Multiple-Stage Knowledge Distillation

Chuanyun Xu 1,2, Nanlan Bai 1,* , Wenjian Gao 1, Tian Li 3, Mengwei Li 1, Gang Li 1,* and Yang Zhang 2

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, 52074 Aachen, Germany
* Correspondence: bainanlan@stu.cqut.edu.cn (N.B.); ligang@cqut.edu.cn (G.L.)

Abstract: Knowledge distillation (KD) is a method in which a teacher network guides the learning of a student network, thereby resulting in an improvement in the performance of the student network. Recent research in this area has concentrated on developing effective definitions of knowledge and efficient methods of knowledge transfer while ignoring the learning ability of the student network. To fully utilize this potential learning ability and improve learning efficiency, this study proposes a multiple-stage KD (MSKD) method that allows students to learn the knowledge delivered by the teacher network in multiple stages. The student network in this method consists of a multi-exit architecture, and the students imitate the output of the teacher network at each exit. The final classification by the student network is achieved through ensemble learning. However, because this results in an unreasonable gap between the number of parameters in the student branch network and those in the teacher branch network, as well as a mismatch in learning capacity between these two networks, we extend the MSKD method to a one-to-one multiple-stage KD method. The experimental results reveal that the proposed method applied to the CIFAR100 and Tiny ImageNet datasets exhibits good performance gain. The proposed method of enhancing KD by changing the style of student learning provides new insight into KD.

Keywords: knowledge distillation; multi-exit architecture; teacher–student network; ensemble learning

1. Introduction

In the past few decades, deep neural network (DNN) models have achieved remarkable success in various computer vision tasks, such as image classification [1], object detection [2], and semantic segmentation [3]. Additionally, this has had an impact on various fields [4]. DNN models can effectively extract features by augmenting “width” and “depth” to different degrees; however, increasing the number of layers in the model will increase the training difficulty. Therefore, other methods must be explored to improve the performance of the model.

Knowledge distillation (KD) is a method of model compression [5–7] that can be applied to improve the performance of a model [8,9]. In the conventional KD method, the teacher network has a deeper layer and more parameters than those of the student network. When the student network mimics the output of the teacher network, the accuracy of the student network is increased and model compression is achieved. KD utilizes the “dark knowledge” of the teacher network to direct and increase the performance of the student network. Existing research reveals that the student network can outperform the teacher network [10]. The current focus of KD research is on studying the categories and methods of knowledge transfer; however, the methods of activating the learning potential of student networks are yet to be investigated.

The most effective teaching method fully utilizes the learning ability of a student. Additionally, to maximize the learning ability of a student, an appropriate learning style must be adopted. Furthermore, the general process of human learning is usually divided into several stages, such as a preview before class, classroom learning, and review after...
class. Each stage contributes to a different level of understanding, which can be revised with the assistance of the teacher and books. Eventually, the imparted knowledge can be correctly understood. Figure 1 describes our idea. The teaching process includes the knowledge imparted by the teacher and learning by the student, both of which are essential. Students’ efforts alone cannot be used to improve the effectiveness of teaching; the teacher’s methods of teaching are equally crucial. Since students have limited ability to accept the knowledge transferred by powerful teachers, teachers must provide targeted instruction according to the different stages of learning of the student and impart knowledge that is easy to understand at the current learning stage to achieve better teaching outcomes.

![Image](image_url)

