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

Zhipeng Yu1* Jiaheng Liu2* Haoyu Qin3 Yichao Wu3 Kun Hu3
Jiayi Tian2 Ding Liang3
1SEECE, UCAS 2Beihang University 3SenseTime 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., L2 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.

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
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 $L_{fcd}$ between $f_s^i$ and $f_t^i$ 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 $L_{sdc}$ loss to align the intra-class similarity distributions between the teacher and student models.

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 $L_{fcd}$ 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 $T$ and $S$, respectively. For each sample $x_i$, the corresponding identity label is $y_i$, and the features extracted by $T$ and $S$ are denoted as $f_t^i$ and $f_s^i$, 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 $S$ with the prediction probabilities from $T$. However, these techniques are usually incompatible with FR. In practice, for FR, we can only obtain a pre-trained $T$ but have no idea about how it was trained (e.g., the training datasets, loss functions). Therefore, the proba-
bility consistency loss is not available when the number of identities of the training dataset for $T$ is different from the current dataset for $S$ or $T$ 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 $S$ and $T$ is more important for FR distillation.

3.2. Feature Consistency Distillation

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

$$L_{fcd} = \frac{1}{2N} \sum_{i=1}^{N} \left\| \frac{f_i^s}{||f_i^s||_2} - \frac{f_i^t}{||f_i^t||_2} \right\|^2,$$

(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 $S$ 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 $M^t \in \mathbb{R}^{Q \times K \times d}$ and a student feature bank $M^s \in \mathbb{R}^{Q \times K \times d}$ 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 $M^s$ as an example to show the details of constructing a feature bank. Specifically, in each iteration, we first update the student feature bank $M^s$ by pushing the features extracted by the student model of the current batch into $M^s$, and then utilize the stored features of $M^s$ 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 $V \in \mathbb{R}^{Q \times K}$ to represent the validness of each feature in the student feature bank $M^s$. Each item in $V$ 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 $V$ as 0 at the beginning of the training process. In each iteration, in ICD-Face, we first extract the features $\{f_i^s\}_{i=1}^N$ of the current batch, where $N$ is the batch size, and $y_i$ is the corresponding label of $f_i^s$. Then, we update the student feature bank $M^s$ based on $\{f_i^s\}_{i=1}^N$. Specifically, for the $i$-th feature $f_i^s$, when the number of the stored features for the corresponding identity $y_i$ is smaller than $K$, we insert $f_i^s$ into $M^s$ 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 $M^s[y_i]$, which is also the index of the smallest value in $V[y_i]$. Then, we replace the oldest feature with the newly extracted feature $f_i^s$ based on the index $idx_i$, which means we set $M^s[y_i][idx_i] = f_i^s$. After that, we set the number of valid steps for $f_i^s$ as $U$, which means we set $V[y_i][idx_i] = U$. After each training step, $V$ is updated by $V = V - 1$, which decreases the valid steps of all stored features in $M^s$.

After the updating process for the student feature bank $M^s$, the number of valid features in the student feature bank $M^s$ is $\sum_{i=1}^{Q} \sum_{j=1}^{K} 1(V[y_i][j] > 0)$, where $1(x)$ is the in-
indicator function. Then, for each feature $f^s_i$ in the current batch, the positive pairs are constructed by using $f^s_i$ 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 $S_D = \{s_i\}_{i=1}^G$, 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 $T_D = \{t_i\}_{i=1}^G$ 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 $T$ and student model $S$, respectively. $T_D$ and $S_D$ 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 [-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 $n_1 = -1, n_2, \ldots, n_R = 1$ uniformly filling $[-1, 1]$ with the step $\Delta = \frac{2}{R-1}$. Then, for the student model $S$, we estimate the value $h^s_r$ of the histogram $H^S$ at the $r$-th node as:

$$h^s_r = \frac{1}{|S_D|} \sum_{i} \delta_{i,r},$$

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

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

where $\gamma$ 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 $D_{KL}$ to constrain the similarity distributions between the student and teacher models, which is defined as the following Similarity Distribution Consistency (SDC) loss:

$$L_{sdc} = D_{KL}(P^s||P^t).$$

### 3.4. Loss Function of ICD-Face

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

$$\mathcal{L} = L_{fed} + \alpha \cdot L_{sdc},$$

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

$$\mathcal{L} = L_{fed} + \alpha \cdot L_{sdc} + \beta \cdot L_{cls},$$

where we call ICD-Face with $L_{cls}$ 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 pair in knowledge distillation. The differences between PACKD and ICD-Face are as follows: (1). Motivation is different. PACKD is used for closed-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...
For student 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 using FCD loss for 50k iterations, and then add the external SDC loss $L_{sdc}$. The loss weights $\alpha$ and $\beta$ 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 $\Delta$ and $\gamma$ 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 $T$ and $S$, 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

