electronics

MDPI

Article

Self-Knowledge Distillation via Progressive Associative Learning

Haoran Zhao [1], Yanxian Bi [2], Shuwen Tian [1], Jian Wang [3], Peiying Zhang [4], Zhaopeng Deng [1,*] and Kai Liu [5,6,*]

[1] School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266520, China; zhaohaoran@qut.edu.cn (H.Z.); tian002680@163.com (S.T.)
[2] China Academy of Electronic and Information Technology, CETC Academy of Electronics and Information Technology Group Co., Ltd., Beijing 100041, China; biyanxian@cetc.com.cn
[3] College of Science, China University of Petroleum (East China), Qingdao 266580, China; wangjiannl@upc.edu.cn
[4] Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China; zhangpeiying@upc.edu.cn
[5] State Key Laboratory of Space Network and Communications, Tsinghua University, Beijing 100084, China
[6] Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
* Correspondence: dengzhaopeng@qut.edu.cn (Z.D.); liukaiv@tsinghua.edu.cn (K.L.)

Abstract: As a specific form of knowledge distillation (KD), self-knowledge distillation enables a student network to progressively distill its own knowledge without relying on a pretrained, complex teacher network; however, recent studies of self-KD have discovered that additional dark knowledge captured by auxiliary architecture or data augmentation could create better soft targets for enhancing the network but at the cost of significantly more computations and/or parameters. Moreover, most existing self-KD methods extract the soft label as a supervisory signal from individual input samples, which overlooks the knowledge of relationships among categories. Inspired by human associative learning, we propose a simple yet effective self-KD method named associative learning for self-distillation (ALSD), which progressively distills richer knowledge regarding the relationships between categories across independent samples. Specifically, in the process of distillation, the propagation of knowledge is weighted based on the intersample relationship between associated samples generated in different minibatches, which are progressively estimated with the current network. In this way, our ALSD framework achieves knowledge ensembling progressively across multiple samples using a single network, resulting in minimal computational and memory overhead compared to existing ensembling methods. Extensive experiments demonstrate that our ALSD method consistently boosts the classification performance of various architectures on multiple datasets. Notably, ALSD pushes forward the self-KD performance to 80.10% on CIFAR-100, which exceeds the standard backpropagation by 4.81%. Furthermore, we observe that the proposed method shows comparable performance with the state-of-the-art knowledge distillation methods without the pretrained teacher network.

Keywords: knowledge distillation; neural network compression; edge computing; image classification; self distillation

Citation: Zhao, H.; Bi, Y.; Tian, S.; Wang, J.; Zhang, P.; Deng, Z.; Liu, K. Self-Knowledge Distillation via Progressive Associative Learning. *Electronics* **2024**, *13*, 2062. https://doi.org/10.3390/electronics13112062

Academic Editor: Claus Pahl

Received: 26 April 2024
Revised: 23 May 2024
Accepted: 23 May 2024
Published: 25 May 2024

1. Introduction

Deep neural networks (DNNs) have made unprecedented advances in a wide range of machine learning tasks such as computer vision [1,2], natural language processing [3,4] and speech recognition [5,6]. However, with the high performance of DNNs, deeper and wider models are proposed at the cost of larger model size and longer inference time. Thus, it is naturally unrealistic to deploy such complex DNNs to some resource-constrained devices such as mobile phones and embedded devices. To solve such issues, many research studies [7–10] propose to compress such complex models into compact ones. Knowledge distillation is one of the most popular and effective methods for model

compression [11]. It is a prominent technology, which trains a portable student network under the supervision of a complicated teacher network by directly mimicking the latter's outputs [12–14].

In recent years, knowledge distillation has emerged as a powerful technique for improving the performance of student networks in a wide range of tasks and domains, such as object detection [15,16], semantic segmentation [17–19], face recognition [20,21] and action recognition [22,23]. Its ability to transfer knowledge from a well-trained teacher network to a smaller student network has made it a popular choice for addressing challenges such as model compression, improving generalization and adapting models to resource-constrained environments.

While conventional knowledge distillation methods have proven effective in improving the accuracy and generalization ability of student networks, they do have some drawbacks that are worth considering. For example, most of these methods have an expensive training process due to the training of a cumbersome teacher model with large parameters. In addition, most of the existing distillation frameworks require substantial efforts and experiments to find the best architecture for the teacher network and student network, which takes a relatively long time.

Self-KD methods [24,25] have been proposed to overcome the above issues by directly optimizing the student network itself. Such methods have shown that a network can distill its own knowledge to teach itself without a pretrained teacher. However, since the parameters of the teacher and student networks are the same in self-KD, it becomes challenging to learn as much useful knowledge as can be obtained from a well-performing teacher network in conventional knowledge distillation. Thus, existing Self-KD techniques frequently incorporate auxiliary architectures or data augmentation methods to capture supplementary knowledge, thereby enhancing the performance of the network. A common characteristic of existing Self-KD methods is that the extracted supervisory signal, such as a soft label, is generated from an individual input sample, which overlooks the knowledge of relationships among categories.

In contrast, we propose a novel self-knowledge distillation method named associative learning for self-distillation (ALSD), which applies human associative learning to transfer the valuable knowledge of relationships among categories. Figure 1 illustrates the process of associative learning, in which new knowledge is acquired based on the relationships observed between various elements of a person's experiences. Since real-world phenomena are interconnected, we can construct a knowledge graph representing these connections from samples. Leveraging the relationships between various elements can significantly enhance human learning. Inspired by this, we introduce associative learning into self-KD, which employs the associational information as dark knowledge during the process of knowledge distillation.

It is widely acknowledged that human associative learning plays a crucial role in many aspects of human cognition and behavior. In our ALSD method, the distillation process is similar to the process of human associative learning. In other words, the distilled model in our framework can effectively learn powerful features by fully exploiting the relationship information between categories. To be specific, the distilled model utilizes the associated samples to comprehensively learn the relationships between categories. It guides itself with its own rich dark knowledge acquired from these associated samples. Following consistent benchmarks, we verify the effectiveness of the ALSD method on several public datasets, including CIFAR-10 [26], CIFAR-100 [26], Tiny-ImageNet-200 [27], CUB200-2011 [28] and Stanford Dogs [29]. In addition, we compare our approach with a variety of knowledge distillation and self-distillation methods. The contributions are summarized as follows:

- We introduce a highly efficient self-KD framework that emulates human associative learning, allowing the distilled network to acquire powerful features through associated samples as its inputs.

