ICD-Face: Intra-class Compactness Distillation for Face Recognition

Zhipeng Yu[1*] Jiaheng Liu[2*] ✉ Haoyu Qin[3] Yichao Wu[3] Kun Hu[3]
Jiayi Tian[2] Ding Liang[3]
[1]SEECE, UCAS [2]Beihang University [3]SenseTime Research

Abstract

Knowledge distillation is an effective model compression method to improve the performance of a lightweight student model by transferring the knowledge of a well-performed teacher model, which has been widely adopted in many computer vision tasks, including face recognition (FR). 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. However, after using FCD, we observe that the intra-class similarities of the student model are lower than the intra-class similarities of the teacher model a lot. Therefore, we propose an effective FR distillation method called ICD-Face by introducing intra-class compactness distillation into the existing distillation framework. Specifically, in ICD-Face, we first propose to calculate the similarity distributions of the teacher and student models, where the feature banks are introduced to construct sufficient and high-quality positive pairs. Then, we estimate the probability distributions of the teacher and student models and introduce the Similarity Distribution Consistency (SDC) loss to improve the intra-class compactness of the student model. Extensive experimental results on multiple benchmark datasets demonstrate the effectiveness of our proposed ICD-Face for face recognition.

1. Introduction

Face recognition (FR) has been well-investigated for many years. Most of the progress is credited to large-scale training datasets [64, 22], resource-intensive networks with millions of parameters [15, 43] and effective loss functions [7, 52]. Although larger FR models usually exhibit better recognition performance, the requirements for huge computational resources are usually prohibitive on mobile and embedded devices. Therefore, how to develop lightweight and effective FR models in resource-limited scenarios has become an emergency problem in recent years. Knowledge distillation [16] is a popular strategy for compressing models, which transfers the "knowledge" from the teacher model to the lightweight student model.

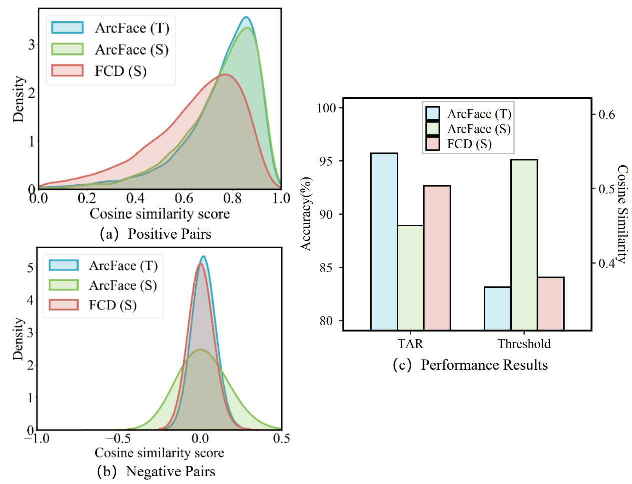


Figure 1: (a) The similarity distributions of positive pairs based on different methods. "ArcFace (T)" and "ArcFace (S)" mean teacher and student models are trained by ArcFace [7]. "FCD (S)" denotes the student model is trained by FCD. (b) The similarity distributions of negative pairs based on different methods. (c) True Accept Rate (TAR) and threshold results of the teacher and student models using different methods.

Most existing knowledge distillation methods usually aim to guide the student to mimic the teacher's behavior by introducing probability constraints (e.g., KL divergence [16]) between the teacher's predictions and student models. However, for FR, the performance is usually evaluated in an open-set setting, where the identities of the testing set are disjoint from the training set. Meanwhile, at the testing phase, the similarities of feature embedding are employed for FR instead of the probabilities of the classifier used in classical close-set classification. Therefore, it is more important to improve the discriminative ability of the feature embedding of the student model for FR. A

* Equal contribution.
✉ Corresponding author: Jiaheng Liu (liujiaheng@buaa.edu.cn).

straightforward and effective FR distillation method is to directly minimize the L_2 distance of the feature embeddings extracted by the teacher and student models [54, 42, 26], which is called Feature Consistency Distillation (FCD). FCD enables the student model to share the same embedding space with the teacher model for similarity comparison, and FCD has been widely used in practice to improve the performance of the lightweight neural networks for FR.

However, we observe that it is unfeasible for student models with low capacities to align the feature space with the teacher model well. As shown in Fig. 1, we use the ResNet-50 and MobileNetV2 as teacher and student models, respectively, and report the performance results of these models using different losses on the IJB-C [56] dataset. Specifically, after using FCD, as shown in Fig. 1(a), when compared with the similarity distribution of the positive pairs using the teacher model, we observe that the similarities of positive pairs using the student model decrease a lot, which means that FCD reduces the intra-class compactness a lot. Meanwhile, in Fig. 1(b), the similarity distributions of the negative pairs between "FCD (S)" and "ArcFace (T)" are very close, which indicates that the student model can maintain inter-class discrepancy well after using FCD. Besides, in Fig. 1(c), we observe that the widely-used FR evaluation metric True Accept Rate (TAR) of the student model will be significantly improved after using FCD, and the similarity threshold value under the False Accept Rate (FAR) of the student model decreases a lot and is close to the threshold calculated by the teacher model, which also represents that the similarities of the negative pairs are effectively decreased to a similar degree with the teacher model and the FCD method can effectively help students to learn the inter-class distribution of the teacher model.

Therefore, FCD cannot preserve the intra-class compactness of the student model well, and it is critical to align the similarity distributions of the positive pairs between the teacher and student models.

