CoupleFace: Relation Matters for Face Recognition Distillation

Jiaheng Liu[1*], Haoyu Qin[2*], Yichao Wu[2], Jinyang Guo[3], Ding Liang[2], and Ke Xu[1]

[1] State Key Lab of Software Development Environment, Beihang University
[2] SenseTime Group Limited
[3] SKLSDE, Institute of Artificial Intelligence, Beihang University
liujiaheng@buaa.edu.cn, {qinhaoyu1,wuyichao,liangding}@sensetime.com

Abstract. Knowledge distillation is an effective method to improve the performance of a lightweight neural network (i.e., student model) by transferring the knowledge of a well-performed neural network (i.e., teacher model), which has been widely applied in many computer vision tasks, including face recognition (FR). Nevertheless, the current FR distillation methods usually utilize the Feature Consistency Distillation (FCD) (e.g., L_2 distance) on the learned embeddings extracted by the teacher and student models for each sample, which is not able to fully transfer the knowledge from the teacher to the student for FR. In this work, we observe that mutual relation knowledge between samples is also important to improve the discriminative ability of the learned representation of the student model, and propose an effective FR distillation method called CoupleFace by additionally introducing the Mutual Relation Distillation (MRD) into existing distillation framework. Specifically, in MRD, we first propose to mine the informative mutual relations, and then introduce the Relation-Aware Distillation (RAD) loss to transfer the mutual relation knowledge of the teacher model to the student model. Extensive experimental results on multiple benchmark datasets demonstrate the effectiveness of our proposed CoupleFace for FR. Moreover, based on our proposed CoupleFace, we have won the first place in the ICCV21 Masked Face Recognition Challenge (MS1M track).

Keywords: Face recognition, Knowledge distillation, Loss function

1 Introduction

Face recognition (FR) has been well investigated for decades. Most of the progress is credited to large-scale training datasets [59,19], resource-intensive networks with millions of parameters [13,38] and effective loss functions [4,47]. In practice, FR models are often deployed on mobile and embedded devices, which are incompatible with the large neural networks (e.g., ResNet-101 [13]). Besides, as shown in Fig. 1(a), the capacities of the lightweight neural networks (e.g.,

*First two authors contributed equally.

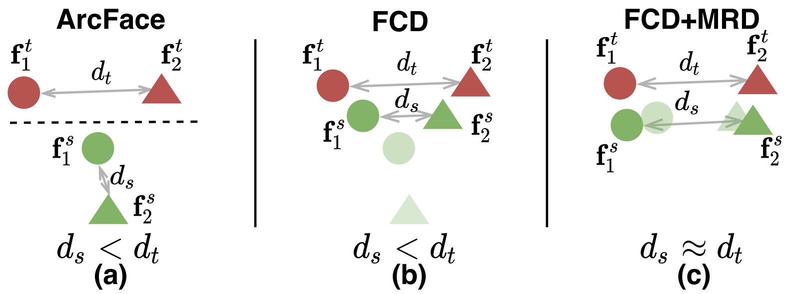

Fig. 1. The illustration of different methods for two samples from different classes. $\mathbf{f}_1^t$ and $\mathbf{f}_2^t$ are extracted by teacher model, and $\mathbf{f}_1^s$ and $\mathbf{f}_2^s$ are extracted by student model. d_t and d_s denote the distances (i.e., $1 - \cos(\mathbf{f}_1^t, \mathbf{f}_2^t)$ and $1 - \cos(\mathbf{f}_1^s, \mathbf{f}_2^s)$) of the embeddings, where $\cos(\cdot,\cdot)$ measures the cosine similarity of two features. Model is better when distance is larger. (a). The teacher and student models are both trained by ArcFace [4]. (b). The teacher model is trained by ArcFace. The student model is trained by using FCD. (c). The teacher model is trained by ArcFace. The student model is trained by using both FCD and MRD in CoupleFace.

MobileNetV2 [35]) cannot be fully exploited when they are only supervised by existing popular FR loss functions (e.g., ArcFace [4]). Therefore, how to develop lightweight and effective FR models for real-world applications has been investigated in recent years. For example, knowledge distillation [14] is being actively discussed to produce lightweight and effective neural networks, which transfers the knowledge from an effective teacher model to a lightweight student model.

Most existing knowledge distillation works usually aim to guide the student to mimic the behavior of the teacher by introducing probability constraints (e.g., KL divergence [14]) between the predictions of teacher and student models, which are not well-designed for FR. In contrast, as improving the discriminative ability of the feature embedding is the core problem for FR, it is important to enable the student model to share the same embedding space with the teacher model for similarity comparison. Thus, a simple and straightforward FR distillation method is to directly minimize the L_2 distance of the embeddings extracted by teacher and student models [49,37], which aims to align the embedding spaces between the teacher and student models. We call this simple method as Feature Consistency Distillation (FCD) as shown in Fig. 1(b), and FCD has been widely used in practice to improve the lightweight neural networks for FR.

In Fig. 1(b), when FCD is used, feature embeddings extracted by student model get close to the corresponding feature embeddings extracted by teacher model. However, there are some cases that the distances of embeddings (i.e., d_s) from different classes of student model are still smaller than d_t of teacher model. Similarly, there are also some cases that d_s from the same class is larger than d_t. In practice, when student model is deployed, smaller d_s for negative pair or larger d_s for positive pair usually leads to false recognition, which indicates that FCD is not sufficient to transfer the knowledge of teacher model to student model. In Fig. 1(c), we define the cosine similarity between samples as mutual relation and observe that $d_s \approx d_t$ when we additionally utilize the mutual relation information of teacher model to distill the corresponding mutual relation information of student model. Thus, we propose to introduce the Mutual Relation Distillation (MRD) into the existing FR distillation by reducing the gap between teacher and student models with respect to the mutual relation information.

Since FCD is able to align the embedding space of student model to teacher model well, the student model can distinguish most image pairs easily, which indicates that the differences of the most mutual relations between teacher and student models are relatively small. In other words, most mutual relations cannot provide valuable knowledge and affect the similarity distribution of all image pairs for FR. Therefore, how to generate sufficient informative mutual relations efficiently in MRD is a challenging issue.