**Figure 1.** Illustration of the proposed methods.

Inspired by the human learning process, a multiple-stage KD method is proposed in this study that modifies the student network in the conventional KD method by dividing the network into several parts according to different depths. Additionally, simplified attention modules, convolutional blocks, and classifiers are added to form a multi-exit architecture student network with a general network backbone. The multiple exits of the student network are used to learn the information delivered by the teacher network in stages for improving network performance. The entire student network is considered to be involved in the learning process with multiple exit branch networks representing various learning stages. Each stage of learning has a particular focus; for example, the early stage of learning focuses on gaining general knowledge of the content, progressing through multiple stages of learning, and finally consolidating the total learning from all the stages. The current KD research focuses on the types of knowledge imparted by the teacher to students while neglecting the exploration of the different learning styles of the student network. Therefore, under the premise of using simple logits as the transmission knowledge, we altered the learning mode of the student network using the multi-exit architecture to fully utilize the learning ability of the student network. However, because the performance of the student network decreases, owing to the lack of learning ability when the capability gap between the teacher and student networks is too large [11], the teacher network possesses a multi-exit architecture similar to that of the student network, in which the early exit of the teacher network is used to guide the corresponding early exit of the student network to achieve targeted guidance. This allows students to adapt to the current stage of learning, and the teacher network can impart comprehensive knowledge, thereby improving teaching efficacy. We validated these methods using two classification datasets, namely CIFAR100 [12] and Tiny ImageNet [13], and analyzed the experimental results.

The main objectives of this study are summarized as follows:
• We propose a multi-stage learning approach where the student can learn the teacher’s understanding of the original knowledge at different stages and the student is able to gradually get an identical understanding with the teacher, even becoming superior to the teacher.

• We propose a multiple-stage KD (MSKD) method for the supervised training of different stages of a multi-exit network. In the MSKD, the output of the teacher network is used to guide the different stages of the student network, and the trained early stage network is trained later in the training process, thereby resulting in phased recursive training.

• We extend the MSKD method to a one-to-one MSKD (OtO) method that allows a multi-exit teacher network to provide one-to-one targeted instruction to a multi-exit student network.

• Extensive experiments were conducted using CIFAR100 and Tiny ImageNet to test different teacher–student network architectures and verify the effectiveness of the proposed method.

2. Related Work

2.1. KD

KD is a common technique for knowledge transfer between two models, which was first discussed in detail by Hinton et al. [14]. In this technique, a temperature was used to soften the output of the teacher network and minimize the Kullback–Leibler divergence of the outputs of the teacher and student networks. Several relevant theoretical explanations [15–18], improved methods [5,19,20], and other applications [21,22] of KD exist. Instead of focusing on the output of the network, Komodakis et al. [5] investigated the feature map of the teacher–student network to minimize the differences in this feature map. Passalis et al. [23] focused on the geometry of the feature space of the teacher network and mapped it effectively to the feature space of the student network. Liu et al. [24] reported the important role of inter-channel correlations in retaining features and aligning the diversity and isomorphism of the feature spaces of the student and teacher networks. Wen et al. [25] adapted the knowledge transferred by the teacher network and used dynamic temperature penalization in case of poor supervision to improve the student network. Tian et al. [26], Xu et al. [27], and Yang et al. [28] combined self-supervision with KD using different definitions of a pretext task, thereby increasing the effectiveness of the knowledge transfer. Notably, these methods usually use data augmentation operations, such as rotating images, to define the pretext task. Unlike the previous model, in which students learned unilaterally from the teacher, Fu et al. [29] utilized an interactive teaching model that increased the effectiveness of KD by the teacher–student network. Most studies have focused on exploring different knowledge categories and how knowledge is transferred while neglecting to explore the different learning styles of student networks. To explore the impact of student learning styles, we chose to use a simple definition of knowledge in our method.

2.2. Multi-Exit Architectures

Teerapittayanon et al. [30] was the first study to propose the use of early exit branch networks to achieve fast inference in DNNs. Subsequently, Huang et al. [31] discovered that added early exits affect the performance of the final exits and the early exits do not possess coarse-scale features; therefore a multi-scale dense network was proposed. Additionally, two experimental conditions were set for image classification under different resource constraints. Since then, many researchers have investigated multi-exit networks from different perspectives. Li et al. [32] explored methods to train multi-exit networks effectively and used gradient equalization algorithms to solve the learning conflicts among the exits and improved the learning speed of the exits. There has been research on combining KD with a multi-exit architecture. Scan [33] and Zhang et al. [34] prevented multiple classifiers of different depths from negatively affecting each other and enabled them to benefit from each other by adding KD and attention mechanisms. Phuong et al. [35] used
KD to encourage early exits to mimic the previous exit, thereby resulting in a new form of KD. Each exit in harmonized dense KD [36] learns from all other exits and uses adaptive weight learning to optimize multiple losses. Such studies ignore teacher guidance, which can reduce resource use but also misses out on more effective guidance.