- The student network not only learns the characteristics of the original samples but also is compelled to acquire knowledge about interclass relations among all categories in a self-distillation manner.
- Our method delivers promising results compared with the state-of-the-art methods. For example, on the CIFAR-100 dataset, with the same network, it achieves a 2.22% higher accuracy rate than the CSKD method. Notably, our method even outperforms conventional distillation with a pretrained teacher.

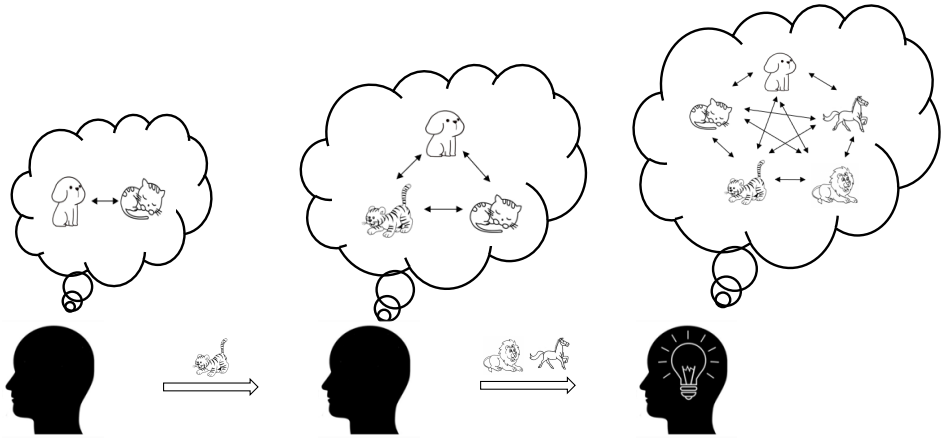


Figure 1. Illustration of human associative learning. Inspired by this, the proposed associative learning for self-distillation (ALSD) method forces the network to learn relationships by knowledge transfer.

The rest of this paper is organized as follows. We briefly introduce the related work in Section 2. Our ALSD method is proposed in Section 3. The results and discussion are conducted in Sections 4 and 5. Finally, our conclusions are offered in Section 6.

2. Related Work

In this section, we first briefly introduce the most related works of knowledge distillation. Then we specifically review recent self-distillation works.

Knowledge distillation is a widely used paradigm for model compression, which transfers knowledge from a complex teacher model to a compact student model. To be specific, the teacher network has high accuracy and huge parameters, while the student network is not as accurate as the teacher network but has fewer parameters. Through knowledge distillation, we hope that the student network can approach or exceed the teacher network as much as possible. In this way, we obtain a compact student network with a similar prediction effect as the teacher network. Ba et al. [30] first proposed a method that uses the teacher's logits before the softmax as the regression target to train the student network, which completes the imitation of the teacher network by forcing the student network to mimic the teacher network's logits. Hinton et al. [12] first proposed to use the soft outputs of the pretrained teacher network as dark knowledge to supervise the training of the student network. They introduced a temperature hyperparameter T and formulated the problem as "knowledge distillation". The student network is forced to learn the soft targets of the teacher network, which are obtained through using a high temperature T on the softmax inputs. In the process of knowledge transfer, soft targets often contain richer information than one-hot targets. Romero et al. [13] extended the knowledge distillation method proposed by Hinton et al. In their method, the student network can be deeper and narrower than the teacher network and improve the performance by learning the outputs of the teacher network and the features of the middle layer. All the above methods are offline distillation methods [31,32], which need a pretrained teacher network.

In contrast to these methods, online knowledge distillation trains the student network under the supervision of a teacher from scratch. For example, Zhang et al. [33] proposed a

mutual learning method, which uses multiple neural networks. Zhao et al. [9] proposed a collaborative training method, which uses both an expert teacher and a from-scratch teacher to supervise the student. To reduce the computational cost, Zhou et al. [34] proposed to employ two different networks which share some low parameters and train separately. Li et al. [24] observed some interesting phenomena through their experiment. First, they found that the performance of the student could also be improved when the weak teacher was used to guide the training process of the student. Second, the performance of the student network could still be improved when the teacher network was worse than the student network.

Self-distillation can be regarded as a special case of knowledge distillation. In self-distillation, the student model uses the same network as the teacher model or the model guides itself through its own knowledge. A significant advantage of the self-distillation framework is that no additional teachers are required. In contrast, traditional distillation first needs to find and train a teacher model with large parameters. Designing a high-quality teacher model requires a lot of experiments. In addition, it takes a long time to train an overparameterized teacher model. These issues can be avoided directly in self-distillation.

Recently, a large number of self-distillation methods have been proposed. Hahn et al. [35] presented a self-knowledge distillation method for NLP, using more information from the soft target probability of the model to train itself. Hou et al. [25] proposed a new method of knowledge distillation for lane detection, called self-attention distillation (SAD), which allows the attention maps in the upper layer of the network to be the learning goal of the lower layer. Similar to the self-attention distillation of the SAD method, Zhang et al. [36] proposed a general self-distillation framework to compress the knowledge of the deeper part of the same network into the shallow part within the network. As a special variant of self-distillation, Yang et al. [37] proposed snapshot distillation by extracting information from earlier epochs of the network (teacher) to supervise later epochs of the network (student), which can effectively prevent the occurrence of the underfitting problem. The self-distillation method is also used for data augmentation [38,39]. The knowledge of augmentation is distilled into the model itself by self-distillation. Xu et al. [40] proposed using different data augmentation methods for each batch, obtaining two batches with the same label. In the training process, the output differences of batches are minimized, which improves the diversity and robustness in the same class. In addition, some researchers pay more attention to combining the self-distillation method with other methods such regularization [41,42] and BiFPN [43] to further enhance the performance of the student network. Yun et al. [44] proposed a regularization method to make the output distribution of two samples with the same label consistent by self-distillation, effectively reducing the differences within the class and overconfidence in false predictions. Recently, Ji et al. [45] innovatively combined BiFPN with self-distillation, which utilizes BiFPN to refine features to construct a self-teaching network and uses an auxiliary self-teaching network to transfer refined knowledge to a classifier network.

