Improving Face Recognition from Hard Samples via Distribution Distillation Loss

Yuge Huang[1*], Pengcheng Shen[1*], Ying Tai[1♯], Shaoxin Li[1♯], Xiaoming Liu[2], Jilin Li[1], Feiyue Huang[1], and Rongrong Ji[3]

[1] Youtu Lab, Tencent
[2] Michigan State University
[3] Xiamen University
{yugehuang, quantshen, yingtai, darwinli, jerolinli, garyhuang}@tencent.com, liuxm@cse.msu.edu, rrji@xmu.seu.cn

Abstract. Large facial variations are the main challenge in face recognition. To this end, previous variation-specific methods make full use of task-related prior to design special network losses, which are typically not general among different tasks and scenarios. In contrast, the existing generic methods focus on improving the feature discriminability to minimize the intra-class distance while maximizing the inter-class distance, which perform well on easy samples but fail on hard samples. To improve the performance on hard samples, we propose a novel Distribution Distillation Loss to narrow the performance gap between easy and hard samples, which is simple, effective and generic for various types of facial variations. Specifically, we first adopt state-of-the-art classifiers such as Arcface to construct two similarity distributions: a teacher distribution from easy samples and a student distribution from hard samples. Then, we propose a novel distribution-driven loss to constrain the student distribution to approximate the teacher distribution, which thus leads to smaller overlap between the positive and negative pairs in the student distribution. We have conducted extensive experiments on both generic large-scale face benchmarks and benchmarks with diverse variations on race, resolution and pose. The quantitative results demonstrate the superiority of our method over strong baselines, *e.g.*, Arcface and Cosface. Code will be available at https://github.com/HuangYG123/DDL.

Keywords: Face Recognition · Loss function · Distribution Distillation

1 Introduction

A primary challenge of large-scale face recognition on unconstrained imagery is to handle the diverse variations on pose, resolution, race and illumination, *etc.*

* equal contribution.
♯ corresponding author.

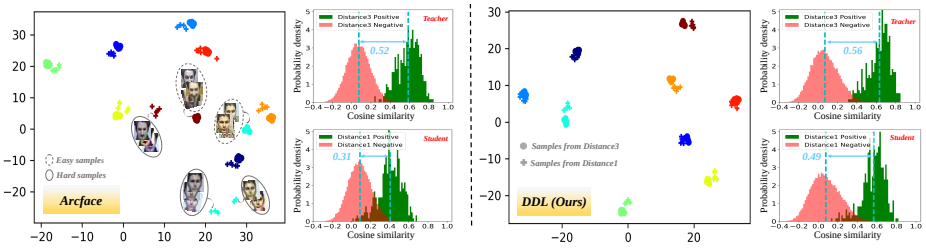

Fig. 1: **Comparisons with Arcface [6] on SCface [10] dataset**. T-SNE [21] visualizations on features, where the same color indicates samples of the same subject. Distance1 (d_1) and Distance3 (d_3) indicate low-resolution and high-resolution images, which were captured at distances of 4.2 and 1.0m, respectively. Each method has two distributions from d_3 and d_1, where there are also two distributions from the positive and negative pairs with a margin indicating the difference of their expectations. With our distribution distillation loss between the teacher and student distributions, our method effectively narrows the performance gap between the easy and hard samples, decreasing the expectation margin from **0.21** (0.52-0.31) to **0.07** (0.56-0.49).

While some variations are easy to address, many others are relatively difficult. As in Fig. 1, State-of-the-Art (SotA) facial classifiers like Arcface [6] well address images with small variations with tight groupings in the feature space. We denote these as easy samples. In contrast, images with large variations are usually far away from the easy ones in the feature space, and are much more difficult to tackle. We denote these as hard samples. To better recognize these hard samples, there are usually two schemes: *variation-specific* and *generic* methods.

Variation-specific methods are usually designed for a specific task. For instance, to achieve pose-invariant face recognition, either handcrafted or learned features are extracted to enhance robustness against pose while remaining discriminative to the identities [33]. Recently, joint face frontalization and disentangled identity preservation are incorporated to facilitate the pose-invariant feature learning [35, 49]. To address resolution-invariant face recognition, a unified feature space is learned in [16, 27] for mapping Low-Resolution (LR) and High-Resolution (HR) images. The works [4, 50] first apply super-resolution on LR images and then perform recognition on the super-resolved images. However, the above methods are specifically designed for the respective variations, therefore their ability to generalize from one variation to another is limited. Yet, it is highly desirable to handle multiple variations in real world recognition systems.

Different from variation-specific methods, generic methods focus on improving the discriminative power of facial features for small intra-class and large inter-class distances. Basically, the prior works fall into two categories, *i.e.*, softmax loss-based and triplet loss-based methods. Softmax loss-based methods regard each identity as a unique class to train the classification networks. Since the traditional softmax loss is insufficient to acquire the discriminative features, several variants [6, 18, 40, 43] are proposed to enhance the discriminability. In contrast, triplet loss-based methods [23, 26] directly learn a Euclidean space embedding

for each face, where faces from the same person form a separate cluster from faces of other people. With large-scale training data and well-designed network structures, both types of methods can obtain promising results.

However, the performance of these methods degrades dramatically on hard samples, such as very large-pose and low-resolution faces. As illustrated in Fig. 1, the features extracted from HR images (*i.e.*, d_3) by the strong face classifier of Arcface [6] are well separated, but the features extracted from LR images (*i.e.*, d_1) cannot be well distinguished. From the perspective of the angle distributions of positive and negative pairs, we can easily observe that Arcface exists more confusion regions on LR face images. It is thereby a natural consequence that such generic methods perform worse on hard samples.