In terms of the overall framework, hierarchical self-supervised augmented KD (HSAKD) [28] is similar to this study. HSAKD adds multiple auxiliary classifiers at different depths of the network to transfer different self-supervised augmented distributions extracted from the hierarchical features. In contrast, the purpose of constructing the teacher-student multi-exit architecture, construction methods, and knowledge transfer in this study are different from those in HSAKD. The purpose of constructing the multi-exit student network is to stimulate the learning ability of the student, to allow multiple stages of learning, and to use the attention mechanism to ensure a different focus for each stage of learning. These can ultimately achieve better ensemble learning. The purpose of constructing a teacher multi-exit network is to transfer knowledge comprehensively and allow the students to adapt to the current stage of learning. Additionally, HSAKD uses a complex knowledge definition, whereas we only use logit, which is a simple knowledge definition in the traditional approach.

3. Proposed Method

In this section, the basics of KD are briefly introduced and the proposed MSKD method is described. Finally, the OtO method is introduced.

3.1. KD

Given a dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$ consisting of $N$ samples with $C$ classes, $y_n \in \{1 \ldots C\}$. The KD method uses the student network to learn the information delivered by the teacher network, thereby improving the performance of the student network. For ease of presentation, $t$ and $s$ are used to distinguish between the teacher and student networks, respectively. The temperature-softened network output is introduced in the conventional KD [14] to obtain supervised information. At this point, the probability of the network output for each category $j \in \{1 \ldots C\}$ is expressed as:

$$p(x, \tau) = \frac{e^{(z_j/\tau)}}{\sum_{i=1}^{C} e^{(z_i/\tau)}}$$

where $x$ denotes the input sample, $z_j$ is the output of the network before the softmax function [37] is applied, and $\sigma(\cdot)$ is the softmax function. $\tau$ is the hyper-parameter temperature, which controls the smoothness of the probability distribution of the output. The conventional KD method attempts to minimize the gap between the output of the teacher and student networks at the same temperature to allow the student network to mimic the teacher network. The distance between the teacher and student, i.e., the distillation loss, is denoted as:

$$\sum_{i=1}^{N} D_{KL}(p_t(x_i, \tau), p_s(x_i, \tau))$$

where $D_{KL}$ is the distance function Kullback–Leibler divergence [38] that measures the distance between the teacher and student networks. $p_t(x_i, \tau)$ and $p_s(x_i, \tau)$ denote the output of the teacher network and that of the student network after temperature smoothing, respectively. The distance between the student network and true label, i.e., the classification loss, is expressed as:

$$\sum_{i=1}^{N} L_{CE}(p_s(x_i, 1), y_i)$$
where \( L_{CE} \) denotes the cross-entropy function and \( y_i \) refers to the true label of this sample. The overall loss of the student network consists of the distillation and classification losses.

\[
\text{Loss}_{KD} = \text{Loss}_{cla} + \text{Loss}_{dis}
\]

\[
= \sum_{i=1}^{N} \left( (1 - \alpha)L_{CE}(p_s(x_i, 1), y_i) + \alpha \tau^2 D_{KL}(p_t(x_i, \tau), p_s(x_i, \tau)) \right)
\]

(4)

where \( \alpha \) is the hyper-parameter used to balance the two losses and since the order of magnitude of the classification loss gradient is approximately \( \tau^2 \) times that of the distillation loss gradient. The distillation loss is multiplied by \( \tau^2 \) to keep the contribution of the two loss gradients consistent.