Motivated by the aforementioned analysis, we propose a new FR distillation framework (**ICD-Face**), which includes FCD and Intra-class Compactness Distillation (**ICD**). The ICD aims to improve the similarity distribution consistency between the teacher and student models. 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 calculate the FCD loss for aligning the embedding spaces of the teacher and student models. In ICD, as it is important to generate sufficient positive pairs to estimate accurate similarity distribution for FR models, inspired by MoCo [14], we propose to construct the teacher and student feature banks and generate the positive pairs using the features from the feature banks and the features from the current batch in our Similarity Distribution Generation module. After that, we estimate the probability distributions of the teacher and student models, and introduce the Similarity Distribution Consistency (SDC) loss to directly align the similarity distributions of the positive pairs between the teacher and student models in the training process.

The contributions of our ICD-Face are as follows:

- In our work, we first investigate the gap of the intra-class similarity distributions between the teacher and student models, and propose a new FR distillation method called as ICD-Face, which additionally introduces Intra-class Compactness Distillation (ICD) into the existing FR distillation method.
- In ICD, we first propose to generate sufficient positive pairs for estimating intra-class similarity distributions of the teacher and student models, and then utilize the similarity distribution consistency loss to align the intra-class similarity distributions between FR models.
- Extensive experiments on multiple benchmark datasets demonstrate the effectiveness and generalization ability of our proposed ICD-Face method.

2. Related work

Knowledge Distillation. Knowledge distillation is a representative method of model compression and acceleration [12, 25, 13], which aims to transfer knowledge from a powerful teacher model trained on a task to a lightweight student model [16]. It has been used in many computer vision tasks [34, 39, 3, 59, 38, 48, 37, 17, 5, 9, 10, 21, 20, 29, 11, 2]. Different kinds of representation have been used as knowledge for better performance by various distillation methods. For instance, FitNet [39] guides the student model training with middle-level hints from hidden layers of the teacher model. CRD [48] uses a contrastive-based objective function for knowledge transfer between deep networks. Some relation-based knowledge distillation methods (e.g., CCKD [38], RKD [37]) improve the student model with relation knowledge. Recently, knowledge distillation has also been applied to enhance the performance of lightweight network (e.g., MobileNetV2 [40]) for FR. For example, EC-KD [54] proposes a position-aware exclusivity strategy to encourage diversity among different filters of the same layer to alleviate the low capability of student models. PACKD [60] first discuss the effect of positive pairs in KD for classification tasks. In contrast to existing methods, our proposed ICD-Face method introduces the intra-class compactness distillation to improve the performance of the student model, which is well-designed for FR distillation.

Face Recognition. Existing FR methods aim to maximize the inter-class discriminability and the intra-class compactness of feature representations. The success of FR depends

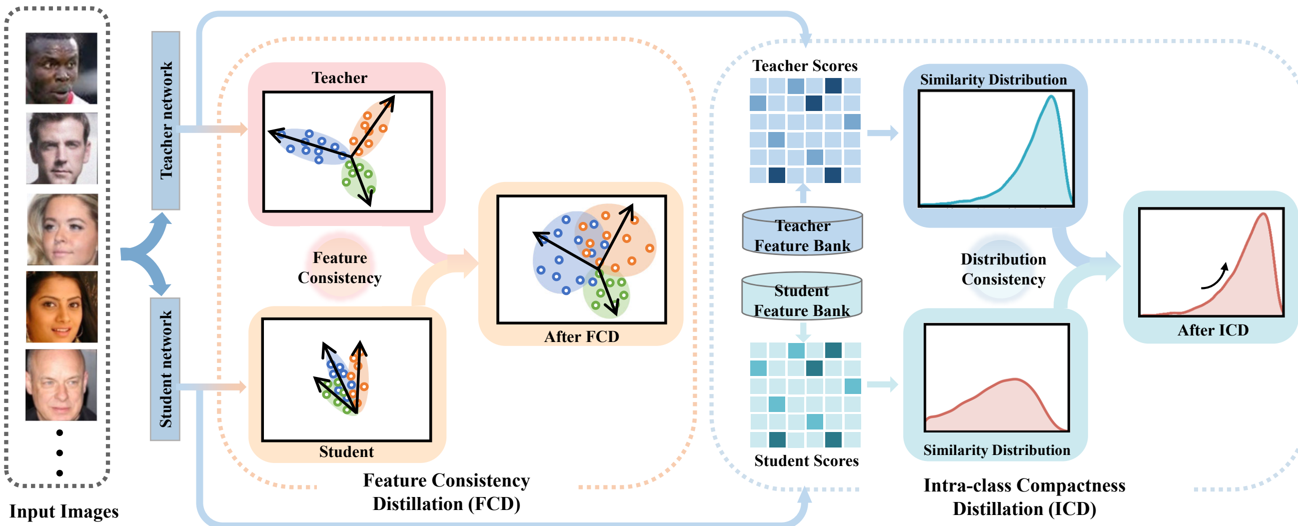


Figure 2: The framework of our proposed ICD-Face for FR distillation. In FCD, we use the Feature Consistency Distillation loss $\mathcal{L}_{fcd}$ between $\mathbf{f}_i^s$ and $\mathbf{f}_i^t$ to align the embedding spaces of the teacher and student models. In ICD, we first generate the teacher and student intra-class similarity distributions using the feature banks for teacher and student models, respectively. Then, we minimize the Similarity Distribution Consistency loss $\mathcal{L}_{sdc}$ loss to align the intra-class similarity distributions between the teacher and student models.