To narrow the performance gap between the easy and hard samples, we propose a novel Distribution Distillation Loss (DDL). By leveraging the best of both the *variation-specific* and *generic* methods, our method is generic and can be applied to diverse variations to improve face recognition in hard samples. Specifically, we first adopt current SotA face classifiers as the baseline (*e.g.*, Arcface) to construct the initial similarity distributions between teacher (*e.g.*, easy samples from d_3 in Fig. 1) and student (*e.g.*, hard samples from d_1 in Fig. 1) according to the difficulties of samples, respectively. Compared to finetuning the baseline models with domain data, our method firstly does not require extra data or inference time (*i.e.*, *simple*); secondly makes full use of hard sample mining and directly optimizes the similarity distributions to improve the performance on hard samples (*i.e.*, *effective*); and finally can be easily applied to address different kinds of large variations in extensive real applications, *e.g.*, women with makeup in fashion stores, surveillance faces in railway stations, and apps looking for missing senior person or children, *etc.*

To sum up, the contributions of this work are three-fold:

- Our method narrows the performance gap between easy and hard samples on *diverse* facial variations, which is simple, effective and general.
- To our best knowledge, it is the first work that adopts similarity distribution distillation loss for face recognition, which provides a new perspective to obtain more discriminative features to better address hard samples.
- Significant gains compared to the SotA Arcface are reported, *e.g.*, 97.0% over 92.7% on SCface, 93.4% over 92.1% on CPLFW, 90.7% over 89.9% (@FAR=$1e{-}4$) on IJB-B and 93.1% over 92.1% (@FAR=$1e{-}4$) on IJB-C.

2 Related Work

Loss Function in FR. Loss function design is pivotal for large-scale face recognition. Softmax is commonly used for face recognition [30, 34, 39], which encourages the separability of features but the learned features are not guaranteed to be discriminative. To address this issue, contrastive [29] and triplet [23, 26] losses are proposed to increase the margin in the Euclidean space. However, both contrastive and triplet losses occasionally encounter training instability due to

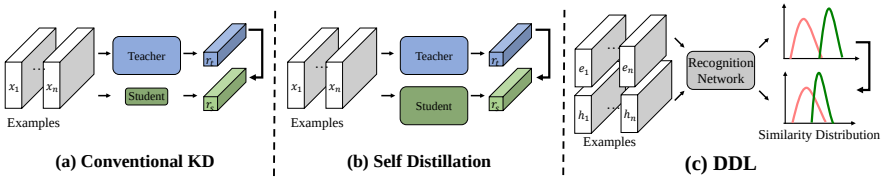

Fig. 2: **Comparisons among conventional knowledge distillation, self-distillation and our DDL**. The student in KD is usually smaller than the teacher. $\{e\}_1^n$ and $\{h\}_1^n$ indicate the easy and hard samples, respectively.

the selection of effective training samples. As a simple alternative, center loss and its variants [7, 43, 52] are proposed to compress the intra-class variance. More recently, angular margin-based losses [6,13,18,19,38] facilitate feature discrimination, and thus lead to larger angular/cosine separability between learned features. The above loss functions are designed to apply constraints either between samples, or between sample and center of the corresponding subject. In contrast, our proposed loss is *distribution* driven. While being similar to the histogram loss [37] that constrains the overlap between the distributions of positive and negative pairs across the training set, our loss differs in that we first separate the training set into a teacher distribution (easy samples) and student distribution (hard samples), and then constrain the student distribution to approximate the teacher distribution via our novel loss, which narrows the performance gap between easy and hard samples.

Variation-Specific FR. Apart from generic solutions [30, 34] for face recognition, there are also many methods designed to handle specific facial variations, such as resolutions, poses, illuminations, expressions and demographics [8]. For example, cross-pose FR [33, 35, 48, 54] is very challenging, and previous methods mainly focus on either face frontalization or pose invariant representations. Low resolution FR is also a difficult task, especially in the surveillance scenario. One common approach is to learn a unified feature space for LR and HR images [11,20,55]. The other way is to perform super resolution [4,31,32] to enhance the facial identity information. Differing from the above methods that mainly deal with one specific variation, our novel loss is a generic approach to improve FR from hard samples, which is applicable to a wide variety of variations.

Knowledge Distillation. Knowledge Distillation (KD) is an emerging topic. Its basic idea is to distill knowledge from a large teacher model into a small one by learning the class distributions provided by the teacher via softened softmax [12]. Typically, Kullback Leibler (KL) divergence [12, 53] and Maximum Mean Discrepancy (MMD) [14] can be adopted to minimize the posterior probabilities between teacher and student models. More recently, transferring mutual relations of data examples from the teacher to the student is proposed [22, 36]. In particular, RKD [22] reported that KD can improve the original performance when the student has the same structure as the teacher (*i.e.*, self-distillation).

Compared to the above distillation methods, our DDL differs in several aspects (see Fig. 2): 1) KD has at least *two* networks, a teacher and a student,

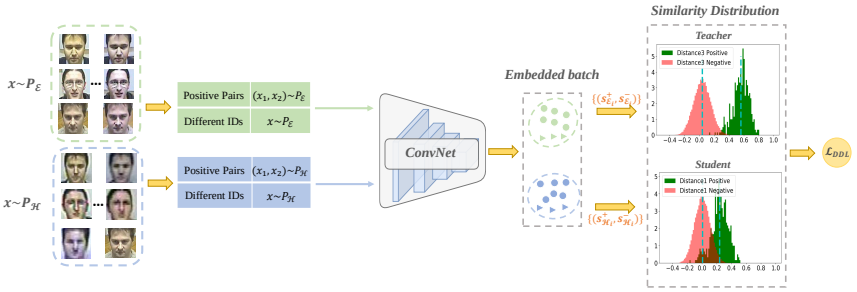

Fig. 3: **Illustration of our DDL**. We sample b positive pairs (*i.e.*, $2b$ samples) and b samples with different identities, for both the teacher $P_{\mathcal{E}}$ and student $P_{\mathcal{H}}$ distributions, to form one mini-batch (*i.e.*, $6b$ in total). $\{(s^+_{\mathcal{E}_i}, s^-_{\mathcal{E}_i})|i = 1, ..., b\}$ indicates we construct b positive and negative pairs from $P_{\mathcal{E}}$ via Eqs. 1 and 2 respectively to estimate the teacher distribution. $\{(s^+_{\mathcal{H}_i}, s^-_{\mathcal{H}_i})|i = 1, ..., b\}$ also indicates we construct b positive and negative pairs from $P_{\mathcal{H}}$ via Eqs. 1 and 2 respectively to estimate the student distribution.

while DDL only learns *one* network. Although in KD the student may have the same structure as the teacher (*e.g.*, self-distillation), they have *different parameters* in training. 2) KD uses sample-wise, Euclidean distance-wise or angle-wise constraints, while DDL proposes a novel cosine similarity distribution-wise constraint which is specifically designed for face recognition. 3) To our best knowledge, currently no KD methods outperform SotA face classifiers on face benchmarks, while DDL consistently outperforms the SotA Arcface classifier.