3.2. MSKD

To make the student network available for multi-stage learning, we divided the general network into several parts according to different depths and set the corresponding early exits. This network with multi-exit architecture and \( K \) exits can be expressed as \( \{c_k(\cdot)\}_{k=1}^{K} \), where \( c_k(\cdot) \) denotes the \( k \)-th exit of the network. The loss of our proposed MSKD method possesses two main components: the classification loss (\( \text{Loss}_{cla} \)), i.e., the loss obtained by comparing the output of the student network with the true labels, and distillation loss (\( \text{Loss}_{dis} \)), i.e., the loss obtained by comparing the output of the student network with the output of the teacher network. The true labels are the actual category of the samples. The output of the teacher network refers to the probability distribution of the teacher network’s judgment on the sample category, and the output of the teacher network is softened by temperature to obtain a smoother probability distribution. The framework is shown in Figure 2 with a multi-exit student network learning true labels in multiple stages, as well as knowledge transferred by the teacher network. The loss corresponding to the \( k \)-th exit of the student network can be expressed as:

\[
\text{Loss}_{MSKD}^k = \text{Loss}_{cla}^k + \text{Loss}_{dis}^k
\]

\[
= \sum_{i=1}^{N} \left( (1 - \alpha)L_{CE}(p^k_s(x_i, 1), y_i) + \alpha \tau^2 D_{KL}(p^k_s(x_i, \tau), p_t(x_i, \tau)) \right)
\]

(5)

where \( p^k_s(x_i, 1) \) denotes the output of the \( k \)-th exit of the student network without temperature softening and \( p^k_s(x_i, \tau) \) represents the output of the \( k \)-th exit of the student network when the temperature is \( \tau \). At this point, the student network output is softened by the temperature to obtain a smoother probability distribution. The value of \( \tau \) ranges from greater than or equal to 1. For a student network with a multi-exit architecture, the overall loss is the combination of the loss at each exit, which can be expressed as:

\[
\text{Loss}_{MSKD} = \sum_{k=1}^{K} \text{Loss}_{MSKD}^k
\]

\[
= \sum_{k=1}^{K} \sum_{i=1}^{N} \left( (1 - \alpha)L_{CE}(p^k_s(x_i, 1), y_i) + \alpha \tau^2 D_{KL}(p^k_s(x_i, \tau), p_t(x_i, \tau)) \right)
\]

(6)
Figure 2. Overall framework diagram of the MSKD. Each of these exits possesses a distillation loss from the teacher network and classification loss from the true labels.

The pseudo-code for this method, which utilizes the standard gradient-based optimization method, is provided in Algorithm 1 below. During testing, we used an ensemble algorithm to combine the knowledge learned at each stage to constitute the final learning.

$$p_s(x) = \frac{1}{K} \sum_{k=1}^{K} p^k_s(x, 1)$$

(7)

Algorithm 1: Multiple-Stage Knowledge Distillation

```
Given: Dataset D, Model t, Multi-exit Model s
for (x, y) in D do
    p_t = Model_t(x)  // The Model_s outputs include multiple exit outputs.
    p_s = Model_s(x)  // Ensure that only students can learn from the teacher.
    p_t.detach()
for k in 1...K do
    loss = loss + (1 - \alpha) \text{CrossEntropy}(p^k_s(x, 1), y) + \alpha \tau^2 \text{KLDivLoss}(p^k_s(x, \tau), p_t(x, \tau))
end
loss.backward()
update()
```

3.3. OtO Method

The multi-stage learning of the final output of the teacher network stimulates the learning potential of the student network; however, the performance improvement of the student network is limited because the teacher network only delivers the probabilistic information available at the final exit. The student network receiving information is a multi-exit branch network, and the gap between the learning ability of the pre-branch
student network and that of the teacher network is too large. This gap affects learning to a certain extent. Therefore, the teacher network is also constructed as a multi-exit architecture network and the teaching style is changed to one-to-one instruction such that the teacher network is targeted to guide the student network through different stages of learning. The overall framework of our proposed OtO method is shown in Figure 3, and the supervisory information consists of two parts: (1) the true label and (2) the branch network of teachers corresponding to that stage. The loss of the $k$-th exit of the student network in this method can be expressed as:

$$\text{Loss}_{oto}^k = \text{Loss}_{cla}^k + \text{Loss}_{dis'}^k = \sum_{i=1}^{N} \left( (1 - \alpha) L_{CE} \left( p_s^k(x_i, 1), y_i \right) + \alpha \tau^2 D_{KL} \left( p_s^k(x_i, \tau), p_t^k(x_i, \tau) \right) \right)$$

