Applied Intelligence (2023) 53:1997–2009
https://doi.org/10.1007/s10489-022-03486-4

Teacher-student collaborative knowledge distillation for image classification

Chuanyun Xu[1,2] · Wenjian Gao[1] · Tian Li[3] · Nanlan Bai[1] · Gang Li[1] · Yang Zhang[2]

Accepted: 7 March 2022 / Published online: 4 May 2022

Abstract

A single model usually cannot learn all the appropriate features with limited data, thus leading to poor performance when test data are used. To improve model performance, we propose a teacher-student collaborative knowledge distillation (TSKD) method based on knowledge distillation and self-distillation. The method consists of two parts: learning in the teacher network and self-teaching in the student network. Learning in the teacher network allows the student network to use knowledge from the teacher network. Self-teaching in the student network is to build a multi-exit network based on self-distillation and provide deep features as supervised information for training. In the inference stage, we use ensembles to vote on the classification results of multiple sub-models in the student network. The experimental results demonstrate the superior performance of our method compared with a traditional knowledge distillation method and a self-distillation-based multi-exit network.

Keywords Knowledge distillation · Self-distillation · Teacher-student collaborative · Ensemble

1 Introduction

With the rapid development of deep learning, convolutional neural networks have exhibited excellent performance in various computer vision tasks [1, 2]. In visual datasets, a category usually contains multi-view features that are easy to categorize, and simple models can only learn some of the features; however, deep neural networks can effectively handle this problem. As the number of network layers and parameters increase, models become prone to overfitting, which affects their performance.

Knowledge distillation is an important method of knowledge transfer; in this process, a lightweight model learns valid information from a heavy model to enhance performance. This model structure is often considered as a teacher-student structure. With an experienced teacher network in place, the inferior student network learns from the valuable information in the teacher network through knowledge distillation and achieves a performance improvement. Similarly, self-distillation allows a model to learn another pretrained network with the same structure. Due to the stochastic nature of feature learning and differences among model initialization methods, models obtain knowledge in different ways. The performance of a network can also be effectively improved by knowledge transfer between models. A model can learn knowledge from other models or itself to improve performance. However, it is still unclear whether a model can achieve performance improvements under the guidance of both a teacher model and itself.

In this paper, our soft label information comes from the teacher network and the output of student network, therefore the student network can be regarded as its own

✉ Wenjian Gao
gwj@2019.cqut.edu.cn

Chuanyun Xu
20210012@cqnu.edu.cn

Tian Li
tian.li@rwth-aachen.de

Nanlan Bai
1354970273@qq.com

Gang Li
ligang@cqut.edu.cn

Yang Zhang
495461428@qq.com

1 School of Artificial Intelligence, Chongqing University of Technology, Chongqing, 400054, China

2 College of Computer and Information Science, Chongqing Normal University, Chongqing, 401331, China

3 Computer Science Department, RWTH Aachen University, Aachen, 52074, Germany

second teacher. Similar to multi-teacher distillation, in our approach, we let a single model learn as many view features as possible from multiple networks. In the process of human learning, students guided by teachers can further improve their abilities with self-reflection. Inspired by this approach, a combination of teacher-student knowledge distillation and student self-distillation is used to enhance the performance of neural networks, and a method called teacher-student collaborative knowledge distillation (TSKD) is proposed. This method not only utilizes the category information from the teacher network but also absorbs student's knowledge. To construct the student self-distillation model, the student network builds multiple exit classifiers from shallow to deep. A shallow classifier can be regarded as a young student, and each classifier is an independent branch of the network, with shared convolutional layers. During training, each classifier receives supervision from the teacher network based on real labels, and the student network deep classifiers also guide the shallow layers through self-distillation. During testing, the category probabilities of multiple weak exits are combined to form a strong network of students. We conduct experiments on the CIFAR100 and Tiny-ImageNet datasets. The experimental results show that the proposed method can significantly improve the classification performance of the model, and the proposed model outperforms existing knowledge distillation methods and multi-exit self-distillation methods. Finally, the experimental results are analyzed in detail.

The main contributions of this paper are summarized as follows:

1. We propose a teacher-student collaborative distillation method. In contrast to traditional knowledge distillation methods in which only the teacher network provides a priori knowledge, our approach also allows the student network to learn from itself. The loss function of collaborative distillation is constructed by combining knowledge distillation and self-distillation, which fully exploits the performance of both optimization methods and thus improves the classification accuracy of the network.
2. We propose a new architecture incorporating a multi-exit network and a teacher-student model. Each exit in the multi-exit network is guided by soft logits from the teacher network, and thereby the classification performance of the multi-exit network is improved. At the same time, the student network benefits from the multi-exit network based on self-distillation.
3. To obtain a strong classifier, we use ensembles to vote on the classification results of multiple sub-models in the student network during the testing stage. Comparative experiments with multiple datasets and different teacher-student frameworks demonstrate the effectiveness and robustness of the proposed method.

2 Related work

2.1 Knowledge distillation

Knowledge distillation (KD), an important method of model compression [3–5], is effective in transferring "dark knowledge" from a larger model to a smaller model, allowing the smaller model to approximate the performance level achieved by the larger model [6–8]. This concept was first proposed in [9], but then was not explicitly explained. In 2014, [10] proposed an approach that enables a student network to learn the soft targets output by a teacher network and defined the method as knowledge distillation. However, conventional knowledge distillation methods only learn the output of the teacher network, which leads to the loss of intermediate layer knowledge. Later approaches attempted to exploit the information contained in middle model layers by designing different knowledge representations rather than just using the output information [11–17]. For example, [11] proposed an approach in which the student network simulates not only the output of the teacher network but also the hidden layer characteristics of the teacher network. [12] used attention transfer mechanisms to significantly improve its performance by forcing the student network to mimic the attention map of the powerful teacher network. Although the above algorithms utilized knowledge from the teacher network, they only consider the output of a specific layer of the teacher network. The relational knowledge distillation (RKD) approach proposed by [15] can transfer the structured relationships associated with the output results obtained by the teacher network to the student network, which alleviates the above problem. The correlations among different categories of probabilities may contain useful information to regularize a learning problem, and [16] found that the generation gap between teacher and student representation of mutual information can be minimized through contrastive representation distillation. Based on an adversarial-based learning strategy as a supervisor to guide and optimize lightweight student networks and recover knowledge from teacher networks, [18] recently proposed a knowledge distillation method for one-stage object detection . [19] constructed a compressed model to learn low-dimensional spatial information from potential representations of teacher networks. Most studies have focused on the representation of feature knowledge or methods of maximizing the transfer of teacher network feature knowledge while ignoring the potential capabilities of student networks. In this paper, we build a multi-exit student model based on the traditional knowledge distillation structure and use the deep feature and category information in the student network as supervision information to guide student network training, which results in improved performance.