3 The Proposed Method

Fig. 3 illustrates the framework of our DDL. We separate the training set into two parts, *i.e.*, $\mathcal{E}$ for easy samples and $\mathcal{H}$ for hard samples to form the teacher and student distributions, respectively. In general, for each mini-batch during training, we sample from both parts. To ensure a good teacher distribution, we use the SotA FR model [6] as our initialization. The extracted features are used to construct the positive and negative pairs (Sec. 3.1), which are further utilized to estimate the similarity distributions (Sec. 3.2). Finally, based on the similarity distributions, the proposed DDL is utilized to train the classifier (Sec. 3.3).

3.1 Sampling Strategy from $P_{\mathcal{E}}$ and $P_{\mathcal{H}}$

First, we introduce the details on how we construct the positive and negative pairs in one mini-batch during training. Given two types of input data from both $P_{\mathcal{E}}$ and $P_{\mathcal{H}}$, each mini-batch consists of four parts, two kinds of positive pairs (*i.e.*, $(x_1, x_2) \sim P_{\mathcal{E}}$ and $(x_1, x_2) \sim P_{\mathcal{H}}$), and two kinds of samples with different identities (*i.e.*, $x \sim P_{\mathcal{E}}$ and $x \sim P_{\mathcal{H}}$). To be specific, we on one hand construct b positive pairs (*i.e.*, $2b$ samples), and on the other hand b samples with different

identities from $P_{\mathcal{E}}$ and $P_{\mathcal{H}}$. As the result, there are $6b = (2b + b) * 2$ samples in each mini-batch (see Fig. 3 for more details).

Positive Pairs. The positive pairs are constructed offline in advance, and each pair consist of two samples with the same identity. As shown in Fig. 3, samples of each positive pair are arranged in order. After embedding data into a high-dimensional feature space by a deep network $\mathcal{F}$, the similarity of a positive pair s^+ can be obtained as follows:

$$s_i^+ =< \mathcal{F}(x_{pos_{i1}}), \mathcal{F}(x_{pos_{i2}}) >, i = 1, ..., b \tag{1}$$

where $x_{pos_{i1}}$, $x_{pos_{i2}}$ are the samples of one positive pair. Note that positive pairs with similarity less than 0 are usually outliers, which are deleted as a practical setting since our main goal is not to specifically handle noise.

Negative Pairs. Different from the positive pairs, we construct negative pairs online from the samples with different identities via hard negative mining, which selects negative pairs with the largest similarities. To be specific, the similarity of a negative pair s^- is defined as:

$$s_i^- = \max_j \Big(\{ s_{ij}^- =< \mathcal{F}(x_{neg_i}), \mathcal{F}(x_{neg_j}) > | j = 1, ..., b \} \Big), \tag{2}$$

where x_{neg_i}, x_{neg_j} are from different subjects. Once the similarities of positive and negative pairs are constructed, the corresponding distributions can be estimated, which is described in the next subsection.

3.2 Similarity Distribution Estimation

The process of similarity distribution estimation is similar to [37], which is performed in a simple and piece-wise differentiable manner using 1D histograms with soft assignment. Specifically, two samples x_i, x_j from the same person form a positive pair, and the corresponding label is denoted as $m_{ij} = +1$. In contrast, two samples from different persons form a negative pair, and the label is denoted as $m_{ij} = -1$. Then, we obtain two sample sets $\mathcal{S}^+ = \{s^+ = \langle \mathcal{F}(x_i), \mathcal{F}(x_j) \rangle | m_{ij} = +1\}$ and $\mathcal{S}^- = \{s^- = \langle \mathcal{F}(x_i), \mathcal{F}(x_j) \rangle | m_{ij} = -1\}$ corresponding to the similarities of positive and negative pairs, respectively.

Let p^+ and p^- denote the two probability distributions of $\mathcal{S}^+$ and $\mathcal{S}^-$, respectively. As in cosine distance-based methods [6], the similarity of each pair is bounded to $[-1, 1]$, which is demonstrated to simplify the task [37]. Motivated by the histogram loss, we estimate this type of one-dimensional distribution by fitting simple histograms with uniformly spaced bins. We adopt R-dimensional histograms $H^+$and H^-, with the nodes $t_1 = -1$, t_2, $\cdots$, $t_R = 1$ uniformly filling $[-1, 1]$ with the step $\triangle = \frac{2}{R-1}$. Then, we estimate the value h_r^+ of the histogram H^+ at each bin as:

$$h_r^+ = \frac{1}{|S^+|} \sum_{(i,j):m_{ij}=+1} \delta_{i,j,r}, \tag{3}$$

where (i, j) spans all the positive pairs. Different from [37], the weights $\delta_{i,j,r}$ are chosen by an exponential function as:

$$\delta_{i,j,r} = exp(-\gamma(s_{ij} - t_r)^2), \tag{4}$$

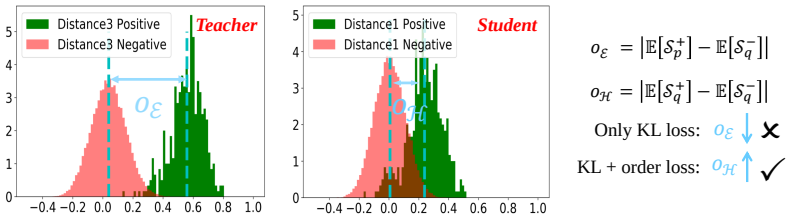

Fig. 4: **Illustration of the effects of our order loss**. Similarity distributions are constructed by Arcface [6] on SCface, in which we have 2 kinds of order distances formed from both of the teacher and student distributions according to Eq. 6.

where γ denotes the spread parameter of Gaussian kernel function, and t_r denotes the rth node of histograms. We adopt the Gaussian kernel function because it is the most commonly used kernel function for density estimation and robust to the small sample size. The estimation of H^- proceeds analogously.

