Student Network Learning via Evolutionary Knowledge Distillation

Kangkai Zhang, Chunhui Zhang, Shikun Li, Dan Zeng, *Member, IEEE,* and Shiming Ge, *Senior Member, IEEE*

***Abstract*—Knowledge distillation provides an effective way to transfer knowledge via teacher-student learning, where most existing distillation approaches apply a fixed pre-trained model as teacher to supervise the learning of student network. This manner usually brings in a big capability gap between teacher and student networks during learning. Recent researches have observed that a small teacher-student capability gap can facilitate knowledge transfer. Inspired by that, we propose an evolutionary knowledge distillation approach to improve the transfer effectiveness of teacher knowledge. Instead of a fixed pre-trained teacher, an evolutionary teacher is learned online and consistently transfers intermediate knowledge to supervise student network learning on-the-fly. To enhance intermediate knowledge representation and mimicking, several simple guided modules are introduced between corresponding teacher-student blocks. In this way, the student can simultaneously obtain rich internal knowledge and capture its growth process, leading to effective student network learning. Extensive experiments clearly demonstrate the effectiveness of our approach as well as good adaptability in the low-resolution and few-sample visual recognition scenarios.**

***Index Terms*—Knowledge distillation, teacher–student learning, visual recognition.**

I. Introduction

DEEP neural networks have proved success in many visual tasks like image classification [1], [2], object detection [3], [4], semantic segmentation [5] and other fileds [6], [7], [8] due to their powerful knowledge extraction capability from massive available data. Beyond these remarkable successes, there is still increasing concern that the development of effective learning approaches for the real-world scenarios where the high-quality data often is not available or insufficient. In this case, network learning will encounter obstacles. Knowledge distillation [9] provides an economic way that can transfer knowledge from a pre-trained teacher network to facilitate the learning of a new student network.

Knowledge distillation approaches mainly follow offline or online strategies. Offline strategies try to design more effective knowledge representation methods to learn from powerful teacher. Park *et al.* [10] and Liu *et al.* [11] focused on the structured knowledge about the instances, while other approaches [12], [13], [14] paid attention to some internal information in the network during distillation process. These offline approaches usually use a fixed pre-trained model as teacher, as shown in Fig. 1 (a), and this manner usually brings in a big capability gap between teacher and student during learning (see Fig. 1 (c)), leading to transfer difficulties. The capability gap refers to the performance difference between teacher and student network, in the image classification task, it specifically refers to the difference in accuracy. By contrast, the online distillation strategies attempt to reduce the capability gap by some training schemes for the absence of pre-trained teachers to improve the learning of student. An on-the-fly native ensemble (ONE) scheme [15] was proposed for one-stage online distillation. Instead of borrowing supervision signals from previous generations, snapshot distillation [16] extracted information from earlier iteration in the same generation. These online practices provide some good solutions to the shackles of capability gap brought by the fixed teacher, resulting in relatively better knowledge transfer, while they need high demands on the efficiency of knowledge representation because they rely more on their relatively less reliable peers' or their own predictions to provide additional supervision [15], [17], [18]. Hence, there is a question: *how to ensure both small capability gap and efficient knowledge representation to facilitate student learning?*

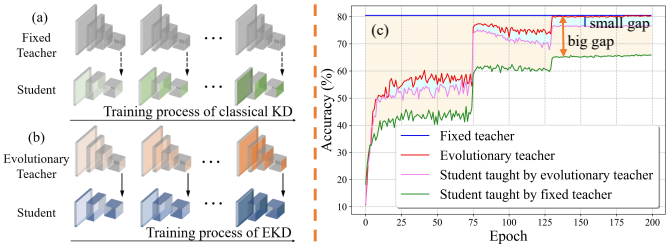


Fig. 1: Unlike classical knowledge distillation (KD) approaches with fixed teacher, an evolutionary teacher can enable more efficient knowledge transfer by minimizing the capability gap between teacher and student. Thus, we propose evolutionary knowledge distillation (EKD) to facilitate student network learning.

Inspired by the recent observations [16], [19], [20] that a small teacher-student capability gap is beneficial to knowledge transfer, we propose an evolutionary knowledge distillation (EKD) approach to improve the learning of student, as shown in Fig. 1 (b). The approach uses an evolutionary teacher whose performance is continuously improved as the training process to constantly transfer intermediate knowledge to the student, the evolutionary teacher could provide richer supervision information for the learning of student and reduce the capability

K. Zhang, C. Zhang, S. Li and S. Ge are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100095, China, and School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China. Email: {zhangkangkai, zhangchunhui, lishikun, geshiming}@iie.ac.cn.

D. Zeng is with the Department of Communication Engineering, Shanghai University, Shanghai 200444, China. E-mail: dzeng@shu.edu.cn.

Shiming Ge is the corresponding author. E-mail: geshiming@iie.ac.cn.

gap. In addition, to improve the knowledge representation ability, the teacher and student networks are both divided into several blocks, and a simple guided module pair is introduced between each corresponding block. In short, the evolutionary teacher not only solves the problem of capability gap between fixed teacher and student of offline knowledge distillation, but also solves the problem of insufficient and relatively unreliable supervision information caused by the absence of a qualified teacher for online knowledge distillation. In addition, the guided modules further promote the representation of intermediate knowledge. In this way, the student can continuously and adequately learn the intermediate knowledge from teacher as well as its growth process.

The main contributions of our work could be summarized in three folds: 1) We propose an evolutionary knowledge distillation approach to facilitate the student network learning by narrowing the teacher-student capability gap; 2) We introduce guided module pairs to enhance the knowledge representation and transfer ability; 3) We conduct extensive experiments to verify that our approach is superior to the state-of-the-arts and exhibits better adaptability in the low-resolution and few-sample scenarios.

II. Related Work

A. The Basic of Knowledge Distillation

Knowledge distillation (KD) provides a concise but effective solution for transferring knowledge from a pre-trained large teacher model to a smaller student model [9]. Since the introduction of knowledge distillation, it has been widely used in image recognition, semantic segmentation, and other fields, especially model compression [9], [21], [22], [23], [24], [25]. In practice, the student model learns the prediction of pre-trained teacher model to make itself more powerful than it's trained alone. Compared to hard ground-truth labels, fine-grained class information in soft predictions helps the small student model to reach flatter local minima, which results in more robust performance and improves generalization ability [26], [27]. Several recent works attempt to further improve that transfer knowledge between varying-capacity network models with offline or online knowledge distillation approaches [13], [15], [18], [28], [29], [30].

B. Offline Knowledge Distillation

The offline knowledge distillation approaches often adopt a two-stage training mode, it first trains the teacher model, and then trains the student network by various distillation strategies. Classical FitNet [12] tried to transfer more supervision information by using the feature map of the teacher network middle layer firstly. Crowley *et al.* [31] proposed structural model distillation for memory reduction using a strategy that produced a student architecture that was a simple transformation of the teacher's: no redesign is needed, and the same hyper-parameters can be used. And some recent approaches [11], [32] attempted to pay more attention to the relationship information of the instances. Tian *et al.* [13] proposed contrastive representation distillation, the main idea is very general: learn a representation that is close in some metric space for "positive" pairs and push apart the representation between "negative" pairs. Furlanello *et al.* [33] interactively absorbed the distilled student models into the teacher model group, through which the better generalization ability on test data is obtained. Sukmin Yun *et al.* [34] proposed a new regularization method that penalizes the predictive distribution between similar samples via self knowledge distillation to mitigate the issue that deep neural networks with millions of parameters may suffer from poor generalization due to overfitting. [35], [36], [37], [38] proposed utilizing multiple teachers to provide more supervision for the learning of student network.

