Learning an Evolutionary Embedding via Massive Knowledge Distillation

Xiang Wu[1,2], Ran He[1,2,3,4*], Yibo Hu[1,2], Zhenan Sun[1,2,4]
[1]Center for Research on Intelligent Perception and Computing, CASIA,
[2]National Laboratory of Pattern Recognition, CASIA,
[3]Center for Excellence in Brain Science and Intelligence Technology, CASIA,
[4]University of Chinese Academy of Sciences
alfredxiangwu@gmail.com, {rhe, znsun}@nlpr.ia.ac.cn, yibo.hu@cripac.ia.ac.cn

Abstract

Knowledge distillation methods aim at transferring knowledge from a large powerful teacher network to a small compact student one. These methods often focus on close-set classification problems and matching features between teacher and student networks from a single sample. However, many real-world classification problems are open-set. This paper proposes an Evolutionary Embedding Learning (EEL) framework to learn a fast and accurate student network for open-set problems via massive knowledge distillation. First, we revisit the formulation of canonical knowledge distillation and make it suitable for the open-set problems with massive classes. Second, by introducing an angular constraint, a novel correlated embedding loss (CEL) is proposed to match embedding spaces between the teacher and student network from a global perspective. Lastly, we propose a simple yet effective paradigm towards a fast and accurate student network development for knowledge distillation. We show the possibility to implement an accelerated student network without sacrificing accuracy, compared with its teacher network. The experimental results are quite encouraging. EEL achieves better performance with other state-of-the-art methods for various large-scale open-set problems, including face recognition, vehicle re-identification and person re-identification.

1. Introduction

With the development of deep learning [31, 32], various deep neural networks have made great improvements on different computer vision tasks. Obviously, given enough data, the network tends to be deeper and wider, which achieves better performance. However, deploying such a larger network requires much more latency and computational resources in practice. Some researchers pay attention to the acceleration and compression of neural networks. Existing efforts can be categorized into three aspects: network pruning [12, 42, 27, 40, 11] and quantization [5, 48], computational efficient network development [17, 22, 71, 51, 79] and knowledge distillation [1, 50, 16, 67, 69, 4]. Network pruning prunes the neurons or weights with low responses based on some criteria. Network quantization utilizes low precision computation to compress and accelerate neural networks. However, the acceleration of these methods highly depends on the dedicated implementations. By considering network development, some efficient architectures with low computation and few parameters are proposed for different computer vision tasks. Meanwhile, the speed also depends on specific implementations or devices. For example, depthwise convolution used in MobileNet [17, 51] and ShuffleNet [71] is fast using a GPU, but not equally efficient due to fewer cores in a CPU.

In contrast, Hinton *et al.* [16] proposed Knowledge Distillation (KD) that directly trains a small student network under the supervision of large teacher networks. Compared with other methods, KD can achieve considerable acceleration and compression without bells and whistles. Based on this idea, many methods [1, 50, 67, 69] employ different forms of knowledge to constrain the student networks. Current KD methods often focus on image classification in a closed-set setting. For this protocol, all the testing identities are the same as ones in training set. However, in practical scenario, the testing identities are often disjointed from the training set. Therefore, it is impossible to classify the testing samples to known identities in training set. This can be defined as open-set problem. For this protocol, the main challenging issue is that we need to map samples to a discriminative feature space rather than achieve perfect classification accuracy. Obviously, it could not be addressed well in the previous KD methods. Another limitation is that the existing KD methods such as attention transfer [69], neuron selectivity transfer [21] and DarkRank [4] only measure the information from a single sample between a teacher and a student network. Obviously, the embedding

space encoded by the teacher network is a valuable knowledge resource, especially for open-set problems. Relational Knowledge Distillation (RKD) [44] proposes distance-wise and angle-wise distillation, penalizing structural differences in relations. Instance Relationship Graph (IRG) [38] exploits knowledge including instance features, instance relationships, and feature space transformation across layers. However, it is hard to feed all the samples into a mini-batch during training. Therefore, a mini-batch SGD still results in providing local structure information (supported by several samples), rather than global one. Besides, it may also lead to a complex batch sampling strategies, especially for large-scale training set.

Moreover, although knowledge is well defined by various forms [67, 69, 4], there is still a performance gap between student networks and teacher networks. Thus there are three questions to be further discussed:

- How can knowledge distillation be applied for open-set problems including face recognition, vehicle re-identification and person re-identification?
- How can the global information of the embedding space (but not subspace supported by several samples) from the teacher network be effectively utilized?
- Is it possible to implement a student network that can be accelerated and also achieve or even outperform the accuracy of the teacher network?

For the first question, we revisit knowledge distillation [16] and reformulate it as a term of matching logits between the teacher and student network. Different from the original KD, the temperature parameter τ is applied to logits in both the soft target and original softmax aimed at the open-set problem. Considering the massive but noisy labeled data in the training set [64], we employ the least absolute deviation regression instead of the least square error for the logit matching term. In this way, the modified knowledge distillation can robustly transfer knowledge from the teacher network, and also achieve discriminative embeddings with intra-class affinity and inter-class separability.

Considering the second question, benefiting from an angular constraint, we propose a novel loss function, named Correlated Embedding Loss (CEL), to match embedding spaces between teacher and student networks. We encode the well-separated embedding space by the class centers of deep features from the teacher network. During training, CEL ensures that embeddings from the student network are close to their own corresponding class center and also far away from other class centers in the encoded embedding space. Therefore, CEL can be aware of the global structure of the embedding space from the teacher network.

Regarding the third question, we should clarify that there are no absolute relationships between computational speed and the number of parameters for any network. A large network with many parameters may not mean slow inference and vice versa. Therefore, in this paper, we focus on model acceleration rather than model compression. By exploring the computational-bound and memory-bound of basic operations in Convolution Neural Network (CNN), we present a paradigm for implementing a fast and accurate student network (compared with its teacher network) in knowledge distillation.

As stated above, in this paper, we propose an Evolutionary Embedding Learning (EEL) framework, which aims to achieve acceleration of the student network without sacrificing accuracy, compared with its teacher, for large-scale open-set problems. The main contributions are as follows:

- We revisit and modify the formulation of the original knowledge distillation for open-set problems. The modified knowledge distillation can efficiently transfer massive knowledge from a teacher network to a student one for open-set problems.
- By introducing an angular constraint, Correlated Embedding Loss (CEL) is proposed to match the embedding spaces between the teacher network and the student one for knowledge distillation from a global perspective.
- By investigating the properties of widely used basic blocks in convolution neural network, we propose a simple yet effective paradigm towards fast and accurate student network development for knowledge distillation.
- We evaluate our approach on various open-set problems with massive classes, including face recognition, vehicle re-identification and person re-identification, and obtain appealing results compared with other state-of-the-art methods.

The rest is organized as follows. We briefly review some related works for knowledge distillation and representation learning in Section 2. In Section 3, we present the details of the proposed Evolutionary Embedding Learning (EEL) framework. Section 4 provides the algorithmic analyses and experimental results on different open-set tasks. Finally, we conclude the paper in Section 5.

2. Related Work

2.1. Knowledge Distillation

Ba *et al.* [1] demonstrate that the shallow feed-forward network can learn the complex functions by mimic learning with regressing logits. Hinton *et al.* propose Knowledge Distillation (KD) [16] to implement information transfer from a large teacher network to a small student one. Fitnet [50] directly matches the embeddings of each sample between the teacher and student network to improve performance. Attention transfer [69] extends Fitnet from embedding matching to attention map matching for different levels of feature maps. Furthermore, Yim *et al.* [67] introduced

the Flow of Solution Procedure (FSP) by matching Gramian matrix across layers for fast optimization and transfer learning. They claimed that FSP can reflect the data flow of how the teachers solve the problems. Instead of knowledge transfer by feed-forward information, Czarnecki *et al.* [6] exploited gradient transfer using Sobolev training. All the above studies focus on close-set problems.

Deep mutual learning [72] learns collaboratively and teaches each other between two networks to boost performance for both close-set and open-set problems. Chen *et al.* [4] propose DarkRank for open-set problems and obtained comparable performance on person re-identification. Luo *et al.* [41] propose a neuron selection method to compress models by the teacher network. However, there are still performance gaps between student and teacher networks.

2.2. Deep Embedding Learning

Metric learning is widely used to obtain the discriminative embedding space in deep learning. Siamese Network [2] is first applied for signature verification to find a discriminative embedding space. Contrastive loss [57] and triplet loss [52, 8, 70] are proposed for face recognition, in order to optimize the distances between positive and negative samples. Benefiting from triplet loss, PDDM [18] proposes quadruplets to enforce the constraints between positive and negative pairs. However, in practice, the performance of metric learning is limited to the effectiveness of positive and negative pairs. Facenet [52] proposes a semi-hard triplet sampling method to learn embedding space efficiently, while lifted structure embedding [56] and n-pair loss [53] define sampling strategies on all the images in each batch during training. HDC [68] learns an ensemble to models with different complexities and finds hard examples adaptively. Beyond these local sampling methods defined on one batch, Song *et al.* [55] propose a new metric learning scheme which is aware of the global structure of the embedding space.