3.3 Distribution Distillation Loss

We make use of SotA face recognition engines like [6], to obtain the similarity distributions from two kinds of samples: easy and hard samples. Here, easy samples indicate that the FR engine performs well, in which the similarity distributions of positive and negative pairs are clearly separated (see the teacher distribution in Fig. 4), while hard samples indicate that the FR engine performs poorly, in which the similarity distributions may be highly overlapped (see the student distribution in Fig. 4).

KL Divergence Loss. To narrow the performance gap between the easy and hard samples, we constrain the similarity distribution of hard samples (*i.e.*, student distribution) to approximate the similarity distribution of easy samples (*i.e.*, teacher distribution). The teacher distribution consists of two similarity distributions of both positive and negative pairs, denoted as P^+ and P^-, respectively. Similarly, the student distribution also consists of two similarity distributions, denoted as Q^+ and Q^-. Motivated by the previous KD methods [12, 53], we adopt the KL divergence to constrain the similarity between the student and teacher distributions, which is defined as follows:

$$\begin{aligned}\mathcal{L}_{KL} &= \lambda_1 \mathbb{D}_{KL}(P^+||Q^+) + \lambda_2 \mathbb{D}_{KL}(P^-||Q^-) \\ &= \underbrace{\lambda_1 \sum_s P^+(s) \log \frac{P^+(s)}{Q^+(s)}}_{KL\ loss\ on\ pos.\ pairs} + \underbrace{\lambda_2 \sum_s P^-(s) \log \frac{P^-(s)}{Q^-(s)}}_{KL\ loss\ on\ neg.\ pairs},\end{aligned} \tag{5}$$

where λ_1, λ_2 are the weight parameters.

Order Loss. However, only using KL loss does not guarantee good performance. In fact, the teacher distribution may choose to approach the student distribution and leads to more confusion regions between the distributions of positive and

negative pairs, which is the opposite of our objective (see Fig. 4). To address this problem, we design a simple yet effective term named *order loss*, which minimizes the distances between the expectations of similarity distributions from the negative and positive pairs to control the overlap. Our order loss can be formulated as follows:

$$\mathcal{L}_{order} = -\lambda_3 \sum_{(i,j)\in(p,q)} (\mathbb{E}[\mathcal{S}_i^+] - \mathbb{E}[\mathcal{S}_j^-]), \tag{6}$$

where $\mathcal{S}_p^+$ and S_p^- denote the similarities of positive and negative pairs of the teacher distribution; $\mathcal{S}_q^+$ and S_q^- denote the similarities of positive and negative pairs of the student distribution; and λ_3 is the weight parameter.

In summary, the entire formulation of our distribution distillation loss is: $\mathcal{L}_{DDL} = \mathcal{L}_{KL} + \mathcal{L}_{order}$. DDL can be easily extended to multiple student distributions varied from one specific variation as follows:

$$\mathcal{L}_{DDL} = \sum_{i=1}^{K} \mathbb{D}_{KL}(P||Q_i) - \lambda_3 \sum_{i,j\in(p,q_1\ldots q_K)} (\mathbb{E}[\mathcal{S}_i^+] - \mathbb{E}[\mathcal{S}_j^-]), \tag{7}$$

where K is the number of student distributions. Further, to maintain the performance on easy samples, we incorporate the loss function of Arcface [6], and thus the final loss is:

$$\mathcal{L}(\Theta) = \mathcal{L}_{DDL} + \mathcal{L}_{Arcface}, \tag{8}$$

where Θ denotes the parameter set. Note that $\mathcal{L}_{Arcface}$ can be easily replaced by any kind of popular losses in FR.

3.4 Generalization on Various Variations

Next, we discuss the generalization of DDL on various variations, which defines our *application scenarios* and how we select easy/hard samples. Basically, we can distinguish the easy and hard samples according to whether the image contains large facial variations that may hinder the identity information, *e.g.*, low-resolution and large pose variation.

Observation from Different Variations. Our method assumes that two or more distributions, each computed from a subset of training data, have differences among themselves, which is a popular phenomenon in face recognition and is demonstrated in Fig. 5. It shows similarity distributions of normal and challenging samples based on Arcface [6] trained on CASIA except CFP, which is trained on VGGFace2. As we can see, 1) since CASIA is biased to Caucasian, Mongolian samples in COX are more difficult and thus relatively regarded as the hard samples, 2) different variations share a common observation that the similarity distributions of challenging samples are usually different from those of easy samples, 3) variations with different extents may have different similarity distributions (*e.g.*, $\mathcal{H}1$ and $\mathcal{H}2$ in Fig. 5(c)). In summary, when a task satisfies that *the similarity distributions differ between easy and hard samples*, our method is a good solution and one can enjoy the performance improvement by properly constructing the positive and negative pairs, as validated in Sec. 4.3.

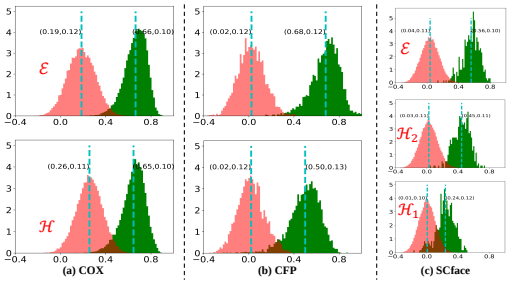

Fig. 5: **Similarity distribution differences** between easy and hard samples on various variations, including race on COX, pose on CFP, and resolution on SCface respectively. $(\cdot,\cdot)$ indicates the mean and standard deviation.