Generally speaking, for offline knowledge distillation, in order to achieve a great effect, the following conditions need to be met: a high-quality pre-trained teacher model, difference between teacher and student, adequate supervision information [14], [16], [39]. In particular, offline knowledge distillation methods rely heavily on a fixed pre-trained model, however, the huge capability gap between fixed teacher and student model will bring great challenges to knowledge transfer. To bridge this gap, Mirzadeh *et al.* [19] introduced multi-step knowledge distillation, which employed an intermediate-sized network (teacher assistant); Jin *et al.* proposed a method named RCO [20], which utilizes the route in parameter space teacher network passed by as a constraint to bring a better optimization to student. None of these methods can completely solve the obstacles of knowledge transfer caused by the capability gap due to the fact that they also rely to some extent on the pre-trained teacher model. Therefore, we need to rethink how to break the shackles of offline distillation approaches to improve the learning of student network more effectively.

C. Online Knowledge Distillation

Different from the conventional two-stage offline distillation approaches, the current approaches increasingly focus on the online distillation strategies, which attempt to reduce the capability gap by some training schemes in the absence of pre-trained teachers to improve the learning of student. A group of networks or sub-networks will be trained almost synchronously, which aims to improve the performance of student by using the predictions of their peers or sub-networks as supervision instead of the high-quality pre-trained teachers'. Deep Mutual Learning (DML) [28] applied distillation losses mutually, treating each other as teachers, and it achieved good results. However, DML lacks an appropriate teacher role, hence provides only limited information to each network. Guocong Song *et al.* [40] introduced collaborative learning in which multiple classifier heads of the same network are simultaneously trained on the same training data to improve generalization and robustness to label noise with no extra inference cost. A similar learning strategy named On-the-fly Native Ensemble (ONE) for one-stage online distillation proposed by Xu Lan *et al.* [15]. Specifically, ONE trains only a single multi-branch network while simultaneously establishing a strong teacher on-the-fly to enhance the learning of target network. Chung *et al.* [41] proposed an online knowledge distillation method that transfers the knowledge of the class

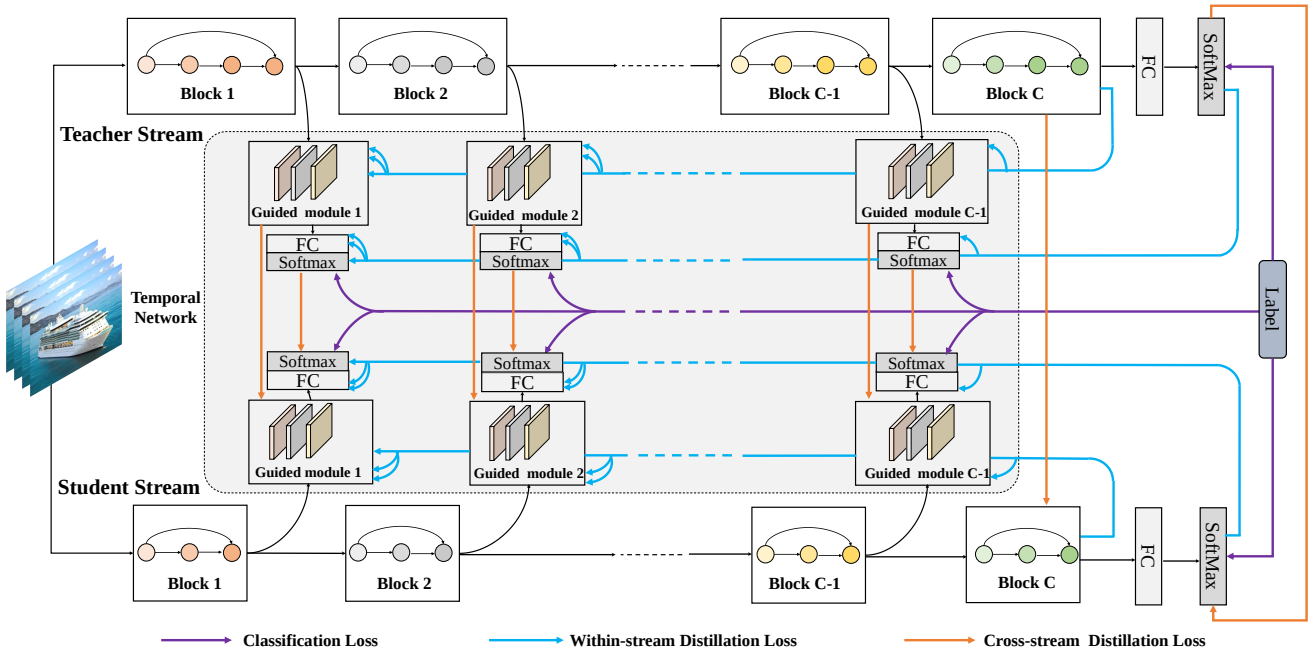


Fig. 2: The network details of each batch data process of our evolutionary knowledge distillation approach. We divide the teacher stream and student stream into C blocks, between each block of teacher and student, a pair of guided modules is introduced to help knowledge representation and transfer, so that the student can effectively capture both **what** knowledge the teacher learns and **how** it grows.

probabilities and the feature map using the adversarial training framework. Zhang *et al.* [42] proposed an online training framework called self-distillation, which forces student to refine its knowledge inside the network, thereby improving itself. And a framework Snapshot Distillation was proposed for teacher-student optimization in one generation [16], which extracted such information from earlier epochs in the same generation, meanwhile made sure that the difference between teacher and student is sufficiently large so as to prevent underfitting. After these, a novel two-level framework OKDDip [18] was proposed to perform distillation during training with multiple auxiliary peers and one group leader for effective online distillation. In OKDDip framework, the first-level distillation works as diversity maintained group distillation with several auxiliary peers, while the second-level distillation transfers the diversity enhanced group knowledge to the ultimate student model called group leader.

These online, timely and efficient training methods are promising and some good progress has been made in narrowing the capability gap between teacher and student model [15], [16], [42]. However, for online distillation, two factors still hold them back. First, although the teachers of online distillation methods are dynamic and have narrowed the capability gap, the gap still exists due to the lack of representation in detail and the process of learning. Second, owing to the absence of a qualified teacher role for online knowledge distillation, the insufficient and relatively unreliable supervision information will restrict the learning of student to some extent. There is still great room for improvement in knowledge transfer and representation. Therefore, we propose the evolutionary knowledge distillation approach to enhance the performance of student network learning by using an evolutionary teacher and focusing on intermediate knowledge of the teaching process dynamically.

III. The Proposed Approach

In our evolutionary knowledge distillation (EKD) approach, the teacher and student network are trained almost synchronously, and an evolutionary teacher can provide supervision information for the learning of student. The performance of teacher is continuously improved as the training process, which can provide richer supervision information and reduce the capability gap. As shown in Fig. 2, EKD consists of teacher stream, student stream and some extra guided modules. The network of each stream is divided into several blocks, depending on the specific network. Previous experiences have shown that using early blocks helps to exploit information inside the network [17], [28], [43], in order to utilize the intermediate knowledge, we improve bottleneck module [43], [44] to form some guided module pairs that are applied to assist the representation and transfer of knowledge. For each corresponding block between teacher and student stream, a guided module pair is inserted to make distillation process effective and efficient within stream and cross stream. Within-stream knowledge distillation aims to enhance knowledge representation, while cross-stream knowledge distillation helps to improve knowledge transfer.