where $p_t^k(x_i, \tau)$ denotes the output of the probability distribution of the $k$-th exit of the teacher network after softening. As can be seen with the above description, the main difference between Equation (5) and Equation (8) is the source of guidance of the teacher network, one is the final output of the teacher network with the specific formula $p_t(x_i, \tau)$ in Equation (5) and the other corresponds to the output of the teacher branch network with the specific formula $p_t^k(x_i, \tau)$ in Equation (8). Therefore, the overall loss of the student network can be expressed as:

$$\text{Loss}_{oto} = \sum_{k=1}^{K} \text{Loss}_{oto}^k = \sum_{k=1}^{K} \sum_{i=1}^{N} \left( (1 - \alpha) L_{CE} \left( p_s^k(x_i, 1), y_i \right) + \alpha \tau^2 D_{KL} \left( p_s^k(x_i, \tau), p_t^k(x_i, \tau) \right) \right)$$

![Figure 3. Overall framework of the OtO method. Each of these exits has two types of losses: a distillation loss from the branch network corresponding to the teacher network and classification loss from the true label.](image-url)
When each exit of the student network corresponds to an exit of the teacher network for one-to-one learning, the difference in the learning capacities between the teacher and student networks becomes small. We can then seek commonalities and differences in the training process with students sharing parameters in the backbone of the network and varying independent learning in the branches. Additionally, information is shared from different stages and relative differences are maintained. The same ensemble approach as that of MSKD is used to integrate multi-stage learning outcomes as the final learning outcome during testing.

4. Experiments

In this section, we first present the datasets used and the relevant setup of the experiment. The experimental data are subsequently reported in relation to the proposed method comparing the benchmark and KD methods.

4.1. Experimental Setup

Datasets. Relevant experiments were conducted on two standard classification datasets, namely CIFAR100 [12] and Tiny ImageNet [13]. CIFAR100 contains a total of 60,000 color images with a size of $32 \times 32$. Additionally, this dataset contains 100 categories with 600 images in each category. We followed the default split using 50,000 images for training and 10,000 images for testing the accuracy of the model. During training, random cropping and horizon flipping are applied to the data augmentation. In contrast, during testing, we do not perform any operations on the images. The Tiny ImageNet dataset contains RGB images of size $64 \times 64$ with 100,000 images in 200 classes for training and 5000 images each for validation and testing. During training, the images are randomly rotated for data augmentation, whereas no operations were performed on the images during testing. The details of the data set are shown in Table 1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total</th>
<th>Category</th>
<th>Size</th>
<th>Data Augmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR100</td>
<td>60k</td>
<td>100</td>
<td>$32 \times 32$</td>
<td>random cropping and horizon flipping</td>
</tr>
<tr>
<td>Tiny ImageNet</td>
<td>120k</td>
<td>200</td>
<td>$64 \times 64$</td>
<td>random rotation</td>
</tr>
</tbody>
</table>

Optimization and hyper-parameters. Following previous research [26], we set the hyper-parameter temperature $\tau$ to 3 and $\alpha$ to 0.3. Each experiment was trained for 240 epochs with a batch size of 128. We use a stochastic gradient descent optimizer with a momentum of 0.9 and the weight decay was set to $5 \times 10^{-4}$. The initial learning rate was 0.1, which was then divided by 10 at the 150th, 180th, and 210th epochs. All the experiments were performed in PyTorch on GPU (RTX2080s) devices.

Model architecture. We selected widely used neural networks including residual network (ResNet) [1], visual geometry group (VGG) [3], and wide residual network (WRN) [39]. Additionally, we followed the most common structure of multi-exit networks constructed by dividing the above networks into four parts as the backbone network, i.e., by adding three early exit classifiers, as shown in Figures 2 and 3. We divide the network according to its own constituent blocks, for example, the ResNet family of networks adds early exits after the ResBlock. Additionally, the FC in the figure indicates the fully connected layer. To prevent the early exits from interfering with the performance of the final exit and ensure that the final exit can benefit from them, we added simplified attention modules [33] at each exit to enable each exit classifier to acquire the unique features of these modules. The attention module mainly consists of a convolutional layer and a deconvolutional layer. Additionally, mutually different exit classifiers ensure the effectiveness of ensemble learning.
4.2. Comparison with the Benchmark Method