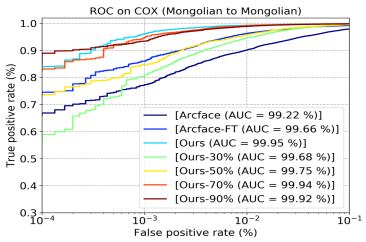

Fig. 6: **Effects of number of training subjects** on COX. Compared to Arcface-FT, DDL achieves comparable results with only *half* the number of training subjects.

Performance Balance Between Easy and Hard Samples. Improving the performance on hard samples while maintaining the performance on easy samples is a trade-off. Two factors in our method help maintain performance on easy samples. First, we incorporate the SotA Arcface loss [6] to maintain feature discriminability on easy samples. Second, our order loss minimizes the distance between the expectations of similarity distributions from the negative and positive pairs, which helps control the overlap between positive and negative pairs.

Discussions on Mixture Variations. As shown in Eq. 7, our method can be easily extended to multiple variations for one task (*e.g.*, low resolution, large pose, *etc*). An alternative is to mix the variations with different extents from one task into *one* student distribution, which, as shown in Sec. 4.2, is not good enough to specifically model the different extents and tends to lead to lower performance. As for different variations from different tasks, one may also construct multiple teacher-student distribution pairs to address the corresponding task respectively, which can be a good future direction.

4 Experiments

4.1 Implementation Details

Datasets. We separately employ SCface [10], COX [15], CASIA-WebFace [47], VGGFace2 [3] and the refined MS1M [6] as our training data to conduct fair comparisons with other methods. We extensively test our method on benchmarks with diverse variations, *i.e.*, COX on race, SCface on resolution, CFP and CPLFW on Pose, as well as generic large-scale benchmarks IJB-B and IJB-C. For COX, the data are collected from two races: Caucasian and Mongolian. Since no race label is given, we manually label 428 Mongolians and 572 Caucasians to conduct experiments, in which half of both races are used for finetuning and the others for testing. For SCface, following [20], 50 subjects are used for finetuning

and 80 subjects are for testing. In the testing stage, we conduct face identification, where HR image is used as the gallery and LR images with three different resolutions form the probe. Specifically, the LR images are captured at three distances: 4.2m for d_1, 2.6m for d_2 and 1.0m for d_3. We split easy and hard samples according to the main variation in each dataset. For race, since the dataset on which the model is pre-trained is biased to Caucasian, Mongolian samples on COX are more difficult and thus relatively regarded as the hard samples. For pose, we estimate the pose of each image [25] on VGGFace2, and images with yaw $< 10°$ and yaw $> 45°$ are used as easy and hard samples respectively. For resolution, images captured under d_3 and d_1/d_2 are used as easy and hard samples respectively on SCface.

Training Setting. We follow [6,40] to generate the normalized faces (112×112) with five landmarks [51]. For the embedding network, we adopt ResNet50 and ResNet100 as in [6]. Our work is implemented in Tensorflow [1]. We train models on 8 NVIDIA Tesla P40 GPUs. On SCface, we set the number of positive/negative pairs as $b = 16$, thus the batch size on one GPU is $3b \times 3 = 144$, including one teacher distribution and two student distributions (see Fig. 5(c)). On other datasets, we set b to 32, thus the batch size per GPU is $3b \times 2 = 192$. The numbers of iterations are 1K, 2K and 20K on SCface, COX and VGGFace2, respectively. The models are trained with SGD, with momentum 0.9 and weight decay $5e^{-4}$. The learning rate is $1e^{-3}$, and is divided by 10 at half of iterations. All of the weight parameters are consistent across all the experiments. λ_1, λ_2 and λ_3 are set to $1e^{-1}$, $2e^{-2}$ and $5e^{-1}$, respectively.

4.2 Ablation Study

Effects of Distance Metric on Distributions. We investigate the effects of several commonly used distribution metrics to constrain the teacher and student distributions in our DDL, including KL divergence, Jensen-Shannon (JS) divergence, and Earth Mover Distance (EMD). Although KL divergence does not qualify as a statistical metric, it is widely used as a measure of how one distribution is different from another. JS divergence is a symmetric version of KL divergence. EMD is another distance function between distributions on a given metric space and has seen success in image synthesis [9]. We incorporate our order loss with the above distance metrics, and report the results in Tab. 1. We choose KL divergence in our DDL since it achieves the best performance, which shares similar conclusion with [53]. To further investigate the effectiveness of each component in our loss, we train the network with each component separately. As shown in Tab. 1, only KL or only Order does not guarantee satisfying performance, while using both components leads to better results.

Effects of Random vs. Hard Mining. To investigate the effect of hard sample mining in our method, we train models on SCface with the corresponding strategy (*i.e.*, negative pairs with the largest similarity are selected), and without the strategy by randomly selecting the negative pairs, respectively. The comparative results are reported in Tab. 1. Comparing with the results of "Random" selecting, it is clear that our hard mining version outperforms the one without.

Table 1: **Extensive ablation studies on SCface dataset. All methods are trained on CASIA with a ResNet50 backbone.** Each color corresponds to a type of ablation study experimental setting.

	EMD	JS	KL	Order	Random Hard Mining	Mixture Specific	d_1	d_2	d_3	Avg.
Arcface-FT							67.3	93.5	98.0	86.3
Distance metric	✓			✓	✓	✓	78.0	97.8	96.8	90.5
		✓		✓	✓	✓	83.0	**98.3**	**99.0**	93.4
			✓		✓	✓	76.0	94.3	98.5	89.6
				✓	✓	✓	80.8	97.5	**99.0**	92.4
Mining strategy			✓	✓	✓	✓	80.3	96.3	95.3	90.6
Mixture training			✓	✓	✓	✓	81.5	97.0	97.8	92.1
DDL (ours)			✓	✓	✓	✓	**86.8**	**98.3**	98.3	**94.4**

Effects of Mixture vs. Specific training. As mentioned in Sec. 3.4, we basically construct different student distributions for samples with different extents of variations on SCface. Here, we mix two variations from d_1 and d_2 into *one* student distribution. The comparison between our specific and mixture training is also shown in Tab. 1. As we expected, the mixture version is worse than the specific version, but is still better than the conventional finetuning (*i.e.*, Avg. being 86.3), which indicates that properly constructing different hard samples for the target tasks may maximize the advantages of our method.