Different from previous self-distillation studies, our method uses associated samples as the input of the network, rather than single original samples. We use the original samples to supervise the probability of associated samples. In such a manner, the network can not only learn the characteristics of the original samples but also fully learn the relationship between classes. The efficacy of our methodology is substantiated through a series of experiments across various public datasets.

3. Materials and Methods

In this section, we describe the proposed ALSD method in detail. As illustrated in Figure 2, the network simulates human associative memory to learn the features of samples. To be specific, the network gradually learns interclass relationships through associated samples and uses this knowledge to guide itself by self-distillation. In the following, we first introduce the motivation of our method. Then we formulate the association distillation loss for ALSD and present the training procedure.

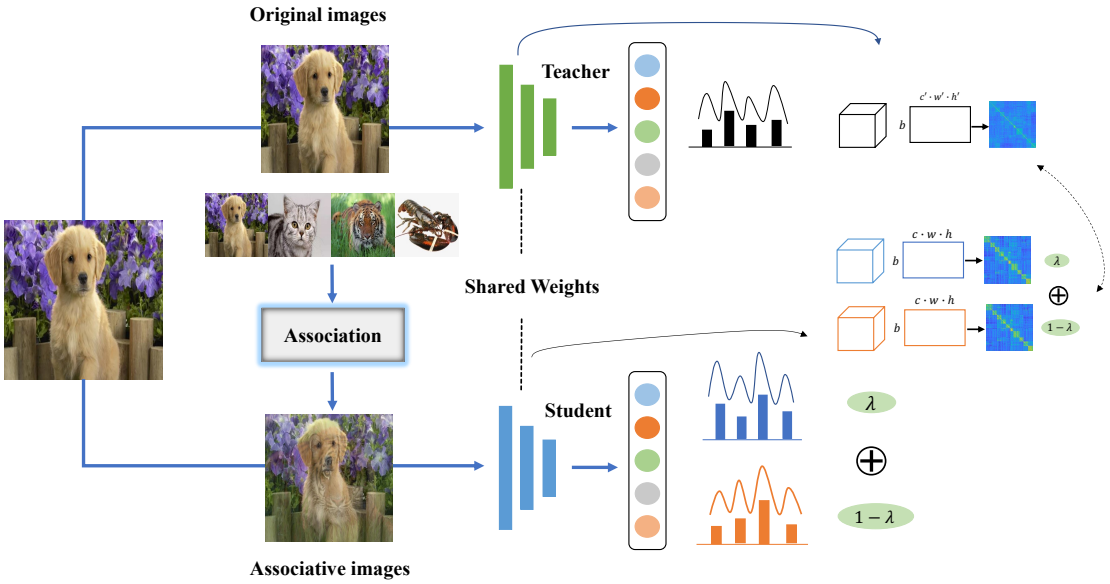


Figure 2. Our ALSD method includes two stages. First, the student network learns the features of classes and interclass relationships through associated samples generated by the original ones. Second, the student network mimics the output distribution of the teacher network. The teacher network shares the weights with the student, i.e., the network utilizes the knowledge to guide itself by self-distillation.

3.1. Motivation

The existing knowledge distillation methods employ diverse strategies to compel the student network to approach the teacher network. One of the most fundamental and effective approaches for the student network is to mimic the teacher network's softmax outputs. It is noteworthy that the teacher network imparts its dark knowledge to the student network, encompassing not only predictions for the correct classes but also predictions for other wrong classes.

In other words, both the correct and wrong soft predictions from the complex teacher model are positive for training the compact student model. This is due to the valuable relationship among categories from the soft prediction information. However, the teacher and student models are the same model in self-KD architecture. Differing from the traditional knowledge distillation frameworks, within the self-KD framework, the teacher network cannot utilize its substantial parameter advantage to distill rich relational information between categories for the student network.

Thus, we propose to improve the self-distillation performance for convolutional neural networks via associative learning. Figure 1 shows the motivation of our method. For example, when humans initially learn about animals, they may only be aware of cats and dogs. At this early stage, humans might only associate cats with dogs because of their similarities in appearance, while lacking knowledge about other animal species. Nevertheless, as humans acquire more knowledge of various animals, they may observe that the patterns of tigers and cats exhibit greater similarity than those of cats and dogs. Whenever people encounter new animals, they tend to link these new creatures with the animals they are already familiar with in order to discern relationships and enhance their memory of the characteristics of these animals.

Inspired by this, we aim to create and distill relationships between categories to improve self-distillation performance by employing associated samples during the training process. The proposed method serves a dual purpose. First, it prevents the loss of class-specific features during the construction of associated samples. Since these samples are generated through associations, the original samples may lose certain characteristics, leading to reduced confidence in the network's output for the original classes. Second,

the network can more effectively convey interclass information acquired from the associated samples to itself through the process of self-distillation, resulting in improved overall performance.

3.2. Associative Learning for Self-Distillation

The purpose of knowledge distillation is to make the distribution of outputs of a teacher network and a student network close enough. Those distributions are obtained through the softmax function. In the general softmax function, index e first enlarges the distance between logits and then normalizes. Its final output is a vector close to one-hot, of which one entity is very large and the others are very small. Output through such a softmax will lead to the loss of the relationship between classes in the process of knowledge transfer, which leads to a student network that cannot fully learn the knowledge from the teacher network. Therefore, we adopted a more general approach [12]. Our task focuses on the fully supervised classification task, which means that $x \in X$ is input and $y \in Y$ is the label. Suppose that a softmax classifier is used to model a posterior predictive distribution with the input x; the predictive distribution is

$$P(y|x;\theta,T) = \frac{exp(f_y(x;\theta)/T)}{\sum_{i=1}^{C} exp(f_i(x;\theta)/T)} \tag{1}$$

where f represents the logits outputs by the classifier parameterized with θ and T is the temperature parameter. This is a concept borrowed from Boltzmann distribution in statistical mechanics. It can be easily proved that the outputs of softmax will converge to a one-hot vector when the temperature T tends to 0, and the outputs of softmax will be softer when the temperature T tends to be infinity. Therefore, we can use a higher T to make the distribution produced by softmax soft enough when training the student model. Furthermore, we let the softmax outputs of the student model approximate the teacher model's. In such a manner, the student model can learn a lot of dark knowledge which cannot be learned from hard targets. The normal temperature T is only in the training phase.