2.2 Self-distillation

Self-distillation is a new approach that was developed from knowledge distillation. Unlike traditional knowledge distillation architectures, the teacher-student architecture of self-distillation uses the same model [20–22], or a network framework without teachers [23–25]. [20] used neural networks to study knowledge distillation from a new perspective; instead of compressing the model, the student network is optimized based on a teacher network with an equivalent parameter settings. [24] proposed a general training framework for self-distillation by constructing a multi-exit network for teacher-free distillation within the student network itself. [26] argued that self-distillation, as a regularization method, mitigates the overconfident predictions of the network and reduces intraclass gaps. All of the above studies found that self-distillation can effectively improve the performance of a student network. Although self-distillation gets rid of the need for a strong network of teacher, it loses the guidance of teacher network. In contrast to [24], we use a shared weighting strategy for the fully connected layer of the applied multi-branch network to reduce the number of model parameters. Moreover, each branch receives guidance from an extra teacher network. In particular, we found that the introduction of new teacher knowledge in the self-distillation network further enhances the effectiveness of self-distillation.

2.3 Ensemble

Ensembles have been widely used to improve model performance [27–29]. Since different models could be complementary, the outputs of multiple models with the same structure and different initialization training ensembles can be used to improve test performance. Several studies have found that ensembles are also effective in improving knowledge distillation performance. [30] found that ensemble teacher networks can effectively improve student network classification performance. To overcome offline-distillation issues, a strong teacher network is needed, [31] combined knowledge distillation and an ensemble approach to train a multi-branch network and then built a strong teacher model based on the branches of the ensemble to enhance the learning capabilities of the target network. However, this approach undoubtedly leads to a complex teacher model. In contrast with multiple teacher network ensembles [32], our student network constructs multiple exit outputs from shallow to deep, and only a small number of parameters are added to achieve the effect of multiple model ensembles. Finally we use ensembles to vote on the classification results of multiple sub-models in the student network, and obtain a strong classifier.

3 Proposed method

In this section, we start by reviewing the classical knowledge distillation algorithms and then introduce the overall framework of the teacher-student collaborative knowledge distillation network proposed in this paper.

3.1 Knowledge distillation

The teacher network function t and the student network function s are defined as follows:

$$t = f^t(x, w_t) \tag{1}$$

$$s = f^s(x, w_s) \tag{2}$$

where x represents the network input, w_t and w_s are the parameters of the teacher network and the student network, respectively. For convenience, t and s also represent the logits of the teacher network and student network outputs. $\mathcal{L}_{KL}$ refers to Kullback–Leibler divergence, $x^{(j)}$ denotes the j-th input image in N data samples. The Kullback–Leibler divergence measures the distance between the student and teacher output logits, which can be measured as:

$$\frac{1}{N}\sum_{j=1}^{N}\mathcal{L}_{\mathrm{KL}}\left(f^s\left(x^{(j)}, w_s\right), f^t\left(x^{(j)}, w_t\right)\right) \tag{3}$$

$\mathcal{L}_{CE}$ refers to cross-entropy loss, and y_j represents the true label of the j-th input image. The distance between the predicted value of the student network and the true label is defined as:

$$\frac{1}{N}\sum_{j=1}^{N}\mathcal{L}_{\mathrm{CE}}\left(f^s\left(x^{(j)}, w_s\right), y_j\right) \tag{4}$$

The optimization goal of knowledge distillation is to minimize the gap between the output of students and the prediction of teacher, as well as that between the output and the true label [10]:

$$\underset{w_s}{argmin}\sum\left(\alpha\tau^2 \cdot \mathcal{L}_{\mathrm{KL}} + (1-\alpha)\cdot\mathcal{L}_{\mathrm{CE}}\right) \tag{5}$$

Where w_s denotes the parameters of the student, α denotes the weight of KL divergence. Here, τ is defined as the distillation temperature, which is used as a hyper-parameter related to the degree of target softening.

3.2 TSKD

The whole framework of the teacher-student collaborative knowledge distillation method proposed in this paper is shown in Fig. 1, and it consists of two parts: a teacher network and a multi-exit output student network, in which the teacher network is usually a large pretrained network and only the student network is involved in training and testing.

Given C categories of N data samples, for the input sample $x \in \{x_i\}_{i=1}^{N}$, z^k represents the output of the fully connected layer about k category, k means category index. the k category probability of the teacher model output is expressed as:

$$t^k = \frac{\exp\left(z^k/\tau\right)}{\sum_k^C \exp\left(z^k/\tau\right)} \tag{6}$$

Similarly, the output of the student model can be represented as s^k. $\tau = 1$ indicates the standard SoftMax function.

Our model constructs n exit classifiers, and the training loss for any $m \in [1, n)$ classifier has two components. The first component is the loss associated with regular knowledge distillation, which is based on the KL divergence between teacher and students and the cross-entropy between student outputs and labels. The second part is related to self-distillation loss. The deepest classifier (exit n) in the multistage classifier is used as the second teacher, and it promotes the use of valuable knowledge of logits and features to guide shallow classifier learning.

$$Loss_{KD}^{m} = \alpha\tau^2 \cdot \mathcal{L}_{\mathrm{KL}}\left(s_m, t\right) + (1-\alpha) \cdot \mathcal{L}_{\mathrm{CE}}\left(s_m, y\right) \tag{7}$$

$$Loss_{SD}^{m} = \alpha\tau^2 \cdot \mathcal{L}_{\mathrm{KL}}\left(s_m, s_n\right) + \beta \cdot \left\|\mu_m\left(F_m\right) - F_n\right\|^2 \tag{8}$$

In the above equation, s_m and t represent the m-th classifier in the student network and the soft logits of the teacher network output based on the temperature τ, respectively. y stands for true label, F_m and F_n denote the feature output before the fully connected layer in the m-th classifier and deepest exit branch of the student network, respectively. To ensure the scales of F_m and F_n are consistent, an adaptive bottleneck layer is added to each exit network, which is similar to the bottleneck layer structure in ResNet50, consisting of a downsampling layer with a 3x3 convolution kernel and a bottleneck structures with 1x1, 3x3, 1x1 convolution kernels. On the one hand, we succeeded in maintaining the scale consistency, and on the other hand, we managed to reduce the number of parameters as much as we could. Adaptability is guaranteed as the use of different numbers of bottleneck modules depends on the size of feature map. we denote the m-th module as $\mu_m\left(F_m\right)$. The L2 loss function is used to minimize the gap between the feature maps of the shallower network and the deepest convolution layer, and α and β is defined as the Kullback–Leibler divergence