Moreover, recent metric learning works [11,53,39] have shown that hard negative samples are crucial for improving the discriminative ability of feature embedding. But these works are not well-designed for mining mutual relations between samples for FR distillation. To this end, we propose to mine informative mutual relations in MRD. Moreover, as the number of positive pairs is usually small, we focus on mining the mutual relations of negative pairs.

Overall, in our work, we propose an effective FR distillation method referred to as CoupleFace, including FCD and MRD. Specifically, we first pre-train the teacher model on the large-scale training dataset. Then, in FCD, we calculate the L_2 distance of the embeddings extracted by the teacher and student models to generate the FCD loss for aligning the embedding spaces of the teacher and student models. In MRD, we first propose the Informative Mutual Relation Mining module to generate informative mutual relations efficiently in the training process, where the informative prototype set generation and memory-updating strategies are introduced to improve the mining efficiency. Then, we introduce the Relation-Aware Distillation (RAD) loss to exploit the mutual relation knowledge by using valid mutual relations and filtering out relations with subtle differences, which aims to better transfer the informative mutual relation knowledge from the teacher model to the student model.

The contributions are summarized as follows:

- In our work, we first investigate the importance of mutual relation knowledge for FR distillation, and propose an effective distillation framework called CoupleFace, which consists of Feature Consistency Distillation (FCD) and Mutual Relation Distillation (MRD).
- In MRD, we propose to obtain informative mutual relations efficiently in our Informative Mutual Relation Mining module, where the informative prototype set generation and memory-updating strategies are used. Then, we introduce the Relation-Aware Distillation (RAD) loss to better transfer the mutual relation knowledge.
- Extensive experiments on multiple benchmark datasets demonstrate the effectiveness and generalization ability of our proposed CoupleFace method.

2 Related Work

Face Recognition. FR aims to maximize the inter-class discriminative ability and the intra-class compactness. The success of FR can be summarized into the three factors: effective loss functions [47,58,28,41,46,4,3,31,5,23,24,21,26], large-scale datasets [59,1,18,57,19], and powerful deep neural networks [43,40,38,42,27]. The loss function design is the main-stream research direction for FR, which improves the generalization and discriminative abilities of the learned feature representation. For example, Triplet loss [36] is proposed to enlarge the distances of negative pairs and reduce the distances of positive pairs. Recently, the angular constraint is applied into the cross-entropy loss function in many angular-based loss functions [29,28]. Besides, CosFace [48] and ArcFace [4] further utilize a margin item for better discriminative capability of the feature representation. Moreover, some mining-based loss functions (e.g., CurricularFace [16] and MV-Arc-Softmax [50]) take the difficulty degree of samples into consideration and achieve promising results. The recent work VPL [5] additionally introduces the sample-to-sample comparisons to reduce the gap between the training and evaluation processes for FR. In contrast, we propose to design an effective distillation loss function to improve the lightweight neural network.

Knowledge Distillation. As a representative type of model compression and acceleration methods [9,10,22], knowledge distillation aims to distill knowledge from a powerful teacher model into a lightweight student model [14], which has been applied in many computer vision tasks [56,54,33,44,32,15,2,6,7,55,25,17]. Many distillation methods have been proposed by utilizing different kinds of representation as knowledge for better performance. For example, FitNet [34] uses the middle-level hints from hidden layers of the teacher model to guide the training process of the student model. CRD [44] utilizes a contrastive-based objective function for transferring knowledge between deep networks. Some relation-based knowledge distillation methods (e.g., CCKD [33], RKD [32]) utilize the relation knowledge to improve the student model. Recently, knowledge distillation has also been applied to improve the performance of lightweight network (e.g., MobileNetV2 [35]) for FR. For example, EC-KD [49] proposes a position-aware exclusivity strategy to encourage diversity among different filters of the same layer to alleviate the low capability of student models. When compared with existing works, CoupleFace is well-designed for FR distillation by considering to mine informative mutual relations and transferring the relation knowledge of the teacher model using Relation-Aware Distillation loss.

3 Method

In this section, we introduce the details of our CoupleFace in Fig. 2, which contains Feature Consistency Distillation (FCD) and Mutual Relation Distillation (MRD) for FR distillation. The overall pipeline is as follows. First, we train the teacher model on a large-scale dataset. Then, in the distillation process of student model, we extract the feature embeddings based on the teacher and student models for each face image. After that, in FCD, we compute the Feature

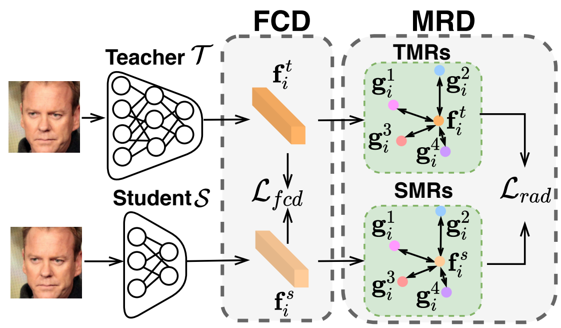

Fig. 2. The framework of CoupleFace for FR distillation. In FCD, we use the $\mathcal{L}_{fcd}$ loss between $\mathbf{f}_i^s$ and $\mathbf{f}_i^t$ to align the embedding spaces of the teacher and student models. In MRD, we first mine the informative features $\{\mathbf{g}_i^k\}_{k=1}^K$ (see Sec. 3.3), where K is the number of features, and calculate the teacher mutual relations (TMRs) and student mutual relations (SMRs) based on $\mathbf{f}_i^s$, $\mathbf{f}_i^t$ and $\{\mathbf{g}_i^k\}_{k=1}^K$. Then, we minimize the $\mathcal{L}_{rad}$ loss to transfer the mutual relation knowledge of the teacher model to student model.

Consistency Distillation (FCD) loss $\mathcal{L}_{fcd}$ based on L_2 distance of the feature embeddings. Meanwhile, in MRD, we first build the informative mutual relations using the feature embeddings of teacher and student models with the mined informative features as shown in Fig. 3, and then calculate the Relation-Aware Distillation (RAD) loss $\mathcal{L}_{rad}$ to transfer the mutual relation knowledge.