We use Mixup [46] to construct the associated samples and take the cross-entropy function used in Mixup as the cross-entropy function of our associated samples.

$$x_{as} = \alpha * x + (1-\alpha) * x' \tag{2}$$

where x is the original sample, x' is the sample generated in the association process and x_{as} is the associated sample.

$$\begin{aligned} \mathcal{L}_{MCE}(x_{as}, y_a, y_b; \theta) = \alpha * \mathcal{L}_{CE}(x_{as}, y_a; \theta) \\ +(1-\alpha) * \mathcal{L}_{CE}(x_{as}, y_b; \theta) \end{aligned} \tag{3}$$

where y_a and y_b are the corresponding labels of x and x'.

In the training process of associated samples, we utilize the associated samples as the input of the network, so that the network can learn the relationship between classes by associated samples. Moreover, as the associated samples may cause feature loss, we use the probability of the original samples to supervise the probability distribution of the associated samples, which forces network to learn features of the original samples and transfer knowledge to itself. The vanilla distillation loss [12] may not achieve our requirements, as we use associated samples as input. Hence, we propose an association distillation loss function of associated samples, which enables the model to fully learn the features of the original samples. The association distillation loss function is shown as follows:

$$\begin{aligned} \mathcal{L}_{Cls}(x, x^{'}, x_{as}; \theta, T) = \alpha * KL(P(y|x;\widetilde{\theta},T)||P(y|x_{as};\theta,T)) \\ +(1-\alpha) * KL(P(y|x^{'};\widetilde{\theta},T)||P(y|x_{as};\theta,T)) \end{aligned} \tag{4}$$

where α is the image-mixing-scale coefficient generated by the beta distribution, KL denotes the Lullback–Leibler (KL) divergence and $\widetilde{\theta}$ is a fixed copy of the parameters θ. The total training loss is defined as follows:

$$\mathcal{L}_{ALSD} = \lambda * \mathcal{L}_{MCE} + \beta * T^2 * \mathcal{L}_{Cls} \tag{5}$$

where λ is the loss weight of cross entropy of associated samples β is the loss weight of distillation of associated samples. The influence of the weight parameter is discussed in detail in the following ablation experiment.

3.3. Training Procedure

The training process can be roughly divided into two stages. The first stage is where the network learns the relationship between the classes. The student network learns the features of classes and the relationship among classes from the associated samples. It then outputs the associative probability distribution. The second stage is where the network self-learns from the original samples. Since the network has fully learned a lot of knowledge in the first stage, we regard it as the teacher network. In the second stage, the distribution obtained from the teacher network with the original samples is applied to supervise the associative probability distribution. As a result, the performance of the network is enhanced by self-distillation.

In the first stage, we minimize associative cross entropy $\mathcal{L}_{MCE}(x_{as}, y_a, y_b; \theta)$ to learn the relationship of classes. In the second stage, Equation (2) is used to optimize the self-distillation process. The total process is realized through optimizing the total loss function as shown in Equation (5). We introduce the detailed procedure in Algorithm 1.

Algorithm 1 Associative Learning for Self-Distillation

Input: image data and label (x, y_a).
Output: parameters θ of student model.
Initialize: θ and training hyper-parameters.
Repeat:
Stage 1: Learning relationship among classes.
1: Sample a batch (x, y_a) from the training dataset.
2: Get a batch (x_{as}, y_a, y_b) by Equation (2).
3: Compute $\mathcal{L}_{MCE}(x_{as}, y_a, y_b; \theta)$ by Equation (3)
Stage 2: Self-distillation.
1: Compute $\mathcal{L}_{Cls}(x, x^{'}, x_{as}; \theta, T)$ by Equation (4)
2: Compute $\mathcal{L}_{ALSD}$ by Equation (5)
Until: θ **converges.**

4. Results

In order to evaluate the performance of our ALSD method, we implement experiments on conventional and fine-grained classification tasks. The classification task is implemented on the CIFAR-10, CIFAR-100 and Tiny-ImageNet-200 datasets, and the fine-grained classification task is implemented on the CUB200-2011 and Stanford Dogs datasets. The fine-grained classification task is also called a subclass classification task, which is a research subject that has been highlightedin the field of computer vision and pattern recognition in recent years. Its purpose is to classify coarse-grained large categories into more detailed subcategories. However, due to subtle differences between subclasses and great interval differences, the fine-grained classification task is more difficult than the conventional classification task.

4.1. Implementation Details

We chose the most commonly used ResNet [1] network for our experiments on multiple datasets. We use stochastic gradient descent (SGD) [47] with momentum 0.9, weight

decay 0.0001. We set initial learning rate as 0.1, which is divided by 10 on epochs 100 and 150, respectively. We set the batch size as 128 and total epochs as 200 for conventional classification tasks and 32 and 200 for fine-grained classification tasks. In our method, the temperature parameter T is set to 4, and the loss weights λ and β are set to 0.1 and 1, respectively. Our experiments are repeated three times in the case of random seeds, and the average value is taken as the final result. All experiments are implemented on GPU using PyTorch.

4.2. Cifar-10 and CIFAR-100

The CIFAR-10 dataset consists of 60,000 32×32 color images in 10 classes, and each class has 6000 images. There are 50,000 training images and 10,000 test images. The dataset is divided into five training batches and one test batch, and each batch has 10,000 images. The test set contains exactly 1000 randomly selected images from each category. The CIFAR-100 dataset is an extension of the CIFAR-10 dataset. It has 20 categories and a total of 100 subclasses. Each subcategory contains 600 images (500 training images and 100 test images), and each image has a small label and one big label. Our experiments use ResNet18 as the basic model. We set comparative experiments with different methods on two datasets, CIFAR-10 and CIFAR-100. The results are shown in Table 1. Note that we first train the ResNet18 network normally on the CIFAR-10 and CIFAR-100 datasets and obtain a pretrained ResNet18 network with an accuracy of 94.86% and 75.30%, respectively. Networks are initialized randomly in the training procedure. From the experimental results in Table 1, we can find our ALSD method improves the generalization ability of the network and shows comparable performance to the existing methods on both the CIFAR-10 and CIFAR-100 datasets.

Table 1. Classification accuracy (%) on CIFAR-10 and CIFAR-100. The best results for each experiment is shown in bold.