Recently, the angular constraints [34, 33, 62] have been introduced into deep embedding learning. Since the intuition of features and weights in softmax can be factorized into amplitude and angular with cosine similarity, large margin softmax [34] encourages the intra-class compactness and inter-class separability in training. A-softmax [33] further improves the angular constraint by normalizing the features and weights onto a hypersphere manifold. Angular loss [62] also enhanced the convergence and performance of triplet loss by constraining the angle in triplet triangles. Besides, normalization also contributes to discriminative embeddings. NormFace [61] and L2-softmax [47] reformulate the sofmtax and explain the necessity of normalization in embedding learning, respectively.

2.3. Efficient Network Architectures

It is important to strike an optimal balance between speed and accuracy and recently, many studies have focused on the efficient convolution neural network architecture developments [17, 22, 71, 51, 79] for mobile and resource constrained environments.

Considering manual architecture search, SqueezeNet [22] achieves an AlexNet-Level accuracy by carefully designing the fire module, including the squeeze and expand layers. The MobileNet [17] based on streamlined architecture employs depthwise separable convolutions, which theoretically reduces the computational cost with little loss of accuracy. MobileNetV2 [51] proposes an inverted residual structure with linear bottleneck in shortcut connections. Compared with MobileNet [17], MobileNetV2 further improves the state-of-the-art performance of mobile models and significantly reduce the memory footprint. Moreover, shuffleNet [71] propose a novel channel shuffle operation to help the information flowing across feature channels in the pointwise group convolutions.

Recently, Neural Architecture Search network (NASNet) [79] brings optimization methods, including reinforcement learning to architectural search. NASNet trains the RNN with reinforcement learning to maximize the expected accuracy on a validation set. In this way, [79] designs novel convolution neural network architectures that are better than most human-invented architectures on ImageNet [7].

Both manual architectures and neural architecture search networks rely on depthwise separate convolutions to implement accelerated models. However, there is not yet an efficient implementation of depthwise separate convolution in most deep learning platforms, especially on devices with few computational cores, such as CPUs.

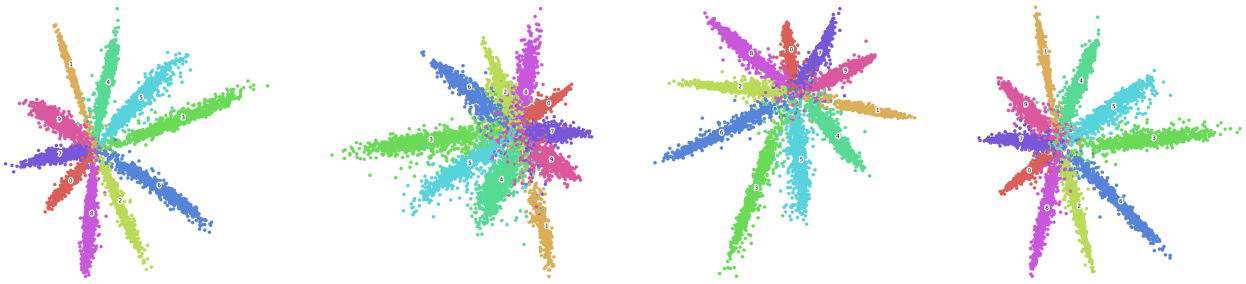
3. Our Approach

3.1. Knowledge Distillation with Massive Labels

Generally, logit can be defined as $z = W^\top x + b$, where W and b are the weight and bias, respectively, and the probability of i-th class can be denoted as $p_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$. Here, let $z^{\mathrm{T}}, z^{\mathrm{S}}, p^{\mathrm{T}}$ and p^{S} denote the logits and probabilities from the teacher and student network, respectively. According to KD [16] that can transfer knowledge from one network to another, we can formulate the loss function as

$$\begin{aligned} L_{\mathrm{KD}} &= -\tfrac{1}{N}\sum \left(\log p_i^{\mathrm{S}} + \alpha \textstyle\sum_k p_k^{\mathrm{T}}(\tau)\log p_k^{\mathrm{S}}(\tau)\right) \\ &= -\tfrac{1}{N}\sum \left(\log \tfrac{\exp(z_i^{\mathrm{S}})}{\sum_j \exp(z_j^{\mathrm{S}})} + \alpha \textstyle\sum_k \tfrac{\exp(z_k^{\mathrm{T}}/\tau)}{\sum_j \exp(z_j^{\mathrm{T}}/\tau)} \log \tfrac{\exp(z_k^{\mathrm{S}}/\tau)}{\sum_j \exp(z_j^{\mathrm{S}}/\tau)}\right) \end{aligned} \tag{1}$$

where the first term is the cross-entropy loss with one-hot labels for the student network, and the second term is the cross-entropy loss with soft targets generated by the



(a) Teacher network. (b) KD (c) MKD (d) EEL

Figure 1. Visualization comparison on MNIST for the learned embeddings of the teacher and the student network with original KD and MKD, respectively. (a) the teacher network with the softmax loss; (b) the student network under the supervision of the teacher network with the original KD loss ($\tau = 10$); (c) the student network under the supervision of the teacher network with the MKD loss ($\tau = 10$). (d) the student network under the supervision of the teacher network with the EEL loss($\tau = 10$, $\lambda = 1$).

teacher network. In KD, the soft target is defined by the logit z_i^{T}/τ from the teacher network. Compared with the original one-hot label, the soft target provides continuous signals, especially for massive labels. The new continuous n-hot label contains the intra-class and inter-class similarities from the teacher network, which can be treated as the knowledge transfer from the teacher network and as an extra regularizer for optimization.

Furthermore, τ is a temperature parameter and using a higher value for τ produces a softer probability distribution over the classes. When τ is a higher temperature than the magnitude of the logits, Eq. (1) can be approximated in [16] as

$$L_{\mathrm{KD}} = \frac{1}{N}\sum\left(-\log\frac{\exp(z_i^{\mathrm{S}})}{\sum_j \exp(z_j^{\mathrm{S}})} + \frac{\alpha}{2}\sum_k(\frac{z_k^{\mathrm{T}}}{\tau} - \frac{z_k^{\mathrm{S}}}{\tau})^2\right) \tag{2}$$

where the scale α is set to τ^2 to balance the magnitudes of the gradients. This formulation can be treated exactly as the matching logits in [1]. However, both KD [16] and Logits [1] are proposed for close-set classification tasks. When considering open-set problems such as face recognition, although the training procedure can be treated as a fine-grained classification with massive labels, the intention is instead to obtain discriminative embeddings. Therefore, we make a simple yet effective modification on Eq. (2) and define modified knowledge distillation as follows:

$$L_{\text{L2-MKD}} = \frac{1}{N}\sum\left(-\log\frac{\exp(z_i^{\mathrm{S}}/\tau)}{\sum_j \exp(z_j^{\mathrm{S}}/\tau)} + \frac{\alpha}{2}\sum_k(\frac{z_k^{\mathrm{T}}}{\tau} - \frac{z_k^{\mathrm{S}}}{\tau})^2\right) \tag{3}$$

Different from the original KD [16], the temperature parameter τ is also applied to the softmax loss for the student network. By selecting a τ greater than 1, Eq. (3) can produce higher gradients for the well-separated samples, which can shrink the intra-class variance and enlarge the inter-class distance. In this way, the modified knowledge distillation is suitable for the open-set problem.

Fig. 1 presents the qualitative comparison for the distributions of the embeddings on MNIST [26]. It is obvious that the embeddings in Fig. 1(d) are more compact and more similar with the embeddings from the teacher net (Fig. 1(a)) than those in Fig. 1(b). The phenomenon indicates that the modified KD loss has better ability to capture the properties of the embeddings from the teacher network.

Moreover, since open-set problems such as face recognition contain massive classes with noisy labels [64], we reformulate the modified knowledge distillation loss functions as follows:

$$L_{\text{L1-MKD}} = \frac{1}{N}\sum\left(-\log\frac{\exp(z_i^{\mathrm{S}}/\tau)}{\sum_j \exp(z_j^{\mathrm{S}}/\tau)} + \alpha\sum_k|\frac{z_k^{\mathrm{T}}}{\tau} - \frac{z_k^{\mathrm{S}}}{\tau}|\right) \tag{4}$$

Similarly, to balance the gradients' magnitudes, we also set scale α to τ^2 in practice. In Eq. (4), we use Least Absolute Deviation (LAD) instead of Least Square Error (LSE) for knowledge transfer. The reasons regarding the modification are as follows. On the one hand, LAD tends to be a higher response than LSE when the value of the difference between z^{T} and z^{S} is located in $[-1, 1]$. Since the training set for open-set problems such as face recognition often contains thousands of classes, employing LAD instead of LSE can produce higher gradients when the network is roughly converged, which is beneficial to avoid saddle points during optimization. On the other hand, considering the noisy labeled data in the training set, LAD is more robust than LSE for the outliers.

3.2. Correlated Embedding Loss

3.2.1 Revisiting N-pair Loss

Since N-pair loss [53] aims to optimize the $(N+1)$-tuples $\{x, x^+, x_1, ..., x_{N-1}\}$, where x^+ is a positive example to x and $\{x_i\}_{i=1}^{N-1}$ are negative examples, the loss can be defined

as follows:

$$L = -\frac{1}{N}\sum \log \frac{\exp(f^\top f^+)}{\exp(f^\top f^+) + \sum_{i=1}^{N-1}\exp(f^\top f_i)} \quad (5)$$

where $f(\cdot) \in \mathbb{R}^d$ is an embedding defined by the convolution neural network. Obviously, N-pair loss can shorten the distance between an embedding f and its positive sample f^+, while it can enlarge the distances between it and multiple negative samples $\{f_i\}_{i=1}^{N-1}$. Specially, if we treat f^+ and $\{f_i\}_{i=1}^{N-1}$ as the weight vectors for each class in a classifier (often defined by the last fully-connected layer in CNN), Eq. (5) is similar to the softmax loss without the bias term.