on three factors: large-scale datasets [64, 1, 19, 58, 22], effective deep neural networks [47, 44, 43, 46, 31] and well-designed loss functions [28, 52, 62, 32, 45, 51, 7, 6, 36, 8, 26, 27, 24, 30]. The loss function design is the main research direction for FR, which investigates improving the generalization and discriminability of the learned feature representations. For example, Triplet loss [41] enlarges the distances of negative pairs and reduces the distances of positive pairs. Recently, many angular-based loss functions [33, 32] introduce the angular constraint into the cross-entropy loss function. Besides, CosFace [53] and ArcFace [7] further use a margin term for better discriminability of feature representations. Moreover, some mining-based loss functions (e.g., MV-Arc-Softmax [55] and CurricularFace [18]) consider the difficulty degree of samples and achieve promising results. The recent work VPL [8] also introduces sample-to-sample comparisons to reduce the gap between training and evaluation processes for FR. When compared to the aforementioned existing methods, we propose to design an effective distillation loss function to improve the intra-class compactness of the lightweight neural network for FR.

3. Method

In this section, we introduce the details of our ICD-Face in Fig. 2, which contains Feature Consistency Distillation (FCD) and Intra-class Compactness Distillation (ICD) 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 the 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 Consistency Distillation (FCD) loss $\mathcal{L}_{fcd}$ based on L_2 distance of the feature embeddings. Meanwhile, in ICD, we first generate the positive pairs based on the features extracted from the current batch and the stored features of the feature banks. Then, we propose to calculate the similarity distribution consistency loss to improve the intra-class compactness of the similarity distribution for the student model.

3.1. Preliminaries

We define some notations in ICD-Face, and discuss the necessity of FR distillation.

Notations. The teacher model and student model are denoted as $\mathcal{T}$ and $\mathcal{S}$, respectively. For each sample x_i, the corresponding identity label is y_i, and the features extracted by $\mathcal{T}$ and $\mathcal{S}$ are denoted as $\mathbf{f}_i^t$ and $\mathbf{f}_i^s$, respectively.

Discussion on Face Recognition Distillation. Most existing knowledge distillation methods are usually proposed for image classification, which utilizing the probability consistency [16] (e.g., KL divergence) to align the prediction probabilities from $\mathcal{S}$ with the prediction probabilities from $\mathcal{T}$. However, these 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 [41]). Moreover, FR models are trained to generate discriminative feature embeddings for similarity comparisons 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. In ICD-Face, FCD is also used to improve the performance of the student model $\mathcal{S}$ for FR distillation.

3.3. Intra-class Compactness Distillation

In this section, we describe the Intra-class Compactness Distillation (ICD) of ICD-Face in detail. First, we describe how to generate sufficient positive pairs based on the feature banks in Similarity Distribution Generation. Then, we introduce the Similarity Distribution Consistency (SDC) loss to transfer the knowledge from positive pairs to the student model.

3.3.1 Similarity Distribution Generation

As it is challenging to construct sufficient positive pairs in each mini-batch for both teacher and student models, we first propose to construct feature banks for teacher and student models and then produce positive pairs using the feature bank and the features of the current mini-batch.

Inspired by MoCo [14] for unsupervised learning, which stores the features from the previous mini-batches to generate sufficient negative samples, we propose to maintain a teacher feature bank $\mathbf{M}^t \in \mathbb{R}^{Q\times K\times d}$ and a student feature bank $\mathbf{M}^s \in \mathbb{R}^{Q\times K\times d}$ in Fig. 2, where Q is the number of identities of the training dataset, K is the maximum number of features for each identity, and d represents the feature dimension of each face image.

As shown in Fig. 3, the procedure of generating similarity distribution for the teacher or student model is similar. Thus, we take the student feature bank $\mathbf{M}^s$ as an example to show the details of constructing a feature bank. Specifically, in each iteration, we first update the student feature bank $\mathbf{M}^s$ by pushing the features extracted by the student model of the current batch into $\mathbf{M}^s$, and then utilize the

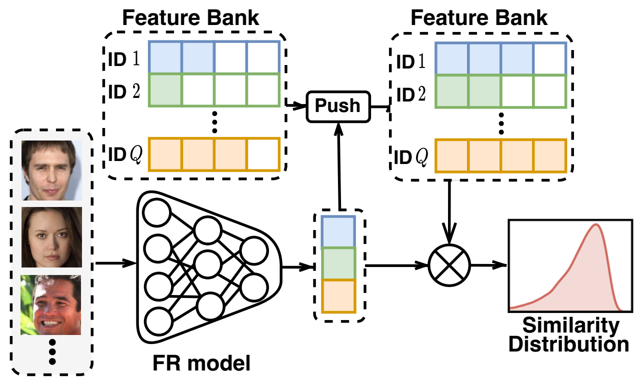


Figure 3: The procedure of generating similarity distribution using feature bank. "ID" means "Identity". Q is the number of identities. "FR model" can represent the teacher model $\mathcal{T}$ or the student model $\mathcal{S}$, and "Feature Bank" can denote the teacher feature bank $\mathbf{M}^t$ or the student feature bank $\mathbf{M}^s$.

stored features of $\mathbf{M}^s$ to construct the positive pairs with the features of the current batch.