Method	CIFAR-10 Acc (%)	CIFAR-100 Acc (%)
Baseline	94.86%	75.30%
KD	95.66%	76.68%
AT	94.89%	75.84%
RKD	95.13%	76.02%
ALSD	**96.04%**	**80.10%**

Specifically, our method has a certain improvement compared with conventional distillation methods such as KD, AT and RKD. We use the vanilla distillation method to train the ResNet network, and it achieves a 95.66% accurate rate. The accuracy of the ResNet18 network trained using our method reaches 96.04% with a 0.38% improvement over the original distillation method. Note that in the KD [12], AT [48] and RKD [49] methods, the teacher networks use the pretrained ResNet18. Furthermore, our method does not use the pretrained ResNet18 as the teacher network. In [24], it is pointed out that the performance of the teacher does not have that great an effect on the student, and experiments have shown that self-training can also achieve similar effects to conventional KD. Although there is no extra teacher for our method, its performance is not worse than the offline distillation method.

On the CIFAR-100 dataset, we set up the same experiment as on the CIFAR-10 dataset. Surprisingly, on the CIFAR-100 dataset, the network trained by our ALSD method is significantly improved compared to the independent ResNet18 network. The ResNet18 trained by our ALSD method achieves an accuracy of 80.10%, which is 4.8% higher than the independent ResNet18. As depicted in Figure 3a, the students obtained by our ALSD method are significantly improved compared to the others. The baseline represents the ResNet18 network trained individually. We can observe that our ALSD method has a

significant improvement in final accuracy and is better than other existing distillation methods. In the following, we discuss the visualization results depicted in Figure 4.

Moreover, we intercept the information of the internal hidden layer and reduce the dimensionality to visualize it, which intuitively shows the difference between our method and other methods. Figure 4 shows the visualization results of different methods using t-SNE [50], which can map high-dimensional data to a low-dimensional space (typically 2D or 3D) to observe the structure and relationships within the data. It can ensure that similar data points in the high-dimensional space remain close in the low-dimensional space, while dissimilar data points move farther apart. From the visualization results, it is evident that our method enhances the clustering of data points belonging to the same category while also increasing the separation between data points from different categories. The result of the visualization highlights the superiority of our method.

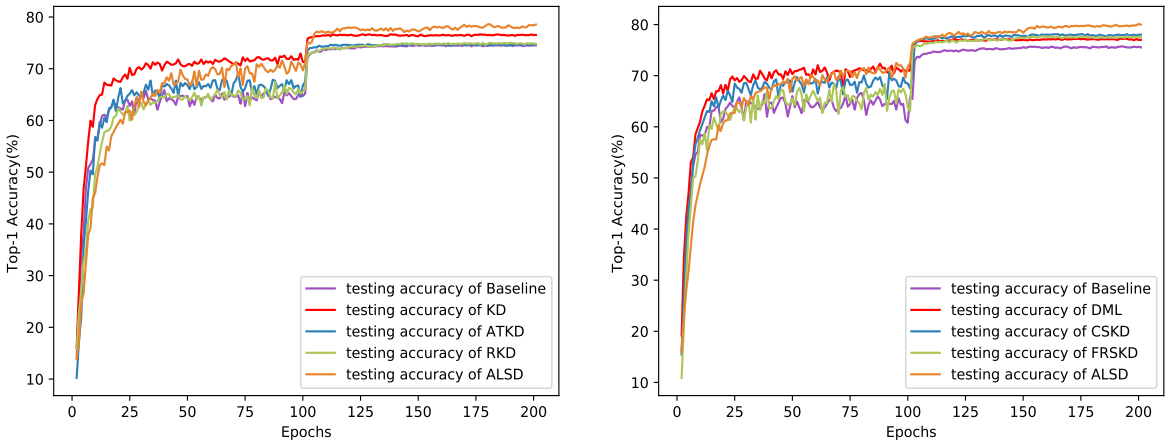
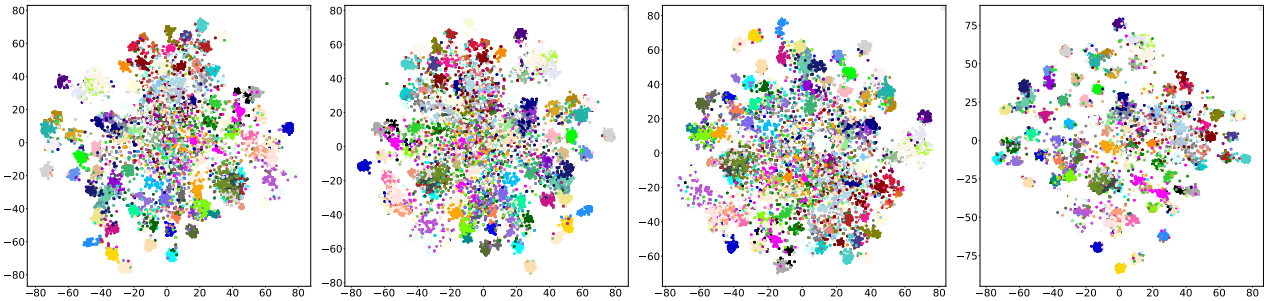


Figure 3. (**a**) Test accuracy of different knowledge distillation methods on the CIFAR-100 dataset. KD, AT and RKD methods all use pretrained ResNet18 as the teacher network. (**b**) The test accuracy of different self-distillation methods on the CIFAR-100 dataset.

In addition, on the CIFAR-100 dataset, we also measure the top-one accuracy rates of our method, ALSD, by comparing with the recent DDGSD, BYOT, DML, CS-KD and FRSKD methods on different classification tasks. Note that although DML is defined as an online distillation method, it is here regarded as a self-distillation method because we use the same network as its teacher and student networks. We can see from Table 2 that our ALSD outperforms others consistently. The CS-KD method is one of the latest self-distillation methods and can be used as a regularization method. This method needs to preprocess the samples according to their labels before sending samples to the network.

In contrast, our method can save the tedious data processing process and replace it with simpler associative processing. Using the CS-KD method, the accuracy of the ResNet18 network is 78.01%, while the accuracy of the ResNet18 network trained by our ALSD method reaches 80.10%, which is 2.09% higher than the latest CS-KD method. We plot the test accuracy curves of various methods in Figure 3b on the CIFAR-100 dataset. As shown in Figure 3b, the students obtained by our ALSD method are significantly improved compared to other methods. From Figure 3b, we can also clearly find that our method quickly surpasses other methods in test accuracy after the last adjustment of the learning rate. At the final epoch, it has a significant improvement in accuracy compared to other methods.



(**a**) KD. (**b**) AT. (**c**) RKD. (**d**) ALSD.