3.2.2 Correlated Embedding Loss

The intention of correlated embedding loss is to match the embedding spaces between the teacher and student networks. Different from previous works [50, 4] that directly match the features from teacher and student networks for each sample, we propose a flexible way to measure the two embedding spaces.

Let $c_j \in$ denote the center of j-th class for deep features. Therefore, we can define the embedding space by a set of centers $\{c_j\}_{j=1}^K$, where K is the number of classes. Hence, we denote $\{c_j^{\mathrm{T}}\}_{j=1}^K$ to represent the embedding space of the teacher network, and the correlated embedding loss can be measured by

$$L_{\mathrm{CEL}} = -\frac{1}{N}\sum_i \log \frac{\exp((f_i^{\mathrm{S}})^\top c_{y_i}^{\mathrm{T}})}{\exp((f_i^{\mathrm{S}})^\top c_{y_i}^{\mathrm{T}}) + \sum_{j\neq y_i}\exp((f_i^{\mathrm{S}})^\top c_j^{\mathrm{T}})} \quad (6)$$

where f_i^{S} is an embedding from the student network with its label y_i. Eq. (6) can be treated as a special form of N-pair loss, where we use the set of centers $c_{y_i}^{\mathrm{T}}$ and $\{c_j^{\mathrm{T}}\}_{j\neq y_i}^{K-1}$ from the teacher network instead of the positive sample f^+ and the negative samples $\{f_i\}_{i=1}^{N-1}$, respectively.

The $\mathcal{G}_j(f_i^{\mathrm{S}}) = (f_i^{\mathrm{S}})^\top c_j^{\mathrm{T}}$ is defined as the correlated similarity, and it can ensure that the distribution of features from the student network is close to the distribution of the teacher network embeddings. From the angular perspective, we can present the formulation of the correlated similarity as follows:

$$\mathcal{G}_j(f_i^{\mathrm{S}}) = \|f_i^{\mathrm{S}}\|\|c_j^{\mathrm{T}}\|\cos(\theta_{j,i}) \quad (7)$$

in which $\theta_{j,i}(0 \leq \theta_{j,i} \leq \pi)$ is the angle between a feature f_i^{S} from the student network and a center c_j^{T} from the teacher network. Here, we set $\|c_j^{\mathrm{T}}\| = 1$. Therefore, Eq. (7) can be written as

$$\mathcal{G}_j(f_i^{\mathrm{S}}) = \|f_i^{\mathrm{S}}\|\cos(\theta_{j,i}) \quad (8)$$

Then, we can reformulate Eq. (6) as

$$L_{\mathrm{CEL}} = -\frac{1}{N}\sum_i \log \frac{\exp(\|f_i^{\mathrm{S}}\|\cos(\theta_{y_i,i}))}{\exp(\|f_i^{\mathrm{S}}\|\cos(\theta_{y_i,i})) + \sum_{j\neq y_i}\exp(\|f_i^{\mathrm{S}}\|\cos(\theta_{j,i}))} \quad (9)$$

By further introducing an angular margin m to Eq. (9), similar to the L-softmax [34] and A-softmax [33], the correlated embedding loss can be written as

$$L_{\mathrm{CEL}} = -\frac{1}{N}\sum_i \log \frac{\exp(\|f_i^{\mathrm{S}}\|\psi(\theta_{y_i,i}))}{\exp(\|f_i^{\mathrm{S}}\|\psi(\theta_{y_i,i})) + \sum_{j\neq y_i}\exp(\|f_i^{\mathrm{S}}\|\cos(\theta_{j,i}))} \quad (10)$$

where

$$\psi(\theta) = (-1)^k\cos(m\theta) - 2k, \theta \in \left[\frac{k\pi}{m}, \frac{(k+1)\pi}{m}\right] \quad (11)$$

Since Eq. (10) is similar to the softmax, it is easy to compute the forward and backward processes. The only difference between the softmax and the correlated embedding loss lies in $\|f_i^{\mathrm{S}}\|\psi(\theta_{y_i,i})$.

If we fix the centers $\{c_j^{\mathrm{T}}\}_{j=1}^K$ from the teacher network and strictly match the embeddings from the student network with their corresponding center, it is hard to optimize and may lead to the collapse of the convergence. To address this problem, we employ a simple way to update the centers using the features of the corresponding classes in each mini-batch as follows:

$$(c_{y_i}^{\mathrm{T}})^{(t+1)} = (c_{y_i}^{\mathrm{T}})^{(t)} + \frac{1}{N_{y_i}}\sum_{j=y_i}\frac{f_j^{\mathrm{S}}}{\|f_j^{\mathrm{S}}\|} \quad (12)$$

3.2.3 Relations with A-softmax

The A-softmax [33] loss imposes discriminative power via the angular margin for deep embedding learning. It incorporates the angular margin in the original softmax loss to restrict the intra-class variance and enlarge the inter-class margins, thus leading to a more discriminative embedding space. Correlated embedding loss (CEL) introduces the angular constraint to match embedding spaces between the teacher network and student one. The main difference is that CEL incorporates two embedding spaces and optimizes their distribution distance using the angular margin, while the A-softmax loss aims to obtain one discriminative embedding space instead.

3.2.4 Relations with Center Loss

By minimizing the distances between the deep features and their corresponding class centers, center loss [63] obtains the discriminative features for deep embedding learning. The main differences between CEL and center loss are as follows: First, the centers in CEL are obtained from teacher network, while center loss is initialized randomly and learned by back-propagation. Second, center loss only considers the intra-class constraints when optimizing, but we utilize both the intra-class and inter-class information for each embedding in CEL.

Layer	LightCNN-9	LightCNN-9-Fast	LightCNN-29	LightCNN-29-Fast
Conv1.x	$[5\times5, 96]\times1$	$[5\times5, 96]\times1$, S2	$[5\times5, 96]\times1$	$[5\times5, 96]\times1$, S2
Pool1	2×2 Max+Ave, S2			
Conv2.x	$\begin{bmatrix}1\times1, 96\\3\times3, 192\end{bmatrix}\times1$		$\begin{bmatrix}3\times3, 96\\3\times3, 96\end{bmatrix}\times1$ $\begin{bmatrix}1\times1, 96\\3\times3, 192\end{bmatrix}\times1$	
Pool2	2×2 Max+Ave, S2			
Conv3.x	$\begin{bmatrix}1\times1, 192\\3\times3, 384\end{bmatrix}\times1$		$\begin{bmatrix}3\times3, 192\\3\times3, 192\end{bmatrix}\times2$ $\begin{bmatrix}1\times1, 192\\3\times3, 384\end{bmatrix}\times1$	
Pool3	2×2 Max+Ave, S2			
Conv4.x	$\begin{bmatrix}1\times1, 384\\3\times3, 256\end{bmatrix}\times1$		$\begin{bmatrix}3\times3, 384\\3\times3, 384\end{bmatrix}\times3$ $\begin{bmatrix}1\times1, 384\\3\times3, 256\end{bmatrix}\times1$	
Conv5.x	$\begin{bmatrix}1\times1, 256\\3\times3, 256\end{bmatrix}\times1$		$\begin{bmatrix}3\times3, 256\\3\times3, 256\end{bmatrix}\times4$ $\begin{bmatrix}1\times1, 256\\3\times3, 256\end{bmatrix}\times1$	
Pool4	2×2 Max+Ave, S2			
Feature	fc-256	Global Ave Pool	fc-256	Global Ave Pool
#Params	5.5 M	1.4 M	10.4 M	8.6 M
#FLOPs	976 M	243 M	3.6 G	922 M

Table 1. The network architectures of LightCNN-9, LightCNN-9-Fast, LightCNN-29 and LightCNN-29-Fast for face recognition. Conv1.x∼Conv5.x denote the convolution units that contain multiple convolution layers. The double-row brackets with the same convolution kernels denote residual units. E.g. $[3\times3, 96]\times2$ means 2 cascaded convolution layers with 96 filters of size 3×3. S2 denotes the stride 2. All the outputs of convolution layers are activated by mfm operation [64]. Pool1∼Pool4 represent the down-sampling units. "Max+Ave" denotes a combination of max-pooling and ave-pooling operation. The "fc-256" means a 256-d fully-connected layer.

3.3. Student Network Development

3.3.1 Development Paradigm

In this section, we investigate a simple yet effective paradigm towards designing fast and accurate student network architectures. First, we present two properties:

- Similar network structures with almost the same number of parameters have similar capacities, thus leading to comparable performance.
- The shallow convolution layers (especially the first one) capture low-level features.

Based on the above properties, we propose our paradigm towards designing a fast and accurate student network:

- We initialize the student network with the same architecture of teacher one.
- For acceleration, we increase the stride of the first convolution layer.

Given the feature map $w\times h\times c_{\text{in}}$ and a convolution kernel $c_{\text{in}}\times k\times k\times c_{\text{out}}$ with stride s and padding p, the size of the output feature map is $(\frac{w-k+2p}{s}+1)\times(\frac{w-k+2p}{s}+1)\times c_{\text{out}}$. Obviously, if we change the stride of the first convolution layer to $2s$, the size of the output feature map is decreased by $1/4$, which leads to nearly $\mathbf{4}\times$ computational acceleration for the whole network. Increasing the stride in first convolution layer seems to down-sample the pixel-level features. Since the pixel-level features may be redundant, if we carefully select the stride of the first convolution layer, it is possible to realize considerable acceleration and preserve the generalization.