3.1 Preliminary on Face Recognition Distillation

In this section, we define some notations in CoupleFace, and discuss the necessity of FR distillation when compared with traditional knowledge distillation.
Notations. We denote the teacher model as $\mathcal{T}$ and the student model as $\mathcal{S}$. For each sample x_i, the corresponding identity label is y_i, and the corresponding features extracted by $\mathcal{T}$ and $\mathcal{S}$ are denoted as $\mathbf{f}_i^t$ and $\mathbf{f}_i^s$, respectively.
Necessity of Face Recognition Distillation. For the traditional knowledge distillation of image classification, the existing methods usually utilize the probability consistency [14] (e.g., KL divergence) to align the prediction probabilities from $\mathcal{S}$ with the prediction probabilities from $\mathcal{T}$. However, the traditional knowledge distillation techniques are usually incompatible with FR. In practice, for FR, we can only obtain a pre-trained $\mathcal{T}$ but have no idea about how it was trained (e.g., the training datasets, loss functions). Therefore, the probability consistency loss is not available when the number of identities of the training dataset for $\mathcal{T}$ is different from the current dataset for $\mathcal{S}$ or $\mathcal{T}$ is trained by other metric learning based loss functions (e.g., triplet loss [36]). Besides, FR models are trained to generate discriminative feature embeddings for similarity comparison in the open-set setting rather than an effective classifier for the close-set classification. Thus, aligning the embedding spaces between $\mathcal{S}$ and $\mathcal{T}$ is more important for FR distillation.

3.2 Feature Consistency Distillation

In Feature Consistency Distillation (FCD), to boost the performance of $\mathcal{S}$ for FR, a simple and effective Feature Consistency Distillation (FCD) loss $\mathcal{L}_{fcd}$ is widely adopted in practice, which is defined as follows:

$$\mathcal{L}_{fcd} = \frac{1}{2N}\sum_{i=1}^{N}\left|\left|\frac{\mathbf{f}_i^t}{||\mathbf{f}_i^t||_2} - \frac{\mathbf{f}_i^s}{||\mathbf{f}_i^s||_2}\right|\right|^2, \tag{1}$$

where N is the number of face images for each mini-batch.

3.3 Mutual Relation Distillation

In this section, we describe the Mutual Relation Distillation (MRD) of CoupleFace in detail. First, we discuss the necessity of MRD for FR distillation. Then, we describe how to generate the informative mutual relations by using our informative mutual relation mining strategy. Finally, we introduce the Relation-Aware Distillation (RAD) loss to transfer the mutual relation knowledge.

Necessity of Mutual Relation Distillation. First, we define a pair of embeddings as a couple. Given a couple $(\mathbf{f}_i^g, \mathbf{f}_j^g)$, $g \in \{s, t\}$ denotes $\mathcal{S}$ or $\mathcal{T}$, the mutual relation $\mathrm{R}(\mathbf{f}_i^g, \mathbf{f}_j^g)$ of this couple is defined as follows:

$$\mathrm{R}(\mathbf{f}_i^g, \mathbf{f}_j^g) = \cos(\mathbf{f}_i^g, \mathbf{f}_j^g), \tag{2}$$

where $\cos(\cdot,\cdot)$ measures the cosine similarity between two features. Given two couples $(\mathbf{f}_i^t, \mathbf{f}_i^s)$ and $(\mathbf{f}_j^t, \mathbf{f}_j^s)$, when FCD loss is applied on $\mathcal{S}$, the mutual relations $\mathrm{R}(\mathbf{f}_i^t, \mathbf{f}_i^s)$ and $\mathrm{R}(\mathbf{f}_j^t, \mathbf{f}_j^s)$ will be maximized. However, in practice, when $\mathcal{S}$ is deployed, the mutual relation $\mathrm{R}(\mathbf{f}_i^s, \mathbf{f}_j^s)$ will be used to measure the similarity of this couple $(\mathbf{f}_i^s, \mathbf{f}_j^s)$ for FR. Thus, if we only use the FCD loss in FR distillation, the optimization on the mutual relation $\mathrm{R}(\mathbf{f}_i^s, \mathbf{f}_j^s)$ is ignored, which limits the further improvement of $\mathcal{S}$ for FR. Meanwhile, $\mathcal{T}$ with superior performance is able to provide an effective mutual relation $\mathrm{R}(\mathbf{f}_i^t, \mathbf{f}_j^t)$ as the ground-truth to distill the mutual relation $\mathrm{R}(\mathbf{f}_i^s, \mathbf{f}_j^s)$ from $\mathcal{S}$. Therefore, we introduce the MRD into the existing FR distillation framework.

However, direct optimization on mutual relation $\mathrm{R}(\mathbf{f}_i^s, \mathbf{f}_j^s)$ may not be a good choice in practice. Due to the batch size limitation and randomly sampling strategy in the training process, the mutual relations for these couples $\{(\mathbf{f}_i^s, \mathbf{f}_j^s)\}_{j=1, j\neq i}^{N}$ constructed across the mini-batch cannot support an effective and efficient mutual relation distillation. To this end, in CoupleFace, we propose to optimize $\mathrm{R}(\mathbf{f}_i^s, \mathbf{f}_j^t)$ instead of $\mathrm{R}(\mathbf{f}_i^s, \mathbf{f}_j^s)$ for following reasons. First, the quantity of mutual relations for these couples $\{(\mathbf{f}_i^s, \mathbf{f}_j^t)\}_{j=1, j\neq i}^{L}$ can be very large, where L is the number of samples of the dataset and $\mathbf{f}_j^t$ can be pre-calculated using teacher model. Second, the quality of the mutual relation for couple $(\mathbf{f}_i^s, \mathbf{f}_j^t)$ can be guaranteed as it can be mined from sufficient couples. Third, $\mathbf{f}_j^s$ is almost the same as $\mathbf{f}_j^t$ from the perspective of $\mathbf{f}_i^s$ when the FCD loss between $\mathbf{f}_j^s$ and $\mathbf{f}_j^t$