\begin{algorithm}
\textbf{Algorithm 1 ICD-Face} \\
\textbf{Input:} Pre-trained teacher model $T$; Randomly initialized student model $S$; 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 $M^s \in \mathbb{R}^{Q \times K \times d}$ and teacher feature bank $M^t \in \mathbb{R}^{Q \times K \times d}$. The validness indicator $V \in \mathbb{R}^{Q \times K}$ for the validness of the stored features in $M^s$ and $M^t$; The maximum valid steps $U$; \\
1: Zero initialize $V$; \\
2: for each iteration in the training process do \\
3: Get features $\{f^t_i\}_{i=1}^N$ and $\{f^s_i\}_{i=1}^N$ by $T$ and $S$; \\
4: Calculate $L_{fed}$ of $\{f^t_i\}_{i=1}^N$ and $\{f^s_i\}_{i=1}^N$ by Eq. (1); \\
5: for features $f^t_i$ and $f^s_i$ in $\{f^t_i\}_{i=1}^N$ and $\{f^s_i\}_{i=1}^N$ do \\
6: Select the index $idx_i$ to insert $f^t_i$ and $f^s_i$ into $M^t$ and $M^s$, respectively; \\
7: $M^t[y_i][idx_i] = f^t_i$, $M^s[y_i][idx_i] = f^s_i$; \\
8: $V[y_i][idx_i] = U$; \\
9: end for \\
10: $V = V - 1$; \\
11: Construct the positive pairs using $\{f^t_i\}_{i=1}^N$ and the valid features $M^t[V > 0]$; \\
12: Construct the positive pairs using $\{f^s_i\}_{i=1}^N$ and the valid features $M^s[V > 0]$; \\
13: Construct the similarity scores of the positive pairs for teacher and student, respectively. \\
14: Calculate SDC loss $L_{sdc}$ based on Eq. (4); \\
15: Calculate $L_{cls}$ based on $\{f^s_i\}_{i=1}^N$; \\
16: Update parameters of $S$ by $L$ in Eq. (5) or Eq. (6); \\
17: end for \\
\textbf{Output:} The optimized student model $S$;
\end{algorithm}
Table 1: Results on IJB-B and IJB-C of different methods.

<table>
<thead>
<tr>
<th>Models</th>
<th>Method</th>
<th>IJB-B(TAR@FAR)</th>
<th>IJB-C(TAR@FAR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MBNet [40]</td>
<td>ArcFace [7]</td>
<td>85.97 1e-4 88.95 1e-5</td>
<td>75.81 1e-5 82.64 1e-5</td>
</tr>
<tr>
<td></td>
<td>KD [16]</td>
<td>86.12 1e-4 89.03 1e-5</td>
<td>75.99 1e-5 82.69 1e-5</td>
</tr>
<tr>
<td></td>
<td>FCD</td>
<td>90.34 1e-4 92.68 1e-5</td>
<td>81.92 1e-5 87.74 1e-5</td>
</tr>
<tr>
<td></td>
<td>AT [61]</td>
<td>90.35 1e-4 92.65 1e-5</td>
<td>82.22 1e-5 87.54 1e-5</td>
</tr>
<tr>
<td>MBNet [40]</td>
<td>CCKD [38]</td>
<td>90.72 1e-4 93.17 1e-5</td>
<td>83.34 1e-5 89.11 1e-5</td>
</tr>
<tr>
<td></td>
<td>RKD [37]</td>
<td>90.32 1e-4 92.33 1e-5</td>
<td>82.45 1e-5 88.12 1e-5</td>
</tr>
<tr>
<td></td>
<td>SP [49]</td>
<td>90.52 1e-4 92.71 1e-5</td>
<td>82.88 1e-5 88.52 1e-5</td>
</tr>
<tr>
<td></td>
<td>EC-KD [54]</td>
<td>90.59 1e-4 92.85 1e-5</td>
<td>83.54 1e-5 88.32 1e-5</td>
</tr>
<tr>
<td></td>
<td>CoupleFace</td>
<td>91.18 1e-4 93.18 1e-5</td>
<td>84.63 1e-5 89.57 1e-5</td>
</tr>
<tr>
<td></td>
<td>ICD-Face</td>
<td>90.52 1e-4 92.71 1e-5</td>
<td>82.88 1e-5 88.52 1e-5</td>
</tr>
<tr>
<td></td>
<td>ICD-Face+</td>
<td>90.59 1e-4 92.85 1e-5</td>
<td>83.54 1e-5 88.32 1e-5</td>
</tr>
</tbody>
</table>

Table 2: Rank-1 accuracy with different distractors.

<table>
<thead>
<tr>
<th>Models</th>
<th>Method</th>
<th>Distractors</th>
</tr>
</thead>
<tbody>
<tr>
<td>MBNet [40]</td>
<td>ArcFace [7]</td>
<td>90.25 5 84.64 5</td>
</tr>
<tr>
<td></td>
<td>KD [16]</td>
<td>90.25 10 84.65 10</td>
</tr>
<tr>
<td></td>
<td>FCD</td>
<td>96.39 10 93.65 10</td>
</tr>
<tr>
<td></td>
<td>AT [61]</td>
<td>90.50 5 93.68 5</td>
</tr>
<tr>
<td>MBNet [40]</td>
<td>CCKD [38]</td>
<td>96.43 10 93.90 10</td>
</tr>
<tr>
<td></td>
<td>RKD [37]</td>
<td>96.41 5 93.84 5</td>
</tr>
<tr>
<td></td>
<td>SP [49]</td>
<td>96.58 10 93.95 10</td>
</tr>
<tr>
<td></td>
<td>EC-KD [54]</td>
<td>96.41 5 93.85 5</td>
</tr>
<tr>
<td></td>
<td>CoupleFace</td>
<td>96.74 5 94.27 5</td>
</tr>
<tr>
<td></td>
<td>ICD-Face</td>
<td>96.87 5 94.55 5</td>
</tr>
<tr>
<td></td>
<td>ICD-Face+</td>
<td>96.90 5 94.58 5</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>K</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAR (%)</td>
<td>92.73</td>
<td>93.39</td>
<td>93.57</td>
<td>93.59</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>U</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAR (%)</td>
<td>93.23</td>
<td>93.45</td>
<td>93.57</td>
<td>93.54</td>
</tr>
</tbody>
</table>

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 $K$ 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 $L_{fcd}$ can maintain the inter-class discrepancy but degrade the intra-class compactness. Here, we further investigate the role of $L_{fcd}$ 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
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.

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


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


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