3.3.2 Implementation Details

Table 1 provides the network architectures for face recognition as an example. The LightCNN-9 and LightCNN-29[1] are used as the teacher networks. Note that there are two differences compared with the original LightCNN [64]: 1) The down-sampling operations employ a combination of

[1]https://github.com/AlfredXiangWu/LightCNN

max-pooling and ave-pooling; 2) The feature layer is a linear projection without any non-linear activation functions, while the original LightCNN is implemented by a 512-d fully-connected layer with the mfm operation.

According to the proposed paradigm in Section 3.3.1, the architectures of the student networks (LightCNN-9-Fast and LightCNN-29-Fast) are shown in Table 1. Obviously, most operations in the teacher and student network are the same except two aspects: 1) The stride of the first convolution layer for the student network is 2. 2) The global average pooling layer is used instead of a 256-d fully-connected layer as the representation. In this way, we observe that the FLOPs of student networks are nearly a quarter of their corresponding teacher networks, which leads to about $4\times$ acceleration in practice.

3.4. Evolutionary Embedding Learning

As stated above, by combining the modified knowledge distillation, the correlated embedding loss and the student network paradigm, we propose Evolutionary Embedding Learning (EEL) for open-set problems via massive knowledge distillation. The loss function can be defined as follows:

$$L_{\text{EEL}} = L_{\text{L1-MKD}} + \lambda L_{\text{CEL}} \tag{13}$$

where λ is a trade-off parameter for CEL and the detailed discussions for all the parameters in EEL are shown in Table 5. Based on the paradigm of designing the student network in Section 3.3, the proposed EEL can realize considerable acceleration without sacrificing accuracy for open-set problems.

The "Evolutionary" means that under the guidance of the teacher network, the student network can be evolved and its performance can reach or even surpass its teacher's. Moreover, different from the previous knowledge distillation works [67, 69, 4], we show that Eq. (13) is effective to transfer knowledge from a high-capacity model to a compact one, and is also suitable for two models with similar capacities (even the same architectures).

4. Experiments

In this section, the proposed EEL framework is evaluated against state-of-the-art methods on one closed-set image classification task and three large-scale open-set tasks, including face recognition, vehicle re-identification and person re-identification.

4.1. Image Classification

We conduct an experiment on the CIFAR-10 dataset to evaluate the proposed EEL method. CIFAR-10 contains 32×32 RGB images totally with 50K training samples and 10K testing samples in 10 classes. We follow [13] for the experimental setting. The input takes a 32×32 random crop from a zero-padding 40×40 image. Horizontal flipping is also used for data augmentation. The teacher and student networks are ResNet-110 and ResNet-20, respectively. For optimization, SGD is used with a mini-batch size of 128. The momentum and weight decay is set to 0.9 and 10^{-4}, respectively. We train 200 epochs. The initial learning rate is 0.1 and is divided by 10 at 100 and 150 epochs.

Method	Teacher	Student	error (%)
ResNet-20	-	-	7.82
ResNet-110			5.96
KD [16]	ResNet-110	ResNet-20	7.18
Logits [1]			**6.67**
Fitnet [50]			7.45
AT [69]			7.16
FSP [67]			7.17
NST [21]			7.49
PKT [46]			7.05
FT [24]			6.80
DML [72]			7.19
RKD [44]			7.11
IRG [38]			7.09
L2-MKD	ResNet-110	ResNet-20	6.66
L1-MKD			6.63
EEL			**6.46**

Table 2. Comparisons with state-of-the-art knowledge distillation methods on the CIFAR-10 dataset. The teacher and student networks are ResNet-110 and ResNet-20, respectively.

Table 2 presents the performance comparisons on CIFAR-10. Comparing with the original ResNet-20 performance (7.82%), EEL achieves **6.46%** error rate, which demonstrates the effectiveness of our method on closed-set problem. Besides, in Table 2, EEL also outperforms other state-of-the-art methods, including KD [16], Logits [1], Fitnet [50], Attention Transfer (AT) [69], Flow of Solution Procedure (FSP) [67], Neural Selective Transfer (NST) [21], Probabilistic Knowledge Transfer (PKT) [46], Factor Transfer (FT) [24], Deep Mutual Learning (DML) [72], Relational Knowledge Distillation (RKD) [44] and Instance Relationship Graph (IRG) [38].

4.2. Face Recognition

4.2.1 Datasets and Protocols

We use the LFW [20] and MegaFace [23] datasets to evaluate EEL. LFW [20] contains 13,233 images of 5,749 people collected from the web and 6,000 pairs evaluation is performed for standard verification protocols. Besides, there is a more challenging and generalized benchmark called BLUFR [29] for the LFW evaluations. There are 10-fold sub-experiments, where each fold contains 156,915 genuine matchings and 46,960,863 impostor matchings for performance evaluation. MegaFace [23] aims to evaluate face recognition algorithms using 1 million distractors. It contains 3,530 images of 80 identities from FaceScrub [43] as the probe set and 1 million images of 690K identities as the distractors. The rank-1 accuracy for face identification and TPR@FAR=1e-6 for face verification are reported.

Method	BLUFR		MegaFace	
	VR@0.1%	DIR@1%	Rank-1	VR@1e-6
Teacher	99.41	94.43	76.02	89.74
Student	98.17	85.82	65.48	79.43
Contrastive [57]	98.20	86.54	65.88	80.23
N-Pair [53]	98.18	86.60	66.07	81.35
Triplet [52]	98.34	88.90	69.03	84.13
Center Loss [63]	98.20	86.97	66.29	80.32
LM-Softmax [34]	98.67	89.80	70.79	84.92
A-Softmax [33]	98.69	90.90	72.14	86.82
KD [16]	98.05	87.84	67.42	80.44
Logits [1]	98.60	88.82	69.29	83.55
Fitnet [50]	98.89	90.97	70.33	83.79
AT [69]	98.85	89.81	69.77	83.60
FSP [67]	98.90	90.97	70.57	83.80
NST [21]	98.90	90.81	70.98	83.98
FT [24]	98.89	90.90	71.03	84.07
L2-MKD	98.97	92.73	72.56	87.19
L1-MKD	99.27	93.48	73.95	89.45
CEL	98.71	91.18	72.15	86.95
L2-MKD+CEL	99.29	93.92	74.32	90.05
L1-MKD+CEL	**99.41**	**94.57**	**76.40**	**90.70**

Table 3. Varying various deep embedding learning and knowledge distillation methods. All the models are trained on MS-Celeb-1M and evaluated on BLUFR for LFW and MegaFace. The LightCNN-29-Fast and the LightCNN-29 are the backbones for student and teacher network, respectively.

τ	BLUFR		MegaFace	
	VR@0.1%	DIR@1%	Rank-1	VR@1e-6
1	99.07	93.04	72.12	86.54
2	**99.27**	**93.48**	**73.95**	**89.45**
4	99.25	91.76	73.83	88.90
6	98.92	90.40	73.75	88.39
8	98.42	90.23	73.48	88.08

Table 4. Varying τ for L1-MKD. All the models are trained on MS-Celeb-1M and evaluated on BLUFR for LFW and MegaFace. The LightCNN-29-Fast and the LightCNN-29 are the backbones for student and teacher network, respectively.

batch size	λ	BLUFR		MegaFace	
		VR@0.1%	DIR@1%	Rank-1	VR@1e-6
128	.5	99.37	94.40	75.79	90.26
256	.5	99.37	94.51	75.70	90.10
512	.5	99.36	94.39	75.54	89.57
128	.75	99.41	94.54	76.29	90.30
256	.75	99.39	94.54	76.04	90.18
512	.75	99.39	94.50	75.86	89.99
128	1	99.40	94.48	**76.54**	90.62
256	1	**99.41**	**94.57**	76.40	**90.70**
512	1	99.39	94.50	76.13	90.30

Table 5. Varying batch size and trade-off λ in EEL.

4.2.2 Implementation Details

We use LightCNN, including LightCNN-9 and LightCNN-29, as the teacher networks, and LightCNN-9-Fast and LightCNN-29-Fast are defined as the student networks according to Section 3.3. All the experiments are trained on the MS-Celeb-1M dataset [10]. During training, the face images are aligned to 144×144 and randomly cropped into 128×128 as the inputs. If not specific, the batch size is 256, the temperature τ is 2, and the trade-off parameter λ is 1. The learning rates for MKD and EEL are 0.005 and 0.0001, respectively. Besides, the student network under the EEL training is initialized by the pre-trained model after the MKD training.

4.2.3 Ablation Study

We conduct ablation studies on different hyper-parameters for the proposed EEL framework. In this section, the default student and teacher networks are LightCNN-29-Fast and LightCNN-29, respectively. The performance of teacher and student networks under the supervision of softmax loss is shown in Table 3.