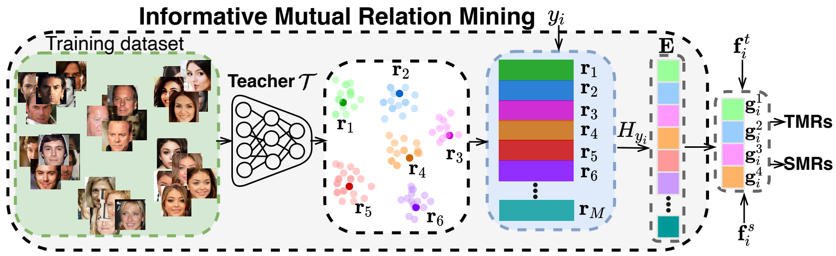

Fig. 3. We first pre-calculate identity prototypes $\{\mathbf{r}_m\}_{m=1}^M$ and generate the informative prototype set H_{y_i} for identity y_i using our informative prototype set generation strategy, where M is the number of identities across the training dataset. Then, we maintain a feature bank $\mathbf{E} \in \mathbb{R}^{M\times d}$ to store the feature embeddings of $\mathcal{T}$, which is updated in each iteration using our memory-updating strategy. Finally, we obtain K informative features $\{\mathbf{g}_i^k\}_{k=1}^K$ based on H_{y_i} and $\mathbf{E}$, and construct the SMRs and the TMRs based on $\mathbf{f}_i^s$, $\mathbf{f}_i^t$ and $\{\mathbf{g}_i^k\}_{k=1}^K$. Here, we set $K = 4$ for better illustration.

is fast converged, which represents that the mutual relation $\mathrm{R}(\mathbf{f}_i^s, \mathbf{f}_j^t)$ of couple $(\mathbf{f}_i^s, \mathbf{f}_j^t)$ is an ideal approximation of $\mathrm{R}(\mathbf{f}_i^s, \mathbf{f}_j^s)$ of couple $(\mathbf{f}_i^s, \mathbf{f}_j^s)$. Therefore, we propose to use the mutual relation $\mathrm{R}(\mathbf{f}_i^t, \mathbf{f}_j^t)$ to distill $\mathrm{R}(\mathbf{f}_i^s, \mathbf{f}_j^t)$.

Note that we call $\mathrm{R}(\mathbf{f}_i^t, \mathbf{f}_j^t)$ and $\mathrm{R}(\mathbf{f}_i^s, \mathbf{f}_j^t)$ as teacher mutual relation (**TMR**) and student mutual relation (**SMR**), respectively. To sum up, in MRD, we propose to utilize TMRs to distill the SMRs in the training process of FR.

Informative Mutual Relation Mining. In this section, we describe how to generate informative mutual relations efficiently in Fig. 3 of CoupleFace.

Intuitively, for $\mathbf{f}_i^s$ and $\mathbf{f}_i^t$, a straightforward way is to construct the couples across all training samples and generate SMRs and TMRs. However, the computation cost is very large in this way, which is not applicable in practice. An alternative way is to generate SMRs and TMRs across the mini-batch. However, the batch size is relatively small and most mutual relations cannot provide valuable knowledge to improve $\mathcal{S}$, as it is easy to distinguish most image pairs and only the hard image pairs will greatly affect the performance of FR model. Meanwhile, recent metric learning works [11,53,39] show that hard negative samples are crucial for improving the discriminative ability of the embeddings. Therefore, in MRD, we propose to mine informative mutual relations among negative pairs to reduce the computation cost and improve $\mathcal{S}$ as shown in Fig. 3. Specifically, we use an informative prototype set generation strategy to find a set of most similar identities called informative prototype set H_{y_i} for identity y_i. Then, we utilize a memory-updating strategy to build the informative mutual relations based on $\mathbf{f}_i^s$, $\mathbf{f}_i^t$ and the features belonging to H_{y_i} efficiently.

Informative prototype set generation. First, we extract the features of the training data by using a well-performed trained model. In our work, we directly use $\mathcal{T}$. The generated features can be denoted as $\{\mathbf{f}_i^t\}_{i=1}^L$ and the corresponding identity label is y_i for $\mathbf{f}_i^t$. For the training dataset, we denote the number of samples as L, the number of identities as M, and the identity label set as $\{m\}_{m=1}^M$. For each identity m, we calculate the identity prototype $\mathbf{r}_m$ as follows:

$$\mathbf{r}_m = \frac{1}{l_m} \sum_{i=1, y_i=m}^{L} \frac{\mathbf{f}_i^t}{||\mathbf{f}_i^t||_2}, \tag{3}$$

where l_m is the number of samples in the training dataset for identity m. Then, all identity prototypes of the dataset can be denoted as $\{\mathbf{r}_m\}_{m=1}^M$. After that, to find the informative prototype set H_m for $\mathbf{r}_m$, we calculate the cosine similarity between $\mathbf{r}_m$ and $\mathbf{r}_n$, where $n \in \{m\}_{m=1}^M$ and $n \neq m$. Afterwards, we select the top K (e.g., $K = 100$) identities with the largest similarities to construct H_m for identity m, where H_m contains K identity labels. Finally, the informative prototype set for identity y_i is denoted as H_{y_i}.

Memory-updating. As the number of samples belonging to H_{y_i} is also relatively large, we further propose a memory-updating strategy to reduce the computation cost while making full use of all samples belonging to H_{y_i} inspired by MoCo [12]. Specifically, we maintain a feature bank $\mathbf{E} \in \mathbb{R}^{M \times d}$ to store feature embeddings extracted by $\mathcal{T}$, where only one embedding is preserved for each identity and d is the dimension (e.g., 512) of the embedding. At the beginning of the training process for $\mathcal{S}$, we initialize the feature bank $\mathbf{E}$ by randomly selecting one feature embedding generated by $\mathcal{T}$ for each identity. Then, in each iteration, we first obtain the feature embeddings $\{\mathbf{f}_i^t\}_{i=1}^N$ extracted by $\mathcal{T}$, where N is the size of mini-batch, and we update the feature bank $\mathbf{E}$ by setting $\mathbf{E}[y_i] = \mathbf{f}_i^t$, where $[\cdot]$ means to obtain features from $\mathbf{E}$ based on identity y_i.