A. Problem Formulation

Denote the training set is $\mathbb{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{|\mathbb{D}|}$, where $\boldsymbol{x}_i$ is the ith sample with label $\boldsymbol{y}_i \in \{1, 2, ..., m\}$. Here m is

Algorithm 1 Student network learning via EKD

Input: Training set $\mathbb{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{|\mathbb{D}|}$;
Output: Optimized student network weight: $\mathbf{w}_s$;

```
Initialize the Teacher and Student;
for The student model did not converge do
    Get a batch of data;
    Feed the data into teacher stream;
    Get the feature map, soften probabilities of the teacher stream;
    Compute the with-stream distillation loss with Eq. (3) and Eq. (4);
    Compute the classification loss with Eq. (9);
    Compute the total loss of teacher with Eq. (11);
    Compute gradient to model parameters w_t and update with the SGD optimizer.
    Feed the data into student stream;
    Get the feature map, soften probabilities of the student stream;
    Compute the with-stream distillation loss with Eq. (3) and Eq. (4);
    Compute the cross-stream distillation loss with Eq. (6) and Eq. (7);
    Compute the classification loss with Eq. (9);
    Compute the total loss of student with Eq. (11);
    Compute gradient to model parameters w_s and update with the SGD optimizer.
end for
Return the optimized student network weight w_s.
```

the number of classes. The objective of distillation is transferring the knowledge of teacher $\phi_t(\boldsymbol{x}; \mathbf{w}_t)$ into the student $\phi_s(\boldsymbol{x}; \mathbf{w}_s)$, where $\mathbf{w}_t$ and $\mathbf{w}_s$ are the model parameters of teacher and student, respectively. In our setting, the network of each stream is divided into C blocks. Between the corresponding teacher and student blocks, a pair of guided modules $\mathcal{M}_b = \left(\phi_g(\boldsymbol{f}_t^{(b)}; \mathbf{w}_t^{(b)}), \phi_g(\boldsymbol{f}_s^{(b)}; \mathbf{w}_s^{(b)})\right)$ is introduced to assist the representation, transfer of knowledge, where $\boldsymbol{f}_k^{(b)}$ and $\mathbf{w}_k^{(b)}$ are the feature maps and model parameters from the bth $(b = 1, ..., C-1)$ block of teacher $(k = t)$ and student $(k = s)$ respectively. Therefore, during student network learning, the problem of EKD is formulated as:

$$\min_{\mathbf{w}_t, \mathbf{w}_s, \{\mathbf{w}_t^{(b)}, \mathbf{w}_s^{(b)}\}} \mathcal{L}(\phi_t(\boldsymbol{x}; \mathbf{w}_t), \{\mathcal{M}_b\}_{b=1}^{C-1}, \phi_s(\boldsymbol{x}; \mathbf{w}_s)), \quad (1)$$

where $\mathcal{L}$ is the function to measure distillation loss. Different from traditional knowledge distillation where ϕ_t is fixed, EKD simultaneously learns both ϕ_t and ϕ_s from $\mathbb{D}$ with the help of several guided module pairs $\{\mathcal{M}_b\}_{b=1}^{C-1}$. After learning, the temporal network within the dashed box (see the Fig. 2) will be discarded and only the parameters of the main model of student stream $\mathbf{w}_s$ will be used for inference.

B. Evolutionary Distillation

The proposed evolutionary knowledge distillation adopts online training of teacher and student model synchronously, the parameters of student and teacher models are updated in each batch data process, which can reduce the teacher-student capability gap. When coupled with Fig. 2, we can see that teacher and student streams are divided into C blocks. Each block followed by a guided module with a fully connected layer constitutes multiple classifiers. We assume a stream with C classifiers. The training process includes two simultaneous stages, within-stream distillation and cross-stream distillation. For within-stream distillation, deeper classifiers provide supervision to help the learning of shallow classifiers, which can improve the ability of the stream itself to represent knowledge. Cross-stream distillation can improve knowledge transfer from the evolutionary teacher to student. The guided modules can help facilitate knowledge representation and transfer.

Within-stream Distillation. For within-stream distillation, we use the shallower classifiers $\phi_g(\boldsymbol{f}_k^{(b)}; \mathbf{w}_k^{(b)}), b \in \{1, ..., C-1\}$ as students, and use the deeper classifiers $\phi_k(\boldsymbol{x}; \mathbf{w}_k)$ and $\phi_g(\boldsymbol{f}_k^{(c)}; \mathbf{w}_k^{(c)})$, here $c \in \{b+1, ..., C-1\}$, as teachers. The students are trained by the supervision provided by teachers, supervision includes Kullback-Leibler (KL) divergence [9] and L2 distance [12] between the feature maps before the final FC layers of teacher and student. In order to simplify the form, here we only show the loss between the backbone network $\phi_k(\boldsymbol{x}; \mathbf{w}_k)$ and each guided modules $\phi_g(\boldsymbol{f}_k^{(b)}; \mathbf{w}_k^{(b)})$.

$$\min_{\mathbf{w}_k^{(b)}, \mathbf{w}_k} \mathcal{L}_w(\phi_g(\boldsymbol{f}_k^{(b)}; \mathbf{w}_k^{(b)}), \phi_k(\boldsymbol{x}; \mathbf{w}_k)), b \in \{1, ..., C-1\}. \quad (2)$$

In practice, we use the classic distillation loss, KL divergence, and L2 distance loss. We compute KL divergence loss between deep classifiers and shallow classifiers within stream. The output of teacher and student is $\phi_k(\boldsymbol{x}; \mathbf{w}_k)$ and $\phi_g(\boldsymbol{f}_k^{(b)}; \mathbf{w}_k^{(b)})$ respectively, $b \in \{1, ..., C-1\}$. And the distillation loss $\mathcal{L}_D$ is calculated by the following.

$$\mathcal{L}_D = \sum_{b=1}^{C-1} \ell(\phi_g(\boldsymbol{f}_k^{(b)}; \mathbf{w}_k^{(b)}), \phi_k(\boldsymbol{x}; \mathbf{w}_k)), \quad (3)$$

where ℓ denotes the KL divergence loss. As multiple different teachers provide different knowledge, we could achieve the more robust and accurate knowledge representation.

We compute feature loss $\mathcal{L}_F$ by L2 distance which can be obtained through computing the feature maps of deep classifier and shallow classifiers.

$$\mathcal{L}_F = \sum_{b=1}^{C-1} \left\| F_k - F_{g,k}^{(b)} \right\|_2^2, \quad (4)$$

where F_k and $F_{g,k}^{(b)}$ represent the feature maps of teacher or student before the FC layer respectively.

Cross-stream Distillation. For the cross stream knowledge transfer, the student will learn under the supervision of the corresponding guided modules of the teacher stream:

$$\min_{\mathbf{w}_t^{(c)}, \mathbf{w}_s^{(c)}} \mathcal{L}_b(\phi_t(\boldsymbol{x}; \mathbf{w}_t^{(c)}), \phi_s(\boldsymbol{x}; \mathbf{w}_s^{(c)})), c \in \{1, ..., C\}. \quad (5)$$

The cross-stream distillation loss contains two types of losses: distillation loss and feature loss. For distillation loss

between backbone networks and the one between each pair of guided modules, they were calculated by the KL divergence.

$$\begin{aligned}\mathcal{L}_{G1} =& \ell(\phi_t(\boldsymbol{x};\mathbf{w}_t),\phi_s(\boldsymbol{x};\mathbf{w}_s)) \\ &+\sum_{b=1}^{C-1}\ell(\phi_g(\boldsymbol{f_t^{(b)}};\mathbf{w}_t^{(b)}),\phi_g(\boldsymbol{f_s^{(b)}};\mathbf{w}_s^{(b)})),\end{aligned} \tag{6}$$

where the outputs of guided module attached to the teacher and student are $\phi_g(\boldsymbol{f_t^{(b)}};\mathbf{w}_t^{(b)})$ and $\phi_g(\boldsymbol{f_s^{(b)}};\mathbf{w}_s^{(b)})$, respectively, $b \in \{1,...,C-1\}$. The ϕ_t and ϕ_s denote the outputs of the backbone networks.

The form of feature loss is as follows,

$$\mathcal{L}_{G2} = \sum_{b=1}^{C-1}\left\|F_{g,t}^{(b)} - F_{g,s}^{(b)}\right\|_2^2, \tag{7}$$

where $F_{g,t}^{(b)}$ and $F_{g,s}^{(b)}$ represent the feature map of classifier before the FC layer respectively. The feature loss $\mathcal{L}_F$ between each pair of guided modules is to measure differences in the feature map, which can promote intermediate knowledge transfer cross stream. However, we divide the stream of teacher and student into several blocks, the output dimensions between each corresponding block may not match. Through the processing of the guided modules, we can ensure the consistency of the dimensions while ensuring the least loss of intermediate knowledge during the transfer process.