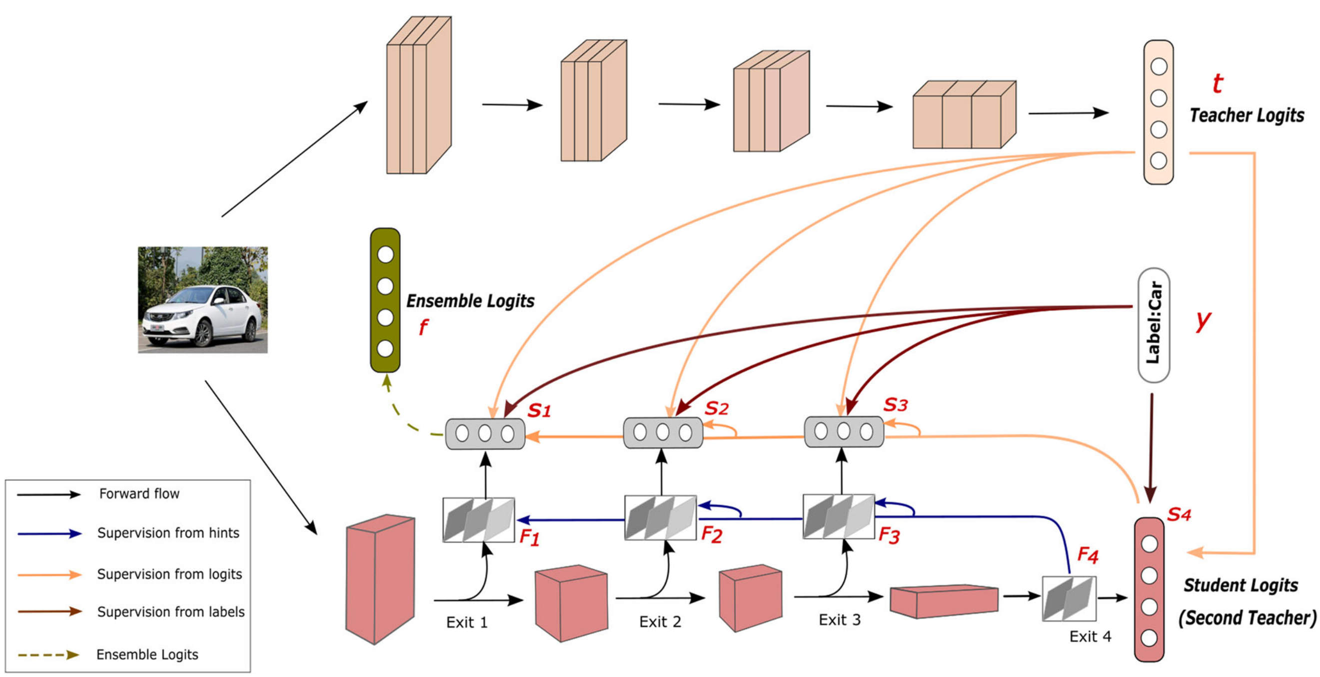


Fig. 1 The details of our approach. The whole framework consists of an offline teacher network and a student network. (i) The teacher network transfers soft logits to guide the student network. (ii) The student network adds a bottleneck layer and a fully connected layer after each block to build a multi-exit network from shallow to deep. (iii) Each shallow classifier receives supervision from the teacher network, its own deepest classifier and the true labels. (iv) Each classifier is combined in an ensemble to form a strong classifier

and L2 loss weights, respectively, $1-\alpha$ is defined as $\mathcal{L}_{CE}$ weights is for weight normalization.

Thus, the total student network loss can be expressed as:

$$Loss = \sum_{m=1}^{n} \left(Loss_{KD}^{m} + Loss_{SD}^{m} \right) \tag{9}$$

For testing, in the student network, we use an average ensemble algorithm to fuse the exits with different classification performance. Different from multi-teacher network ensembles and multi-student collaborative ensembles, we averagely integrate the multi-exit outputs of the student network without introducing additional models, which can effectively reduce model complexity. S_m represent the output of the m-th classifier exit, and f represents the final output of the model.

$$f = \frac{1}{n} \sum_{m=1}^{n} S_m \tag{10}$$

4 Experiments

This section first introduces the datasets and hyper-parameter settings used in the experiments. Then we compare the benchmark method, the traditional knowledge distillation method and a multi-exit network.

4.1 Benchmark datasets and implementation details

1. CIFAR100 [33]: This dataset was collected by Alex Krizhevsky, Vinod Nair and Geoffrey Hinton, with a total of 60k color images of size 32x32 divided into 100 categories; additionally, the dataset includes 50k training samples and 10k test samples. The data preprocessing method used was based on the CRD [16] processing method. The training set images were filled with 4 pixels on each side and then randomly cropped to 32x32 with random horizontal flipping at a probability of 0.5. For testing, the original images were used for evaluation. The experiments were performed using SGD optimization, and the weight decay and momentum were set to 0.0001 and 0.9, respectively. The batch size was set to 128, the initial learning rate was 0.1, the epoch was reduced to 0.1 times the previous value at 150, 180 and 210 epochs, and the training ended after 240 rounds. The temperature (T) for computing soft targets was set to 3.0. We set α = 0.3 and β= 0.03 in knowledge distillation loss function. The exit number (n) was set to 4. All the experiments were implemented in PyTorch on GPU (RTX2080s) devices.
2. TinyImageNet [34]: A subset of ImageNet released by Stanford University in 2016 was used in this study. A total of 120k RGB images of size 64x64 were divided into 200 categories, there are 100k training samples, 10k validation samples and 10k test samples were used. Preprocessing involved a simple random horizontal flip, and training and testing were performed at the original image size. The optimization approach and hyper-parameter settings were the same as those for the CIFAR dataset.

4.2 Comparison with the benchmark method

The traditional ResNet [2], VGG [35], WRN [36] and ShuffleNet [37, 38] were chosen as the backbone networks for the experiments. To fuse the different stages of knowledge learning in the teacher network and student network, we constructed a multi-exit output student network under regular teacher guidance. For convenience, three independent classifier branches were inserted between blocks with decreasing feature space resolution, and each branch contained a bottleneck layer and a fully connected layer. The bottleneck layer ensured that the output feature map size remained consistent and mitigated the impact of variations among shallow classifiers. Table 1 shows the performance result of each branch of the student network on CIFAR100, and we found that semantic features were captured differently due to the different depths of the networks. Comparatively, the deep classifier possessed higher classification accuracy than the shallow classifier. An ensemble was applied during testing, and the highest weights were assigned to the classification exits with high classification accuracy. The experimental results show that the final test accuracies of our method are all improved by 4%-7% compared to those of the baseline methods. In addition, we found that our teacher-student collaborative knowledge distillation method outperformed the baseline methods in the early stage.

