Hong Kong Polytechnic University

**Abstract.** Recent years have witnessed remarkable progress in deep face recognition due to the advancement of softmax-based methods. In this work, we first provide the analysis to reveal the working mechanism of softmax-based methods from the geometry view. Margin-based softmax methods enhance the feature discrimination by the extra margin. Mining-based softmax methods pay more attention to hard samples and try to enlarge their diversity during training. Both closeness and diversity are essential for discriminative features learning; however, we observe that most previous works dealing with hard samples fail to balance the relationship between closeness and diversity. Therefore, we propose a novel approach to tackle the above issue. Specifically, we design a two-branch cooperative network: the Elementary Representation Branch (ERB) and the Refined Representation Branch (RRB). ERB employs the margin-based softmax to guide the network to learn elementary features and measure the difficulty of training samples. RRB employs the proposed sampling strategy in conjunction with two loss terms to enhance closeness and diversity simultaneously. Extensive experimental results on popular benchmarks demonstrate the superiority of our proposed method over state-of-the-art methods.

**Keywords:** Deep face representation · Closeness and diversity · Difficulty measure · Two-branch cooperative network

## 1 Introduction

Deep face recognition (FR) has witnessed tremendous progress during recent years, mainly attributed to the growing scale of publicly available datasets, the development of convolutional neural network architectures, and the improvement of loss functions. In 2014, DeepFace [30] closely reached the human-level performance in unconstrained face recognition based on a nine-layer deep neural network. Subsequently, several successful FR systems such as DeepID2 [26], DeepID3 [27], VGGFace [20], and FaceNet [23] demonstrate that well-designed deep architectures can obtain promising performance.

Besides, the major advance comes from the evolution of loss functions for training deep convolutional neural networks. Early FR works often adopt methods based on metric learning, such as Contrastive Loss [3] and Triplet Loss [23]. However, most of them suffer from the combinatorial explosion, especially on large-scale datasets. Current deep FR approaches are typically based on margin-based softmax loss functions, such as L-Softmax [16], SphereFace [15], CosFace [32], AM-Softmax [31], and ArcFace [4]. These methods share a common idea of introducing a margin penalty between different classes to encourage feature discrimination. Subsequently, mining-based methods such as MV-Softmax [34] and CurricularFace [10] demonstrate that margin-based softmax methods fail to make good use of hard samples. They introduce the hard sample mining strategy and enlarge the distance between a misclassified sample and its negative class centers. By contrast, MagFace [18] learns the well-structured within-class feature distribution by loosening the margin constraint for uncertain samples.

As is well analyzed in several works [8,29], softmax aims to optimize $(s_n - s_p)$ to achieve the decision boundary $s_n - s_p = -m(m$ is the margin), where $s_p$ is the intra-class similarity and $s_n$ is the inter-class similarity. Moreover, the optimization of $s_p$ and $s_n$ are highly coupled. When mining-based methods enlarge the optimization strength of $s_n$ for hard samples, their $s_p$ will also get an extra tendency to be maximized (see Sec. 3.2). This naturally poses a problem: mining-based methods can indirectly enhance the within-class closeness while ignoring that hard samples usually contain much uncertainty and thus should lie on the edge of the intra-class distribution as is claimed in [18]. Consequently, the intra-class distribution structure in high dimensional space will be vulnerable, and the model will tend to overfit on noisy samples.

Based on the observation above, this paper first analyzes the working mechanism of softmax-based methods from the view of closeness and diversity in geometry space. Then we introduce the embedding feature constraint to enhance the closeness $s_p$ for easy samples to establish robust class centers quickly, and meanwhile increase the diversity for hard samples to shift the optimization emphasis from their $s_p$ to $s_n$ and thus prevent overfitting.

To summarize, our key contributions are as follows:

- We analyze the working mechanism of softmax-based methods from the geometry view and claim that both closeness and diversity should be simultaneously emphasized for discriminative feature learning.
- We propose a two-branch cooperative network to simultaneously learn closeness and diversity so that the model can improve both discrimination and generalization ability.
- We conduct extensive experiments on several publicly available benchmarks. Experimental results demonstrate the superiority of our proposed method over state-of-the-art methods.

# 2 Related Work

## 2.1 Margin-based Methods

In deep face recognition, softmax is widely applied to supervise the network for promoting features' separability. However, it exceeds softmax's ability when facing challenging tasks where intra-class variations get larger. Several margin-based methods [4,15,16,31,32] are then proposed. L-Softmax [16] first introduces margin penalty into traditional softmax. SphereFace [15] further normalizes weight vectors by $l2$-normalization to learn face representations on a hypersphere. Subsequently, CosFace [32], AM-Softmax [31], and ArcFace [4] introduce an additive margin penalty on cosine/angle space to further improve the discriminative power of learned face representations. AdaCos [41] employs the adaptive scale parameter to promote the training supervision in dealing with various facial samples. MagFace [18] improves the performance of previous margin-based methods by keeping ambiguous samples away from class centers.

## 2.2 Mining-based Methods

Hard sample mining strategy is also a critical step to enhance the feature representation ability [1,25]. OHEM [25] automatically indicates and emphasizes hard samples within a mini-batch according to their loss values. Focal Loss [13] reduces the weight for easy samples during training by introducing the re-weighting factor into the standard cross-entropy loss. MV-Softmax [34] emphasizes hard samples to guide the networks for learning discriminative features by introducing an extra margin penalty when a sample is misclassified. CurricularFace [10] employs the Curriculum Learning (CL) strategy to focus on easy samples in the early training stage and concentrate on hard ones later.