Then, we combine the two parts of the guided loss,

$$\mathcal{L}_G = \mathcal{L}_{G1} + \mathcal{L}_{G2}. \tag{8}$$

For each classifier, we compute cross entropy loss ℓ_{ce} between $\phi(\boldsymbol{x};\mathbf{w}^{(c)})$ and $\boldsymbol{y}$. In this way, the label $\boldsymbol{y}$ directs each classifier's probability as possible. There are multiple classifiers and we sum each cross entropy loss as follows:

$$\mathcal{L}_L = \sum_{c=1}^{C}\ell_{ce}\left(\phi(\boldsymbol{x};\mathbf{w}^{(c)}),\boldsymbol{y}\right). \tag{9}$$

As the parameters of teacher and student are constantly updated in each iteration, the evolutionary knowledge distillation process can be formulated as:

$$\min_{\mathbf{w}_t,\mathbf{w}_s,\mathbf{w}_t^{(b)},\mathbf{w}_s^{(b)}} \mathcal{L}^e = \sum\mathcal{L}_w^e + \sum\mathcal{L}_b^e, \tag{10}$$

where e denotes the number of iterations. Specifically, for teacher stream and student stream, we used different loss functions, $\mathcal{L}^{(t)}$ and $\mathcal{L}^{(s)}$, respectively.

$$\mathcal{L}^{(k)} = \mathcal{L}_D + \mathcal{L}_F + \mathcal{L}_L + \delta(k=t)\mathcal{L}_G, \tag{11}$$

where $\delta(k=t)$ is an indicator function which equals 1 if the stream k is teacher and 0 otherwise. This means that the training of teacher and student streams are similar but different. The student learns from evolutionary teacher constantly with the supervision provided by teacher network and guided modules. For each iteration, we first optimize a teacher network to provide supervision to the learning of student, after that, the optimization of teacher and student will be carried out synchronously. In general, the evolutionary teacher reduces the capability gap between teacher and student. The guided modules not only facilitate the knowledge representation of teacher and student stream but also promote the transfer of more intermediate and process knowledge from teacher to student.

C. Implementation Details

The detailed training of the EKD is shown in Algorithm 1. We introduce several identical guided modules on each stream, and each module is followed by a fully connected layer and a softmax layer. In training process, the two networks are trained at the same time, and they will adopt a certain degree of randomization in order to ensure they are not completely synchronized. This kind of incomplete synchronization can provide enough second-hand knowledge to promote student's learning [39], [16]. Once the training is completed, we only keep the student's parameters of the backbone network for inference.

The experimental results of other distillation methods are based on the open source code of CRD [13], we implement experiments according to their hyper-parameters settings. The networks are trained with SGD optimizer [51]. The hyper-parameters are set as follows, batch size is 64, the number of threads is 8, the initial learning rate is 0.1, and it will be multiplied by 0.1 when the epoch is equal to 75,130 and 180, respectively. We fix the random seed to 5 and the temperature T of distillation [9] to 4. All the experiments are implemented by PyTorch on a NVIDIA TITAN GPU.

IV. Experiments

A. Experimental Settings

To verify the effectiveness and adaptability of our EKD approach, we conduct experiments on five benchmarks, including CIFAR10 [52], CIFAR100 [52], Tiny-ImageNet [53], UMDFaces [54] and UCCS [55]. We first compare EKD with several state-of-the-arts offline distillation approaches, including KD [9], Fitnet [12], Attention [45], Factor [46], PKT [47], RKD [10], Similarity [48], Correlation [49], VID [**?**], Abound [50], CRD [13], Res-KD [56] and AFD [57]. We also compare with several online distillation approaches that do not use pre-trained teacher, including DML [28], CL-ILR [40], ONE [15], Snapshot-KD [16], OKDDip [18] and Self-KD [42]. Then, ablation experiments are conducted to study the impact of evolutionary teacher, guided modules and component of loss function. Finally, we further conduct the experiments on low-resolution and few-sample scenarios to verify the adaptability of our approach. In our experiments, we use VGG(bn) [58], ResNet[43], Wide ResNet [59], ShuffleNetV1 [60] and ShuffleNetV2 [61] as the backbone networks.

CIFAR10. The dataset consists of 60,000 32×32 colour images in 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images.

CIFAR100. This dataset is just like the CIFAR10, except it has 100 classes containing 600 images each. There are 500 training images and 100 testing images per class.

Tiny-ImageNet. It is an image classification dataset, which contains about 64×64 sized 100,000 training images and

TABLE I: Classification accuracy with peer-architecture setting on CIFAR100.

Teacher Network	VGG19(bn)	VGG11(bn)	ResNet50	ResNet18	ResNet101	ResNet50	WRN50-2
Student Network	VGG11(bn)	VGG11(bn)	ResNet18	ResNet18	ResNet50	ResNet50	WRN50-2
Teacher Acc. (%)	74.17	70.76	78.71	76.61	80.05	78.71	80.24
Student Acc. (%)	70.76	70.76	76.61	76.61	78.71	78.71	80.24
KD [9]	73.28	72.12	77.57	78.97	79.16	79.54	80.41
Fitnet [12]	71.51	71.00	76.59	77.58	79.76	79.18	80.84
Attention [45]	72.94	70.09	76.82	77.37	80.01	78.81	80.58
Factor [46]	68.12	68.12	75.26	73.81	77.54	73.07	80.12
PKT [47]	73.15	72.43	78.44	78.51	79.51	79.66	80.85
RKD [10]	72.81	71.88	78.21	78.10	80.07	78.73	80.40
Similarity [48]	73.49	72.10	78.71	78.55	80.39	79.44	80.87
Correlation [49]	70.90	71.35	77.48	77.83	78.17	77.18	80.37
VID [**?**]	72.83	72.32	77.71	78.16	78.17	76.76	80.88
Abound [50]	71.88	71.07	77.70	77.47	79.02	78.46	80.95
CRD [13]	71.73	73.00	78.92	78.18	80.25	79.45	81.14
EKD (Ours)	**74.12**	**74.52**	**82.05**	**80.74**	**81.91**	**82.48**	**82.53**

TABLE II: Classification accuracy with cross-architecture setting on CIFAR100.

Teacher Network	ResNet18	VGG11(bn)	ResNet18	WRN50-2	WRN50-2
Student Network	VGG8(bn)	ShuffleNetV1	ShuffleNetV2	ShuffleNetV1	VGG8(bn)
Teacher Acc. (%)	76.61	70.76	76.61	80.24	80.24
Student Acc. (%)	69.21	66.18	70.48	66.18	69.21
KD [9]	71.17	72.40	75.03	71.78	70.31
Fitnet [12]	70.59	70.50	72.24	70.46	70.04
Attention [45]	71.62	69.64	73.83	70.55	69.78
Factor [46]	68.06	68.16	69.99	70.51	70.12
PKT [47]	72.74	72.06	74.31	69.80	69.76
RKD [10]	71.03	70.92	73.26	70.58	70.41
Similarity [48]	73.07	72.31	74.95	70.70	70.00
Correlation [49]	69.82	70.70	72.21	70.66	69.96
VID [**?**]	71.75	70.59	72.07	71.61	71.00
Abound [50]	70.42	72.56	74.64	**74.28**	69.81
CRD [13]	73.17	72.38	74.88	71.08	72.50
EKD (Ours)	**73.82**	**73.18**	**75.26**	73.61	**74.05**

10,000 verification images with 200 classes. It is more difficult than CIFAR100 dataset. On Tiny-ImageNet, we first randomly adjust the crop size to 224×224, apply random horizontal flip and normalize it finally . For testing, we only resize the pictures to 224×224.

UMDFaces. This face dataset is collected from Internet, and contains 367,888 face annotations for 8,277 subjects. In our Experiments IV-F *Adaptability Analysis*, UMDFaces is as a training dataset.