Meanwhile, as discussed in VPL [8], features drift slowly for FR models, which indicates that features extracted previously can be considered as an approximation of the output of the current network within a certain number of training steps. Therefore, we also create a validness indicator $\mathbf{V} \in \mathbb{R}^{Q\times K}$ to represent the validness of each feature in the student feature bank $\mathbf{M}^s$. Each item in $\mathbf{V}$ is a scalar value, which denotes the remaining valid steps for the corresponding feature in the feature bank. The maximum valid step for each feature is U, and we initialize all items of $\mathbf{V}$ as 0 at the beginning of the training process. In each iteration, in ICD-Face, we first extract the features $\{\mathbf{f}_i^s\}_{i=1}^N$ of the current batch, where N is the batch size, and y_i is the corresponding label of $\mathbf{f}_i^s$. Then, we update the student feature bank $\mathbf{M}^s$ based on $\{\mathbf{f}_i^s\}_{i=1}^N$. Specifically, for the i-th feature $\mathbf{f}_i^s$, when the number of the stored features for the corresponding identity y_i is smaller than K, we insert $\mathbf{f}_i^s$ into $\mathbf{M}^s$ based on the identity y_i. When the number of the stored features is equal to K for identity y_i, we first find out the index idx_i of the most oldest feature in $\mathbf{M}^s[y_i]$, which is also the index of the smallest value in $\mathbf{V}[y_i]$. Then, we replace the oldest feature with the newly extracted feature $\mathbf{f}_i^s$ based on the index idx_i, which means we set $\mathbf{M}^s[y_i][idx_i] = \mathbf{f}_i^s$. After that, we set the number of valid steps for $\mathbf{f}_i^s$ as U, which means we set $\mathbf{V}[y_i][idx_i] = U$. After each training step, $\mathbf{V}$ is updated by $\mathbf{V} = \mathbf{V} - 1$, which decreases the valid steps of all stored features in $\mathbf{M}^s$.

After the updating process for the student feature bank $\mathbf{M}^s$, the number of valid features in the student feature bank $\mathbf{M}^s$ is $\sum_{i=1}^{Q}\sum_{j=1}^{K}\mathbb{1}(\mathbf{V}[i][j] > 0)$, where $\mathbb{1}(x)$ is the indicator function. Then, for each feature $\mathbf{f}_i^s$ in the current batch, the positive pairs are constructed by using $\mathbf{f}_i^s$ and the stored valid features of the corresponding identity y_i in the student feature bank $\mathbf{M}^s$. Then, we compute the cosine similarity of the positive pairs to generate the student similarity distribution $\mathcal{S}_D = \{s_i\}_{i=1}^G$, where G is the number of positive pairs and s_i is the similarity of the i-th positive pair.

Similarly, in Fig. 3, we can also generate teacher similarity distribution $\mathcal{T}_D = \{t_i\}_{i=1}^G$ based on the aforementioned scheme by replacing the student model and student feature bank with the teacher model and teacher feature bank. Note that t_i is the similarity of the i-th positive pair.

3.3.2 Similarity Distribution Consistency

After obtaining the similarity distributions for teacher and student models, we define the following similarity distribution consistency loss.

Let P^t and P^s denote the two probability distributions of teacher model $\mathcal{T}$ and student model $\mathcal{S}$, respectively. $\mathcal{T}_D$ and $\mathcal{S}_D$ can be regarded as as similarities of all positive pairs, which denotes the sample set of P^t and P^s. Given finite data samples, we can use existing statistical methods to estimate P^t and P^s. It is critical to accurately estimate P^s and P^t, and the estimated distribution needs to be differentiable.

For cosine distance-based methods [7], these distribution of similarity score are one-dimensional and bounded to is bounded to [-1, 1], which is shown to simplify the task [50]. Thus, we first estimate this type of one-dimensional distribution by fitting simple histograms with uniformly spaced bins, and we adopt R-dimensional histograms H^T and H^S with the nodes $n_1 = -1, n_2, \cdots, n_R = 1$ uniformly filling $[-1, 1]$ with the step $\Delta = \frac{2}{R-1}$. Then, for the student model $\mathcal{S}$, we estimate the value h_r^s of the histogram H^S at the r-th node as:

$$h_r^s = \frac{1}{|\mathcal{S}_D|} \sum_i^G \delta_{i,r}, \tag{2}$$

where $r \in \{1, 2, ..., R\}$, and the weights $\delta_{i,r}$ are based on an exponential function as follows:

$$\delta_{i,r} = exp(-\gamma(s_i - n_r)^2), \tag{3}$$

where γ denotes the spread parameter of the Gaussian kernel function [17], and n_r denotes the r-th node of histograms. We adopt the Gaussian kernel function as it is the most commonly used kernel function for density estimation. Compared to other non-continuous or discrete surrogate functions, the continuous kernel function can prevent probabilities from being 0 even if the samples do not appear in r-th bin. Then the estimated P^s can be calculated by a simple normalization function $P^s = \frac{H^S}{\sum(H^S)}$. The estimation of P^t proceeds analogously.

To narrow the distribution gap between the teacher and student models, we constrain the student distribution P^s to approximate the teacher distribution P^t. Motivated by the previous KD methods [63, 16], we adopt the KL divergence $\mathbb{D}_{KL}$ to constrain the similarity distributions between the student and teacher models, which is defined as the following Similarity Distribution Consistency (SDC) loss:

$$\mathcal{L}_{sdc} = \mathbb{D}_{KL}(P^t || P^s). \tag{4}$$

3.4. Loss Function of ICD-Face

The overall loss function of ICD-Face is as follows:

$$\mathcal{L} = \mathcal{L}_{fcd} + \alpha \cdot \mathcal{L}_{sdc}, \tag{5}$$

where α is the loss weight for the ICD loss $\mathcal{L}_{sdc}$. Meanwhile, we can also add the classification loss $\mathcal{L}_{cls}$ (e.g., ArcFace [7]) as follows:

$$\mathcal{L} = \mathcal{L}_{fcd} + \alpha \cdot \mathcal{L}_{sdc} + \beta \cdot \mathcal{L}_{cls}, \tag{6}$$

where we call ICD-Face with $\mathcal{L}_{cls}$ as ICD-Face+.