Figure 4. The t-SNE of students trained by different methods on the CIFAR-100 dataset. A color represents a class in CIFAR-100. The visualization results of our method are obviously better than the other three methods.

4.3. Tiny-Imagenet-200

Tiny-imagenet-200 is an image classification dataset provided by Stanford University. The Tiny-imagenet-200 dataset is a popular subset of the ImageNet dataset. It contains 200 categories, and each category contains 500 training images, 50 verification images and 50 test images. On the Tiny-imagenet-200 dataset, we use ResNet18 as the student network to conduct comparative experiments on self-distillation methods.

Table 2 shows the experimental results comparing our method with different self-distillation methods. As shown in Table 2, we verify the effectiveness of our ALSD method on the Tiny-Imagenet-200 dataset. On the Tiny-imagenet-200 dataset, the accuracy of ResNet18 for independent training is 56.63%, and our ALSD method reaches 59.70%, which is an increase of 2.23% compared to the baseline. Compared with the CS-KD method, our result is 1.32% higher than the CS-KD method. FRSDK is the latest proposed method, and its accuracy is 0.3% lower than our method. The above experimental results show that our proposed ALSD method is better than the existing self-distillation methods.

Table 2. Classification accuracy (%) of various methods on CIFAR-100 and Tiny-Imagenet-200 datasets. The best result for each experiment is shown in bold.

Dataset	Baseline	DDGSD	BYOT	DML	CS-KD	FRSKD	ALSD
CIFAR-100	75.30%	76.15%	76.19%	77.51%	78.01%	77.76%	**80.10%**
Tiny-Imagenet-200	56.63%	58.52%	55.98%	58.65%	58.38%	59.60%	**59.70%**

Table 3 shows the comparison between our method and conventional distillation methods on the Tiny-imagenet-200 dataset, and the effect is improved compared with conventional distillation methods. Note that in Table 3, all distillation methods except our method use the pretrained ResNet18 teacher network. In this group of experiments, we process the image resolution to 32×32 and then perform other data augmentation operations. The other experimental settings are the same as those in IV-A.

Table 3. Classification accuracy (%) on Tiny-ImageNet-200 dataset. ACC is calculated as the median of 3 runs of random seeds. The best result for each experiment is shown in bold.

Method	Acc (%)
Baseline	56.63%
KD	58.44%
AT	57.49%
RKD	57.45%
ALSD	**59.70%**

4.4. Cub200-2011 and Stanford Dogs

In this section, we verify the effectiveness of our method by performing fine-grained classification tasks on the CUB200-2011 and Stanford Dogs datasets.

The CUB200-2011 dataset has 11788 bird images, including 200 bird subclasses. The training dataset has 5994 images, and the test set has 5794 images. Each image provides image tagging information, the bounding box of the bird in the image, the key-part information about the bird and the attribute information of the bird. The Stanford Dogs dataset contains images of 120 kinds of dogs from around the world. This dataset is built using images and annotations from ImageNet for fine-grained classification tasks. On these two datasets, we use ResNet18 as the student network to implement fine-grained classification tasks on the self-distillation methods.During image data processing, we first process the image resolution to 224×224 and then send it to the network after random rotation and horizontal flip. The other experimental settings are the same as in IV-A.

As can be seen from Table 4, our method on the CUB200-2011 dataset greatly improves the performance of the student network and shows comparable performance with other SOTA methods. The accuracy of the individually trained ResNet18 network reaches 54%, and the accuracy of our method reaches 70.09%. The FRSKD method, with the highest accuracy of the previous self-distillation methods, is 12.72% higher than the baseline. Our method continues to improve by 3.37% on the latest self-distillation method. Similarly on the Stanford Dogs dataset, our method also has better performance than the latest self-distillation method. The latest FRSKD method has an accuracy of 69.15%, while our method has an accuracy of 71.51%. The overall results show that our method, ALSD, outperforms the SOTA self-distillation methods.

Table 4. Classification accuracy (%) of various methods on CUB200-2011 and Stanford Dogs datasets. Baseline means the ResNet18 network trained individually. The best result for each experiment is shown in bold.

Dataset	Baseline	DDGSD	BYOT	DML	CS-KD	FRSKD	ALSD
CUB200-2011	54.00%	58.53%	59.24%	54.15%	66.72%	67.52%	**70.09%**
Stanford Dogs	62.71%	68.47%	65.98%	63.24%	69.15%	70.75%	**71.51%**

5. Discussion

5.1. Improvements on Various Architectures

As shown in Table 5, we verify the effectiveness of our method on several network structures. We not only consider the complex network structure (e.g., DenseNet-121) but also the lightweight network (e.g., MobileNetV2). The experimental results show that our method is obviously improved compared with the common training methods. For example, the precision of DenseNet-121 on CIFAR-100 is improved by 3.69% compared with the vanilla method.

Table 5. Demographic prediction performance comparison by three evaluation metrics.

Dateset	Method	Architecture				
CIFAR-100		Vgg13	Vgg8	MobileNetV2	ShuffleNetV2	DenseNet-121
	Vanilla	72.93%	68.99%	52.41%	65.47%	77.77%
	Our ALSD	75.36%	70.27%	62.83%	72.83%	81.46%
Tiny-Imagenet-200		Vgg13	Vgg8	MobileNetV2	ShuffleNetV2	DenseNet-121
	Vanilla	59.90%	55.68%	49.27%	56.67%	60.78%
	Our ALSD	62.61%	58.80%	54.89%	59.36%	63.36%

5.2. Ablation Experiment

The influence of the hyperparameters. In this part, we investigate the influence of the hyperparameters λ and β on the experimental results. With the ResNet18 network on the CIFAR-100 dataset, we use our self-distillation method to implement the ablation experiments. We constantly adjust the values of λ and β to observe the influence of the ratio of λ to β on the experimental results. From Table 6, we can observe that the cross-entropy loss plays a decisive role in the entire backpropagation process. When λ is set to 0, the network accuracy is only 1.29%. When we set β to 0, it is equivalent to only that of Mixup. We find that appropriately reducing the ratio of λ to β can improve the accuracy. Interestingly, when the ratio is lower than 0.1, the accuracy decreases instead.

Table 6. Classification accuracy (%) on CIFAR-100 dataset. In Equation (4), λ is the loss weight of cross entropy of associated samples. β is the loss weight of distillation of associated samples.