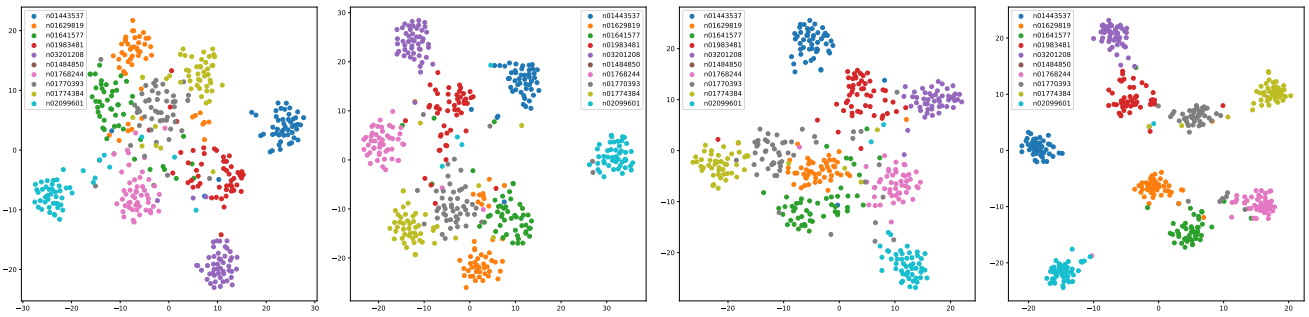
UCCS. This dataset contains16,149 images in 1,732 subjects in the wild conditions. It is a very challenging benchmark with various levels of challenges, including blurry image, occluded appearance and bad illumination. We follow the setting as [62], randomly select a 180-subject subset, and separate the images into 3,918 training images and 907 testing images, and report the results with the standard top-1 accuracy.

B. Comparisons with Offline Distillation Approaches

We conduct experiments on CIFAR100 dataset and perform the comparisons on the offline distillation with peer-architecture setting and cross-architecture setting. The experimental results of other offline distillation methods are based on the open source code *RepDistiller* of CRD [13], we implement experiments according to their hyper-parameters setting.

Peer-Architecture Setting. In this setting, the teacher and student networks share similar or the same structures. From the results shown in Tab. I, we get some meaningful observations. First, our approach outperforms all other offline distillation approaches, implying its remarkable effectiveness in improving student network learning, such as, when the teacher and student is VGG19(bn) and VGG11(bn) respectively, we achieve 74.12% accuracy on CIFAR100 which is 0.63% higher than the state-of-art method Similarity [48], and we gain 3.13% improvement when the teacher and student is ResNet50 and ResNet18 compared to CRD [13]. Second, the effect is also obvious when the teacher stream and student stream adopt the same structure. For example, when both the teacher and student are VGG11(bn), we gain 1.20% improvement even higher than learning from VGG19(bn), and the same is true when teacher and student are ResNet50. We speculate that it's due to the smaller capability gap between two identical backbone networks, which facilitates the knowledge transfer from evolutionary teacher to student.

Cross-Architecture Setting. According to the experimental



(a) Baseline (b) KD (c) CRD (d) EKD

Fig. 3: t-SNE Visualisation of ResNet18 trained with KD [9], CRD [13] and EKD on the Tiny-ImageNet dataset. Different numbers indicate different classes.

TABLE III: Classification accuracy (%) on CIFAR100 with different online distillation approaches.

Network	Baseline	DML [28]	CL-ILR [40]	Snapshot-KD [16]	ONE [15]	Self-KD [42]	OKDDip [18]	EKD (Ours)
VGG16(bn)	73.81	74.67	74.38	-	74.37	75.38	75.12	**76.86**
ResNet110	75.88	77.50	78.44	73.51	78.33	77.04	78.91	**79.28**

TABLE IV: Performance comparison on Tiny-ImageNet. The teacher is ResNet34 and the student is ResNet18.

Method	Student	KD [9]	FitNet [12]	Attention [45]	RKD [10]	CRD [13]	Res-KD [56]	AFD [57]	EKD (Ours)
Accuracy (%)	65.30	68.18	67.79	67.82	67.72	68.19	68.62	68.80	**69.46**

results of the previous section, we can find that the effect of proposed approach in the same or similar network architecture is superior to the conventional offline knowledge distillation methods. Then, to verify the generalization performance of our approach, we conduct further experiments on more challenging knowledge transfer tasks, cross-architecture setting, which means the architecture of teacher and student network are completely different.

We verify the performance of cross-architecture setting on CIFAR100 dataset, and learning rate, update strategy and simple data augmentation methods keep the same settings as before. The results are shown in Tab. II. We achieve the best results under the condition that the teacher and student networks are completely different. Specifically, when the teacher is ResNet18 and the student is VGG8(bn), our approach is at least 0.65% higher than the other offline distillation strategies, and for the teacher is WRN50-2, the student (VGG8 (bn)) gains a clear advantage, its performance improvement is more than 1.55%. When the teacher and student network are WRN50-2 and ShuffleNetV1 respectively, we achieve an improvement of at least 1.83% except when compared to Similarity [48]. These are due to the evolutionary teacher reduces the shackles caused by the capability gap between teacher and student, provides richer supervision information, and the guided modules ensure the efficiency of knowledge representation and transfer. Results on cross-architecture setting clearly demonstrate that our method does not rely on architecture-specific cues.

Comparisons on Tiny-ImageNet. In order to further verify the effectiveness of our proposed method, we conduct experiments on the more challenging Tiny-ImageNet. We mainly compare our approach with some typical methods, including KD [9], FitNet [12], Attention [45], RKD [10], CRD [13], Res-KD [56], AFD [57]. Here, Res-KD used the knowledge gap between teacher and student as a guide to train a more lightweight student network, which we call "res-student", and AFD is an attention-based feature matching distillation method utilizing all the feature levels of the teacher. As illustrated with Tab. IV, the proposed method EKD achieves better performance on large-scale datasets than other baseline knowledge distillation methods, for example, our method achieves a 0.66% performance improvement compared with the previous best results of AFD.

C. Comparisons with Online Distillation Approaches

We also compare the proposed EKD to several recent online knowledge distillation approaches, including Deep Mutual Learning (DML) [28], Collaborative Learning for Deep Neural Networks (CL-ILR) [40], Knowledge Distillation by On-the-Fly Native Ensemble (ONE) [15], Snapshot-KD [16], Self Distillation (Self-KD) [42] and Online knowledge distillation with diverse peers (OKDDip) [18]. The results in Tab. III are based on the code of OKDDip [18], in which the results of Snapshot-KD are based data from [16], and the results of Self-KD are obtained by re-implementing the framework of original research paper [42].

As shown in the Tab. III, the "Baseline" approach trains a model by ground-truth labels only. Compared with the other online knowledge distillation methods, our EKD shows some advantages, when the backbone network is VGG16(bn), our method is 1.48% higher in accuracy than the current best approach Self-KD [42]. And when the backbone network is ResNet110, our method has a 0.37% improvement than OKDDip [18], which is an increase of 3.40% from the

"Baseline". We suspect that the capability gap still exits due to the lack of representation in detail and the process of learning for previous online knowledge distillation methods. So that the insufficient and relatively unreliable supervision information from peers or itself will restrict the learning of student network to some extent. Our evolutionary teacher not only reduces the capability gap brought by fixed pre-trained teacher, but also provides a more flexible teacher role for online knowledge distillation instead of themselves or their peers. Above results illustrate that our method can transfer knowledge more adequately and effectively by combining guided modules and evolutionary teacher.

D. Visualisation

Previous experimental results quantitatively demonstrate the superiority of our proposed evolutionary knowledge distillation. In order to further visually demonstrate the advantages of our approach, we use the t-SNE [63] for visualization. t-SNE is a tool to visualize high-dimensional data. It converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. The Fig. 3 gives the visualisation results of ResNet18 trained with KD [9], CRD [13] and EKD on the Tiny-ImageNet dataset. The "Baseline" in Fig. 3 denotes that we train ResNet18 without any distillation methods, for KD, CRD and EKD, we train the student with the helps of teacher ResNet50. We randomly select ten classes in this dataset for visualization experiment, with 50 samples for each class, different numbers indicate different classes in Fig. 3. To begin with, it is obvious that our approach achieves more concentrated clusters than KD and CRD. In addition, as demonstrated in Fig. 3, the changes of the distances in classifiers of KD and CRD are more severe than that in classifier of EKD. We speculate that the evolutionary teacher can facilitate the student network learning by narrowing the teacher-student capability gap and the guided modules can enhance intermediate knowledge representation to improvement of student network learning.