For better clarification, we also provide an algorithm of our proposed ICD-Face in Alg. 1.

3.5. Discussion

Differences between PACKD and ICD-Face. Both PACKD and ICD-Face discuss the effectiveness of positive pairs in knowledge distillation. The differences between PACKD and ICD-Face are as follows: (1). Motivation is different. PACKD is used for close-set classification and aims to transfer inter-class and intra-class relation knowledge, where the quality of the classifier is important. In contrast, ICD-Face is well-designed for open-set FR. For the student model, increasing the discrepancy of similarity distributions between positive and negative pairs is more important for FR distillation. (2). Method is also different a lot. PACKD uses a mixup augmentation policy to obtain extra positive samples. In contrast, we construct sufficient positive pairs by using a feature bank and mixup is not used. Meanwhile, PACKD uses an optimal transport strategy to adjust weights of the pair-wise losses for different positive pairs, while we use SDC loss to directly align the positive score distributions of teacher and student.

4. Experiments

Datasets. In our ICD-Face, we evaluate our proposed ICD-Face method on the following benchmark datasets.

- IARPA Janus Benchmark (IJB) [56, 35] is designed to evaluate the performance of unconstrained face recognition, which is a challenging template-based benchmark. IJB-B [56] has 67K face images, 7K face videos

Algorithm 1 ICD-Face

Input: Pre-trained teacher model $\mathcal{T}$; Randomly initialized student model $\mathcal{S}$; Current batch with N images; The feature dimension d; The maximum number of features for each identity K; The number of identities Q; The student feature bank $\mathbf{M}^s \in \mathbb{R}^{Q\times K\times d}$ and teacher feature bank $\mathbf{M}^t \in \mathbb{R}^{Q\times K\times d}$; The validness indicator $\mathbf{V} \in \mathbb{R}^{Q\times K}$ for the validness of the stored features in $\mathbf{M}^s$ and $\mathbf{M}^t$; The maximum valid steps U;

Zero initialize $\mathbf{V}$;
for each iteration in the training process **do**
  Get features $\{\mathbf{f}_i^t\}_{i=1}^N$ and $\{\mathbf{f}_i^s\}_{i=1}^N$ by $\mathcal{T}$ and $\mathcal{S}$;
  Calculate $\mathcal{L}_{fcd}$ of $\{\mathbf{f}_i^s\}_{i=1}^N$ and $\{\mathbf{f}_i^t\}_{i=1}^N$ by Eq. (1);
  for features $\mathbf{f}_i^t$ and $\mathbf{f}_i^s$ in $\{\mathbf{f}_i^t\}_{i=1}^N$ and $\{\mathbf{f}_i^s\}_{i=1}^N$ **do**
    Select the index idx_i to insert $\mathbf{f}_i^t$ and $\mathbf{f}_i^s$ into $\mathbf{M}^t$ and $\mathbf{M}^s$, respectively;
    $\mathbf{M}^t[y_i][idx_i] = \mathbf{f}_i^t$, $\mathbf{M}^s[y_i][idx_i] = \mathbf{f}_i^s$;
    $\mathbf{V}[y_i][idx_i] = U$;
  end for
  $\mathbf{V} = \mathbf{V} - 1$;
  Construct the positive pairs using $\{\mathbf{f}_i^t\}_{i=1}^N$ and the valid features $\mathbf{M}^t[\mathbf{V} > 0]$;
  Construct the positive pairs using $\{\mathbf{f}_i^s\}_{i=1}^N$ and the valid features $\mathbf{M}^s[\mathbf{V} > 0]$;
  Calculate the similarity scores of the positive pairs for teacher and student, respectively;
  Calculate SDC loss $\mathcal{L}_{sdc}$ based on Eq. (4);
  Calculate $\mathcal{L}_{cls}$ based on $\{\mathbf{f}_i^s\}_{i=1}^N$;
  Update parameters of $\mathcal{S}$ by $\mathcal{L}$ in Eq. (5) or Eq. (6);
end for

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

and 10K non-face images. When compared with IJB-B, IJB-C [35] contains new individuals with increased occlusion and diversity of geographic origin and is composed of 138K face images, 11K face videos and 10K non-face images. For IJB-B and IJB-C, we report the TAR results under the FAR of 1e-4 and 1e-5.

- MegaFace [22] contains more than 1M images from 690K identities to evaluate the face recognition performance. For MegaFace, we report the identification accuracy results under different distractors.

For training, the mini version of Glint360K [1] named as Glint-Mini [19] is used. Glint-Mini [19] contains 5.2M images of 91K identities. For testing, we use four benchmark datasets (i.e., IJB-B [56], IJB-C [35], and MegaFace [22]).
Experimental setting. For the pre-processing of the training data, we follow the recent works [7, 23, 6] 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 [15]). For student models, we use MobileNetV2 [40] and MobileFaceNet [4]. 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, and 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 90k, 140k, 180k iterations. Note that we first pretrain the student model only using FCD loss $\mathcal{L}_{fcd}$ for 50k iterations, and then add the external SDC loss $\mathcal{L}_{sdc}$. The loss weights α and β are set as 0.5 and 0.1, respectively. For the feature bank, the maximum number of features for each identity K and the maximum number of valid steps U are set as 5 and 200, respectively. For the similarity distribution consistency, following [17], the Δ and γ in Eq. (3) are set as 0.001 and 50, respectively. The batch size and the total iteration are set as 512 and 200k, respectively. In the following experiments, by default, we use the ResNet-50 (**R-50**), MobileNetV2 (**MBNet**) as $\mathcal{T}$ and $\mathcal{S}$, respectively, and use Glint-Mini [19] as the training dataset to achieve competitive results and reduce the GPU resource consumption.