## 2.3 Contrastive Learning

Traditional contrastive loss functions [6,35] are generally found in early FR works. However, most of them suffer from the combinatorial explosion when dealing with large-scale datasets [23,26,28]. Center Loss [36] proposes a joint supervision signal based on softmax to penalize the distances between the samples and their corresponding class centers. Range Loss [40] is proposed to address the long-tail problem by reducing within-class variances and enlarging inter-class differences in each mini-batch. Modern contrastive approaches [2,7,33] show promising results in the field of unsupervised representation learning. SimCLR [2] learns deep representations by minimizing the distance between multiple augmented views of the same image in the latent space. MoCo [7] proposes a dynamic dictionary and a moving-averaged encoder to learn visual representations. Wang et al. [33] analyze alignment and uniformity on the feature hypersphere to guide the unsupervised representation learning. Supervised contrastive learning [12] significantly outperforms the traditional contrastive approaches by incorporating label information to construct genuine and imposter pairs.

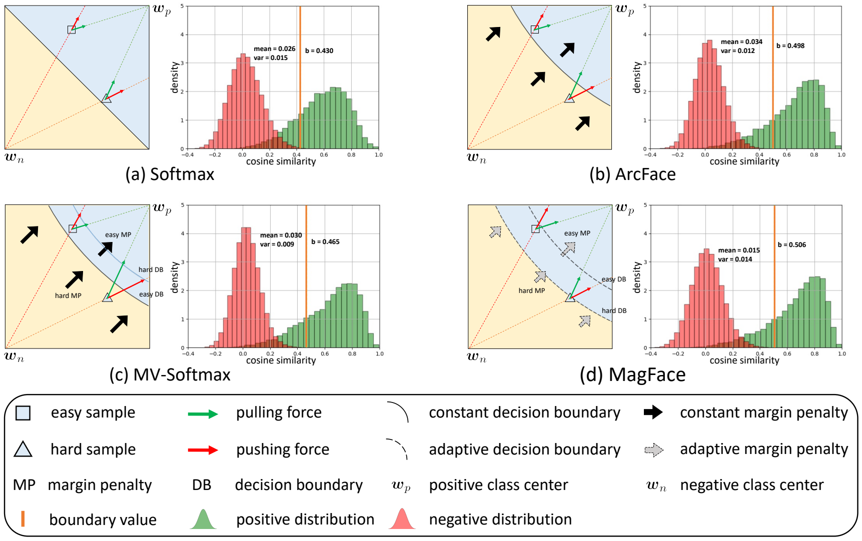


**Fig. 1.** The comparison between softmax and its variants from the geometry view(left) and the distribution view(right). The explanations of the components are listed under the plot. The distribution view is obtained by training a ResNet18 and evaluated on IJB-B. Closeness is higher when the boundary value('b' in the figure) is larger. The boundary value asides the top 80% of positive scores on its right to indicate closeness. The diversity is higher when the mean of the red distribution is closer to zero and the variance is smaller.

# 3 Preliminary

## 3.1 Understanding Softmax-based Methods

The softmax cross-entropy loss function(denoted as 'softmax' for short) can be formulated as follows:

$$\mathcal{L}_{CE} = -\log \frac{e^{f_y}}{e^{f_y} + \sum_{k=1,k\neq y}^{C} e^{f_k}} \tag{1}$$

The development of softmax variants in FR is mainly attributed to the advancement of the positive logit $f_y$ and the negative logit $f_k$.

Let $\boldsymbol{x}$, $\boldsymbol{w}_i$, and $b_i$ denote the feature vector of the input sample, the $i$-th class center, and the bias term, respectively. Then the traditional logit is calculated as $f_i = \boldsymbol{w}_i^T \boldsymbol{x} + b_i$. It is a common practice to ignore the bias term $b_i$ in deep FR works [4,15,22,32]. The logit can be transformed as $\boldsymbol{w}_i^T \boldsymbol{x} = \|\boldsymbol{w}_i\| \, \|\boldsymbol{x}\| \cos\theta_i$, where $\theta_i$ is the angle distance between the feature $\boldsymbol{x}$ and the class center $\boldsymbol{w}_i$. Several works [4,32] further fix $\|\boldsymbol{w}_i\| = \|\boldsymbol{x}\| = 1$ and scale $\|\boldsymbol{x}\|$ to $s$. Then the logit can be reformulated as $f_i = s\cos\theta_i$.

As well-discussed in several works [8,29,38], the softmax's constraint can be decoupled into the pulling force from the same class center and the pushing forces from the negative ones. Softmax can thus increase intra-class similarity $s_p$ with the pulling force while reducing the inter-class similarity $s_n$ with the pushing forces. In addition, if the pulling force is enlarged, then the pushing forces will be amplified as well, and vice versa. To clarify this phenomenon, we provide a toy example in Sec. 3.2.

Original softmax simply optimizes $(s_n - s_p)$ to the decision boundary $s_n - s_p = 0$. However, the discriminative ability of facial features learned by the original softmax is limited. Therefore, several works introduce various margin penalties into softmax to improve feature discrimination. Generally speaking, they employ the decision boundary $s_n - s_p = -m$, and $m$ is called margin.