E. Further Analysis

The approach we proposed mainly includes two parts, evolutionary teacher and guided modules. In this section, we conduct further experiments to explore the specific influence of evolutionary teacher, guided modules and components of loss function. In addition, we compare the influence of our approach in training time with other distillation approaches. All ablation experiments were performed in the same setting as previous ones.

Influence of Evolutionary Teacher. Here, we make ResNet50 and ResNet18 as teacher and student, respectively. As illustrated in Tab. V, for the classic knowledge distillation method KD [9], the accuracy of student on the CIFAR100 dataset is 77.57%. When we use an evolutionary teacher to teach student network, and other settings remain unchanged, the performance of our student network will increase by 2.14%, reach 79.71%. We believe it is because the evolutionary teacher provides more sufficient supervision information to promote student network learning. What's more, the guided modules play a positive role in within-stream distillation and cross-stream distillation, so, when the guided modules are working, can the evolutionary teacher still play an active role in the learning of student? In order to further verify the influence of evolutionary teacher, we added guided modules to teacher stream and student stream respectively, and compared the performance differences of student model before and after adopting evolutionary teacher strategy. As shown in Tab. V, Introducing the guided modules for the student stream only, the evolutionary teacher strategy improves the accuracy by 2.04%, while the introduction of the guided modules for the teacher and student streams, the evolutionary teacher strategy improves the performance by 2.49%. It's necessary to point out that the teacher model also has a good performance with the help of its guided modules, when our evolutionary teacher without guided modules to supervise the learning of student, it can still increase by 3.77% compared to KD (81.37% VS. 77.57%). The above results demonstrate that evolutionary teacher can promote student network learning by reducing the capability gap between teacher and student.

TABLE V: Classification accuracy on CIFAR100. The teacher is ResNet50 and the student is ResNet18. "ET" denotes evolutionary teacher without guided modules; "T_G" denotes teacher stream with guided modules; "S_G" denotes student stream with guided modules.

Approach	Accuracy (%)
KD	77.57
KD+ET	79.71 (+2.14)
KD+S_G	78.85
KD+S_G+ET	81.34 (+2.49)
KD+T_G+S_G	80.01
KD+T_G+S_G+ET (EKD)	82.05 (+2.04)

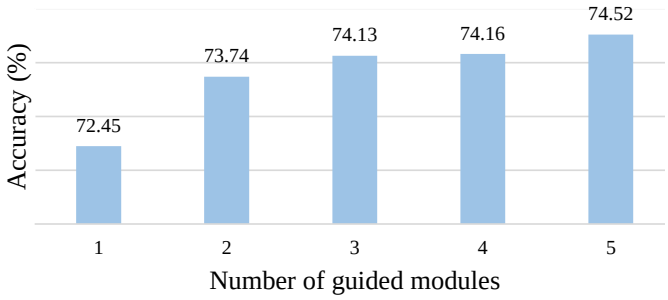


Fig. 4: The influence of guided modules. We train student models with different numbers of guided module pairs on CIFAR100 dataset.

Influence of Guided Modules. In order to explore the influence of introduced guided modules, we conduct further experiments. As shown in Tab. V, The effect of the guided modules to improve the performance of the student network is obvious. When only the student stream uses the guided modules, the performance of student is improved by 1.28% compared to KD (78.85% VS. 77.57%). When the teacher

stream and student stream use the guided modules at the same time, the performance is improved by 2.44% (80.01% VS. 77.57%). These results show that the guided modules promote the effective transfer of knowledge by making full use of the intermediate information of the network, and improve the performance of student.

In addition, the number of guided modules also has an important impact on knowledge transfer, so, we conduct experiments base on VGG11(bn). For the network structure, we use up to four guided modules and at least zero module. When the number of guided modules is zero, our approach essentially degenerates into classic knowledge distillation with evolutionary teacher. As shown in Fig. 4, experimental results show that guided modules play a vital role. When the number of guided modules increases, the effect is continuously improved, but as the number increases, the improvement becomes less obvious. It should be pointed out that due to the extremely simple structure of guided modules, their consumption of computing resources is negligible.

Influence of Loss Function. The training process of EKD includes two simultaneous stages, within-stream distillation and cross-stream distillation. The loss function mainly includes within-stream distillation loss ($\mathcal{L}_D$, $\mathcal{L}_F$) and cross-distillation loss ($\mathcal{L}_G$), $\mathcal{L}_L$ is classification loss of each classifier. We conduct related experiments to explore the influence of components of loss function (Eq. (11)). The results are shown in the Fig. 5, it is obvious that the loss function $\mathcal{L}_G$ has a promotion effect on the improvement of student network, which shows that the guided modules effectively promote the evolutionary teacher to transfer knowledge to student. $\mathcal{L}_D$ and $\mathcal{L}_F$ are the distillation losses that act on the within-stream feature level and predicted distribution level, respectively. All of them play a positive role in the learning of the student network, and the effect of $\mathcal{L}_D$ is more obvious, which indicate that the softened label distribution and feature maps that contain rich information about image intensity and spatial correlation provide sufficient supervision for the learning of students in within-stream and cross-stream distillation. In general, for within-stream distillation, deeper classifiers provide supervision to help the learning of shallow ones, thereby bringing about the performance improvement of the stream itself. Cross-stream distillation can improve knowledge transfer from the evolutionary teacher with the help of guided modules.

Training Time Analysis. We compare training time with other knowledge distillation methods, all the comparative experiments are implemented on NVIDIA TITAN GPU with identical hyper-parameters. In fact, the guided modules won't increase the amount of calculation too much because of their extremely simple structure, what's more, since the traditional distillation method needs to train the teacher and student model separately in two stages, our approach trains them at the same time, so the training time will not increase significantly. Specifically, when the teacher and student are ResNet50 and ResNet18 respectively, the training time of our approach is 277.5s/epoch, the traditional knowledge distillation method KD [9] training time for teacher and student are 205.29s/epoch and 171.07s/epoch. The memory of GPU occupied during training is slightly higher than that of traditional distillation methods. In general, our training time is not significantly higher than traditional methods, but brings considerable performance improvements.

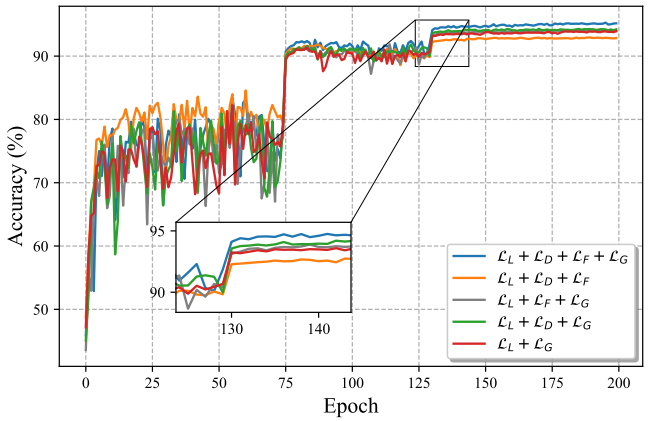


Fig. 5: Ablation study about components of loss function on CIFAR10 dataset. The teacher and student are VGG19(bn) and VGG11(bn) respectively.

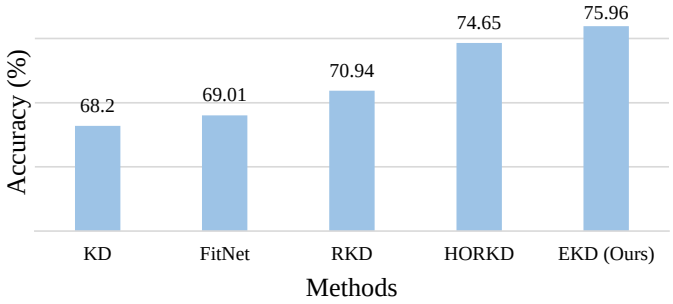


Fig. 6: Low-resolution (16×16) image classification accuracy on CIFAR100 dataset.