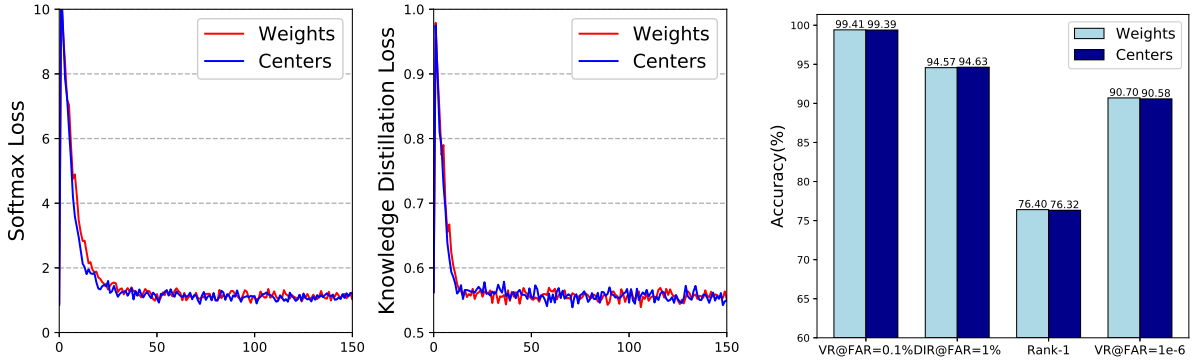
First, we attempt to discuss the influence of knowledge distillation methods. Table 3 lists different knowledge distillation methods including KD [16], Logits [1], Fitnet [50], Attention Transfer (AT) [69], Flow of Solution Procedure (FSP) [67], Neural Selective Transfer (NST) [21], Factor Transfer (FT) [24], L2-MKD in Eq.(3) and L1-MKD in Eq.(4). Compared with the baseline (directly training the student network only with the softmax loss), all the knowledge distillation methods achieve better performance. Since Logits [1], Fitnet [50], AT [69], FSP [67] and NST [21] directly match features as a regression model, the student network has only a marginal improvement over the baseline that is trained by the softmax loss. The reason is that the student model struggles to match features without relaxation due to limited capacity. Meanwhile L2-MKD can make significant improvements compared with KD [16], Logits [1] and Fitnet [50]. Furthermore, compared with L2-MKD, L1-MKD gains 1.39% on Rank-1 and 2.26% on VR@FAR=1e-6 for MegaFace. This finding indicates that our proposed L1-MKD in Eq (4) is suitable for transferring knowledge under massive class training for open-set problems. Therefore, we choose L1-MKD as the default knowledge distillation method in EEL.

Second, we compare the proposed EEL with other well-known deep embedding learning methods such as Contrastive [57], N-pair [53], Triplet [52], Center loss [63], LM-softmax [34] and A-softmax [33], respectively, in Table 3. Obviously, although various deep embedding learning methods contribute to performance improvements compared with baseline, EEL outperforms all of them. It indicates that benefiting the prior knowledge from the strong teacher network, EEL has more powerful potential of performance improvements than conventional deep embedding learning methods.

Third, in order to understand the knowledge distillation for better deep embeddings, we analyze the influence of the temperature τ in the L1-MKD loss, as shown in Table 4. Especially, we take our default LightCNN-29-Fast model trained by varying τ for the L1-MKD method under the LightCNN-29 teacher network. As the temperature τ increases, the convergence of the student net-

Backbone	Method	BLUFR		MegaFace		Performance	
		VR@0.1%	DIR@1%	Rank-1	VR@1e-6	Speed	Size
SqueezeNet [22]	Softmax	93.69	71.88	49.84	56.00	17 ms	3.26 MB
	L1-MKD	95.58	76.16	54.89	66.91		
	EEL	96.30	79.79	57.93	69.05		
MobileNet [17]	Softmax	97.97	85.35	66.18	79.22	45 ms	13.33 MB
	L1-MKD	98.58	88.71	68.62	81.57		
	EEL	98.84	90.52	71.61	85.28		
ShuffleNet [71]	Softmax	96.54	78.64	56.64	66.59	25 ms	5.02 MB
	L1-MKD	96.69	78.28	57.73	70.53		
	EEL	97.05	83.13	61.23	72.76		
MobileNetV2 [51]	Softmax	97.99	84.73	65.56	78.76	67 ms	9.89 MB
	L1-MKD	98.35	87.54	68.03	81.01		
	EEL	98.76	90.20	71.16	85.03		
LightCNN-9-Fast	Softmax	95.11	75.78	54.92	59.69	7.6 ms	5.22 MB
	L1-MKD	95.99	79.17	60.95	70.71		
	EEL	96.54	82.62	62.78	72.54		
LightCNN-29-Fast	Softmax	98.17	85.82	65.48	79.43	26 ms	33.10 MB
	L1-MKD	99.27	93.48	73.95	89.45		
	EEL	99.41	94.57	76.40	90.70		
LightCNN-29 [64]	Softmax	99.41	94.43	76.02	89.74	92 ms	39.97 MB
	L1-MKD	99.44	95.01	76.75	90.61		
	EEL	99.57	95.91	77.14	92.01		

Table 6. Varing different backbones for the student networks. The teacher network is LightCNN-29. The time cost is evaluated on i7-4790. Note that the depthwise convolution operations used in Mobilenet, ShuffleNet and MobileNetV2 are not parallel for each group due to the fair comparisons of a single thread implementation.



(a) Training curves (b) Evaluation performance

Figure 2. Training and evaluation performance by different center initialization. The "Weights" denotes initializing the centers by the weights of the classifier in the teacher network. The "Centers" denotes averaging the features of each class from the teacher network. (a) shows the training curves of the softmax loss and knowledge distillation loss with EEL training. (b) presents the performance on LFW BLUFR and MegaFace.

work becomes difficult. Compared with the baseline, L1-MKD under $\tau = 2$ yields 1.10% and 7.66% improvements on VR@FAR=0.1% and DIR@FAR=1% for BLUFR,respectively, and 8.47% and 10.02% improvements on Rank-1 and VR@FAR=1e-6 for MegaFace, respectively.

Next, since the aim of correlated embedding loss is to constrain the embedding spaces between the teacher and student networks from a global perspective, it is important to investigate the influences of batch size and loss trade-off λ. The results are presented in Table 5. When fixing λ and varying the batch size for CEL, we find there is nearly no difference among the results. The phenomenon demonstrates that, different from some metric learning [57, 52] as a local view optimization, the proposed correlated embedding loss is not sensitive to the batch size. And then, we explore the influence of the loss trade-off parameter λ and find the best λ ranges in $[0.5, 1.0]$ (we test $\lambda \in [0.25, 2.0]$). Note that if λ is too large, the hard optimization tends to collapse the convergence.

Finally, we present the effectiveness of our proposed EEL framework on different student structures, including SqueezeNet [22], MobileNet [17], ShuffleNet [71], MobileNetV2 [51], LightCNN-9-Fast, LightCNN-29-Fast and LightCNN-29 [64]. The results are tabulated in Table 6. Obviously, the performance of all the student backbones trained by EEL is significantly improved, comparing with the results of L1-MKD and softmax. This phenomenon indicates that the proposed EEL is applicatble for various

Method	Acc on LFW	BLUFR		MegaFace		Performance	
		VR@0.1%	DIR@1%	Rank-1	VR@1e-6	Speed	Size
DeepID2+ [58]	99.47	-	-	-	-	-	-
WebFace [66]	97.73	80.26	28.90	-	-	-	-
FaceNet [52]	99.63	-	-	70.49	86.47	-	-
VGG Face [45]	98.95	-	-	-	-	-	-
CenterLoss [63]	99.28	-	-	65.23	76.51	-	-
SphereFace [33]	99.42	-	-	72.73	85.56	-	-
LightCNN [64]	99.40	98.88	92.29	73.75	85.13	121 ms	50.30 MB
VGG Face* [45]	97.27	73.34	36.56	-	-	581 ms	524 MB
CenterLoss*	98.70	94.64	70.66	63.10	74.66	160 ms	76.54 MB
T1	98.70	96.80	83.06	65.78	76.28	59 ms	21.70 MB
T2	99.43	99.41	94.43	76.02	89.74	92 ms	39.97 MB
S1 (softmax)	98.50	95.11	75.78	54.92	59.69	7.6 ms	5.22 MB
S1 (T1 EEL)	**98.70**	96.15	**82.71**	**62.81**	**73.03**		
S1 (T2 EEL)	**98.70**	**96.54**	82.62	62.78	72.54		
S2 (softmax)	99.27	98.17	85.82	65.48	79.43	26 ms	33.10 MB
S2 (T2 EEL)	**99.47**	**99.41**	**94.57**	**76.40**	**90.70**		

Table 7. Comparisons with other state-of-the-art methods on LFW for standard, BLUFR protocols, and MegaFace protocols. * denotes that we evaluate the performance according to the released models (or features). S1 and S2 denote the LightCNN-9-Fast and LightCNN-29-Fast models for student networks, respectively. T1 and T2 denote the LightCNN-9 and LightCNN-29 models for teacher networks, respectively.

student networks and improves the performance by a large margin. When the teacher network is LightCNN-29, we find that the best advancement comes from LightCNN-29-Fast, which gains 8.75% on DIR@FAR=1% for BLUFR, 10.92% on Rank-1 and 11.37% on VR@FAR=1e-6 for MegaFace. It also certificates that the proposed paradigm for the student network development is effective to implement a fast and accurate network without bells and whistles.

Besides, we employ the LightCNN-29 as the student network, which means both the teacher and the student networks are the same architectures. Surprisingly, the student network outperforms its teacher network by a large margin (99.57% vs 99.41% on VR@FAR=0.1% and 95.91% vs 94.43% on DIR@FAR=1% for BLUFR as well as 77.14% vs 76.02% on Rank-1 and 92.01% vs 89.74% on VR@FAR=1e-6 for MegaFace, respectively), as shown in Table 6. The results demonstrate that knowledge distillation should not be limited to the model acceleration or model compression. It also contributes to obtain better generalization as a prior to regularize convolutional neural networks.