Based on the informative prototype set H_{y_i} for y_i, we can obtain K informative negative features $\mathbf{G}_i = \mathbf{E}[H_{y_i}]$, where $\mathbf{G}_i \in \mathbb{R}^{K \times d}$. Meanwhile, we denote each feature in $\mathbf{G}_i$ as $\mathbf{g}_i^k$, where $k \in \{1, ..., K\}$. Finally, a set of couples $\{(\mathbf{f}_i^s, \mathbf{g}_i^k)\}_{k=1}^K$ is constructed for $\mathbf{f}_i^s$ and we can calculate the informative SMRs using these couples. Similarly, we can also generate the informative TMRs based on a set of couples $\{(\mathbf{f}_i^t, \mathbf{g}_i^k)\}_{k=1}^K$ as the ground-truth of these SMRs.

Relation-Aware Distillation Loss. Based on the mined TMRs and SMRs, the Relation-Aware Distillation (**RAD**) loss can be easily defined as follows:

$$\mathcal{L}_{rad} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} |\cos(\mathbf{f}_i^s, \mathbf{g}_i^k) - \cos(\mathbf{f}_i^t, \mathbf{g}_i^k)|. \tag{4}$$

However, the teacher model is not always better than the student model for each case in the training dataset. As illustrated in Fig. 4 of Sec. 4.4, we observe that $\cos(\mathbf{f}_i^t, \mathbf{g}_i^k) > \cos(\mathbf{f}_i^s, \mathbf{g}_i^k)$ does exist between the mined TMRs and SMRs. Therefore, if we directly use the Eq. (4) to transfer the mutual relation knowledge of $\mathcal{T}$, $\mathcal{S}$ will be misled in some cases, which may degrade the performance of $\mathcal{S}$.

To this end, we propose only to use the valid mutual relations when $\cos(\mathbf{f}_i^t, \mathbf{g}_i^k) < \cos(\mathbf{f}_i^s, \mathbf{g}_i^k)$, and we reformulate the RAD loss of Eq. (4) as follows:

$$\mathcal{L}_{rad} = \frac{1}{N'} \sum_{i=1}^{N} \sum_{k=1}^{K} \max(\cos(\mathbf{f}_i^s, \mathbf{g}_i^k) - \cos(\mathbf{f}_i^t, \mathbf{g}_i^k), 0), \tag{5}$$

where $N^{'}$ is the number of valid mutual relations across the mini-batch. Thus, RAD loss of Eq. (5) will only affect the gradient when $\cos(\mathbf{f}_i^t, \mathbf{g}_i^k) < \cos(\mathbf{f}_i^s, \mathbf{g}_i^k)$ and transfer the accurate mutual relation knowledge from $\mathcal{T}$ to $\mathcal{S}$ in MRD.

We mine informative mutual relations for each identity. But there exists some identities, which can be easily distinguished from other identities, which indicates that the differences between the mined SMRs and TMRs for these identities are still subtle. Inspired by the hinge loss [8], we further propose a more effective variant of our RAD loss by introducing a margin q as follows:

$$\mathcal{L}_{rad} = \frac{1}{N'} \sum_{i=1}^{N} \sum_{k=1}^{K} \max(\cos(\mathbf{f}_i^s, \mathbf{g}_i^k) - \cos(\mathbf{f}_i^t, \mathbf{g}_i^k) - q, 0). \tag{6}$$

Intuitively, the RAD loss in Eq. (6) further filters out the mutual relations with subtle differences between the SMRs and TMRs and pays attention to these SMRs, which are far away from their corresponding TMRs.

3.4 Loss Function of CoupleFace

The overall loss function of CoupleFace is defined as follows:

$$\mathcal{L} = \mathcal{L}_{fcd} + \alpha \cdot \mathcal{L}_{rad} + \beta \cdot \mathcal{L}_{ce}, \tag{7}$$

where α and β are the weights of RAD loss $\mathcal{L}_{rad}$ and recognition loss $\mathcal{L}_{ce}$ (e.g., ArcFace [4]), respectively. We also provide an algorithm in Alg. 1.

4 Experiments

Datasets. For training, the mini version of Glint360K [1] named as Glint-Mini [18] is used, where Glint-Mini [18] contains 5.2M images of 91k identities. For testing, we use four datasets (i.e., IJB-B [51], IJB-C [30], and MegaFace [19]).
Experimental setting. For the pre-processing of the training data, we follow the recent works [4,20,3] to generate the normalized face crops (112×112). For teacher models, we use the widely used large neural networks (e.g., ResNet-34, ResNet-50 and ResNet-100 [13]). For student models, we use MobileNetV2 [35] and ResNet-18 [13]. For all models, the feature dimension is 512. For the training process of all models based on ArcFace loss, the initial learning rate is 0.1 and divided by 10 at the 100k, 160k, 180k iterations. The batch size and the total iteration are set as 512 and 200k, respectively. For the distillation process, the initial learning rate is 0.1 and divided by 10 at the 45k, 70k, 90k iterations. The batch size and the total iteration are set as 512 and 100k, respectively. In the informative mutual relation mining stage, we set the number of most similar identities (i.e., K) as 100. In Eq. (6), we set the margin (i.e., q) as 0.03. The loss weight α is set as 1, where β is set as 0 in the first 100k iterations, and is set as 0.01 in CoupleFace+ of Table 1. In the following experiments of different distillation methods, by default, we use the ResNet-50 (R-50), MobileNetV2 (MBNet) as $\mathcal{T}$ and $\mathcal{S}$, respectively.