Compared with original softmax, the positive logit is reformulated as $f_y = s\cos(\theta_y + m)$ in ArcFace [4]. As shown on the left side in Fig. 1(b), ArcFace shrinks the decision boundary towards the positive direction with a constant $m$. In this way, it can learn more discriminative face representations, as shown on the right side in Fig. 1(b). In MV-Softmax, the misclassified sample's negative logit is further reformulated as $f_k = s(t\cos\theta_k + t - 1)$, where $t > 1$. Based on ArcFace, MV-Softmax makes the decision boundary more rigorous by introducing an extra margin penalty on the negative logit for handling hard samples, as shown on the left side in Fig. 1(c). The right of Fig. 1(c) shows that (1) the negative distribution of MV-Softmax is more compact than ArcFace, which illustrates that MV-Softmax improves the diversity with the extra margin; (2) MV-Softmax has an inferior boundary value, indicating its insufficient closeness. A possible explanation is that MV-Softmax intends to improve the diversity for misclassified samples by the extra margin, however, it will indirectly enforce hard samples containing large uncertainty and noise to get closer to their positive class centers(see analysis in Sec. 3.2), leading to the overfitting and the inferior generalization ability.

Besides, another line of works exists, e.g., AdaCos [41], AdaptiveFace [14], and MagFace [18]. They substitute the constant margin penalty with the adaptive one to generate more effective supervision during training. MagFace learns well-structured intra-class features by dynamically adjusting the decision boundaries based on the feature magnitude. As shown on the left side in Fig. 1(d), MagFace relaxes the decision boundary for hard samples with large uncertainty and tightens the decision boundary for easy ones with high quality. The right side of Fig. 1(d) shows that the negative distribution in MagFace is not so compact as that in ArcFace. A probable reason is that MagFace prevents hard samples from obtaining excessive $s_p$ during training by reducing the margin. Considering that the optimization of $s_n$ and $s_p$ are highly coupled in softmax, a suitable $s_n$ is not well-learned.

In summary, MV-Softmax and MagFace adopt different strategies to deal with hard samples. MV-Softmax enhances the constraint strength for hard samples. Because MV-Softmax indirectly emphasizes $s_p$ for uncertain samples, it fails to generalize well on challenging tasks during testing. By contrast, MagFace relaxes the constraint for hard samples. Although MagFace can prevent uncertain samples from obtaining excessive $s_p$, it can not ensure a desirable $s_n$

for discriminative features learning. Therefore, both of the above methods fail to emphasize closeness and diversity simultaneously.

### 3.2 Derivative Analysis

In this subsection, we investigate how the margin penalty affects the pulling force and the pushing forces of softmax. Specifically, we demonstrate that if the pulling force is enlarged, the pushing forces will also get increased, and vice versa. A toy example is additionally provided to explain this phenomenon.

Let us start with the gradient of softmax with respect to the logit $f_i$, which is calculated as:

$$\frac{\partial \mathcal{L}_{CE}}{\partial f_i} = \underbrace{\mathbb{1}(i = y) \cdot (p_i - 1)}_{\text{pull}} + \underbrace{\mathbb{1}(i \neq y) \cdot p_i}_{\text{push}} \tag{2}$$

where $p_i = \frac{e^{f_i}}{\sum_{k=1}^{C} e^{f_k}}$ is the predicted probability of the $i$-th class, and satisfies $\sum_{i=1}^{C} p_i = 1$. Eq. 2 contains two parts: (1) the first part aims to pull a sample towards its positive class center; (2) the second part aims to push a sample away from its negative class centers. The differences between the two parts lie in that the pulling force can quickly establish the class center; however, it can hurt the generalization ability by pulling a noise sample near its class center. The pushing forces can help to enhance the discrimination ability via encouraging diversity, but they can not be directly employed to establish the class centers.

Additionally, the pulling and pushing forces are equipped with the opposite signs due to the different relative directions. The gradient summation with respect to each class always equals to the constant zero:

$$\sum_{i=1}^{C} \frac{\partial \mathcal{L}_{CE}}{\partial f_i} = \underbrace{p_y - 1}_{\text{pull}} + \underbrace{\sum\nolimits_{i=1, i \neq y}^{C} p_i}_{\text{push}} = \sum\nolimits_{i=1}^{C} p_i - 1 = 0 \tag{3}$$

Therefore, if we enlarge either of the two forces by a margin, the other one will inevitably get increased at the same time.

Fig. 2 exhibits a simplified example to depict the above issue. We assume the feature vector $\boldsymbol{x}$ and the classifier W are identical among all four cases. Therefore, we only need to care about the relative changes of the logits and the gradients. In Fig. 2(a), we assume the positive logit $f_y$ takes two units and the negative logits $f_{k_1}$ and $f_{k_2}$ take one unit. The classifier W makes the right classification during training with softmax and thus generates limited gradients. Fig. 2(b) shows that the positive logit $f_y$ gets smaller in ArcFace with the margin penalty. The pulling force is directly enhanced, and the pushing forces are indirectly enlarged according to Eq. 3.