4.3 Comparison with traditional knowledge distillation methods

To show the effectiveness and robustness of the teacher-student collaborative distillation method proposed in this paper, we chose seven different teacher-student architectures with both homogeneous and heterogeneous models and compared them with some mainstream knowledge distillation methods. Most of the experimental methods were implemented based on the original open-source codes, and a few methods were based on the information in [16] for both the CIFAR100 and TinyImageNet datasets. The classification accuracy and number of parameters were used as evaluation metrics, and the classification accuracy is shown in Tables 2 and 3. The number of model parameters is shown in Table 4. Since we construct a multi-exit network

Table 1 Comparison of the accuracy of the proposed method and benchmark methods

Neural Networks(T/S)	Baseline(T/S)	Classifier1/4	Classifier2/4	Classifier3/4	Classifier4/4	Ensemble
ResNet152/ResNet50	80.88/77.98	80.30	81.18	81.40	81.63	83.09
ResNet152/ResNet18	80.88/77.09	75.01	76.87	79.35	80.46	81.27
ResNet152/ResNet10	80.88/75.37	75.39	76.22	76.44	77.87	79.34
ResNet50/VGG8	79.34/70.36	68.98	69.49	70.12	73.23	75.63
VGG13/VGG8	76.64/70.36	68.05	69.13	69.58	72.65	75.42
WRN40-2/ShuffleNetV1	75.61/70.50	72.43	76.80	75.21	73.31	77.98
ResNet32x4/ShuffleNetV2	79.42/71.82	70.63	74.27	75.56	74.51	78.16

based on the student network, which leads to a slightly higher number of parameters than considered in the traditional KD algorithm but considerably fewer parameters than considered in the teacher network, we can also achieve a good model compression effect. Moreover, in terms of classification accuracy, our method is slightly lower than the SOTA HSAKD on the ResNet series networks. However, our approach exhibits outstanding performance on VGG and ShuffleNet. **Bold** and <u>underline</u> denote the best and the second best results, respectively.

4.4 Comparison with the multi-exit networks

Our student network can also be considered a kind of multi-exit classification network based on knowledge distillation. The main difference from the past multi-classifier networks proposed in [24] is that each of our classifiers receives supervision from the teacher network rather than just the deep classifier. Deeply supervised net (DSN) [45], on the other hand, constrains the intermediate layer with real labels to improve classification accuracy by mitigating gradient explosion or gradient disappearance. To verify the effectiveness of the proposed method, the two methods were experimentally compared. ResNet152 was selected as the teacher network, and ResNet18 and ResNet50 were used as the backbone networks of the multi-exit model. The experimental results are shown in Table 5. For both the shallow classifier and the final output of the model, the multi-exit student network based on collaborative distillation in this paper exhibited superior performance. In particular, the output of the first classifier is improved by 7.78% and 7.16% for ResNet18 and 12.43% and 12.07% for ResNet50, respectively. Knowledge distillation allows the multi-exit network to learn effectively knowledge from an additional teacher network. Our method effectively demonstrates the potential of shallow networks, thus enabling flexible deployment with limited hardware resources. **Bold** denote the best results.

Table 2 Comparison of the accuracy of the proposed method and knowledge distillation methods on CIFAR100

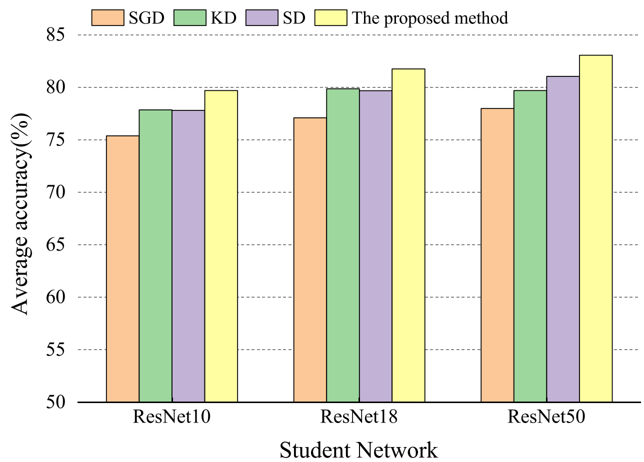
Teacher Student	ResNet152 ResNet50	ResNet152 ResNet18	ResNet152 ResNet10	ResNet50 VGG8	VGG13 VGG8	WRN40-2 ShuffleNetV1	ResNet32x4 ShuffleNetV2
Baseline	80.91	80.91	80.91	79.35	74.64	75.61	79.42
	77.98	77.09	75.37	70.36	70.36	70.50	71.82
KD [10]	79.69	79.86	77.85	73.81	72.98	74.83	74.45
FIT [11]	80.51	79.24	78.02	73.24	73.22	73.73	73.54
AT [12]	80.41	80.19	78.45	74.01	73.48	73.32	72.73
SP [13]	80.72	79.87	78.25	73.52	73.49	74.52	74.56
CC [39]	79.89	79.82	78.06	73.48	73.04	71.38	71.29
VID [14]	79.24	79.67	77.80	73.46	72.97	73.63	73.40
RKD [15]	80.22	79.60	77.90	73.51	73.19	72.21	73.21
PKT [40]	80.57	79.44	78.41	73.61	73.25	73.89	74.69
AB [17]	81.21	79.50	78.18	73.65	73.35	73.34	74.31
FT [41]	80.37	79.26	77.53	72.98	73.44	72.03	72.50
CRD [16]	80.53	79.81	78.60	74.58	74.29	76.27	76.05
SSKD [42]	80.29	80.36	78.60	<u>75.36</u>	<u>75.27</u>	77.32	77.45
HSAKD[43]	**83.33**	**82.17**	**79.75**	75.20	75.07	<u>77.51</u>	<u>77.89</u>
TSKD(ours)	<u>83.09</u>	<u>81.27</u>	<u>79.34</u>	**75.63**	**75.42**	**77.98**	**78.16**

Table 3 Comparison of the accuracy of the proposed method and knowledge distillation methods on TinyImageNet