Algorithm 1 CoupleFace

Input: Pre-trained teacher model $\mathcal{T}$; Randomly initialized student model $\mathcal{S}$; Current batch with N images; The dimension of feature representation d; The number of identities M; The training dataset with L images; The feature bank $\mathbf{E} \in \mathbb{R}^{M \times d}$;

Extract all features $\{\mathbf{f}_i^t\}_{i=1}^{L}$ of the dataset using $\mathcal{T}$;
Generate all identity prototypes $\{\mathbf{r}_m\}_{m=1}^{M}$ based on $\{\mathbf{f}_i^t\}_{i=1}^{L}$ according to Eq. (3);
Based on $\{\mathbf{r}_m\}_{m=1}^{M}$, calculate informative prototype set $\{H_{y_i}\}_{i=1}^{L}$ for each sample according to the label y_i;
Initialize feature bank $\mathbf{E}$ using $\{\mathbf{f}_i^t\}_{i=1}^{L}$;
for each iteration in the training process **do**
    Get features $\{\mathbf{f}_i^t\}_{i=1}^{N}$ extracted by $\mathcal{T}$ from $\{\mathbf{f}_i^t\}_{i=1}^{L}$;
    Get features $\{\mathbf{f}_i^s\}_{i=1}^{N}$ extracted by $\mathcal{S}$;
    Calculate $\mathcal{L}_{fcd}$ of $\{\mathbf{f}_i^s\}_{i=1}^{N}$ and $\{\mathbf{f}_i^t\}_{i=1}^{N}$ by Eq. (1);
    Update feature bank $\mathbf{E}$ using $\{\mathbf{f}_i^t\}_{i=1}^{N}$;
    for each feature $\mathbf{f}_i^s$ in $\{\mathbf{f}_i^s\}_{i=1}^{N}$ **do**
        Get K informative negative features $\{\mathbf{g}_i^k\}_{k=1}^{K}$ from $\mathbf{E}$ using H_{y_i};
        Build TMRs and SMRs by $\{(\mathbf{f}_i^s, \mathbf{g}_i^k)\}_{k=1}^{K}$ and $\{(\mathbf{f}_i^t, \mathbf{g}_i^k)\}_{k=1}^{K}$, respectively;
    end for
    Calculate $\mathcal{L}_{rad}$ using TMRs and SMRs by Eq. (6);
    Update the parameters of $\mathcal{S}$ by minimizing the loss function $\mathcal{L} = \mathcal{L}_{fcd} + \alpha \cdot \mathcal{L}_{rad} + \beta \cdot \mathcal{L}_{ce}$;
end for

Output: The optimized student model $\mathcal{S}$.

4.1 Results on the IJB-B and IJB-C datasets

As shown in Table 1, the first two rows represent the performance of models trained by using the ArcFace loss function [4]. We compare our method with classical KD [14], FCD, CCKD [33], SP [45], RKD [32], EC-KD [49]. For FCD, we only use the FCD loss of Eq. (1) to align the embedding space of the student and teacher models, which is a very strong baseline to improve the performance of student model for FR. For these methods (i.e., CCKD [33], SP [45] and RKD [32]), we combine these methods with FCD loss instead of the classical KD loss to achieve better performance. For EC-KD proposed for FR, we reimplement this method. In Table 1, FCD is much better than classical KD, which indicates the importance of aligning embedding space for FR when compared with classical KD. Moreover, we observe that CoupleFace achieves significant performance improvements when compared with existing methods, which demonstrates the effectiveness of CoupleFace. For the CoupleFace+, we first pretrain student by CoupleFace, and then train student by CoupleFace with ArcFace by setting β in Eq.(7) as 0.01 for another 100k iterations, better results are obtained.

4.2 Results on the MegaFace dataset

In Table 2, we also provide the results of CoupleFace on MegaFace [52], and we observe that CoupleFace is better than other methods. For example, when

Table 1. Results (TAR@FAR) on IJB-B and IJB-C of different methods.

Models	Method	IJB-B		IJB-C	
		1e-4	1e-5	1e-4	1e-5
R-50 [13]	ArcFace [4]	93.89	89.61	95.75	93.44
MBNet [35]	ArcFace [4]	85.97	75.81	88.95	82.64
MBNet [35]	KD [14]	86.12	75.99	89.03	82.69
	FCD	90.34	81.92	92.68	87.74
	CCKD [33]	90.72	83.34	93.17	89.11
	RKD [32]	90.32	82.45	92.33	88.12
	SP [45]	90.52	82.88	92.71	88.52
	EC-KD [49]	90.59	83.54	92.85	88.32
	CoupleFace	91.18	84.63	93.18	89.57
	CoupleFace+	**91.48**	**85.12**	**93.37**	**89.85**

compared with the FCD baseline, our method improves the rank-1 accuracy by 0.62% on MegaFace under the distractor size as 10^6.

Table 2. Rank-1 accuracy with different distractors on MegaFace.

Models	Method	Distractors		
		10^4	10^5	10^6
R-50 [13]	ArcFace [4]	99.40	98.98	98.33
MBNet [35]	ArcFace [4]	94.56	90.25	84.64
MBNet [35]	KD [14]	94.46	90.25	84.65
	FCD	97.81	96.39	93.65
	CCKD [33]	98.07	96.43	93.90
	RKD [32]	98.06	96.41	93.84
	SP [45]	98.01	96.58	93.95
	EC-KD [49]	98.00	96.41	93.85
	CoupleFace	**98.09**	**96.74**	**94.27**

4.3 Ablation Study

The effect of different variants of RAD loss. In MRD, we propose three variants of RAD loss (i.e., Eq. (4), Eq. (5) and Eq. (6)). To analyze the effect of different variants, we also perform additional experiments based on Eq. (4) and Eq. (5) and report the results of MBNet on IJB-B and IJB-C. In Table 3, for CoupleFace-A, we replace the RAD loss of Eq. (6) with Eq. (4), which means that we transfer the mutual relations by distilling all TMRs to the corresponding SMRs without any selection process. For CoupleFace-B, we replace the RAD loss of Eq. (6) with Eq. (5) without using the margin item. In Table 3, we observe that CoupleFace-B outperforms CoupleFace-A a lot, which shows the effectiveness of only using the mutual relations when $\cos(\mathbf{f}_i^t, \mathbf{g}_i^k) < \cos(\mathbf{f}_i^s, \mathbf{g}_i^k)$. Moreover,

Table 3. Results on IJB-B and IJB-C of different methods.