4.2.4 Detailed Analysis of Centers

In this section, we present a detailed analysis of center initialization and online updating strategy. The teacher and student networks are LightCNN-29 and LightCNN-29-Fast, respectively.

Since we denote a set of centers $\{c_j^{\mathrm{T}}\}_{j=1}^{N}$ to represent the embedding space of the teacher network, there are two simple methods to initialize centers for correlated embedding loss: (1) Using the weights of the classification layer in the teacher network; (2) Averaging deep features of each class extracted from the teacher network. As shown in Fig. 2, these two initializations obtain similar performance, indicating that the proposed correlated embedding loss is insensitive to the center initialization during training. Therefore, for convenience, we use the weights of the classification layer in the teacher network as the initialization for the correlated embedding loss.

According to Eq. (12), the centers of each identity are updated online. However, for mini-batch SGD, it is hard to put all the identities into one mini-batch during training, especially for massive identities. Therefore, we simply employ a moving average to update centers in the training phase. We conduct experiments with two updating forms, including Cumulative Moving Average (CMA) and Exponential Moving Average (EMA). For the CMA updating strategy, EEL obtains 99.38% and 94.52% on VR@FAR=0.1% and DIR@1% for BLUFR, respectively. Considering EMA, we achieve 99.41% and 94.57% on VR@FAR=0.1% and DIR@1% for LFW BLUFR, respectively.

4.2.5 Comparisons with State-of-the-arts

Table 7 summarizes the comparison results about the complexity of the networks and the performance on LFW and MegaFace. We use LightCNN-9 (T1) and LightCNN-29 (T2) as the teacher networks, and LightCNN-9-Fast (S1) and LightCNN-29-Fast (S2) as the student networks. It is obvious that the proposed EEL framework can significantly improve the performance of the student networks, which can achieve or even outperform their teacher networks. For example, the LightCNN-29-Fast under the LightCNN-29 EEL training (S2+T2 EEL) beats its teacher network on DIR@FAR=1% (94.57% vs 94.43%) for BLUFR, as well as Rank-1 accuracy (76.40% vs 76.02%) and VR@FAR=1e-6 (90.70% vs 89.74%) for MegaFace. Considering the computational time cost, LightCNN-29-Fast is approximately $\mathbf{3.5}\times$ faster (26 ms vs 92 ms) than LightCNN-29 on a CPU with a single thread implementation. The results indicate that it is possible to implement both a fast and accurate stu-

Method	CASIA NIR-VIS 2.0		Oulu-CASIA NIR-VIS			BUAA-VisNir		
	Rank-1	FAR=0.1%	Rank-1	FAR=1%	FAR=0.1%	Rank-1	FAR=1%	FAR=0.1%
TRIVET [37]	95.7	91.0	92.2	67.9	33.6	93.9	93.0	80.9
IDR [14]	97.3	95.7	94.3	73.4	46.2	94.3	93.4	84.7
CDL [65]	98.6	**98.5**	94.3	81.6	53.9	96.9	95.9	90.1
W-CNN [15]	**98.7**	98.4	98.0	81.5	54.6	97.4	96.0	91.9
Teacher	98.1	97.4	**100.0**	**96.6**	78.3	99.1	99.1	**97.7**
Student	91.6	87.7	99.0	93.1	58.7	96.9	97.1	89.6
EEL	98.5	97.6	**100.0**	95.1	**83.2**	**99.3**	**99.2**	97.4

Table 8. Comparisons with other state-of-the-art heterogeneous face recognition methods on the CASIA NIR-VIS 2.0 dataset, the Oulu-CASIA NIR-VIS dataset and the BUAA-VisNir dataset. The teacher network and the student network are LightCNN-29 and LigthCNN-29-Fast, respectively.

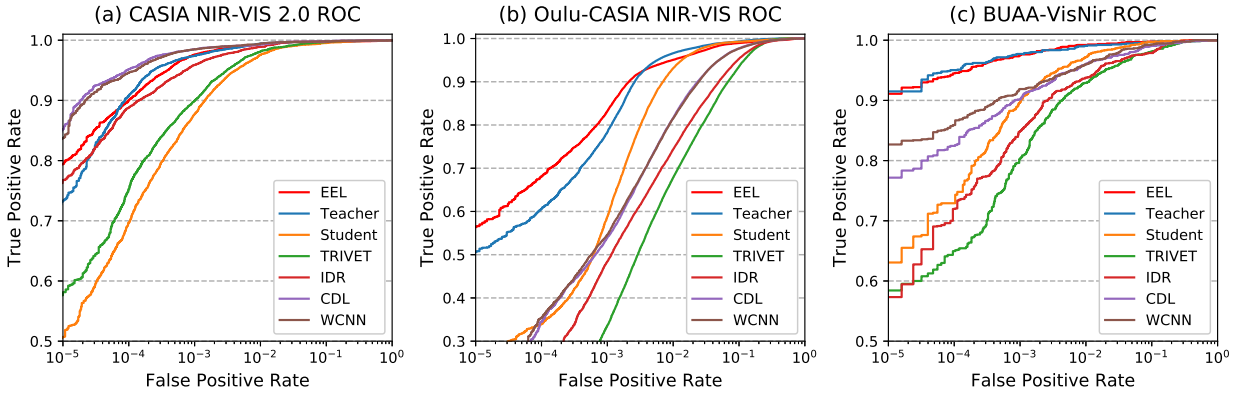


Figure 3. ROC curves of different methods on the three heterogeneous face recognition datasets, including the CASIA NIR-VIS 2.0, the Oulu-CASIA NIR-VIS and the BUAA-VisNir databases.

dent network and compete with the high-capacity teacher network, under our proposed EEL framework.

Further comparisons with current state-of-the-art methods including DeepID2+ [58], WebFace [66], FaceNet [52], CenterLoss [63], SphereFace [33], VGGFace [45] and LightCNN [64] are also shown in Table 7. The LightCNN-29-Fast (S2) network under EEL achieves **99.41%** on VR@FAR=0.1% and **94.57%** on DIR@FAR=1% for BLUFR, as well as **76.40%** on Rank-1 and **90.70%** on VR@FAR=1e-6 for MegaFace, which significantly outperforms other state-of-the-art methods. We also design LightCNN-9-Fast (S1) as a student network by referring to LightCNN-9. It can achieve 98.70% on LFW and the speed of inference is only 7.6 ms on a CPU with only 5.22 MB parameters.

In addition, in terms of the LightCNN-9-Fast (S1) shown in Table 7, we assess whether the optimal teacher network is LightCNN-9 or LightCNN-29, since their performances are similar. Here, we clarify that EEL is able to improve the performance of the student network. However, if the capacity of the student network has a large gap compared with the teacher's, the improvements are limited. It may be the main reason that the performance of the student network could not reach that of the teacher's in previous works [67, 69, 4]. When the student network is changed to the LightCNN-29-Fast, designed by referring to the LightCNN-29, the results outperform its teacher network. This phenomenon demonstrates that the proposed paradigm is significant towards designing a student network, which can accelerate the model and also preserve accuracy.

4.2.6 Generalization on Heterogeneous Face Recognition

This section aims to verify the generalization of EEL. We directly evaluate the proposed LightCNN-29-Fast model with EEL training on heterogeneous face recognition tasks. Three heterogeneous face recognition datasets, including CASIA NIR-VIS 2.0 Face database [28], Oulu-CASIA NIR-VIS database [3] and BUAA-VisNir face database [19], are used for evaluations.

The CASIA NIR-VIS 2.0 database [28] is the largest public and most challenging NIR-VIS dataset due to the large variations of lighting, expression and pose. For testing, it contains 10-fold experiments and each fold contains 358 identities. For each fold, the gallery set is constructed from 358 identities and each identity only has one VIS image. The probe set has over 6,000 NIR face images from the same 358 identities. All the NIR images in the probe set are to be matched against the VIS images in the gallery set, resulting in a 6000×358 similarity matrix. The Rank-1 accuracy and VR@FAR=0.1% are reported.

Method	Backbone	Speed	Vehicle-ID						Vehicle-1M					
			top-1			top-5			top-1			top-5		
			Small	Medium	Large	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
DRDL [30]	VGG-M	154 ms	49.0	42.8	38.2	73.5	66.8	61.6	-	-	-	-	-	-
FACT [35]	GoogleNet	92 ms	49.5	44.6	39.9	67.9	64.1	60.4	-	-	-	-	-	-
NuFACT [36]	GoogleNet	92 ms	48.9	43.6	38.6	69.5	65.3	60.7	-	-	-	-	-	-
GoogleNet [9]	GoogleNet	92 ms	46.6	42.5	38.1	62.2	58.9	55.4	54.5	50.8	40.9	60.9	58.6	50.5
C2F-Rank [9]	GoogleNet	92 ms	61.1	56.2	51.4	81.7	76.2	72.2	67.1	62.0	52.8	70.3	67.1	60.1
Teacher	LightCNN-29	65 ms	67.0	66.4	64.0	77.5	75.5	72.8	89.9	87.8	83.9	94.3	93.3	91.1
Student	LightCNN-29-Fast	34 ms	61.4	60.0	57.3	72.1	70.4	67.1	86.3	83.0	77.9	90.8	89.0	86.7
EEL	LightCNN-29-Fast	34 ms	**70.2**	**69.4**	**66.3**	**83.1**	**79.7**	**77.1**	**92.2**	**88.7**	**85.7**	**96.9**	**95.1**	**94.2**

Table 9. Comparisons with other state-of-the-art methods of Vehicle Re-Identification on the Vehicle-ID and Vehicle-1M datasets. The time cost is evaluated on i7-4790.