Based on ArcFace, MV-Softmax introduces a margin penalty on the negative logits $f_{k_1}$ and $f_{k_2}$ for misclassified samples. The margin penalty directly enlarges $f_{k_1}$ and $f_{k_2}$, leading to the enhanced pushing forces. Although MV-Softmax

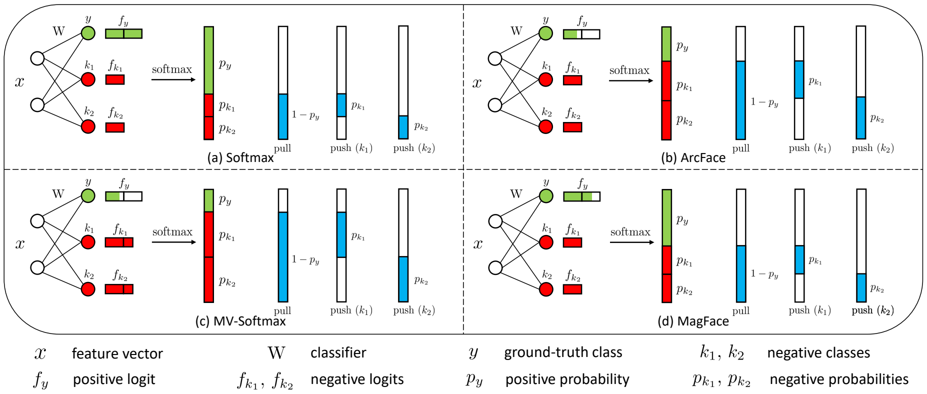


**Fig. 2.** The comparison between softmax and its variants on a toy example. The blue bars denote the magnitude of gradients generated by predictions.

intrinsically enhance the diversity for hard samples, it also indirectly magnifies the pulling force, as shown in Fig. 2(c). By contrast, MagFace adopts a smaller penalty on $f_y$ than ArcFace, leading to both kinds of forces getting smaller, as shown in Fig. 2(d).

Therefore, the pulling and pushing forces are highly coupled in softmax-based loss functions. Due to this phenomenon, softmax-based methods dealing with hard samples tend to get overfitting or underdiscriminative as discussed in Sec. 3.1.

## 4 Proposed Method

Based on the analysis in Sec. 3, we propose a two-branch cooperative framework to enhance closeness and diversity simultaneously. The proposed framework contains three parts: (1) the Hard Sample Mining Scheme for dynamically computing difficulty scores of training samples; (2) the Elementary Representation Branch (ERB) for learning initial face representations; (3) the Refined Representation Branch (RRB) for simultaneous closeness and diversity learning.[1]

### 4.1 Hard Sample Mining

Hard samples play an important role in guiding DCNNs to learn discriminative features. The previous works [10,34] indicate the misclassified samples as hard ones, but they can not measure the hardness quantitatively. MagFace [18] employs feature magnitude to determine the difficulty degree, but it lacks intuitive interpretability. Different from the above works, we employ cosine similarity to characterize the difficulty degree for its simplicity and effectiveness.

[1] Code is available at: https://github.com/Zacharynjust/FR-closeness-and-diversity.

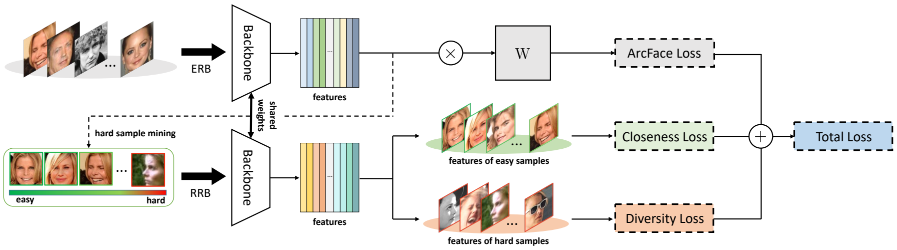


**Fig. 3.** The overview of the proposed framework. Overall, it contains three parts: (1) the Hard Sample Mining Scheme to maintain difficulty scores; (2) the Elementary Representation Branch(ERB) to learn basic face representations; (3) the Refined Representation Branch(RRB) to learn closeness and diversity.

To smooth out short-term fluctuations caused by the sampling sequence, we calculate the moving average of cosine similarity to characterize the difficulty degree for each sample:

$$d^{(t)} = \alpha d^{(t-1)} + (1-\alpha)\cos\theta^{(t)} \tag{4}$$

where $d$ represents the moving averaged similarity, when $d$ is smaller the difficulty degree is higher. $t$ stands for the $t$-th iteration. $\theta$ is the angle between a feature vector and its positive class center. $\alpha$ is the weight factor.

Based on Eq. 4, we propose two schemes to indicate hard samples: the Hard Mining Scheme(HMS) and the Soft Mining Scheme(SMS). HMS explicitly divides all training samples into easy/hard groups. SMS does not need the explicit division; by contrast, it adopts the difficulty degree to balance the weights of closeness and diversity for each sample in the training stage. In this way, easy samples can help the network establish robust class centers quickly. By contrast, hard samples are beneficial for the network to further improve feature discriminations.

### 4.2 Loss Design

Both closeness and diversity are indispensable for achieving better results; therefore, they should be simultaneously and properly emphasized for specific samples. Overall, we need to ensure the following conditions: (1) Easy samples should get enough closeness to ensure robust class centers [18]; (2) Hard samples should keep a suitable distance from their positive class centers to ensure generalization ability [18]; (3) Hard samples should also gain enough distance away from their negative class centers to ensure discrimination ability [34].

Equipped with the above three conditions, we design the following two loss functions in RRB to emphasize closeness and diversity simultaneously during training.