Methods	IJB-B		IJB-C	
	1e-4	1e-5	1e-4	1e-5
CoupleFace	91.18	84.63	93.18	89.57
CoupleFace-A	90.73	83.68	92.75	88.35
CoupleFace-B	90.88	84.23	93.02	88.99
CoupleFace-C	90.65	83.18	92.52	88.89
CoupleFace-D	90.85	83.73	92.86	89.04
CoupleFace-E	90.78	83.32	92.78	88.79

CoupleFace also outperforms CoupleFace-B, so it is necessary to emphasize the distillation on these SMRs, which are far from their corresponding TMRs.

The effect of Informative Mutual Relation Mining. To demonstrate the effect of mining the informative mutual relations, we further propose three alternative variants of CoupleFace (i.e., CoupleFace-C, CoupleFace-D, CoupleFace-E). Specifically, for CoupleFace-C, we propose to directly optimize mutual relation $\mathrm{R}(\mathbf{f}_i^s, \mathbf{f}_j^s)$ without the process of mining mutual relations, where only mutual relations from limited couples $\{(\mathbf{f}_i^s, \mathbf{f}_j^s)\}_{j=1, j\neq i}^{N}$ across the mini-batch are constructed. For CoupleFace-D, we randomly select $K = 100$ identities to construct the H_{y_i} for each identity y_i, while for CoupleFace-E, we propose to build the TMRs and SMRs across the mini-batch without mining process, and compute the RAD loss from them. As shown in Table 3, we observe that CoupleFace is much better than three alternative variants, which demonstrates that it is beneficial to mine the informative mutual relations in CoupleFace.

The effect of informative prototype set generation when using different models. In our work, we directly use $\mathcal{T}$ (i.e., R-50) to generate the informative prototype set H_{y_i} for each identity y_i. Here, we propose to use other pre-trained models (i.e., **MBNet** and ResNet-100 (**R-100**)) trained by ArcFace loss to generate H_{y_i} for y_i, and we call these alternative methods as CoupleFace-MBN and CoupleFace-RN100, respectively. In Table 4, we report the results of MBNet on IJB-B and IJB-C after using CoupleFace, CoupleFace-MBN and CoupleFace-RN100. We observe that CoupleFace achieves comparable performance with CoupleFace-RN100, and higher performance than CoupleFace-MBN. For this phenomenon, we assume that when using a more effective model, we will generate more discriminative identity prototypes $\{\mathbf{r}_m\}_{m=1}^{M}$, which leads to generating a more accurate informative prototype set H_{y_i}. Thus, it is beneficial to use effective models for obtaining the informative prototype set.

4.4 Further Analysis

Visualization on the differences of SMRs and TMRs. To further analyze the effect of CoupleFace, we visualize the distributions of the differences between SMRs and TMRs for both FCD and CoupleFace in Fig. 4. Specifically, we use the models of MBNet in the 20,000th, 60,000th, 100,000th iterations for both

Table 4. Results on IJB-B and IJB-C of different methods.

Methods	IJB-B		IJB-C	
	1e-4	1e-5	1e-4	1e-5
CoupleFace	91.18	84.63	93.18	89.57
CoupleFace-MBN	91.06	83.91	92.95	88.94
CoupleFace-RN100	91.15	84.65	93.16	89.58

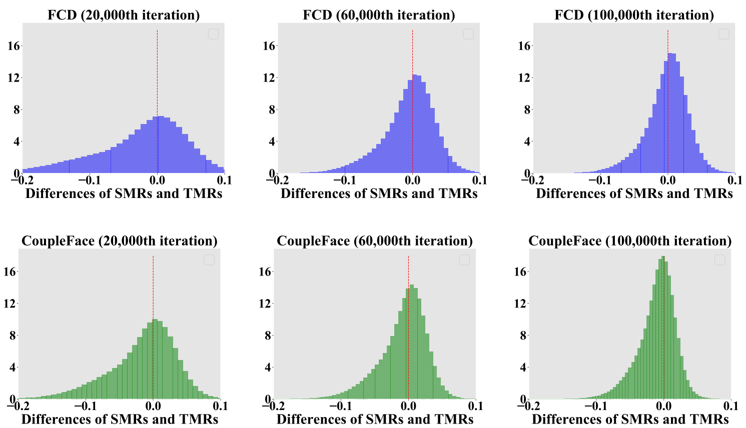

Fig. 4. The distributions of differences between the SMRs and TMRs of different iterations for FCD and CoupleFace.

FCD and CoupleFace. The first and the second rows show the results of FCD and CoupleFace, respectively. In Fig. 4, during training, for CoupleFace, we observe that the differences of SMRs and TMRs gradually decrease at the right side of the red line, which demonstrates the effect of minimizing the RAD loss of Eq. (6). Besides, when compared with FCD, most SMRs are less than or approximate to TMRs in CoupleFace, which indicates that CoupleFace transfers the mutual relation knowledge of the teacher model to the student model well.

Visualization on the distributions of similarity scores. We visualize the distributions of similarity scores on IJB-C of **MBNet** based on different methods in Fig. 5. Specifically, we still use the **R-50** as $\mathcal{T}$ to distill **MBNet** in FCD and CoupleFace. As shown in Fig. 5, when compared with FCD, the similarity distributions of positive pairs and negative pairs in CoupleFace are more compact and separable, which further shows the effectiveness of CoupleFace.

Comparison with existing relation-based knowledge distillation methods (RB-KDs). The differences between CoupleFace and existing RB-KDs (e.g., CCKD, RKD [33,32]) are as follows. (1) General KDs are usually incompatible with FR. Existing RB-KDs are proposed for general vision tasks (e.g., close-set classification). In contrast, it is non-trivial to transfer relation knowledge for open-set FR well and CoupleFace is well-designed for FR distillation.