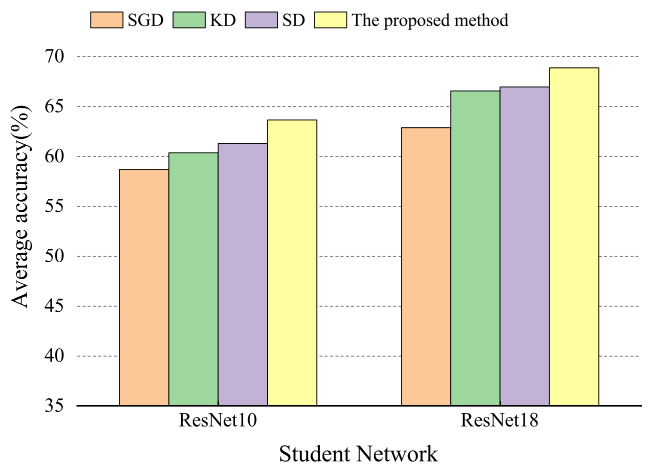
Teacher Student	ResNet34 ResNet18	ResNet50 ResNet10
Baseline	65.64	66.26
	62.86	58.70
KD [10]	66.54	60.34
FIT [11]	67.18	61.30
AT [12]	66.66	61.94
SP [13]	67.56	62.18
CC [39]	66.80	61.90
VID [14]	67.56	62.32
RKD [15]	66.92	61.84
AB [17]	65.42	62.58
FT [41]	65.92	62.26
CRD [16]	67.66	61.96
AFD[44]	68.10	62.52
TSKD(ours)	**68.86**	**63.64**

Table 4 Comparison of the number of parameters (M) in the student network (ResNet152 and VGG13 were used as teacher networks with no change in the number of parameters)

Methods	KD	TSKD
ResNet152	58.348	58.348
VGG13	9.923	9.923
ResNet50	23.712	37.812
ResNet34	21.798	22.055
ResNet18	11.227	12.334
ResNet10	4.957	5.859
VGG8	4.426	5.383



(a) CIFAR100



(b) TinyImageNet

Fig. 2 Ablation experiments on CIFAR100 and TinyImageNet datasets

Table 5 Comparison of the accuracy of the proposed method and multi-exit networks on CIFAR100

Network	Method	Classifier1/4	Classifier2/4	Classifier3/4	Classifier4/4	Ensemble
ResNet18	Baseline	-	-	-	77.09	-
	DSN [45]	67.23	73.80	77.75	78.38	79.27
	SD [24]	67.85	75.57	78.23	78.64	79.67
	TSKD(ours)	**75.01**	**76.87**	**79.35**	**80.46**	**81.27**
ResNet50	Baseline	-	-	-	77.98	-
	DSN[45]	67.87	73.80	75.54	80.27	80.67
	SD [24]	68.23	74.21	75.23	80.56	81.04
	TSKD(ours)	**80.30**	**81.18**	**81.40**	**81.63**	**83.09**

Table 6 The results of ablation experiments with different strategies on CIFAR100

Method	Model	Accuracy (%)
Teacher	ResNet152	80.91
Student	ResNet18	77.09
Logits(T)	ResNet152- ResNet18	79.37
Logits(T)+Logits(S)	ResNet152- ResNet18	80.37
Logits(T)+Logits(S)+Features(S)	ResNet152- ResNet18	81.03
Logits(T)+Logits(S)+Features(S)+Ensemble	ResNet152- ResNet18	81.27

5 Analysis

In this section, we further analyze the observations from the experiment. Firstly, the performance of each strategy is examined with ablation experiments. Secondly multi-exit network features for dimensionality reduction visualization. Finally, the effectiveness of ensemble modules are analyzed.

5.1 Ablation study

Since our method is implemented based on knowledge distillation between teacher and student and the self-distillation of the student network, it is unclear whether the improvement is associated with knowledge distillation or self-distillation. Different networks and datasets are selected, and three methods, including stochastic gradient descent (SGD), knowledge distillation (KD) and self-distillation (SD), are implemented for comparison, with classification accuracy as the evaluation metrics. The experimental results are shown in Fig. 2, and the proposed method significantly outperforms conventional knowledge distillation and self-distillation.

In addition, the teacher-student collaborative distillation method proposed in this paper incorporates three types of

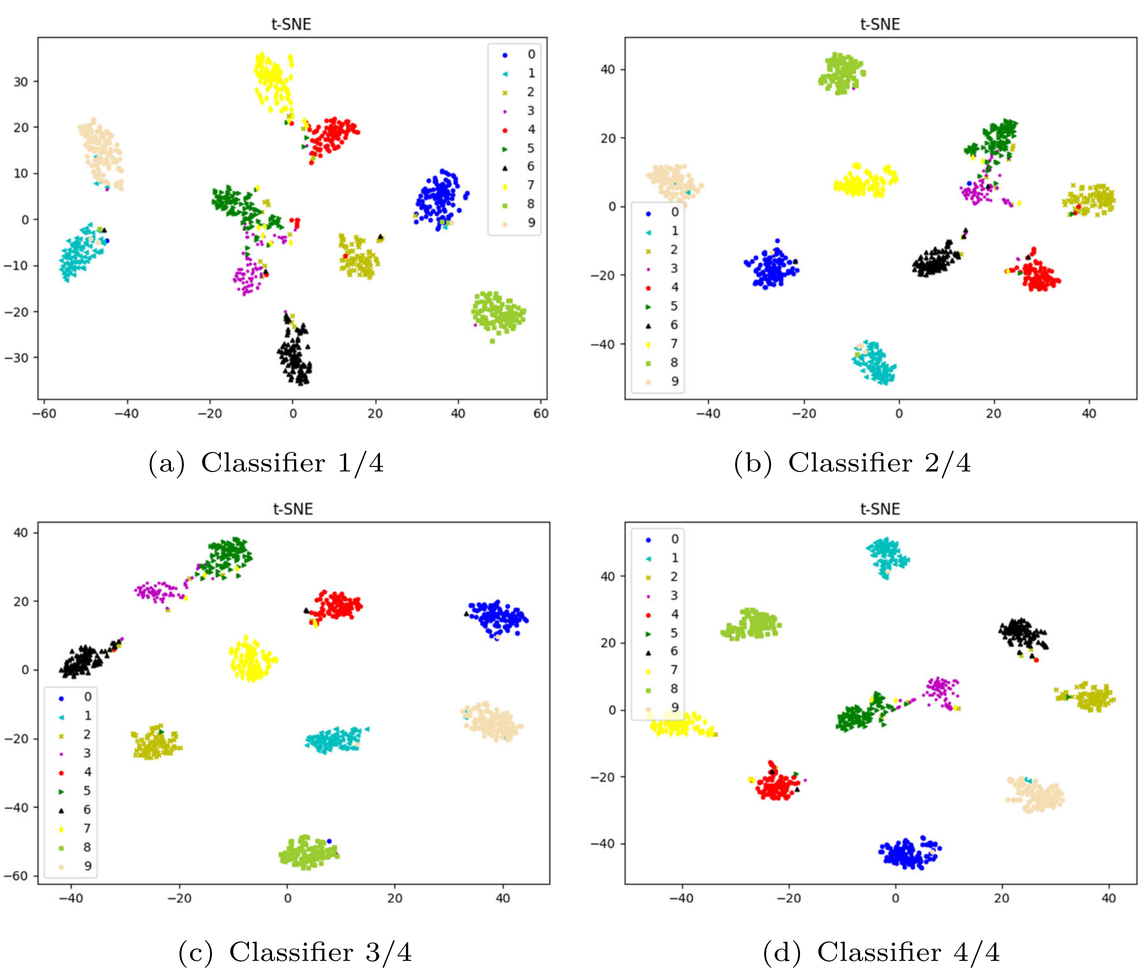


Fig. 3 Feature dimensionality reduction visualization. **a** - **d** represent the feature output of exit 1 - exit 4 in the student network