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 [7]. We compare our method with classical KD [16], FCD, AT [61], CCKD [38], SP [49], RKD [37], EC-KD [54], and CoupleFace [26]. 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 the student model for FR. For these methods (i.e., AT [61], CCKD [38], SP [49] and RKD [37]), we follow CoupleFace [26] to combine these methods with FCD loss instead of the classical KD loss for better performance. 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 ICD-Face achieves significant performance improvements when compared with existing methods. Besides, our ICD-Face+ with ArcFace loss can further improve the results of ICD-Face, which demonstrates the effectiveness of our proposed method.

4.2. Results on the MegaFace dataset

In Table 2, we provide the results on MegaFace [57], and we observe that ICD-Face is better than other methods. For example, when compared with the state-of-the-art CoupleFace, our method improves the rank-1 accuracy by 0.28% on MegaFace under the distractor size as 10^6.

4.3. Ablation study

The effect of the hyper-parameters in the feature banks. To investigate the performance variation of our method with

Table 1: Results on IJB-B and IJB-C of different methods.

Models	Method	IJB-B(TAR@FAR)		IJB-C(TAR@FAR)	
		1e-4	1e-5	1e-4	1e-5
R-50 [15]	ArcFace [7]	93.89	89.61	95.75	93.44
MBNet [40]	ArcFace [7]	85.97	75.81	88.95	82.64
MBNet [40]	KD [16]	86.12	75.99	89.03	82.69
	FCD	90.34	81.92	92.68	87.74
	AT [61]	90.35	82.22	92.65	87.54
	CCKD [38]	90.72	83.34	93.17	89.11
	RKD [37]	90.32	82.45	92.33	88.12
	SP [49]	90.52	82.88	92.71	88.52
	EC-KD [54]	90.59	83.54	92.85	88.32
	CoupleFace [26]	91.18	84.63	93.18	89.57
	ICD-Face	91.46	85.32	93.57	89.90
	ICD-Face+	**91.66**	**85.42**	**93.81**	**90.10**

Table 2: Rank-1 accuracy with different distractors.

Models	Method	Distractors	
		10^5	10^6
R-50 [15]	ArcFace [7]	98.98	98.33
MBNet [40]	ArcFace [7]	90.25	84.64
MBNet [40]	KD [16]	90.25	84.65
	FCD	96.39	93.65
	AT [61]	96.50	93.68
	CCKD [38]	96.43	93.90
	RKD [37]	96.41	93.84
	SP [49]	96.58	93.95
	EC-KD [54]	96.41	93.85
	CoupleFace [26]	96.74	94.27
	ICD-Face	96.87	94.55
	ICD-Face+	**96.90**	**94.58**

Table 3: TAR(@FAR=1e-4) on IJB-C of ICD-Face.

K	1	3	5	7
TAR (%)	92.73	93.39	93.57	93.59

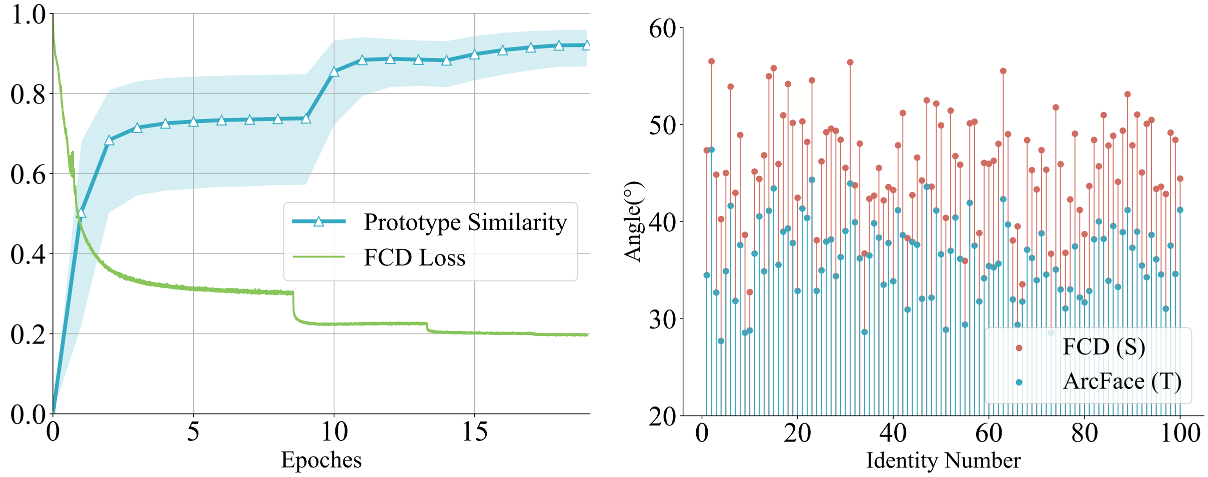
Table 4: TAR(@FAR=1e-4) on IJB-C of ICD-Face.

U	50	100	200	500
TAR (%)	93.23	93.45	93.57	93.54

respect to the hyper-parameters of the feature banks (i.e., the maximum number of features for each identity K and the maximum number of valid steps U), we perform ICD-Face using different values of K and U, and reporting the results of MBNet on IJB-C dataset. In Table 3, we set U as 200, and use different values of K. When K increases from 1 to 5, our method achieves better performance. We suppose that when K is larger, more positive pairs are generated, which leads to more accurate similarity distribution estimation. However, when we continue to increase the value of K, the improvement of performance becomes relatively stable and the GPU memory usage also increases. Thus, we set K as 5 to achieve better performance with acceptable memory usage. When U increases from 50 to 200, our method achieves better performance, which indicates that it is effective to generate more pairs for our ICD-Face. However, when U continues to increase, the performance begins to gradually degrade. It is reasonable that the quality of feature representations begins to decrease when U is larger, which causes inaccurate similarity distribution estimation. Therefore, by default, we set K as 5 and U as 200, respectively.