F. Adaptability Analysis

Low-resolution Scenario. To verify the performance of EKD in low-resolution scenario, we conduct experiments on challenging low-resolution image classification and low-resolution face recognition tasks.

For image classification task, we first train ResNet50 as teacher, and then make ResNet18 as student. The dataset used is down-sampled to reduce the resolution to 16×16, and the classification loss is classic cross-entropy loss.The experimental results are shown in Fig. 6, our approach shows good adaptability on more challenging low-resolution image classification task, which is 1.31% higher than the current best method HORKD [64]. In addition, HORKD needs to consume more computing resources due to the introduction of the assistant network. These results indicate the effectiveness of evolutionary teacher supervision and the introduced guided modules in representing and transferring knowledge in low-resolution scenario.

Then we conduct experiments on the low-resolution face recognition task which is very helpful in many real-world ap-

TABLE VI: The performance of various methods on UCCS benchmark. Our student model achieves great performance when working at low resolution (16×16) and costing less parameters.

Method	Top-1 Accuracy (%)	Parameters	Year
VLRR [65]	59.03	-	2016
SphereFace [66]	78.73	37M	2017
CosFace [67]	91.83	37M	2018
SKD [62]	67.25	0.79M	2019
AGC-GAN [68]	70.68	-	2019
VGGFace2 [69]	84.56	26M	2019
ArcFace [70]	88.73	37.8M	2019
LRFRW [71]	93.40	4.2M	2019
HORKD [64]	92.11	7.8M	2020
EKD (Ours)	**93.85**	**0.61M**	-

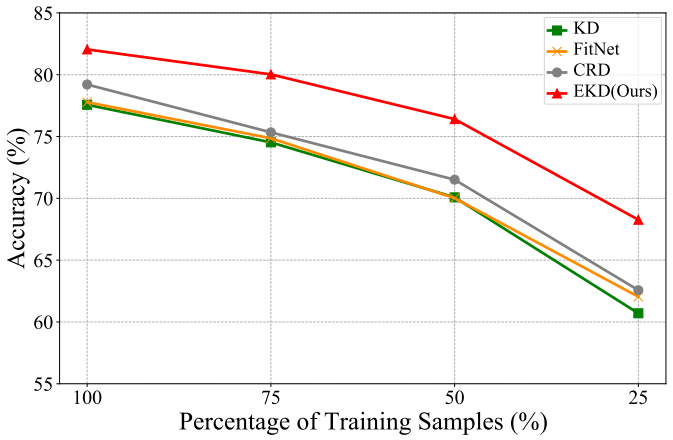


Fig. 7: Few-sample image classification results on CIFAR100 dataset. EKD trains teacher and student with subsets. KD, FitNet and CRD train the teacher models on full set and train the students with subsets.

plications, *e.g.*, recognizing low-resolution surveillance faces in the wild. In our experiments, the teacher uses a recent face recognizer VGGFace2 [69] with ResNet50 structure and the student network is based on streamlined ResNet18 with only 0.61M parameters, they are trained on UMDFaces [54] and tested on UCCS dataset [55]. In order to verify the validity of our low-resolution models, we emphatically check the accuracy when the input resolution is 16×16. As shown in Tab. VI, our student model achieves better low-resolution face recognition performance and costs less parameters. Specifically, we achieve a top-1 accuracy of 93.85% on the UCCS benchmark, which is 0.45% higher than the state-of-art LRFRW [71], and the amount of parameters is reduced by nearly ten times. Classical face recognition methods, such as ArcFace and CosFace, do not perform well in low-resolution scenario, with their highest recognition accuracy only reaching 88.73%. Moreover, their models have more parameters, which will lead to a significant increase in the computing cost for inference. It is worth mentioning that whether it is VLRR [65], SKD [62], HORKD [64] or LRFRW [71] in the distillation process, high-resolution images corresponding to low-resolution faces are necessary to provide more information, but such high-resolution images are not always easy to obtain, our approach only uses low-resolution face images for training, which adopts real-world application scenarios.

Few-sample Scenario. In practical scenario (*e.g.*, in the wild), the number of samples available for training is usually limited to a certain extent. In order to study the adaptability of our approach in a few-sample scenario, we conduct experiments on different subsets of CIFAR100 dataset. We randomly select images of each class to form new subsets, and use the newly designed training set to train the student models while maintaining the same test set. ResNet50 and ResNet18 are chosen as teacher and student, respectively. We compare the performance of student models with several typical distillation methods include KD [9], FitNet [12] and CRD [64]. The percentages of retained samples are 100%, 75%, 50% and 25%. For a fair comparison, we use the same data in different distillation approaches to train student models and train their teacher models on full dataset.

The results in Fig. 7 show that our approach remarkably surpasses other distillation approaches under few-sample scenario, even when the teacher stream of EKD is learned and distilled knowledge from less training data. Specifically, as the amount of training data decreases, the performance of distillation methods represented by KD, FitNet, and CRD in the figure will decrease significantly, while the downward trend of EKD proposed by us is significantly milder, even the advantage is more obvious when the training samples are less. When the percentage of training samples is 25%, our approach is nearly 10% higher than CRD.

V. Conclusion

In this paper, we propose an evolutionary knowledge distillation and show its superiority by comparing it with the state-of-the-art offline distillation and online distillation approaches. Our approach uses an evolutionary teacher to supervise the learning of student from scratch, which can reduce the teacher-student capability gap to promote knowledge transfer. What's more, through the introduction of some simple guided module pairs between corresponding teacher-student blocks, the efficiency of intermediate knowledge representation is improved. We believe this evolutionary knowledge transfer manner and simple guided mechanism are very promising in knowledge distillation community. In the future, we will explore the potential of our approach in combining with existing knowledge distillation schemes, and performing more practical tasks.

Acknowledgement. This work was partially supported by grants from the National Key Research and Development Plan (2020AAA0140001), National Natural Science Foundation of China (61772513), Beijing Natural Science Foundation (19L2040), and the project from Beijing Municipal Science and Technology Commission (Z191100007119002). Shiming Ge is also supported by the Youth Innovation Promotion Association, Chinese Academy of Sciences.

References

[1] M. R. Minar and J. Naher, "Recent advances in deep learning: An overview," *International Journal of Machine Learning and Computing*, pp. 747–750, 2018.

[2] R. Takahashi, T. Matsubara, and K. Uehara, "Data augmentation using random image cropping and patching for deep cnns," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2917–2931, 2020.

[3] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 742–751.

[4] Y. Chen, J. Wang, B. Zhu, M. Tang, and H. Lu, "Pixelwise deep sequence learning for moving object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 9, pp. 2567–2579, 2017.

[5] S. Ghosh, N. Das, I. Das, and U. Maulik, "Understanding deep learning techniques for image segmentation," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–35, 2019.

[6] M. Zhao, C. Zhang, J. Zhang, F. Porikli, B. Ni, and W. Zhang, "Scale-aware crowd counting via depth-embedded convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3651–3662, 2019.

[7] Y. Hu, J. Li, Y. Huang, and X. Gao, "Channel-wise and spatial feature modulation network for single image super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3911–3927, 2019.

[8] H. Zhai, S. Lai, H. Jin, X. Qian, and T. Mei, "Deep transfer hashing for image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 742–753, 2021.

[9] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Deep Learning and Representation Learning Workshop on Neural Information Processing Systems*, 2015.

[10] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[11] Y. Liu, J. Cao, B. Li, C. Yuan, W. Hu, Y. Li, and Y. Duan, "Knowledge distillation via instance relationship graph," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7096–7104.

[12] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *International Conference on Learning Representations*, 2015.

[13] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *International Conference on Learning Representations*, 2020.

[14] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4133–4141.

[15] X. Lan, X. Zhu, and S. Gong, "Knowledge distillation by on-the-fly native ensemble," in *Conference on Neural Information Processing Systems*, 2018, pp. 7528–7538.