Effects of Number of Training Subjects. Here, we conduct tests on COX dataset to show the effects of using different numbers of training subjects. Specifically, we adopt 10%, 30%, 50%, 70%, 90% and 100% of training subjects, respectively. A pre-trained Arcface on CASIA is used as the baseline. For fair comparison, we also compare our method against Arcface with conventional finetuning (*i.e.*, Arcface-FT). From Fig. 6 we see that: 1) Compared to Arcface-FT, our method clearly boosts the performance on Mongolian-Mongolian verification tests with comparable training data. 2) Our method can have comparable performance with the only *half* of the entire training subjects, which demonstrate the superiority of utilizing the global similarity distributions.

4.3 Comparisons with SotA Methods

Resolution on SCface. SCface mimics the real-world surveillance watch-list problem, where the gallery contains HR faces and the probe consists of LR faces captured from surveillance cameras. We compare our method with SotA low-resolution face recognition methods in Tab. 2. Most results are directly cited from [20], while the results of Arcface come from our re-implementation. From Tab. 2, we have some observations: 1) The baseline Arcface achieves much better results than the other methods without finetuning, especially on the relatively high-resolution images from d_3. 2) Our (CASIA+ResNet50)-FT version already outperforms all of the other methods, including Arcface (MS1M+ResNet100)-FT, which uses a larger model that is trained by a much larger dataset. 3) We achieve significant improvement on d_1 setting, which is the hardest. This demon-

Table 2: **Rank-1 performance (%) of face identification on SCface test-ing set**. '-FT' represents finetuning with training set from SCface.

Distance →	d_1	d_2	d_3	avg.
LDMDS [46]	62.7	70.7	65.5	66.3
Center Loss [43]	36.3	81.8	94.3	70.8
Arcface (CASIA+R50)	48.0	92.0	99.3	79.8
Arcface (MS1M+R100)	58.9	98.3	**99.5**	85.5
Center Loss-FT	54.8	86.3	95.8	79.0
DCR-FT [20]	73.3	93.5	98.0	88.3
Histogram (CASIA+R50)-FT [37]	74.3	95.0	97.3	88.8
OHEM (CASIA+R50)-FT [28]	82.5	97.3	97.5	92.7
Focal (CASIA+R50)-FT [17]	76.8	95.5	96.8	89.7
Triplet (CASIA+R50)-FT [6]	84.2	97.2	99.2	93.5
Arcface (CASIA+R50)-FT	67.3	93.5	98.0	86.3
Arcface (MS1M+R100)-FT	80.5	98.0	99.5	92.7
Ours (CASIA+R50)	86.8	98.3	98.3	94.4
Ours (MS1M+R100)	**93.2**	**99.2**	98.5	**97.0**

Table 3: **Verification comparisons with SotA methods** on LFW and two popular pose benchmarks, including CFP-FP and CPLFW.

Methods (%)	LFW	CFP-FP	CPLFW
Triplet Loss (CVPR'15)	98.98	91.90	–
Center Loss (ECCV'16) [43]	98.75	–	77.48
SphereFace (CVPR'17) [18]	99.27	–	81.40
DRGAN (CVPR'17) [35]	–	93.41	–
Peng *et al.* (ICCV'17) [24]	–	93.76	–
Yin *et al.* (TIP'17) [48]	98.27	94.39	–
VGGFace2 (FG'18) [3]	99.43	–	84.00
Dream (CVPR'18) [2]	–	93.98	–
Deng *et al.* (CVPR'18) [5]	99.60	94.05	–
SV-Arc-Softmax (arXiv'19) [42]	99.78	98.28	92.83
CO-Mining (ICCV'19) [41]	–	95.87	87.31
Arcface (MS1M+R100)-Official [6][1]	**99.82**	98.37	92.08
Arcface (MS1M+R100)	99.80	98.29	92.52
Arcface (VGG+R100)	99.62	98.30	93.13
Ours (VGG+R100)	99.68	**98.53**	**93.43**

strates the effectiveness of our novel loss. 4) Histogram loss performs poorly, which demonstrates the effects of our constraint between teacher and student distributions.

Moreover, different to the prior hard mining methods [17, 26, 28] where the hard samples are mined based on the loss values during the training process, we *pre-define* hard samples according to human prior. Penalizing *individual* samples or triplets as in previous hard mining methods does not leverage sufficient contextual insight of the overall distribution. DDL minimizes the difference of *global* similarity distributions between the easy and hard samples, which is more robust for tackling hard samples and against the noisy samples. The word "global" means our method leverages sufficient contextual insight of the overall distribution in a mini-batch, rather than focusing on a sample.

Fig. 7 illustrates the estimated similarity distributions of various SotA methods. To quantify the differences among these methods, we introduce two statistics for evaluation, the expectation margin and histogram intersection (*i.e.*, $\sum_{r=1}^{R} \min(h_r^+, h_r^-)$) between the two distributions from positive and negative pairs. Typically, smaller histogram intersection and larger expectation margin indicate better verification/identification performance, since it means more discriminative embeddings are learned [37]. Our DDL achieves the closest statistics to the teacher distribution, and thus obtains the best performance.

Pose on CFP-FP and CPLFW. We compare our method with SotA pose-invariant methods [2, 5, 24, 35, 48] and generic solutions [3, 6, 18, 41–43]. Since VGGFace2 includes comprehensive pose variations, we use it to pre-train a ResNet100 with Arcface. Next, we construct teacher and student distributions to finetune the model with our loss. From Tab. 3, we can see that: 1) Our Arcface re-implementations achieve comparable results against the official version, with similar results on LFW and CFP-FP, as well as better performance on CPLFW.

[1] Results are from the official model: https://github.com/deepinsight/insightface, which is trained on MS1M and adopts ResNet100 as the backbone.