To demonstrate the performance improvement of the proposed method over the network in the training approach of the benchmark method, we display the performance of our OTO method for CIFAR100 and Tiny ImageNet datasets in Tables 2 and 3, respectively. The benchmark method trains the student using standard cross-entropy. As shown in the two tables, the final accuracy obtained from testing the proposed OTO method is at least 4% higher than that of the benchmark method for CIFAR100 and Tiny ImageNet for a variety of teacher–student architectures. Selecting VGG13 and VGG8 [3] as the teacher and student networks, respectively, improves the performance of the student network by 5.91% for CIFAR100. ResNet64 and ResNet18 [1] are selected as the teacher and student networks, respectively, for Tiny ImageNet, and the student network performance is improved by 5.46%. In addition, the proposed method can outperform the benchmark method in the earlier exit. For example, when VGG13–VGG8 is used as the teacher–student architecture for CIFAR100, VGG8 outperforms the benchmark method at the first exit. This demonstrates that the proposed method can achieve better performance than that of the benchmark method. According to the accuracy of each exit branch network shown in the table, the performance of the later branch network is not necessarily better than that of the earlier branch network; however, the performance of each exit exhibits a gradually increasing trend. This is because the proposed method jointly optimizes multiple exit branch networks to improve each of them. The final result of the total approach is obtained using an ensemble to combine the knowledge learned in multiple exit branch networks.

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Baseline</th>
<th>Student</th>
<th>Baseline</th>
<th>Exit1</th>
<th>Exit2</th>
<th>Exit3</th>
<th>Exit4</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet152</td>
<td>80.88</td>
<td>ResNet50</td>
<td>79.34</td>
<td>79.32</td>
<td>80.80</td>
<td>81.78</td>
<td>81.74</td>
<td>83.20</td>
</tr>
<tr>
<td>ResNet152</td>
<td>80.88</td>
<td>ResNet18</td>
<td>78.47</td>
<td>75.06</td>
<td>77.40</td>
<td>80.15</td>
<td>80.50</td>
<td>81.73</td>
</tr>
<tr>
<td>ResNet152</td>
<td>80.88</td>
<td>ResNet10</td>
<td>76.19</td>
<td>73.19</td>
<td>75.12</td>
<td>77.52</td>
<td>78.02</td>
<td>79.74</td>
</tr>
<tr>
<td>ResNet50</td>
<td>79.34</td>
<td>ResNet18</td>
<td>78.47</td>
<td>74.37</td>
<td>77.62</td>
<td>80.09</td>
<td>80.65</td>
<td>81.42</td>
</tr>
<tr>
<td>ResNet50</td>
<td>79.34</td>
<td>VGG8</td>
<td>70.36</td>
<td>71.20</td>
<td>69.79</td>
<td>70.95</td>
<td>72.98</td>
<td>76.05</td>
</tr>
<tr>
<td>VGG13</td>
<td>74.64</td>
<td>VGG8</td>
<td>70.36</td>
<td>71.17</td>
<td>70.59</td>
<td>71.10</td>
<td>72.65</td>
<td>76.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Baseline</th>
<th>Student</th>
<th>Baseline</th>
<th>Exit1</th>
<th>Exit2</th>
<th>Exit3</th>
<th>Exit4</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet34</td>
<td>66.40</td>
<td>ResNet34</td>
<td>66.40</td>
<td>59.52</td>
<td>63.80</td>
<td>68.65</td>
<td>70.38</td>
<td>71.38</td>
</tr>
<tr>
<td>ResNet34</td>
<td>66.40</td>
<td>ResNet18</td>
<td>64.40</td>
<td>59.86</td>
<td>62.30</td>
<td>65.30</td>
<td>68.04</td>
<td>69.86</td>
</tr>
</tbody>
</table>