**Loss for Closeness.** To enhance the closeness, we directly minimize the feature differences of within-class samples. The loss term with HMS can be formulated as follows:

$$\mathcal{L}^{H}_{closeness} = \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim P_{pos}} \|g(\boldsymbol{x}) - g(\boldsymbol{y})\|^2 \tag{5}$$

where $P_{pos}$ stands for the distribution of positive pairs constructed from the mini-batch. $g(\cdot)$ is the feature encoder with $l2$-normalization in its output layer. The loss term with SMS is actually the re-weighted version of Eq. 5:

$$\mathcal{L}^{S}_{closeness} = \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim P_{pos}} \phi\left(d_x, d_y\right) \|g(\boldsymbol{x}) - g(\boldsymbol{y})\|^2 \tag{6}$$

where $d_x$ and $d_y$ are the difficulty degrees of $\boldsymbol{x}$ and $\boldsymbol{y}$. $\phi(\cdot)$ is a monotonically non-decreasing function of $d_x$ and $d_y$.

**Loss for Diversity.** In this part, we design diversity loss to enhance the discrimination and generalization ability. Inspired by the uniform loss format used in [33], we design loss for enhancing diversity which enlarges the distance between a sample and its negative class centers. The proposed diversity loss for a single sample can be formulated as follows:

$$\mathcal{L}^{i}_{diversity} = \log \mathop{\mathbb{E}}_{\boldsymbol{w}_j\sim \mathrm{W}^{(i)}_{\mathrm{sub}}} \left[e^{\max(0,\mathrm{sgn}(\cos\theta_j))\cdot s(\cos\theta_j)^2}\right] \tag{7}$$

where $\mathrm{W}^{(i)}_{\mathrm{sub}}$ stands for the subset of the negative class centers for $i$-th sample. $s$ is the scale parameter. $\mathrm{sgn}(\cdot)$ is the sign function and $\max(0, \mathrm{sgn}(\cos\theta_j))$ is used to truncate the gradient when $\cos\theta_j$ is smaller than 0. $\cos\theta_j = \boldsymbol{w}_j^T\boldsymbol{x}_i$ is the similarity between a sample and its negative class $j$. For a mini-batch hard samples, we unite all their subsets of negative class centers and then exclude their positive labels to construct the final negative class centers for a mini-batch W, which can be formulated as follows:

$$\mathrm{W} = \bigcup_{i=1}^{N} \mathrm{W}^{(i)}_{\mathrm{sub}} - \overline{\mathrm{W}} \tag{8}$$

where $N$ stands for the total number of hard samples within a mini-batch. $\overline{\mathrm{W}}$ represents the positive class centers of hard samples in a mini-batch. Then the diversity loss function within a mini-batch using HMS can be further calculated as follows:

$$\mathcal{L}^{H}_{diversity} = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim \mathrm{X}} \left[\log \mathop{\mathbb{E}}_{\boldsymbol{w}_j\sim \mathrm{W}} \left[e^{\max(0,\mathrm{sgn}(\cos\theta_j))\cdot s(\cos\theta_j)^2}\right]\right] \tag{9}$$

where X stands for the sample set in a mini-batch. W represents the final negative class centers for a mini-batch. The proposed diversity loss in conjunction with SMS is formulated as follows:

$$\mathcal{L}^{S}_{diversity} = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim \mathrm{X}} \left[\psi(d_x) \log \mathop{\mathbb{E}}_{\boldsymbol{w}_j\sim \mathrm{W}} \left[e^{\max(0,\mathrm{sgn}(\cos\theta_j))\cdot s(\cos\theta_j)^2}\right]\right] \tag{10}$$

where $d_x$ is the difficulty degree of $x$. $\psi(\cdot)$ is a monotonically non-increasing function of $d_x$.

Because the proposed diversity loss minimizes the cosine similarity between a sample feature and its negative class centers, it can directly push hard samples away from their negative class centers. Margin-based softmax actually seeks to minimize $(s_n - s_p)$ to achieve the decision boundary $s_n - s_p = -m$, when we reduce $s_n$ via diversity loss, we actually shift the optimization emphasis of hard samples on reducing $s_n$. Therefore, diversity loss can somewhat prevent hard samples from achieving excessive $s_p$.

### 4.3 Sampling Strategy

In this subsection, we introduce the proposed sampling strategy for RRB in detail. We construct a mini-batch depending on class labels and difficulty degrees. The sampling strategy can be formulated as follows:

$$\mathrm{X} = \mathcal{S}(\boldsymbol{y}, \boldsymbol{d}) \tag{11}$$

where $\boldsymbol{y}$ denotes the ground truth labels. $\boldsymbol{d}$ stands for the difficulty degrees of the whole dataset as introduced in Sec. 4.1.