λ	β	Acc (%)
1	0	78.65%
0	1	1.29%
1	1	79.10%
0.1	1	80.09%
0.1	1.5	79.85%
0.1	2	79.54%

Comparison with traditional knowledge distillation. Traditional knowledge distillation methods utilize a pretrained teacher network's refined feature map or its soft label. Thus, we compare our ALSD approach and the traditional knowledge distillation approaches. For our experiments, we pretrain ResNet-34 as the teacher network, and we use ResNet-18 as the student network. For fair comparisons, all methods in our experiments employ the feature distillation as well as the soft label distillation. Our ALSD approach uses ResNet-18 as a classifier network to meet the identical conditions. Table 7 shows that our ALSD method outperforms the traditional knowledge distillation methods with a pretrained teacher network on most datasets.

Compared to the noisy student algorithm, our approach based on a self-distillation framework is simpler and more straightforward. Thus, our approach does not rely on additional unlabeled data; it only uses the original labeled data for training, making it suitable for scenarios with limited data availability. The objective of our approach is to improve model generalization and robustness by learning from the soft labels generated by the model itself. In the majority of cases, self-training performs well on datasets of different sizes and complements pretraining. However, its effectiveness depends on data quality, reliability of labeled data, accuracy of pseudolabels and choice of model and adjustment strategies. For our approach based on a self-distillation framework, the performance improvement of the self-distillation method depends on how efficiently effective information is distilled from the teacher model.

Table 7. Classification accuracy (%) of knowledge distillation. ResNet-18 is used as student network. The best result for each experiment is shown in bold.

Method	CIFAR-100	Tiny-Imagenet-200	CUB200-2011	Stanford Dogs	MIT67
Baseline	75.30%	56.63%	54.00%	62.71%	55.91%
FitNet	76.67%	58.92%	58.97%	67.18%	59.18%
AT	75.89%	59.52%	59.28%	67.65%	59.38%
Overhaul	74.51%	59.42%	59.58%	66.43%	58.88%
FRSKD	77.76%	59.60%	67.52%	70.75%	**61.18%**
our ALSD	**80.10%**	**59.70%**	**70.09%**	**71.51%**	61.11%

6. Conclusions

We present a novel method for self-distillation and show the performance of our framework by comparing with state-of-the-art distillation methods. The proposed method is similar to human associative learning, which uses associated samples as inputs instead of the original samples, aiming to fully improve the efficiency of knowledge transfer in the process of knowledge distillation. In detail, we employ associated samples as the inputs, so that the network can fully learn the relationships among categories. Moreover, we combine self-distillation to constantly guide the learned knowledge so as to better learn the features of the original samples. Our method achieved excellent performance on several public datasets such as the CIFAR-10, CIFAR-100, Tiny-ImageNet-200, CUB200-2011 and Stanford Dogs datasets, which proves the effectiveness of our method. Unlike traditional distillation methods used to compress and accelerate models, our method is more of a training technique to improve the model performance. Although most of the previous studies focus on the knowledge transfer of original samples, we believe that the knowledge transfer method of associated samples is also promising.

Author Contributions: Conceptualization, H.Z. and Z.D.; methodology, H.Z.; software, H.Z., S.T. and J.W.; validation, H.Z., Z.D. and K.L.; formal analysis, Y.B.; investigation, H.Z.; resources, H.Z.; data curation, H.Z.; writing—original draft preparation, H.Z.; writing—review and editing, H.Z.; visualization, H.Z.; supervision, Z.D.; project administration, P.Z.; funding acquisition, P.Z. Furthermore, Z.D. and K.L. contributed to the manuscript equally. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Shandong Province (No. ZR2023QF043, ZR2023LZH017, ZR2022LZH015), Natural Science Foundation of Qingdao (No. 23-2-1-109-zyyd-jch), National Natural Science Foundation of China (No. 62001263, 62173345).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available from the corresponding authors upon request.

Conflicts of Interest: Author Yanxian Bi was employed by the company TCETC Academy of Electronics and Information Technology Group Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
2. Chen, Z.; Li, J.; You, X. Learn to focus on objects for visual detection. *Neurocomputing* **2019**, *348*, 27–39. [CrossRef]
3. Noh, H.; Hongsuck Seo, P.; Han, B. Image Question Answering Using Convolutional Neural Network With Dynamic Parameter Prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 30–38.
4. Gong, X.; Rong, Z.; Wang, J.; Zhang, K.; Yang, S. A hybrid algorithm based on state-adaptive slime mold model and fractional-order ant system for the travelling salesman problem. *Complex Intell. Syst.* **2023**, *in press*. [CrossRef]