[16] C. Yang, L. Xie, C. Su, and A. L. Yuille, "Snapshot distillation: Teacher-student optimization in one generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2859–2868.

[17] R. Anil, G. Pereyra, A. Passos, R. Ormándi, G. E. Dahl, and G. E. Hinton, "Large scale distributed neural network training through online distillation," in *International Conference on Learning Representations*, 2018.

[18] D. Chen, J.-P. Mei, C. Wang, Y. Feng, and C. Chen, "Online knowledge distillation with diverse peers," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 3430–3437.

[19] S.-I. Mirzadeh, M. Farajtabar, A. Li, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher," *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5191–5198, 2020.

[20] X. Jin, B. Peng, Y. Wu, Y. Liu, J. Liu, D. Liang, J. Yan, and X. Hu, "Knowledge distillation via route constrained optimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1345–1354.

[21] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 535–541.

[22] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Conference on Neural Information Processing Systems*, 2014, pp. 2654–2662.

[23] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," in *International Conference on Learning Representations*, 2018.

[24] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2604–2613.

[25] T. Liu, K.-M. Lam, R. Zhao, and G. Qiu, "Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[26] G. Pereyra, G. Tucker, J. Chorowski, L. Kaiser, and G. E. Hinton, "Regularizing neural networks by penalizing confident output distributions," in *International Conference on Learning Representations*, 2017.

[27] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," in *International Conference on Learning Representations*, 2017.

[28] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4320–4328.

[29] S. Ahn, S. X. Hu, A. C. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9163–9171.

[30] S. Chen, C. Zhang, and M. Dong, "Coupled end-to-end transfer learning with generalized fisher information," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4329–4338.

[31] J. Crowley, G. Gavin, and A. Storkey, "Moonshine: Distilling with cheap convolutions," in *Conference on Neural Information Processing Systems*, 2018, pp. 2893–2903.

[32] N. Passalis, M. Tzelepi, and A. Tefas, "Heterogeneous knowledge distillation using information flow modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2336–2345.

[33] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *International Conference on Machine Learning*, 2018, pp. 1607–1616.

[34] S. Yun, J. Park, K. Lee, and J. Shin, "Regularizing class-wise predictions via self-knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 876–13 885.

[35] S. You, C. Xu, C. Xu, and D. Tao, "Learning from multiple teacher networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1285–1294.

[36] S. Ruder, P. Ghaffari, and J. G. Breslin, "Knowledge adaptation: Teaching to adapt," *arXiv preprint arXiv:1702.02052*, 2017.

[37] Z. Shen, Z. He, and X. Xue, "Meal: Multi-model ensemble via adversarial learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4886–4893.

[38] Z. Shen and M. Savvides, "Meal v2: Boosting vanilla resnet-50 to 80%+ top-1 accuracy on imagenet without tricks," *arXiv preprint arXiv:2009.08453*, 2020.

[39] C. Yang, L. Xie, S. Qiao, and A. Yuille, "Knowledge distillation in generations: More tolerant teachers educate better students," *arXiv preprint arXiv:1805.05551*, 2018.

[40] G. Song and W. Chai, "Collaborative learning for deep neural networks," in *Conference on Neural Information Processing Systems*, 2018, pp. 1837–1846.

[41] I. Chung, S. Park, J. Kim, and N. Kwak, "Feature-map-level online adversarial knowledge distillation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 2006–2015.

[42] L. Zhang, J. Song, A. Gao *et al.*, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3712–3721.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[44] X. Cheng, Z. Rao, Y. Chen, and Q. Zhang, "Explaining knowledge distillation by quantifying the knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 925–12 935.

[45] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *International Conference on Learning Representations*, 2017.

[46] K. Jangho, P. Seonguk, and K. Nojun, "Paraphrasing complex network: Network compression via factor transfer," in *Conference on Neural Information Processing Systems*, 2018, pp. 2765–2774.

[47] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 268–284.

[48] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1365–1374.

[49] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5007–5016.

[50] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3779–3787.

[51] L. Bottou, "Stochastic gradient descent tricks," in *Neural networks: Tricks of the trade - Second Edition*, 2012, vol. 7700, pp. 421–436.

[52] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[53] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, p. 7, 2015.

[54] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa, "Umdfaces: An annotated face dataset for training deep networks," in *International Joint Conference on Biometrics*, 2017, pp. 464–473.

[55] M. Günther, P. Hu, C. Herrmann, C. H. Chan, M. Jiang, S. Yang, A. R. Dhamija, D. Ramanan, J. Beyerer, J. Kittler, M. A. Jazaery, M. I. Nouyed, G. Guo, C. Stankiewicz, and T. E. Boult, "Unconstrained face detection and open-set face recognition challenge," in *International Joint Conference on Biometrics*, 2017.

[56] X. Li, S. Li, B. Omar, and X. Li, "Reskd: Residual-guided knowledge distillation," *arXiv preprint arXiv:2006.04719*, 2020.

[57] M. Ji, B. Heo, and S. Park, "Show, attend and distill: Knowledge distillation via attention-based feature matching," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[59] N. K. S. Zagoruyko, "Wide residual networks," in *The British Machine Vision Conference*, 2016.

[60] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.

[61] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision*, 2018, pp. 116–131.

[62] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *TIP*, pp. 2051–2062, 2018.

[63] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[64] S. Ge, K. Zhang, H. Liu, Y. Hua, S. Zhao, X. Jin, and H. Wen, "Look one and more: Distilling hybrid order relational knowledge for cross-resolution image recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 10 845–10 852.

[65] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4792–4800.

[66] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[67] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[68] V. Talreja, F. Taherkhani, M. C. Valenti, and N. M. Nasrabadi, "Attribute-guided coupled gan for cross-resolution face recognition," in *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2019, pp. 1–10.

[69] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pp. 67–74.

[70] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[71] P. Li, L. Prieto, D. Mery, and P. J. Flynn, "On low-resolution face recognition in the wild: Comparisons and new techniques," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 8, pp. 2000–2012, 2019.

Kangkai Zhang received his B.S. degree in Electronical Information Science and Technology from the School of Electronic Information and Optical Engineering in Nankai University, Tianjin, China. He is now a Master student at the Institute of Information Engineering at Chinese Academy of Sciences and the School of Cyber Security at the University of Chinese Academy of Sciences, Beijing. His major research interests are deep learning and computer vision.

Chunhui Zhang received the B.S. degree from the School of Mathematics and Computational Science, Hunan University of Science and Technology, Xiangtan, China. He is currently pursuing the master's degree with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, and the School of Cyber Security, University of Chinese Academy of Sciences, Beijing. His major research interests include machine learning and visual tracking.

Shikun Li received the B.S. degree from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing, China. He is currently pursuing the Ph.D. degree with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, and the School of Cyber Security, University of Chinese Academy of Sciences, Beijing. His research interests include machine learning, data analysis, and computer vision.

Dan Zeng received her Ph.D. degree in circuits and systems, and her B.S. degree in electronic science and technology, both from University of Science and Technology of China, Hefei. She is a full professor and the Dean of the Department of Communication Engineering at Shanghai University, directing the Computer Vision and Pattern Recognition Lab. Her main research interests include computer vision, multimedia analysis, and machine learning. She is serving as the Associate Editor of the IEEE Transactions on Multimedia, the TC Member of IEEE MSA and Associate TC member of IEEE MMSP.

Shiming Ge (M'13-SM'15) is a Professor with the Institute of Information Engineering, Chinese Academy of Sciences. He is also the member of Youth Innovation Promotion Association, Chinese Academy of Sciences. Prior to that, he was a senior researcher and project manager in Shanda Innovations, a researcher in Samsung Electronics and Nokia Research Center. He received the B.S. and Ph.D degrees both in Electronic Engineering from the University of Science and Technology of China (USTC) in 2003 and 2008, respectively. His research mainly focuses on computer vision, data analysis, machine learning and AI security, especially efficient learning solutions towards scalable applications. He is a senior member of IEEE, CSIG and CCF.