For HMS, we further select a specific number of easy samples $\mathrm{X}_e$ and hard samples $\mathrm{X}_h$ according to the divided easy/hard groups within each class. Then, the total loss can be formulated as follows:

$$\mathcal{L}^{H}_{total}(\mathrm{X}) = \mathcal{L}_{CE}(\mathrm{X}) + \lambda_1 \mathcal{L}^{H}_{closeness}(\mathrm{X}_e) + \lambda_2 \mathcal{L}^{H}_{diversity}(\mathrm{X}_h) \tag{12}$$

where $\mathcal{L}_{CE}$ is ArcFace in our model. We can also choose other margin-based softmax loss functions. $\lambda_1$ and $\lambda_2$ are weight parameters for closeness and diversity respectively. For SMS, we randomly select $N$ samples within each class and use their difficulty degrees to calculate the weight functions. Then, the total loss can be formulated as follows:

$$\mathcal{L}^{S}_{total}(\mathrm{X}, \boldsymbol{d}) = \mathcal{L}_{CE}(\mathrm{X}) + \lambda_1 \mathcal{L}^{S}_{closeness}(\mathrm{X}, \boldsymbol{d}) + \lambda_2 \mathcal{L}^{S}_{diversity}(\mathrm{X}, \boldsymbol{d}) \tag{13}$$

where $\boldsymbol{d}$ is the difficulty degrees of samples in a mini-batch.

## 5 Experiment

### 5.1 Implementation Details

**Datasets.** We employ MS1MV2 [4] as our training data for a fair comparison with other methods. MS1MV2 is a semi-automatic refined version of the MS1M [5], containing about 5.8M images of 85K different identities. For testing, we extensively evaluate our proposed method and the competed methods on several popular benchmarks, including LFW [9], CFP-FP [24], CPLFW [42], AgeDB [19], CALFW [43], IJB-B [37], IJB-C [17], and MegaFace [11].

**Table 1.** Performance comparisons between the proposed method and state-of-the-art methods on various benchmarks. * denotes our re-implement results on ResNet100. [**Best**, Second Best]

| Methods | Verification Accuracy | | | | | IJB | | MegaFace | |
|---|---|---|---|---|---|---|---|---|---|
| | LFW | CFP-FP | CPLFW | AgeDB | CALFW | IJB-B | IJB-C | Id | Ver |
| Focal Loss* (CVPR16) | 99.73 | 98.19 | 92.80 | 98.13 | 96.01 | 93.60 | 95.19 | 98.09 | 98.60 |
| SphereFace (CVPR17) | 99.42 | - | - | 97.16 | 94.55 | - | - | - | - |
| CosFace* (CVPR18) | 99.78 | 98.12 | 92.28 | 98.11 | 95.76 | 94.10 | 95.51 | 98.20 | 98.32 |
| ArcFace* (CVPR19) | 99.80 | 98.27 | 92.75 | 98.00 | 95.96 | 94.26 | 95.73 | 98.34 | 98.55 |
| MV-Softmax* (AAAI20) | 99.80 | 98.30 | 92.93 | 97.98 | 96.10 | 94.01 | 95.59 | 98.22 | 98.28 |
| Circle Loss (CVPR20) | 99.73 | 96.02 | - | - | - | - | 93.95 | 98.50 | 98.73 |
| CurricularFace* (CVPR20) | 99.82 | 98.30 | 93.05 | 98.32 | 96.05 | 94.75 | 96.04 | 98.65 | 98.70 |
| MagFace* (CVPR21) | **99.83** | 98.23 | 92.93 | 98.27 | 96.12 | 94.42 | 95.81 | 98.51 | 98.64 |
| Ours, HMS | **99.83** | **98.44** | 93.05 | 98.20 | 96.05 | 94.86 | 96.25 | 98.60 | 98.75 |
| Ours, SMS, Linear | 99.82 | 98.27 | 93.03 | 98.18 | 96.12 | 94.72 | 96.03 | 98.58 | 98.73 |
| Ours, SMS, Non-Linear | 99.82 | 98.40 | **93.12** | **98.37** | **96.15** | **95.02** | **96.35** | **98.72** | **98.84** |

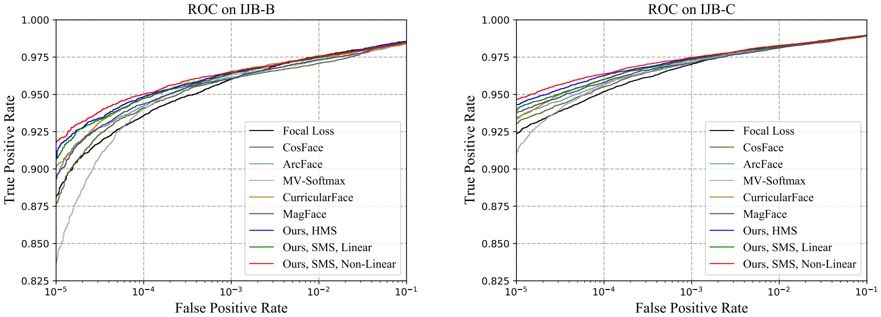
**Experimental Setting.** We follow the setting in ArcFace [4] to align the images with five facial key points [39] and normalize the face images to 112 × 112. ResNet100 is used as the backbone network in our model. We implement our framework with PyTorch [21]. The models are trained by stochastic gradient descent. The batchsize is set to 512. The weight decay is set to $5e-4$, and the momentum is 0.9.

To obtain the difficulty degree of samples, the backbone is firstly trained with ERB for 4 epochs with the learning rate 0.1. The backbone is then trained with the joint supervisions of ERB and RRB for extra 21 epochs. The learning rate is set to 0.1 initially and divided by 10 when the extra epoch is 6, 12 and 18. For the sampling strategy, we choose 64 unique classes for a mini-batch. In addition, 500 negative class centers of each hard sample are selected to construct the imposter pairs.