The Oulu-CASIA NIR-VIS database [3] contains 80 identities with 6 expression variations. Following the setting in [15], 20 identities are used for testing. We randomly select 8 images for each expression from NIR and VIS images, respectively. All the VIS images of the 20 identities are used as the gallery set and all the NIR images are treated as the probe. The similarity matrix between the probe set and the gallery set is 960×960. The rank-1 accuracy, VR@FAR=1% and VR@FAR=0.1% are reported for comparisons.

The BUAA-VisNir face database [19] is composed of 150 identities with 9 NIR and 9 VIS images. Following the setting in [15], we randomly select 100 identities for testing. Only one VIS image of each identity is selected in the gallery set and the probe set contains 900 NIR images. We report the rank-1 accuracy, VR@FAR=1% and VR@FAR=0.1% by the 900×100 similarity matrix.

Table 8 and Fig. 3 show the results of the proposed EEL method with other state-of-the-art heterogeneous face recognition methods. **Note that** the three datasets are strictly independent from the MS-Celeb-1M training dataset [10] and we don't fine-tune our models on the training data for these three datasets. The results are promising. As the student network, the LightCNN-29-Fast with EEL training obtains comparable performance against its teacher network LightCNN-29 and other state-of-the-art methods, including TRIVET [37], IDR [14], CDL [65] and W-CNN [15]. On the most challenging CASIA NIR-VIS 2.0 database, the student network with EEL training outperforms its teacher network in terms of both Rank-1 accuracy (98.5% vs 98.1%) and VR@FAR=0.1% (97.6% vs 97.4%). The performance of EEL is a little lower than CDL [65] and W-CNN [15], because the LigthCNN-29-Fast doesn't fine-tune on the heterogeneous face data, and other state-of-the-art methods employ cross domain matching on the CASIA NIR-VIS 2.0 database instead. Besides, considering the Oulu-CASIA NIR-VIS database and the BUAA-VisNir face database, the LigthCNN-29-Fast with EEL training outperforms state-of-the-art methods by a large margin. The experimental results on heterogeneous face recognition show that the proposed EEL method has good generalization ability for open-set problems.

4.3. Vehicle Re-Identification

4.3.1 Datasets and Protocols

We evaluate EEL on Vehicle-ID [30] and Vehicle-1M [9], the top two largest vehicle re-identification datasets. Vehicle-ID contains 110,178 images of 13,134 vehicles for training and 11,1585 images of 13,133 vehicles for testing. The testing data are divided into small (7,332 images of 800 vehicles), medium (12,995 images of 1,600 vehicles) and large (20,038 images of 2,400 vehicles) testing sets. Vehicle-1M is even larger, including 884,571 images from 50,000 vehicles for training and 91,480 images from 5,527 vehicles for testing. Similarly, the testing data are also divided into three subsets. The small set contains 16,123 images of 1,000 vehicles, the medium includes 32,539 images of 2,000 vehicles and the large one covers 49,259 images of 3,000 vehicles, respectively. For Vehicle-ID and Vehicle-1M, we randomly select one image of each vehicle as the gallery set and the others are all probe queries. The results are measured by the Cumulative Matching Characteristic (CMC) curve.

4.3.2 Implementation Details

The teacher and student networks have the same architectures as LightCNN-29 and LightCNN-29-Fast, respectively. The main differences are as follows: 1) the inputs are RGB vehicle images with the size of 224×224, and 2) the strides of the first convolution layer for the teacher and student network are 2 and 4, respectively. During training, we resize images to 256×256 and randomly crop them to 224×224. The batch size is 128, the learning rate is 0.005 for MKD and 0.0001 for EEL, and λ and τ are set to 1 and 2, respectively. Moreover, the horizontal flipping and color jittering are used as the data augmentation.

4.3.3 Comparisons

Table 9 illustrates the top-1 and top-5 match rates of our proposed method and other competitors on the Vehicle-ID and Vehicle-1M datasets. It is obvious that the student network under EEL outperforms state-of-the-art methods

Method	Backbone	Speed	Market-1501				DukeMTMC-reID	
			Single Query		Multi. Query			
			Rank-1	mAP	RanK-1	mAP	Rank-1	mAP
DarkRank [4]	NIN-BN	23 ms	86.7	68.2	91.4	76.4	-	-
Basel+LSRO [77]	ResNet-50	180 ms	83.9	66.0	88.4	76.1	67.6	47.1
PAN [76]	ResNet-50*	>180 ms	86.6	69.3	90.8	76.3	71.5	51.5
SVDNet [59]	ResNet-50	180 ms	82.3	62.1	-	-	76.7	56.8
MGCAM [54]	ResNet-50*	>180 ms	83.7	74.3	-	-	-	-
IDE [78]	ResNet-50*	>180 ms	88.1	68.7	-	-	75.2	57.6
Mancs[60]	ResNet-50*	>180 ms	93.1	82.3	-	-	84.9	71.8
DSA-reID [73]	ResNet-18*	>43 ms	**95.7**	**87.6**	-	-	86.2	74.8
DG-Net [75]	ResNet-50	180 ms	94.8	86.0	-	-	**86.6**	**74.8**
Teacher	LightCNN-29	123 ms	90.6	74.6	92.8	78.9	75.6	56.3
Student	LightCNN-29-Fast	42 ms	88.5	71.5	90.6	74.4	71.1	51.0
EEL	LightCNN-29-Fast	42 ms	**91.3**	**78.3**	**93.3**	**83.4**	**76.4**	**57.5**
Teacher	ResNet-101	209 ms	94.5	85.9	95.9	90.9	87.5	76.6
Student	ResNet-50	180 ms	94.4	85.8	95.9	89.9	86.4	76.0
EEL	ResNet-50	180 ms	**94.9**	**87.5**	**96.2**	**91.1**	**88.2**	**78.3**

Table 10. Comparisons with other state-of-the-art methods of Person Re-Identification on the Market-1501 and DukeMTMC-reID datasets. The time cost is evaluated on i7-4790. * denotes that these methods incorporate other branches on the backbone, thus the time cost takes longer than the speed of original backbone.

including DRDL [30] and C2F-Rank [9] by a large margin. Especially, considering that the large-scale Vehicle-1M dataset contains 50,000 classes approximately 880,000 images, EEL significantly improves the generalization from 67.1%, 62.0%, and 52.8% to **92.2%**, **88.7%** and **85.7%** on the small, medium and large testing set, respectively. Moreover, both MKD and EEL improve the performance of the fast student network, and the student under EEL also beats its respective teacher on all the small (92.2% vs 89.9%), medium (88.7% vs 87.8%) and large (85.7% vs 83.9%) testing sets. The results indicate that EEL facilitates the bridge of the performance gap between teacher and student networks on the vehicle re-identification task.

4.4. Person Re-Identification

4.4.1 Datasets and Protocols

For person re-identification, we evaluate EEL on the Market-1501 [74] and DukeMTMC-reID [49] datasets. Market-1501 dataset is collected from six cameras, including five high-resolution cameras and one low-resolution camera. It contains 32,688 images of 1,501 identities, where the 12,936 images of 750 identities are used for training, and the others are used for testing. DukeMTMC-reID dataset is collected from eight cameras, and is a subset for cross-camera tracking. Zheng *et al.* [77] selected 1,404 identities from eight cameras, where a total of 16,522 images of 702 IDs are used for training, and the other 702 IDs (including 2,282 query images and 17,611 gallery images) are used for testing.

4.4.2 Implementation Details

We conduct two experiments for person re-identification. For the first one, the teacher and student networks are LightCNN-29 and LightCNN-29-Fast, respectively, following the same structures of face recognition. There are some trivial differences. Instead of gray-scale inputs for face recognition, we use RGB images as the input since the color information is important for person re-identification tasks. During training, each input image is resized to 256×128 and then randomly cropped to 224×112. Moreover, horizontal flipping is used as the data augmentation. The batch size is 128, the learning rates are 0.005 for MKD and 0.0001 for EEL, and λ and τ are set to 1 and 2, respectively. For the second one, we utilize ResNet-101 and ResNet-50 as the teacher and student networks, respectively. Following [39], the input size is set to 256×128. Horizontal flipping and random erasing are used as the data augmentation. The networks are initialized from ImageNet pretrained models. The optimization method is Adam [25] and the learning rate is 0.00035 for MKD and EEL. The hyper-parameter λ and τ is set to 1 and 2, respectively.

4.4.3 Comparisons

The results on the Market-1501 and DukeMTMC-reID datasets are shown in Table 10.

For LightCNN-29-Fast as the student backbone, EEL achieves **Rank-1=91.3, mAP=78.3** in the single query mode and **Rank-1=93.3, mAP=83.4** in the multiple query mode on Market-1501, and **Rank-1=76.4**, **mAP=57.5** on DukeMTMC-reID. Further, as shown in Table 10, when utilizing more powerful student backbone (ResNet-50), we achieve comparable results on Market-1501 (**87.5** vs 87.6 on mAP and 94.9 vs 95.7 on Rank-1) and DukeMTMC-reID (**78.3** vs 74.8 on mAP and **88.2** vs 86.6 on Rank-1), respectively.