4.4. Further analysis

Necessity on Intra-class Compactness Distillation. In Section 1, we have discussed that $\mathcal{L}_{fcd}$ can maintain the inter-class discrepancy but degrade the intra-class compactness. Here, we further investigate the role of $\mathcal{L}_{fcd}$ in face recognition, as shown in Fig. 4(a), we visualize the changes of the prototype similarities and the FCD loss in the training process, respectively. Note that we adopt the mean feature of all samples belonging to the same identity as the prototype of this identity, and the prototype similarity is computed based on the student prototype and the corresponding teacher prototype with the same identity. In Fig. 4(a), we observe that the prototype similarities between the teacher and student models increase a lot in the training process, which indicates that the class centers of the



(a) (b)

Figure 4: (a). The prototype similarity and FCD loss $\mathcal{L}_{fcd}$ along the training process impact and limitation of $\mathcal{L}_{fcd}$. (b) The intra-class compactness bias degree of different methods.

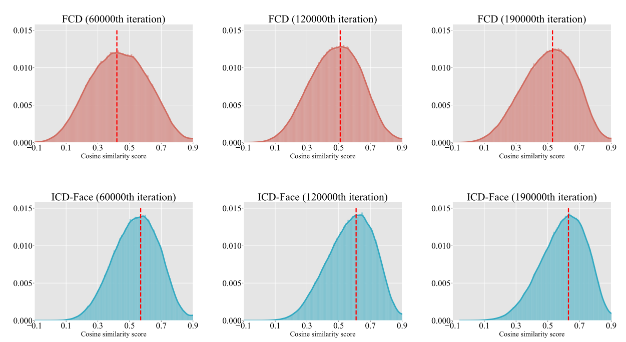


Figure 5: The intra-class similarity distributions of different iterations for FCD and ICD-Face on the training dataset.

student are very close to the corresponding centers of the teacher. In other words, FCD can align the prototypes of the teacher and student well in the feature space. However, in Fig. 4(b), we randomly sample 100 identities to visualize the intra-class compactness bias degree for teacher and student models. We take the student model trained by FCD (i.e., "FCD (S)" in Fig 4(b)) as an example to compute the intra-class compactness bias degree. Specifically, the number of images for the k-th identity is N_k, where $k \in [1, 100]$. For the k-th identity, we compute the angles between $\{\mathbf{f}_i^s\}_{i=1}^{N_k}$ with the corresponding prototype $\mathbf{f}_p^s$, and obtain the mean angle as the intra-class compactness bias degree (i.e., $\frac{1}{N_k}\sum_{i=1}^{N_k}\arccos(\cos(\mathbf{f}_i^s, \mathbf{f}_p^s))$. Similarly, we can also compute the intra-class compactness bias degree for the teacher model (i.e., "ArcFace (T)" in Fig 4(b)) Therefore, larger mean angle represents lower intra-class compactness. As shown in Fig. 4(b), we observe that "FCD (S)" is higher than "ArcFace (T)" a lot, which means that the intra-class compactness after using FCD degrades a lot for the student model. Therefore, intra-class compactness distillation is necessary for FR distillation.

Visualization on the intra-class similarity distributions. To further analyze the effect of ICD-Face, we visualize the intra-class similarity distributions for FCD and ICD-Face in Fig. 5 on the training dataset. Specifically, we use the models of MBNet in the 60,000th, 120,000th, 190,000th iterations for both FCD and ICD-Face. The first and the second rows show the results of FCD and ICD-Face, respectively. In Fig. 5, during training, when compared with the FCD baseline method, we observe that the intra-class similarity distribution is more compact and higher, which further demonstrates the effectiveness of our ICD-Face.

5. Conclusion

In our work, we first investigate the problems of existing FR distillation methods. Then, we propose a new FR distillation method called ICD-Face, which additionally introduces Intra-class Compactness Distillation (ICD) into the existing methods. Specifically, we first estimate the similarity distributions of the teacher and student models, and then utilize the similarity distribution consistency loss to align the intra-class similarity distributions between the teacher and student models. Extensive experiments on multiple FR benchmark datasets demonstrate the effectiveness of our ICD-Face.

References

[1] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, and Ying Fu. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1445–1449, October 2021. 3, 6

[2] Jiaqi Bai, Hongcheng Guo, Jiaheng Liu, Jian Yang, Xinnian Liang, Zhao Yan, and Zhoujun Li. Griprank: Bridging the gap between retrieval and generation via the generative knowledge improved passage ranking, 2023. 2

[3] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017. 2

[4] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*, pages 428–438. Springer, 2018. 6

[5] Svitov David and Alyamkin Sergey. Margindistillation: Distillation for face recognition neural networks with margin-based softmax. *International Journal of Computer and Information Engineering*, 15(3):206–210, 2021. 2

[6] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. 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. 3, 6

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1, 3, 5, 6, 7

[8] Jiankang Deng, Jia Guo, Jing Yang, Alexandros Lattas, and Stefanos Zafeiriou. Variational prototype learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11906–11915, June 2021. 3, 4

[9] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. {SEED}: Self-supervised distillation for visual representation. In *International Conference on Learning Representations*, 2021. 2