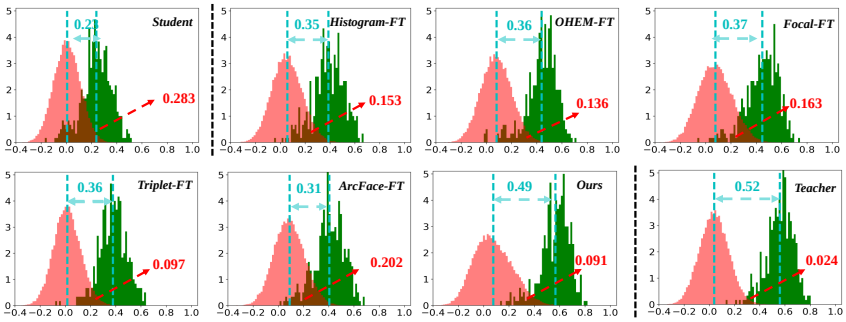

Fig. 7: **Illustrations of similarity distributions of different SotA methods**, which are all pre-trained by CASIA with ResNet50 and then finetuned on SCface. The leftmost and rightmost are the student and teacher distributions estimated from a pre-trained Arcface model on d_1 and d_3 settings, respectively. The similarity distributions in the middle are obtained by various methods finetuned on SCface. The red number indicates the *histogram intersection* between the estimated similarity distributions from the positive and negative pairs.

Arcface is also much better than other methods, including those pose-invariant face recognition methods. 2) Our method achieves the best performance on both pose benchmarks, while also maintaining the performance on LFW (*i.e.*, 99.68% vs. 99.62%).

Note that when using the model pre-trained on MS1M, and finetuning it with easy/hard samples from VGGFace2, our method can further push the performance to a higher level (**99.06**% on CFP-FP and **94.20**% on CPLFW), which is the *first* method that exceeds 99.0% on CFP-FP and 94.0% on CPLFW using images cropped by MTCNN. Besides, we also train our DDL on the smaller training set CASIA with a smaller backbone ResNet50. Again, our DDL outperforms the competitors. Please refer to our supplementary material for details.

Large-Scale Benchmarks: IJB-B and IJB-C. On IJB-B/C datasets, we employ VGGFace2 with ResNet50 for a fair comparison with recent methods. We first construct the teacher and student distributions according to the pose of each image, and then follow the testing protocol in [6] to take the *average of the image features* as the corresponding template representation without bells and whistles. Tabs. 4 and 5 show the 1:1 verification and 1:N identification comparisons with the recent SotA methods, respectively. Note that our method is *not* a set-based face recognition method, and the experiments on these two datasets are just to prove that our DDL can obtain more discriminate features than generic methods like Arcface, even on all-variations-included datasets. Please refer to our supplementary material for the detailed analysis.

Comparisons with SotA KD Methods. We further conduct fair comparisons between our DDL and the recent SotA KD/self-distillation methods, *i.e.*, SP [36] and RKD [22]. Note that since both SP and RKD have not reported SOTA results on face recognition tasks, we re-implement the two methods under the same experimental setting on VGGFace2, using their officially released

Table 4: 1:1 **verification TAR** on the IJB-B and IJB-C datasets. All methods are trained on VGGFace2 with ResNet50.

Methods (%)	IJB-B			IJB-C		
	FAR=$1e{-}5$	FAR=$1e{-}4$	FAR=$1e{-}3$	FAR=$1e{-}5$	FAR=$1e{-}4$	FAR=$1e{-}3$
VGGFace2 [3]	67.1	80.0	88.7	74.7	84.1	90.9
MN [45]	70.8	83.1	90.9	77.1	86.2	92.7
DCN [44]	–	84.9	93.7	–	88.5	94.7
Arcface [6]	80.5	89.9	94.5	86.1	92.1	96.0
SP [36]	79.4	89.8	94.9	85.9	92.3	96.2
RKD [22]	78.4	89.6	94.7	85.5	92.1	96.1
Ours	**83.4**	**90.7**	**95.2**	**88.4**	**93.1**	**96.3**

Table 5: **1:N (mixed media) Identification** on IJB-B/C. All methods are trained on VGGFace2 with ResNet50. VGGFace2 is cited from the paper, and Arcface is from its official released model.

Methods (%)	IJB-B				IJB-C			
	FPIR=0.01	FPIR=0.1	Rank 1	Rank 5	FPIR=0.01	FPIR=0.1	Rank 1	Rank 5
VGGFace2 [3]	70.6	83.9	90.1	94.5	74.6	84.2	91.2	94.9
Arcface [6]	73.1	88.2	93.6	96.5	79.6	89.5	94.8	96.9
SP [36]	72.4	88.0	93.8	96.6	79.9	89.5	94.7	97.0
RKD [22]	70.6	87.6	93.4	96.5	79.3	89.1	94.6	96.9
Ours	**76.3**	**89.5**	**93.9**	**96.6**	**85.4**	**91.1**	**95.4**	**97.2**

code. Specifically, we first train a ResNet50 with Arcface on VGGFace2 as the teacher model, and then train a student ResNet50 via combining the knowledge distillation method (*e.g.*, SP or RKD) and Arcface loss under the guidance of the teacher model. As in Tabs. 4 and 5, our DDL outperforms the SotA KD/self-distillation methods, which achieve similar results to vanilla Arcface.

5 Conclusion

In this paper, we propose a novel framework Distribution Distillation Loss (DDL) to improve various variation-*specific* tasks, which comes from the observations that state-of-the-art methods (*e.g.*, Arcface) witness significant performance gaps between easy and hard samples. The key idea of our method is to construct a teacher and a student distribution from easy and hard samples, respectively. Then, the proposed loss drives the student distribution to approximate the teacher distribution to reduce the overlap between the positive and negative pairs. Extensive experiments demonstrate the effectiveness of our DDL on a wide range of recognition tasks compared to the state-of-the-art face recognition methods. In subsequent work, we can try to extend our method to multiple teacher-student distribution pairs for the corresponding task respectively.