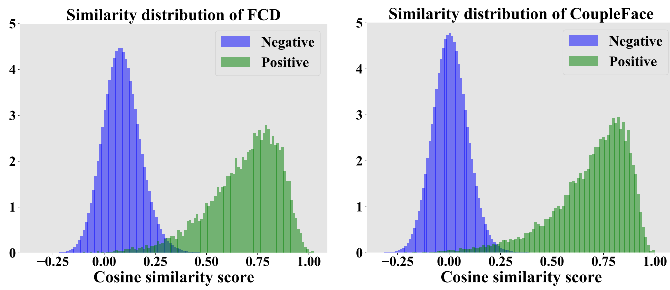

Fig. 5. Cosine similarity distributions of the positive pairs and negative pairs.

(2) Mining is considered. We observe that most mutual relations cannot provide valuable knowledge and affect the similarity distribution for FR, so how to produce sufficient informative mutual relations efficiently is a challenging issue. In CoupleFace, we propose to mine informative mutual relations in MRD, while existing RB-KDs have not discussed the mining process. (3) Loss function is intrinsically different. The RAD loss in Eq. (6) aims to better exploit the mutual relation knowledge by using valid mutual relations and filtering out the mutual relations with subtle differences, which are not discussed in existing works. (4) Better performance. CoupleFace outperforms existing RB-KDs a lot.
Computation costs. When compared with FCD, training time and GPU memory use of CoupleFace are 1.056 times and 1.002 times, respectively. which further demonstrates the efficiency of our proposed CoupleFace.

5 Conclusion

In our work, we investigate the importance of mutual relation knowledge for FR distillation and propose an effective FR distillation method named as CoupleFace. When compared with existing methods using Feature Consistency Distillation (FCD), CoupleFace further introduces the Mutual Relation Distillation (MRD), where we propose to mine the informative mutual relations and utilize the Relation-Aware Distillation (RAD) loss to transfer the mutual relation knowledge from the teacher model to the student model. Extensive experiments on multiple FR benchmark datasets demonstrate the effectiveness of CoupleFace. In our future work, we will continue to explore what kind of information is important for FR distillation and develop more effective distillation methods.

6 Acknowledgments

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

References

1. 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)
2. David, S., Sergey, A.: Margindistillation: Distillation for face recognition neural networks with margin-based softmax. International Journal of Computer and Information Engineering **15**(3), 206–210 (2021)
3. 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)
4. 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)
5. 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)
6. Fang, Z., Wang, J., Wang, L., Zhang, L., Yang, Y., Liu, Z.: {SEED}: Self-supervised distillation for visual representation. In: International Conference on Learning Representations (2021)
7. Feng, Y., Wang, H., Hu, H.R., Yu, L., Wang, W., Wang, S.: Triplet distillation for deep face recognition. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 808–812. IEEE (2020)
8. Gentile, C., Warmuth, M.K.: Linear hinge loss and average margin. Advances in neural information processing systems **11**, 225–231 (1998)
9. Guo, J., Liu, J., Xu, D.: Jointpruning: Pruning networks along multiple dimensions for efficient point cloud processing. IEEE Transactions on Circuits and Systems for Video Technology **32**(6), 3659–3672 (2021)
10. Guo, J., Ouyang, W., Xu, D.: Multi-dimensional pruning: A unified framework for model compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1508–1517 (2020)
11. Harwood, B., Kumar BG, V., Carneiro, G., Reid, I., Drummond, T.: Smart mining for deep metric learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2821–2829 (2017)
12. 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)
13. 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)
14. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
15. Huang, Y., Shen, P., Tai, Y., Li, S., Liu, X., Li, J., Huang, F., Ji, R.: Improving face recognition from hard samples via distribution distillation loss. In: European Conference on Computer Vision. pp. 138–154. Springer (2020)
16. 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)

17. Huang, Y., Wu, J., Xu, X., Ding, S.: Evaluation-oriented knowledge distillation for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18740–18749 (June 2022)
18. InsightFace: Glint-mini face recognition dataset. [online]. https://github.com/deepinsight/insightface/tree/master/recognition/_datasets_ (2021)
19. 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)
20. Kim, Y., Park, W., Shin, J.: Broadface: Looking at tens of thousands of people at once for face recognition. ECCV (2020)
21. 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)
22. Liu, J., Guo, J., Xu, D.: Apsnet: Towards adaptive point sampling for efficient 3d action recognition. IEEE Transactions on Image Processing (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., Yu, T., Peng, H., Sun, M., Li, P.: Cross-lingual cross-modal consolidation for effective multilingual video corpus moment retrieval. In: NAACL-HLT (2022)
26. Liu, J., Yu, Z., Qin, H., Wu, Y., Liang, D., Zhao, G., Xu, K.: Oneface: One threshold for all. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
27. 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)
28. 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)
29. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: ICML. vol. 2, p. 7 (2016)
30. 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)
31. 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)
32. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3967–3976 (2019)
33. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: ICCV (October 2019)
34. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
35. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)

36. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. pp. 815–823 (2015)
37. Shi, W., Ren, G., Chen, Y., Yan, S.: Proxylesskd: Direct knowledge distillation with inherited classifier for face recognition. arXiv preprint arXiv:2011.00265 (2020)
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
39. Suh, Y., Han, B., Kim, W., Lee, K.M.: Stochastic class-based hard example mining for deep metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7251–7259 (2019)
40. 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)
41. 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)
42. 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)
43. 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)
44. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: ICLR (2020)
45. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
46. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Processing Letters **25**(7), 926–930 (2018)
47. 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)
48. 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)
49. Wang, X., Fu, T., Liao, S., Wang, S., Lei, Z., Mei, T.: Exclusivity-consistency regularized knowledge distillation for face recognition. In: ECCV. pp. 325–342. Springer (2020)
50. 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)
51. 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)
52. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. IEEE (2011)
53. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2840–2848 (2017)

54. Yang, C., An, Z., Cai, L., Xu, Y.: Hierarchical self-supervised augmented knowledge distillation. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. pp. 1217–1223 (2021)
55. Yang, C., An, Z., Cai, L., Xu, Y.: Knowledge distillation using hierarchical self-supervision augmented distribution. IEEE Transactions on Neural Networks and Learning Systems (2022)
56. Yang, C., Zhou, H., An, Z., Jiang, X., Xu, Y., Zhang, Q.: Cross-image relational knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12319–12328 (2022)
57. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
58. 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)
59. 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)