Table 7 (a) Example 'lizard' on CIFAR100

Category	Kangaroo	Rocket	**lizard**	Snail	Dolphin	Shark	Rabbit	Baby	Ray
Classifier1/4	0.529	0.188	0.083	0.044	0.029	0.011	0.009	0.007	0.006
Category	rocket	**lizard**	ray	shark	rabbit	snail	kangaroo	turtle	crocodile
Classifier2/4	0.512	0.098	0.048	0.044	0.034	0.030	0.029	0.018	0.016
Category	shrew	mouse	shark	**lizard**	snail	rocket	ray	trout	chair
Classifier3/4	0.510	0.199	0.069	0.031	0.030	0.025	0.023	0.020	0.015
Category	shark	**lizard**	seal	trout	shrew	snail	rocket	whale	turtle
Classifier4/4	0.406	0.264	0.050	0.044	0.316	0.289	0.017	0.011	0.008
Category	**lizard**	rocket	shark	snail	shrew	mouse	kangaroo	ray	trout
Ensemble	0.211	0.180	0.136	0.073	0.042	0.038	0.036	0.034	0.024

supervised learning: (i) the logits output from the teacher network to the student network (Logits(T)), (ii) the logits transferred from the deepest layer of the student network to the shallow classifier (Logits(S)), and (iii) the features from the shallow layer of the student network matched to deep features (Features(S)). In addition, the average ensemble strategy is used. To evaluate the effectiveness of each type of supervised learning, we chose ResNet152 and ResNet18 as the teacher and student networks, respectively, and conducted ablation experiments on CIFAR100. The experimental results are summarized in Table 6. It is found that each strategy has different degrees of improvement for classification accuracy, and has a large improvement over the traditional knowledge distillation method using only Logits(T). Notably, our method even outperforms the teacher network.

5.2 Multi-exit network features for dimensionality reduction visualization

In this paper, we construct a student network with multiple exits based on self-distillation. Similar to multi-teacher distillation, in this approach, the deepest output of the backbone network is considered as the second teacher, different networks learn different view features, and the student network matches the feature representation knowledge of multiple models through knowledge distillation and self-distillation. Finally we use ensembles to vote on the classification results of multiple sub-models in the student network, and obtain a strong classifier. We visualize the high-dimensional features input into the fully connected layer in the three branch networks and the backbone network by dimensionality reduction. As shown in Fig. 3, the classification effect of each exit of the student model is remarkable, and the classification accuracy of the shallow layer even approaches that of the deep layer.

5.3 Ensemble validity and sensitivity analysis

In this section, we discuss the validity of multiple exit ensembles and how the number of exits affects the results. In the student network, we construct multiple output channels, each of which is a separate classification network. Although our sub-models are uniformly optimized by the same teacher and share some weights, the structure of each sub-model is different. Firstly, the depth of the backbone network is different, secondly, the bottleneck layer is different. Due to the different depths of sub-networks and the different number of nonlinear functions introduced, each sub-model has different ability to fit data, so the classification

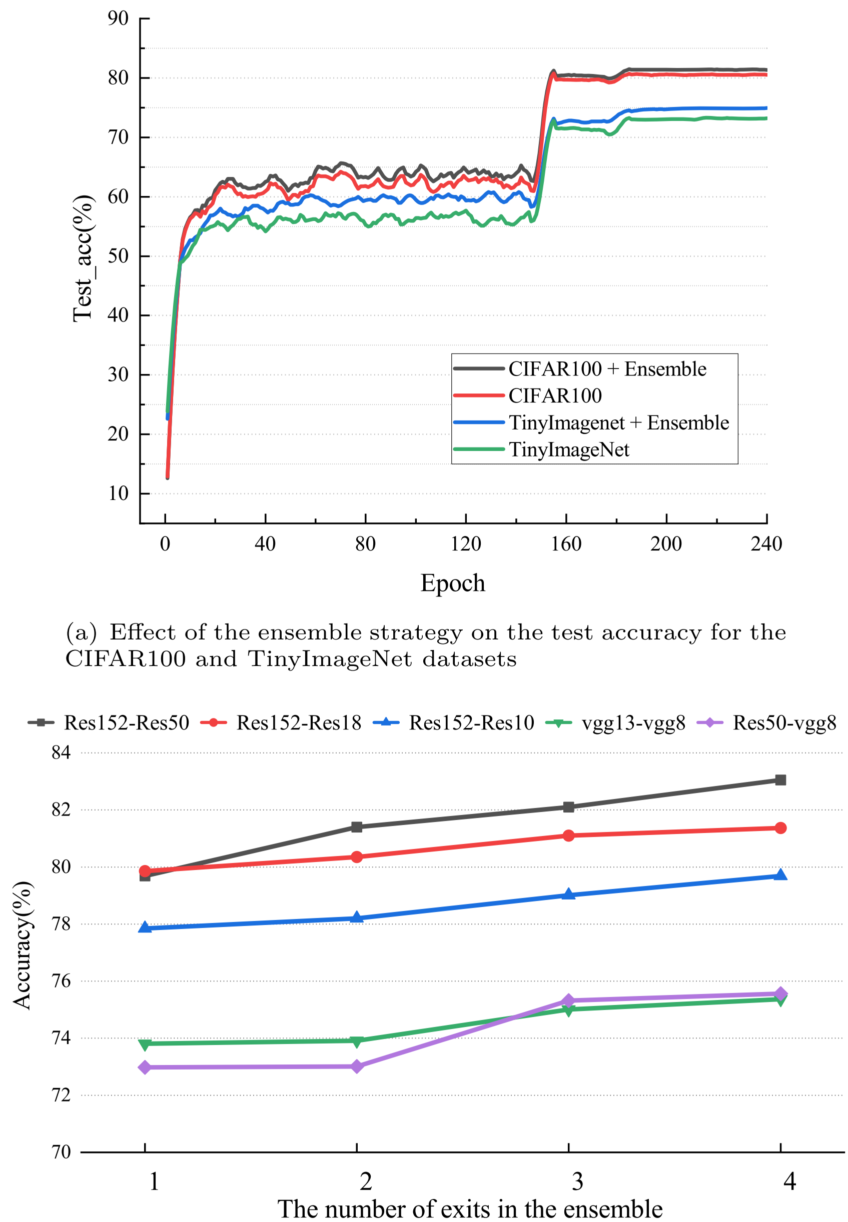
Table 8 (b) Example 'castle' on CIFAR100