References

1. et al., M.A.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), http://tensorflow.org/, software available from tensorflow.org
2. Cao, K., Rong, Y., Li, C., Tang, X., Change Loy, C.: Pose-robust face recognition via deep residual equivariant mapping. In: CVPR. pp. 5187–5196 (2018)
3. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: FG. pp. 67–74. IEEE (2018)
4. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. In: CVPR. pp. 2492–2501 (2018)
5. Deng, J., Cheng, S., Xue, N., Zhou, Y., Zafeiriou, S.: Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In: CVPR. pp. 7093–7102 (2018)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR. pp. 4690–4699 (2019)
7. Deng, J., Zhou, Y., Zafeiriou, S.: Marginal loss for deep face recognition. In: CVPR Workshops. pp. 60–68 (2017)
8. Gong, S., Liu, X., Jain, A.: Jointly de-biasing face recognition and demographic attribute estimation. In: ECCV (2020)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. pp. 2672–2680 (2014)
10. Grgic, M., Delac, K., Grgic, S.: Scface–surveillance cameras face database. Multimedia Tools and Applications **51**(3), 863–879 (2011)
11. Hennings-Yeomans, P.H., Baker, S., Kumar, B.V.: Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In: CVPR. pp. 1–8. IEEE (2008)
12. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Workshop (2014)
13. 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: CVPR (2020)
14. Huang, Z., Wang, N.: Like what you like: Knowledge distill via neuron selectivity transfer. arXiv:1707.01219v2 (2017)
15. Huang, Z., Shan, S., Wang, R., Zhang, H., Lao, S., Kuerban, A., Chen, X.: A benchmark and comparative study of video-based face recognition on cox face database. IEEE Transactions on Image Processing **24**(12), 5967–5981 (2015)
16. Lei, Z., Ahonen, T., Pietikäinen, M., Li, S.Z.: Local frequency descriptor for low-resolution face recognition. In: FG. pp. 161–166. IEEE (2011)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)
18. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: CVPR. pp. 212–220 (2017)
19. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: ICML. vol. 2, p. 7 (2016)
20. Lu, Z., Jiang, X., Kot, A.: Deep coupled resnet for low-resolution face recognition. IEEE Signal Processing Letters **25**(4), 526–530 (2018)
21. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(Nov), 2579–2605 (2008)

22. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3967–3976 (2019)
23. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: BMVC. vol. 1, p. 6 (2015)
24. Peng, X., Yu, X., Sohn, K., Metaxas, D.N., Chandraker, M.: Reconstruction-based disentanglement for pose-invariant face recognition. In: ICCV. pp. 1623–1632 (2017)
25. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. In: CVPR Workshops. pp. 2074–2083 (2018)
26. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. pp. 815–823 (2015)
27. Shekhar, S., Patel, V.M., Chellappa, R.: Synthesis-based recognition of low resolution faces. In: IJCB. pp. 1–6. IEEE (2011)
28. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: CVPR. pp. 761–769 (2016)
29. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: NIPS. pp. 1988–1996 (2014)
30. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: CVPR. pp. 1891–1898 (2014)
31. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: CVPR. pp. 3147–3155 (2017)
32. Tai, Y., Yang, J., Liu, X., Xu, C.: Memnet: A persistent memory network for image restoration. In: ICCV. pp. 4539–4547 (2017)
33. Tai, Y., Yang, J., Zhang, Y., Luo, L., Qian, J., Chen, Y.: Face recognition with pose variations and misalignment via orthogonal procrustes regression. IEEE Transactions on Image Processing **25**(6), 2673–2683 (2016)
34. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: CVPR. pp. 1701–1708 (2014)
35. Tran, L., Yin, X., Liu, X.: Disentangled representation learning gan for pose-invariant face recognition. In: CVPR. pp. 1415–1424 (2017)
36. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1365–1374 (2019)
37. Ustinova, E., Lempitsky, V.: Learning deep embeddings with histogram loss. In: NIPS. pp. 4170–4178 (2016)
38. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Processing Letters **25**(7), 926–930 (2018)
39. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: l 2 hypersphere embedding for face verification. In: ACMMM. pp. 1041–1049. ACM (2017)
40. 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: CVPR. pp. 5265–5274 (2018)
41. Wang, X., Wang, S., Wang, J., Shi, H., Mei, T.: Co-mining: Deep face recognition with noisy labels. In: ICCV. pp. 9358–9367 (2019)
42. Wang, X., Wang, S., Zhang, S., Fu, T., Shi, H., Mei, T.: Support vector guided softmax loss for face recognition. arXiv:1812.11317 (2018)
43. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: ECCV. pp. 499–515. Springer (2016)
44. Xie, W., Shen, L., Zisserman, A.: Comparator networks. In: ECCV. pp. 782–797 (2018)
45. Xie, W., Zisserman, A.: Multicolumn networks for face recognition. In: BMVC (2018)

46. Yang, F., Yang, W., Gao, R., Liao, Q.: Discriminative multidimensional scaling for low-resolution face recognition. IEEE Signal Processing Letters **25**(3), 388–392 (2017)
47. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv:1411.7923 (2014)
48. Yin, X., Liu, X.: Multi-task convolutional neural network for pose-invariant face recognition. IEEE Transactions on Image Processing **27**(2), 964–975 (2017)
49. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Towards large-pose face frontalization in the wild. In: ICCV. pp. 3990–3999 (2017)
50. Zhang, K., Zhang, Z., Cheng, C.W., Hsu, W.H., Qiao, Y., Liu, W., Zhang, T.: Super-identity convolutional neural network for face hallucination. In: ECCV. pp. 183–198 (2018)
51. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
52. Zhang, X., Fang, Z., Wen, Y., Li, Z., Qiao, Y.: Range loss for deep face recognition with long-tailed training data. In: ICCV. pp. 5409–5418 (2017)
53. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: CVPR. pp. 4320–4328 (2018)
54. Zhao, J., Xiong, L., Cheng, Y., Cheng, Y., Li, J., Zhou, L., Xu, Y., Karlekar, J., Pranata, S., Shen, S., et al.: 3d-aided deep pose-invariant face recognition. In: IJCAI. vol. 2, p. 11 (2018)
55. Zou, W.W., Yuen, P.C.: Very low resolution face recognition problem. IEEE Transactions on Image Processing **21**(1), 327–340 (2011)