Although Mancs[60], DSA-reID [73] and DG-Net [75] also obtain good performance, our EEL only utilizes a single network to learn embeddings under the supervision of Eq. (13). Mancs introduces attention mechanism to capture the spatial information, while DSA-reID utilizes two

stream networks that consist of a full image stream and a densely semantically-aligned stream. DG-Net utilizes generated data from a generative module as data augmentation to improve the person re-identification performance.

Here, we should clarify that the motivation of EEL is to train a fast and accurate student network which can reach the performance of its teacher. As shown in Table 10, the performance on mAP is promising since LightCNN-29-Fast improves it by approximate 3.7% on Market-1501 and 1.2% on DukeMTMC-reID. For ResNet-50 as the student backbone, comparing with its teacher (ResNet-101), we also gain 1.6% and 1.7% on mAP for Market-1501 and DukeMTMC-reID, respectively. These results indicate that the knowledge provided by EEL can be transferred and treated as regularizer to alleviate the overfitting with a limited number of training samples. Unquestionably, there are similar phenomena with vehicle re-identification in which the student network under EEL training outperforms its teacher network.

5. Conclusion

In this paper, we develop an efficient knowledge distillation framework called Evolutionary Embedding Learning (EEL) for open-set problems. It provides a simple yet effective way to implement a model acceleration for embedding learning without sacrificing accuracy. First, EEL reformulates the knowledge distillation for open-set problems with massive classes. Second, we introduce an angular constraint and formulate the Correlated Embedding Loss (CEL) to match the embedding spaces between teacher and student networks from the global perspective for knowledge transfer. Third, EEL proposes a paradigm towards fast and accurate student network developments, which can hold the capacity of embeddings, and lead to a fast model without any dedicated implementation on specific devices. Thus, our proposed EEL can achieve comparable or even better results on different open-set tasks, including face recognition, vehicle re-identification and person re-identification.

6. Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61622310, Grant 61721004, and in part by the Beijing Natural Science Foundation Grant JQ18017.

References

[1] Ba, J., Caruana, R.: Do deep nets really need to be deep? In: NeurIPS (2014)

[2] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a" siamese" time delay neural network. In: NeurIPS (1994)

[3] Chen, J., Yi, D., Yang, J., Zhao, G., Li, S.Z., Pietikainen, M.: Learning mappings for face synthesis from near infrared to visual light images. In: CVPR (2009)

[4] Chen, Y., Wang, N., Zhang, Z.: Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In: AAAI (2018)

[5] Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., Bengio, Y.: Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. In: NeurIPS (2016)

[6] Czarnecki, W.M., Osindero, S., Jaderberg, M., Swirszcz, G., Pascanu, R.: Sobolev training for neural networks. In: NeurIPS (2017)

[7] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)

[8] yong Ding, S., Lin, L., Wang, G., Chao, H.Y.: Deep feature learning with relative distance comparison for person re-identification. PR (2015)

[9] Guo, H., Zhao, C., Liu, Z., Wang, J., Lu, H.: Learning coarse-to-fine structured feature embedding for vehicle re-identification. In: AAAI (2018)

[10] Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: ECCV (2016)

[11] Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In: ICLR (2016)

[12] Han, S., Pool, J., Tran, J., Dally, W.J.: Learning both weights and connections for efficient neural network. In: NeurIPS (2015)

[13] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

[14] He, R., Wu, X., Sun, Z., Tan, T.: Learning invariant deep representation for nir-vis face recognition. In: AAAI (2017)

[15] He, R., Wu, X., Sun, Z., Tan, T.: Wasserstein CNN: learning invariant features for NIR-VIS face recognition. TPAMI (2018)

[16] Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NeurIPS Workshop (2015)

[17] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR **abs/1704.04861** (2017)

[18] Huang, C., Loy, C.C., Tang, X.: Local similarity-aware deep feature embedding. In: NeurIPS (2016)

[19] Huang, D., Sun, J., Wang, Y.: The BUAA-VisNir face database instructions. Tech. Rep. IRIP-TR-12-FR-001, Beihang University, Beijing, China (2012)

[20] Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., Technical Report 07-49, University of Massachusetts, Amherst (2007)

[21] Huang, Z., Wang, N.: Like what you like: Knowledge distill via neuron selectivity transfer. CoRR **abs/1707.01219** (2017)

[22] Iandola, F.N., Moskewicz, M.W., Ashraf, K., Han, S., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. CoRR **abs/1602.07360** (2016)

[23] Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: CVPR (2016)

[24] Kim, J., Park, S., Kwak, N.: Paraphrasing complex network: Network compression via factor transfer. In: NeurIPS (2018)

[25] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2014)

[26] LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit database. AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist (2010)

[27] Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H.P.: Pruning filters for efficient convnets. In: ICLR (2017)

[28] Li, S.Z., Yi, D., Lei, Z., Liao, S.: The casia nir-vis 2.0 face database. In: CVPR Workshops (2013)

[29] Liao, S., Lei, Z., Yi, D., Li, S.Z.: A benchmark study of large-scale unconstrained face recognition. In: IJCB (2014)

[30] Liu, H., Tian, Y., Wang, Y., Pang, L., Huang, T.: Deep relative distance learning: Tell the difference between similar vehicles. In: CVPR (2016)

[31] Liu, L., Chen, J., Fieguth, P.W., Zhao, G., Chellappa, R., Pietikäinen, M.: From bow to cnn: Two decades of texture representation for texture classification. IJCV (2018)

[32] Liu, L., Ouyang, W., Wang, X., Fieguth, P.W., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: A survey. IJCV (2018)

[33] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: CVPR (2017)

[34] Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: ICML (2016)

[35] Liu, X., Liu, W., Ma, H., Fu, H.: Large-scale vehicle re-identification in urban surveillance videos. In: ICME (2016)

[36] Liu, X., Liu, W., Mei, T., Ma, H.: Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. TMM (2018)

[37] Liu, X., Song, L., Wu, X., Tan, T.: Transferring deep representation for nir-vis heterogeneous face recognition. In: ICB (2016)

[38] Liu, Y., Cao, J., Li, B., Yuan, Y., Hu, W., Li, Y., Duan, Y.: Knowledge distillation via instance relationship graph. In: CVPR (2019)

[39] Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: CVPR Workshops (2019)

[40] Luo, J.H., Wu, J., Lin, W.: Thinet: A filter level pruning method for deep neural network compression. In: ICCV (2017)

[41] Luo, P., Zhu, Z., Liu, Z., Wang, X., Tang, X.: Face model compression by distilling knowledge from neurons. In: AAAI (2016)

[42] Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J.: Pruning convolutional neural networks for resource efficient inference. In: ICLR (2017)

[43] Ng, H., Winkler, S.: A data-driven approach to cleaning large face datasets. In: ICIP (2014)

[44] Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: CVPR (2019)

[45] Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC (2015)

[46] Passalis, N., Tefas, A.: Learning deep representations with probabilistic knowledge transfer. In: ECCV (2018)

[47] Ranjan, R., Castillo, C.D., Chellappa, R.: L2-constrained softmax loss for discriminative face verification. CoRR **abs/1703.09507** (2017)

[48] Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A.: Xnor-net: Imagenet classification using binary convolutional neural networks. In: ECCV (2016)

[49] Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV Workshop (2016)

[50] Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: ICLR (2015)

[51] Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. CoRR **abs/1801.04381** (2018)

[52] Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR (2015)

[53] Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: NeurIPS (2016)

[54] Song, C., Huang, Y., Ouyang, W., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: CVPR (2018)

[55] Song, H.O., Jegelka, S., Rathod, V., Murphy, K.: Deep metric learning via facility location. In: CVPR (2017)

[56] Song, H.O., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: CVPR (2016)

[57] Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: NeurIPS (2014)

[58] Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: CVPR (2015)

[59] Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: ICCV (2017)

[60] Wang, C., Zhang, Q., Huang, C., Liu, W., Wang, X.: Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In: ECCV (2018)

[61] Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: L_2 hypersphere embedding for face verification. In: ACM MM (2017)

[62] Wang, J., Zhou, F., Wen, S., Liu, X., Lin, Y.: Deep metric learning with angular loss. In: ICCV (2017)

[63] Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: ECCV (2016)

[64] Wu, X., He, R., Sun, Z., Tan, T.: A light CNN for deep face representation with noisy labels. TIFS (2018)

[65] Wu, X., Song, L., He, R., Tan, T.: Coupled deep learning for heterogeneous face recognition. In: AAAI (2018)

[66] Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. CoRR **abs/1411.7923** (2014)

[67] Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: CVPR (2017)

[68] Yuan, Y., Yang, K., Zhang, C.: Hard-aware deeply cascaded embedding. In: ICCV (2017)

[69] Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: ICLR (2017)

[70] Zhang, R., Lin, L., Zhang, R., Zuo, W., Zhang, L.: Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. TIP (2015)

[71] Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. CoRR **abs/1707.01083** (2017)

[72] Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: CVPR (2018)

[73] Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Densely semantically aligned person re-identification. In: CVPR (2019)

[74] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)

[75] Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J.: Joint discriminative and generative learning for person re-identification. In: CVPR (2019)

[76] Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification. CoRR **abs/1707.00408** (2017)

[77] Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV (2017)

[78] Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: CVPR (2018)

[79] Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: CVPR (2018)