5. Zhang, Y.; Pezeshki, M.; Brakel, P.; Zhang, S.; Laurent, C.; Bengio, Y.; Courville, A.C. Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks. In Proceedings of the Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; ISCA: Singapore, 2016; pp. 410–414.
6. Alam, M.; Samad, M.D.; Vidyaratne, L.; Glandon, A.; Iftekharuddin, K.M. Survey on Deep Neural Networks in Speech and Vision Systems. *Neurocomputing* **2020**, *417*, 302–321. [CrossRef]
7. Bian, C.; Feng, W.; Wan, L.; Wang, S. Structural Knowledge Distillation for Efficient Skeleton-Based Action Recognition. *IEEE Trans. Image Process.* **2021**, *30*, 2963–2976. [CrossRef]
8. Zhao, H.; Sun, X.; Dong, J.; Dong, Z.; Li, Q. Knowledge distillation via instance-level sequence learning. *Knowl. Based Syst.* **2021**, *233*, 107519. [CrossRef]
9. Zhao, H.; Sun, X.; Dong, J.; Chen, C.; Dong, Z. Highlight Every Step: Knowledge Distillation via Collaborative Teaching. *IEEE Trans. Cybern.* **2020**, *52*, 1–12. [CrossRef]
10. Ding, F.; Luo, F.; Hu, H.; Yang, Y. Multi-level Knowledge Distillation. *Neurocomputing* **2020**, *415*, 106–113. [CrossRef]
11. Wu, S.; Wang, J.; Sun, H.; Zhang, K.; Pal, N.R. Fractional Approximation of Broad Learning System. *IEEE Trans. Cybern.* **2023**, *in press*. [CrossRef] [PubMed]
12. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *Comput. Sci.* **2015**, *14*, 38–39.
13. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Bengio, Y. FitNets: Hints for Thin Deep Nets. In Proceedings of the 3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA, 7–9 May 2015.
14. Zhao, H.; Sun, X.; Dong, J.; Yu, H.; Wang, G. Multi-instance semantic similarity transferring for knowledge distillation. *Knowl. Based Syst.* **2022**, *256*, 109832. [CrossRef]
15. Liu, T.; Lam, K.M.; Zhao, R.; Qiu, G. Deep Cross-modal Representation Learning and Distillation for Illumination-invariant Pedestrian Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 315–329. [CrossRef]
16. Chen, L.; Jiang, Z.; Tong, L.; Liu, Z.; Zhao, A.; Zhang, Q.; Dong, J.; Zhou, H. Perceptual underwater image enhancement with deep learning and physical priors. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 3078–3092. [CrossRef]
17. Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; Wang, J. Structured Knowledge Distillation for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 2604–2613. [CrossRef]
18. Zhao, H.; Sun, X.; Dong, J.; Yu, H.; Zhou, H. Dual Discriminator Adversarial Distillation for Data-free Model Compression. *arXiv* **2021**, arXiv:2104.05382.
19. Lateef, F.; Ruichek, Y. Survey on semantic segmentation using deep learning techniques. *Neurocomputing* **2019**, *338*, 321–348. [CrossRef]
20. Guo, P.; Du, G.; Wei, L.; Lu, H.; Chen, S.; Gao, C.; Chen, Y.; Li, J.; Luo, D. Multiscale face recognition in cluttered backgrounds based on visual attention. *Neurocomputing* **2022**, *469*, 65–80. [CrossRef]
21. Ge, S.; Zhao, S.; Li, C.; Zhang, Y.; Li, J. Efficient Low-Resolution Face Recognition via Bridge Distillation. *IEEE Trans. Image Process.* **2020**, *29*, 6898–6908. [CrossRef]
22. Xue, G.; Wang, J.; Yuan, B.; Dai, C. DG-ALETSK: A High-Dimensional Fuzzy Approach With Simultaneous Feature Selection and Rule Extraction. *IEEE Trans. Fuzzy Syst.* **2023**, *31*, 3866–3880. [CrossRef]
23. Tang, Y.; Wei, Y.; Yu, X.; Lu, J.; Zhou, J. Graph Interaction Networks for Relation Transfer in Human Activity Videos. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 2872–2886. [CrossRef]
24. Yuan, L.; Tay, F.E.H.; Li, G.; Wang, T.; Feng, J. Revisit Knowledge Distillation: A Teacher-free Framework. *arXiv* **2019**, arXiv:1909.11723.
25. Hou, Y.; Ma, Z.; Liu, C.; Loy, C.C. Learning Lightweight Lane Detection CNNs by Self Attention Distillation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: New York, NY, USA, 2019; pp. 1013–1021.
26. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images. Technical Report. 2009. Available online: http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf (accessed on 22 May 2024).
27. Le, Y.; Yang, X. Tiny Imagenet Visual Recognition Challenge. Stanford Class CS 231N. 2015. Available online: http://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf (accessed on 22 May 2024).
28. Welinder, P.; Branson, S.; Wah, C.; Schroff, F.; Belongie, S.; Perona, P. *Caltech-UCSD Birds 200*; California Institute of Technology: Pasadena, CA, USA, 2010. Available online: https://www.florian-schroff.de/publications/CUB-200.pdf (accessed on 22 May 2024).
29. Khosla, A.; Jayadevaprakash, N.; Yao, B.; Li, F.L. Novel dataset for fine-grained image categorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011.
30. Ba, J.; Caruana, R. Do Deep Nets Really Need to be Deep? In Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 2654–2662.
31. Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; Liang, J. Decoupled knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11953–11962.
32. Huang, T.; Zhang, Y.; Zheng, M.; You, S.; Wang, F.; Qian, C.; Xu, C. Knowledge diffusion for distillation. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 65299–65316.

33. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep mutual learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4320–4328.
34. Zhou, G.; Fan, Y.; Cui, R.; Bian, W.; Zhu, X.; Gai, K. Rocket Launching: A Universal and Efficient Framework for Training Well-Performing Light Net. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, LA, USA, 2–7 February 2018; pp. 4580–4587.
35. Hahn, S.; Choi, H. Self-Knowledge Distillation in Natural Language Processing. In Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, Varna, Bulgaria, 2–4 September 2019; pp. 423–430.
36. Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; Ma, K. Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: New York, NY, USA, 2019; pp. 3712–3721.
37. Yang, C.; Xie, L.; Su, C.; Yuille, A.L. Snapshot Distillation: Teacher-Student Optimization in One Generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 2859–2868.
38. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and Flexible Image Augmentations. *Information* **2020**, *11*, 125. [CrossRef]
39. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, 7–12 February 2020; AAAI Press: Washington, DC, USA, 2020; pp. 13001–13008.
40. Xu, T.B.; Liu, C.L. Data-Distortion Guided Self-Distillation for Deep Neural Networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 5565–5572.
41. Nowlan, S.J.; Hinton, G.E. Simplifying Neural Networks by Soft Weight-Sharing. *Neural Comput.* **1992**, *4*, 473–493. [CrossRef]
42. Srivastava, N.; Hinton, G.E.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
43. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; IEEE: New York, NY, USA, 2020; pp. 10778–10787.
44. Yun, S.; Park, J.; Lee, K.; Shin, J. Regularizing Class-Wise Predictions via Self-Knowledge Distillation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; IEEE: New York, NY, USA, 2020; pp. 13873–13882.
45. Ji, M.; Shin, S.; Hwang, S.; Park, G.; Moon, I. Refine Myself by Teaching Myself: Feature Refinement via Self-Knowledge Distillation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
46. Zhang, H.; Cissé, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018,Vancouver, BC, Canada, 30 April–3 May 2018.
47. Bottou, L. Stochastic Gradient Descent Tricks. In *Neural Networks: Tricks of the Trade*, 2nd ed.; Montavon, G., Orr, G.B., Müller, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7700, pp. 421–436.
48. Zagoruyko, S.; Komodakis, N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In Proceedings of the 5th International Conference on Learning Representations, ICLR, Toulon, France, 24–26 April 2017.
49. Park, W.; Kim, D.; Lu, Y.; Cho, M. Relational Knowledge Distillation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 3967–3976.
50. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.