Category	**Castle**	Skyscraper	Rocket	House	Mountain	Mushroom	Lamp	Pear	Forest
Classifier1/4	0.700	0.210	0.034	0.013	0.007	0.004	0.003	0.003	0.002
Category	**castle**	skyscraper	pear	pine_tree	rocket	mountain	house	orchid	road
Classifier2/4	0.424	0.110	0.069	0.059	0.051	0.040	0.024	0.017	0.016
Category	pear	pine_tree	forest	**castle**	skyscraper	house	orchid	road	lamp
Classifier3/4	0.530	0.138	0.134	0.074	0.031	0.021	0.020	0.010	0.005
Category	skyscraper	pear	**castle**	mountain	pine_tree	orchid	forest	rocket	house
Classifier4/4	0.338	0.124	0.104	0.093	0.074	0.042	0.030	0.024	0.023
Category	**castle**	skyscraper	pear	pine_tree	house	forest	rocket	mountain	orchid
Ensemble	0.0376	0.214	0.108	0.058	0.034	0.033	0.029	0.027	0.018

Table 9 (c) Example 'dinosaur' on CIFAR100

Category	House	Tank	**dinosaur**	Elephant	Bridge	Tractor	Train	Castle	Pickup_truck
Classifier1/4	0.208	0.110	0.080	0.060	0.059	0.045	0.042	0.029	0.028
Category	tank	tractor	lawn_mower	cattle	bus	**dinosaur**	palm_tree	motorcycle	train
Classifier2/4	0.299	0.121	0.062	0.051	0.050	0.042	0.038	0.031	0.027
Category	**dinosaur**	tractor	tank	lobster	castle	palm_tree	willow_tree	house	bus
Classifier3/4	0.512	0.164	0.079	0.028	0.021	0.017	0.016	0.013	0.013
Category	**dinosaur**	tank	castle	house	tractor	bridge	willow_tree	palm_tree	orchid
Classifier4/4	0.716	0.067	0.052	0.035	0.027	0.012	0.011	0.009	0.007
Category	**dinosaur**	tank	tractor	house	castle	palm_tree	rocket	pickup_truck	willow_tree
Ensemble	0.295	0.181	0.111	0.058	0.039	0.030	0.021	0.019	0.018

Fig. 4 Verification of the effectiveness of ensemble strategies and a sensitivity analysis



(a) Effect of the ensemble strategy on the test accuracy for the CIFAR100 and TinyImageNet datasets

(b) Effect of the number of ensemble exits on the experiment results are also different. Several sets of experiments were added to illustrate the differences in classification performance of each submodel and the effectiveness of multi-exit integration. We chose ResNet152-ResNet18 as the teacher-student network on CIFAR100, and we counted the number of error samples for each sub-model as well as after integration. We found that the wrong sample was classified differently for each exit. According to the statistics, the samples with incorrect predictions at the deepest exit were correctly predicted at the first, second and third exits approximately 22%, 20% and 14% of the total errors at the deepest exit. This suggests that although the deepest classifier predicts incorrectly, it can be predicted correctly at the shallow level. When integrated, these samples with incorrect predictions at the deep level may also be predicted correctly. For example, The experimental results are shown in Table 7, 8 and 9, we give three cases of integration validity: (a) All four classifiers predicted incorrectly at first but then predicted correctly after integration. (b) The first two classifiers predicted correctly, the last two classifiers predicted incorrectly, then predicted correctly after integration. (c) The first two classifiers predicted incorrectly, the last two classifiers predicted correctly, then predicted correctly after integration. Red denotes real label.

In addition, we also compare the use of ensemble with the non-application of ensemble, the experimental results are shown in Fig. 4. Our experiments were conducted based on the CIFAR100 and TinyImageNet datasets, and different teacher-student architectures were used to verify the effectiveness of the ensemble strategy. Furthermore, we explored the effect of the number of ensemble exits on the accuracy of classification, and the results showed that within a certain range, a higher number of ensemble exits can improve the final performance of the network.

6 Conclusion

In this paper, we propose a teacher-student collaborative distillation approach. Unlike traditional transfer learning, our approach fuses knowledge distillation and self-distillation, allowing the student model to learn new knowledge from the teacher network and from itself. During test stage, we vote on the different classification results of multiple sub-models in the student network. Through extensive experiments, the effectiveness of our proposed method and each component is verified, and this approach can be used to guide both knowledge distillation and multi-exit networks. Since the multiple exits in the student network can be constructed in any distillation network, we only consider a traditional distillation structure to ensure that the method is representative and can be further tested in other distillation cases in the future. Finally, the balance between model complexity and classification accuracy should be assessed in future research.

Acknowledgements This work was supported in part by the China Chongqing Science and Technology Commission under Grant cstc2020jscx-msxmX0086, cstc2019jscx-zdztzx0043, cstc2019jcyj-msxmX0442. China Chongqing Banan District Science and Technology Commission project under Grant 2020QC413, and China Chongqing Municipal Education Commission under Grant KJQN202001137.

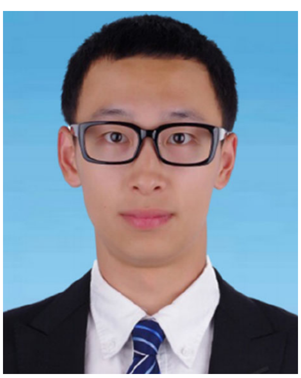
References

1. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
2. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
3. Tan M, Le Q (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp 6105–6114
4. Cheng Y, Wang D, Zhou P, Zhang T (2018) Model compression and acceleration for deep neural networks: The principles, progress, and challenges. IEEE Signal Proc Mag 35(1):126–136
5. Bashir D, Montañez GD, Sehra S, Segura PS, Lauw J (2020) An information-theoretic perspective on overfitting and underfitting. In: Australasian Joint Conference on Artificial Intelligence. Springer, pp 347–358
6. Yim J, Joo D, Bae J, Kim J (2017) A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4133–4141
7. Kim Y, Rush AM (2016) Sequence-level knowledge distillation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp 1317–1327
8. Gou J, Yu B, Maybank SJ, Tao D (2021) Knowledge distillation: A survey. Int J Comput Vis 129(6):1789–1819
9. Bucilua C, Caruana R, Niculescu-Mizil A (2006) Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 535–541
10. Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network
11. Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y (2015) Fitnets: Hints for thin deep nets. ICLR