For HMS, $\lambda_1$ and $\lambda_2$ are set to 0.5 and 1.0. We divide the top 20% of training samples into hard groups and collect seven easy samples and one hard sample within each class. For SMS, $\lambda_1$ and $\lambda_2$ are set to 0.5 and 2.0. In addition, we conduct experiments on both linear and non-linear weight functions. For the linear functions, we set $\phi(d_x, d_y) = \frac{d_x+d_y}{2}$ and $\psi(d_x) = 1 - d_x$. For the non-linear functions, we employ the sigmoid-like function $\sigma(x; \mu, \gamma) = \frac{1}{1+e^{-\gamma(x-\mu)}}$ to conduct non-linear transformation, where we fix $\mu = 0.5$ and $\gamma = 10$. We set $\phi(d_x, d_y) = \sigma(\frac{d_x+d_y}{2})$, and $\psi(d_x) = 1 - \sigma(d_x)$. We set $\alpha$ to 0.9 to calculate the difficulty degree.

### 5.2 Comparisons with SOTA Methods

**Results on Small Benchmarks.** In this subsection, we conducted experiments on various benchmarks, including LFW [9] for unconstrained face verifi-



(a) ROC for IJB-B (b) ROC for IJB-C

**Fig. 4.** ROC of 1:1 verification protocol on IJB-B/C.

cation, CFP-FP [24] and CPLFW [42] for cross-pose variations, AgeDB [19] and CALFW [43] for cross-age variations.

The 1:1 verification accuracy among different methods on the above five benchmarks is listed in Table 1. According to Table 1, the proposed models achieve promising results, especially when they are integrated with HMS or SMS (non-linear). Note that our models give slight improvements on LFW since the performance of LFW is nearly saturated. Besides, for age/pose invariant face verification, our proposed method can achieve better results than our competitors.

**Results on IJB-B and IJB-C.** In this part, we compare our method with the state-of-the-art methods on IJB. Both IJB-B/C datasets are challenging tasks containing a considerable number of face images clipped from videos.

Table 1 lists the comparisons on TAR@FAR=$1e-4$ between our models and the competed methods. Without bells and whistles, our models achieve the leading results among all methods and improve the performance of IJB-B/C clearly. Among our three models, SMS with non-linearity achieves the top results on both IJB-B/C. The reason for the improvements is that both closeness and diversity should be simultaneously emphasized. As well analyzed in Sec 3, both MV-Softmax and MagFace fail to emphasize closeness and diversity in a proper way. In addition, our models outperform the other competitors under most FPR variations, as shown in Fig. 4.

**Results on MegaFace.** In this subsection, we evaluate the proposed method and the competed methods in terms of the identification and verification on MegaFace [11]. In our experiment, MegaFace is used as the gallery set, and FaceScrub is employed as the probe set.

The results of compared methods are listed in Table 1. “Id” refers to the rank-1 face identification accuracy with 1M distractors, and “Ver” refers to the

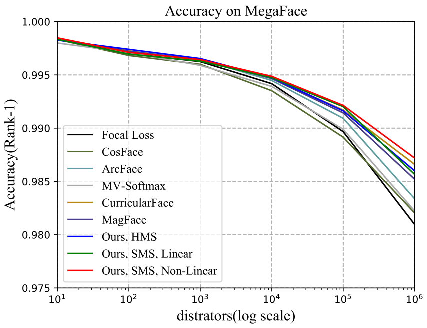


**Fig. 5.** The rank-1 face identification accuracy with different distractors on MegaFace.

**Table 2.** Verification comparisons on several benchmarks under different loss terms and mining schemes. We conduct ablation experiments on the MS1MV2's subset containing 10K unique identites with ResNet34.[**Best**, $\underline{\text{Second Best}}$]

| Model | Closeness | Diversity | Mining Scheme | LFW | CFP-FP | AgeDB | IJB-B | IJB-C |
|---|---|---|---|---|---|---|---|---|
| 1 | | | - | 99.20 | 90.94 | 94.08 | 82.61 | 86.32 |
| 2 | ✓ | | HMS, $T = 20\%$ | $\underline{99.30}$ | 90.79 | 93.96 | 82.14 | 85.91 |
| 3 | | ✓ | HMS, $T = 20\%$ | 99.20 | **91.68** | 94.46 | **83.64** | **86.82** |
| 4 | ✓ | ✓ | HMS, $T = 10\%$ | 99.25 | 91.55 | 94.38 | 83.50 | 86.71 |
| 5 | ✓ | ✓ | HMS, $T = 20\%$ | **99.32** | 91.62 | **94.63** | $\underline{83.62}$ | 86.78 |
| 6 | ✓ | ✓ | HMS, $T = 30\%$ | 99.26 | $\underline{91.65}$ | $\underline{94.60}$ | 83.53 | $\underline{86.87}$ |
| 7 | ✓ | | SMS, $\gamma = 10$ | $\underline{99.28}$ | 90.92 | 94.01 | 82.33 | 86.10 |
| 8 | | ✓ | SMS, $\gamma = 10$ | 99.13 | 91.46 | 94.25 | **83.98** | **87.20** |
| 9 | ✓ | ✓ | SMS, Linear | 99.25 | 91.01 | 94.10 | 82.88 | 86.53 |
| 10 | ✓ | ✓ | SMS, $\gamma = 5$ | 99.27 | $\underline{91.55}$ | 94.23 | 83.80 | 86.95 |
| 11 | ✓ | ✓ | SMS, $\gamma = 10$ | **99.30** | **91.80** | $\underline{94.55}$ | $\underline{83.92}$ | 87.04 |
| 12 | ✓ | ✓ | SMS, $\gamma = 15$ | 99.25 | 91.37 | **94.61** | 83.85 | $\underline{87.10}$ |

face verification on TAR@FPR=$1e-6$. Table 1 shows that our models obtain the overall best results on both identification and verification tasks. Specifically, our model with SMS (non-linear) obtains the highest identification/verification results among all methods. In addition, although the performance will degrade with the increasing number of distractors, our model can achieve overall superiority over other methods, as shown in Fig. 5.