[10] Yushu Feng, Huan Wang, Haoji Roland Hu, Lu Yu, Wei Wang, and Shiyan Wang. Triplet distillation for deep face recognition. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 808–812. IEEE, 2020. 2

[11] Jinyang Guo, Jiaheng Liu, Zining Wang, Yuqing Ma, Ruihao Gong, Ke Xu, and Xianglong Liu. Adaptive contrastive distillation for bert compression. In *ACL*, 2023. 2

[12] Jinyang Guo, Jiaheng Liu, and Dong Xu. 3d-pruning: A model compression framework for efficient 3d action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8717–8729, 2022. 2

[13] Jinyang Guo, Jiaheng Liu, and Dong Xu. Jointpruning: Pruning networks along multiple dimensions for efficient point cloud processing. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3659–3672, 2022. 2

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2, 4

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 6, 7

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 3, 5, 6, 7

[17] Yuge Huang, Pengcheng Shen, Ying Tai, Shaoxin Li, Xiaoming Liu, Jilin Li, Feiyue Huang, and Rongrong Ji. Improving face recognition from hard samples via distribution distillation loss. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 138–154. Springer, 2020. 2, 5, 6

[18] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5901–5910, 2020. 3

[19] InsightFace. Glint-mini face recognition dataset. [online]. https://github.com/deepinsight/insightface/tree/master/recognition/_datasets_, 2021. 3, 6

[20] Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. Knowledge distillation via route constrained optimization. In *ICCV*, 2019. 2

[21] Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12115–12124, 2021. 2

[22] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016. 1, 3, 6

[23] Yonghyun Kim, Wonpyo Park, and Jongju Shin. Broadface: Looking at tens of thousands of people at once for face recognition. *ECCV*, 2020. 6

[24] Zhenmao Li, Yichao Wu, Ken Chen, Yudong Wu, Shunfeng Zhou, Jiaheng Liu, and Junjie Yan. Learning to auto weight: Entirely data-driven and highly efficient weighting framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4788–4795, 2020. 3

[25] Jiaheng Liu, Jinyang Guo, and Dong Xu. Apsnet: Toward adaptive point sampling for efficient 3d action recognition. *IEEE Transactions on Image Processing*, 31:5287–5302, 2022. 2

[26] Jiaheng Liu, Haoyu Qin, Yichao Wu, Jinyang Guo, Ding Liang, and Ke Xu. Coupleface: relation matters for face

recognition distillation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pages 683–700. Springer, 2022. 2, 3, 6, 7

[27] Jiaheng Liu, Haoyu Qin, Yichao Wu, and Ding Liang. Anchorface: Boosting tar@far for practical face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 3

[28] Jiaheng Liu, Yudong Wu, Yichao Wu, Chuming Li, Xiaolin Hu, Ding Liang, and Mengyu Wang. Dam: Discrepancy alignment metric for face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3814–3823, 2021. 3

[29] Jiaheng Liu, Tan Yu, Hanyu Peng, Mingming Sun, and Ping Li. Cross-lingual cross-modal consolidation for effective multilingual video corpus moment retrieval. In *NAACL*, 2022. 2

[30] Jiaheng Liu, Zhipeng Yu, Haoyu Qin, Yichao Wu, Ding Liang, Gangming Zhao, and Ke Xu. Oneface: One threshold fornbsp;all. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, page 545–561, 2022. 3

[31] Jiaheng Liu, Shunfeng Zhou, Yichao Wu, Ken Chen, Wanli Ouyang, and Dong Xu. Block proposal neural architecture search. *IEEE Transactions on Image Processing*, 30:15–25, 2020. 3

[32] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 3

[33] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7, 2016. 3

[34] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. 2

[35] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018. 5, 6

[36] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021. 3

[37] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 2, 6, 7

[38] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *ICCV*, October 2019. 2, 6, 7

[39] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2

[40] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 2, 6, 7

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 3, 4

[42] Weidong Shi, Guanghui Ren, Yunpeng Chen, and Shuicheng Yan. Proxylesskd: Direct knowledge distillation with inherited classifier for face recognition. *arXiv preprint arXiv:2011.00265*, 2020. 2

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 3

[44] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014. 3

[45] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020. 3

[46] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 3

[47] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 3

[48] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020. 2

[49] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 6, 7

[50] Evgeniya Ustinova and Victor Lempitsky. Learning deep embeddings with histogram loss. *Advances in Neural Information Processing Systems*, 29, 2016. 5

[51] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 3

[52] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017. 1, 3

[53] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface:

Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 3

[54] Xiaobo Wang, Tianyu Fu, Shengcai Liao, Shuo Wang, Zhen Lei, and Tao Mei. Exclusivity-consistency regularized knowledge distillation for face recognition. In *ECCV*, pages 325–342. Springer, 2020. 2, 6, 7

[55] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. Mis-classified vector guided softmax loss for face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12241–12248, 2020. 3

[56] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus benchmark-b face dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 90–98, 2017. 2, 5, 6

[57] Lior Wolf, Tal Hassner, and Itay Maoz. *Face recognition in unconstrained videos with matched background similarity*. IEEE, 2011. 6

[58] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 3

[59] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017. 2

[60] Zhipeng Yu, Qianqian Xu, Yangbangyan Jiang, Haoyu Qin, and Qingming Huang. Pay attention to your positive pairs: Positive pair aware contrastive knowledge distillation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5862–5870, 2022. 2

[61] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 6, 7

[62] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5409–5418, 2017. 3

[63] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018. 5

[64] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *CVPR*, pages 10492–10502, June 2021. 1, 3