12. Komodakis N, Zagoruyko S (2017) Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations
13. Tung F, Mori G (2019) Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 1365–1374
14. Ahn S, Hu SX, Damianou A, Lawrence ND, Dai Z (2019) Variational information distillation for knowledge transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9163–9171
15. Park W, Kim D, Lu Y, Cho M (2019) Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3967–3976
16. Tian Y, Krishnan D, Isola P (2019) Contrastive representation distillation. In: International Conference on Learning Representations
17. Heo B, Lee M, Yun S, Choi JY (2019) Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 3779–3787
18. Dong N, Zhang Y, Ding M, Xu S, Bai Y (2021) One-stage object detection knowledge distillation via adversarial learning. Appl Intell:1–17
19. Oyedotun OK, Shabayek AER, Aouada D, Ottersten B (2021) Deep network compression with teacher latent subspace learning and lasso. Appl Intell 51(2):834–853
20. Furlanello T, Lipton Z, Tschannen M, Itti L, Anandkumar A (2018) Born again neural networks. In: International Conference on Machine Learning. PMLR, pp 1607–1616
21. Yuan L, Tay FrancisEH, Li G, Wang T, Feng J (2020) Revisiting knowledge distillation via label smoothing regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3903–3911
22. Mobahi H, Farajtabar M, Bartlett P (2020) Self-distillation amplifies regularization in hilbert space. Neural Information Processing Systems (NeurIPS). https://papers.nips.cc/paper/2020/file/2288f691b58edecadcc9a8691762b4fd-Paper.pdf
23. Phuong M, Lampert CH (2019) Distillation-based training for multi-exit architectures. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 1355–1364
24. Zhang L, Song J, Gao A, Chen J, Bao C, Ma K (2019) 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, pp 3713–3722
25. Ji M, Shin S, Hwang S, Park G, Moon I-C (2021) Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10664–10673
26. Yun S, Park J, Lee K, Shin J (2020) Regularizing class-wise predictions via self-knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 13876–13885
27. Dietterich TG (2000) Ensemble methods in machine learning. In: International workshop on multiple classifier systems. Springer, pp 1–15
28. Zhou Z-H, Wu J, Tang W (2002) Ensembling neural networks: many could be better than all. Artif Intell 137(1-2):239–263
29. Rokach L (2010) Ensemble-based classifiers. Artif Intell Rev 33(1):1–39
30. Fukuda T, Suzuki M, Kurata G, Thomas S, Cui J, Ramabhadran B (2017) Efficient knowledge distillation from an ensemble of teachers. In: Interspeech, pp 3697–3701
31. Lan X, Zhu X, Gong S (2018) Knowledge distillation by on-the-fly native ensemble. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp 7528–7538
32. Liu Y, Zhang W, Wang J (2020) Adaptive multi-teacher multi-level knowledge distillation. Neurocomputing 415:106–113
33. Krizhevsky A (2009) Learning multiple layers of features from tiny images. Master's thesis, University of Tront
34. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 248–255
35. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Bengio Y, LeCun Y (eds) 3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings, San Diego
36. Zagoruyko S, Komodakis N (2016) Wide residual networks. In: British Machine Vision Conference 2016. British Machine Vision Association
37. Zhang X, Zhou X, Lin M, Sun J (2018) Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6848–6856
38. Ma N, Zhang X, Zheng H-T, Sun J (2018) Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV), pp 116–131
39. Peng B, Jin X, Liu J, Li D, Wu Y, Liu Y, Zhou S, Zhang Z (2019) Correlation congruence for knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 5007–5016
40. Passalis N, Tefas A (2018) Learning deep representations with probabilistic knowledge transfer. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 268–284
41. Kim J, Park S, Kwak N (2018) Paraphrasing complex network: network compression via factor transfer. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp 2765–2774
42. Xu G, Liu Z, Li X, Loy CC (2020) Knowledge distillation meets self-supervision. In: European Conference on Computer Vision. Springer, pp 588–604
43. Yang C, An Z, Cai L, Xu Y (2021) Hierarchical self-supervised augmented knowledge distillation. In: Zhou Z-H (ed) Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21, pp 1217–1223
44. Ji M, Heo B, Park S (2021) Show, attend and distill: Knowledge distillation via attention-based feature matching. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 35, pp 7945–7952
45. Lee C-Y, Xie S, Gallagher P, Zhang Z, Tu Z (2015) Deeply-supervised nets. In: Artificial intelligence and statistics. PMLR, pp 562–570

Chuanyun Xu is currently an Associate Professor in School of Artificial Intelligence,Chongqing University of Technology and College of Computer and Information Science, Chongqing Normal University. He received the M.S. degree in Software Engineering from Chongqing University, in 2006, the Ph.D. degree in computer science from the Chongqing University in 2014. He worked as a project scientist at University of California, Riverside between 2016 and 2018. His research interests include artificial intelligence, machine learning and image processing.

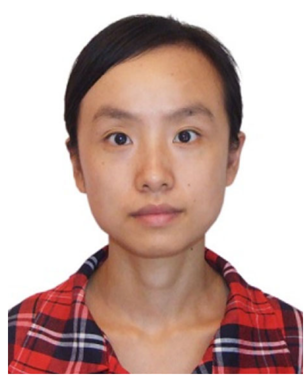
Wenjian Gao received the B.S. degree in electronic information engineering from Chongqing University of Technology in 2019. He is persuing M.S. degree in signal and information processing from Chongqing University of Technology. His research is focused on knowledge distillation and image processing.



Tian Li enrolled postgraduate of computer science at RWTH Aachen. Passionate about artificial intelligence and machine learning, with strong technical, business and interpersonal skills for working in a team and successfully completing several projects. Currently focus on theory and practice of data science In process mining and data mining.

Nanlan Bai received the B.S. degree in computer science and technology from Yangtze Normal University in 2020. She is persuing her M.S. degree in computer technology from Chongqing University of Technology. Her research is focused on knowledge distillation and Image classification.

Gang Li is currently a Professor in School of Artificial Intelligence, Chongqing University of Technology. He received the Ph.D. degree in computer science from the Chongqing University. He is a member of China Computer Federation (CCF) and Chinese Association for Artificial Intelligence (CAAI). His research interests include pattern recognition, image processing, and computer vision.



Yang Zhang is currently an Associate Professor in College of Computer and Information Science, Chongqing Normal University. She received the Ph.D. degree in computer science from the Chongqing University. Her research interests include software measurement, services computing and trusted computing.