### 5.3 Ablation Study

In this part, we conduct experiments under different settings to investigate the effectiveness of the proposed components.

**Loss Terms.** In Table 2, Model 2, 3, 7, and 8 illustrate that it is difficult to obtain promising results if either closeness or diversity is absent. Model 2 and 7 pay more attention to closeness and achieve desirable results on the simple benchmark(e.g., LFW). However, their performances are degraded when dealing with the challenging datasets(e.g., IJB-B/C). Model 3 and 8 enforce diversity and perform better results on IJB-B/C. However, they have no improvements compared with Model 1 on LFW.

**Division Thresholds for HMS.** Here, we evaluate the proposed hard sample division scheme on several benchmarks. The experimental results of different division percentages for hard samples are listed in Table 2 (Model 4-6). Table 2 shows that our model achieves the best overall performance by taking 20% of the training data as hard samples. Besides, our model can also give competitive results when the hard samples occupy 10% and 30% in training samples.

**Non-Linearity Magnitudes $\gamma$ for SMS.** Model 9-12 in Table 2 provide the experimental results with different magnitudes of non-linearity. Model 9 achieves limited improvements compared with Model 1, indicating that it is difficult to achieve closeness and diversity by using the proposed SMS with linear function. Model 10-12 demonstrate that a suitable non-linearity by adjusting $\gamma$(i.e., 10) is helpful to achieve closeness and diversity. In addition, SMS with a properly $\gamma$ can achieve better results on IJB datasets than HMS. The possible reason is that non-linear function can adjust the weight of samples adaptively based on different difficulty degrees.

## 6 Conclusion

This paper has proposed a two-branch cooperative network to learn discriminative features according to the understanding of margin-based softmax methods from the geometry view. Softmax-based methods can be considered as the pulling force from the corresponding class center and the pushing force from the negative class centers. Based on this, our model further enlarges the pulling force to enhance closeness and employ pushing force to enforce diversity. Several experimental results demonstrate the superiority of our proposed method over other competitors.

**Acknowledgements.** This work was supported by the National Science Fund of China under Grant Nos. 61876083,62176124.

## References

1. Chen, B., Liu, W., Yu, Z., Kautz, J., Shrivastava, A., Garg, A., Anandkumar, A.: Angular visual hardness. In: International Conference on Machine Learning. pp. 1637–1648. PMLR (2020)

2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
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., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
5. 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)
6. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742. IEEE (2006)
7. 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)
8. He, L., Wang, Z., Li, Y., Wang, S.: Softmax dissection: Towards understanding intra-and inter-class objective for embedding learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10957–10964 (2020)
9. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition (2008)
10. 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)
11. 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)
12. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
14. Liu, H., Zhu, X., Lei, Z., Li, S.Z.: Adaptiveface: Adaptive margin and sampling for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11947–11956 (2019)
15. 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)
16. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: ICML. vol. 2, p. 7 (2016)
17. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 International Conference on Biometrics (ICB). pp. 158–165. IEEE (2018)

18. 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)
19. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 51–59 (2017)
20. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition (2015)
21. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
22. Ranjan, R., Castillo, C.D., Chellappa, R.: L2-constrained softmax loss for discriminative face verification. arXiv preprint arXiv:1703.09507 (2017)
23. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
24. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9. IEEE (2016)
25. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 761–769 (2016)
26. Sun, Y.: Deep learning face representation by joint identification-verification. The Chinese University of Hong Kong (Hong Kong) (2015)
27. Sun, Y., Liang, D., Wang, X., Tang, X.: Deepid3: Face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2015)
28. Sun, Y., Wang, X., Tang, X.: Hybrid deep learning for face verification. In: Proceedings of the IEEE international conference on computer vision. pp. 1489–1496 (2013)
29. 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)
30. 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)
31. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Processing Letters **25**(7), 926–930 (2018)
32. 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)
33. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning. pp. 9929–9939. PMLR (2020)
34. 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)
35. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. Journal of machine learning research **10**(2) (2009)
36. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. pp. 499–515. Springer (2016)

37. 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)
38. Zeng, D., Shi, H., Du, H., Wang, J., Lei, Z., Mei, T.: Npcface: Negative-positive collaborative training for large-scale face recognition. arXiv preprint arXiv:2007.10172 (2020)
39. 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)
40. 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)
41. 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/CVF Conference on Computer Vision and Pattern Recognition. pp. 10823–10832 (2019)
42. Zheng, T., Deng, W.: Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Beijing University of Posts and Telecommunications, Tech. Rep **5**, 7 (2018)
43. Zheng, T., Deng, W., Hu, J.: Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. arXiv preprint arXiv:1708.08197 (2017)