4.3. Comparison with Other KD Methods

To demonstrate the effectiveness and robustness of the proposed method, we compared it with other mainstream KD methods for the CIFAR100 and Tiny ImageNet datasets, and the experimental results are shown in Tables 4 and 5. All KD methods can improve network performance to some extent, as seen by the statistics presented in the tables. However, not all improved KD methods consistently outperform conventional KD methods under different teacher–student structures. The proposed method consistently outperforms the traditional KD methods under a variety of teacher–student architectures. Specifically, for a ResNet18 student trained by a ResNet50 teacher on CIFAR100, the proposed method results in a 2.7% improvement in the student network performance, and the accuracy of the student network is 1.83% higher than that of the teacher network. The OTO method, which is an extension of MSKD, further improves the accuracy, its performance is slightly worse than the SOTA distillation method HSAKD.
when the teacher network is ResNet152, but is excellent with the rest of the teacher–student architectures. The SOTA HSAKD [28] method augments data for self-supervised learning, and data augmentation aids its effectiveness. Furthermore, we used the AutoAugment [40] strategy in OtO, and Table 4 shows that the proposed method superior to the previous KD method at this time.

Table 4. Comparison of the accuracy (%) of our method with those of other KD methods for CIFAR100. The best and second-best results are shown in bold and underlined, respectively. “OtO∗” denotes the results after data augmentation.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>KD [14]</td>
<td>79.69</td>
<td>79.86</td>
<td>77.85</td>
<td>79.72</td>
<td>73.81</td>
<td>72.98</td>
<td>74.92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FitNet [41]</td>
<td>80.51</td>
<td>79.24</td>
<td>78.02</td>
<td>76.27</td>
<td>73.24</td>
<td>72.32</td>
<td>73.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT [5]</td>
<td>80.41</td>
<td>80.19</td>
<td>78.45</td>
<td>78.28</td>
<td>74.01</td>
<td>73.48</td>
<td>73.97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP [42]</td>
<td>80.72</td>
<td>79.87</td>
<td>78.25</td>
<td>79.32</td>
<td>73.52</td>
<td>73.49</td>
<td>74.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC [43]</td>
<td>79.89</td>
<td>79.82</td>
<td>78.06</td>
<td>77.42</td>
<td>73.48</td>
<td>73.04</td>
<td>73.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VID [44]</td>
<td>79.24</td>
<td>79.67</td>
<td>77.80</td>
<td>76.79</td>
<td>73.46</td>
<td>72.97</td>
<td>74.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RKD [45]</td>
<td>80.22</td>
<td>79.60</td>
<td>77.90</td>
<td>78.75</td>
<td>73.51</td>
<td>73.19</td>
<td>73.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PKT [23]</td>
<td>80.57</td>
<td>79.44</td>
<td>78.41</td>
<td>78.89</td>
<td>73.61</td>
<td>73.25</td>
<td>74.42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB [46]</td>
<td>81.21</td>
<td>79.50</td>
<td>78.18</td>
<td>76.77</td>
<td>73.65</td>
<td>73.35</td>
<td>72.65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FT [47]</td>
<td>80.37</td>
<td>79.26</td>
<td>77.53</td>
<td>76.29</td>
<td>72.98</td>
<td>73.44</td>
<td>73.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRD [26]</td>
<td>80.53</td>
<td>79.81</td>
<td>78.60</td>
<td>74.58</td>
<td>74.58</td>
<td>74.29</td>
<td>75.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSKD [27]</td>
<td>80.29</td>
<td>80.36</td>
<td>78.60</td>
<td>79.72</td>
<td>75.36</td>
<td>75.27</td>
<td>76.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSAKD [28]</td>
<td>83.33</td>
<td>82.17</td>
<td>79.75</td>
<td>81.22</td>
<td>75.20</td>
<td>75.07</td>
<td>78.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSKD [48]</td>
<td>83.09</td>
<td>81.27</td>
<td>79.34</td>
<td>81.28</td>
<td>75.63</td>
<td>75.42</td>
<td>78.31</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Teacher Student</th>
<th>ResNet34</th>
<th>ResNet34</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSKD (ours)</td>
<td>82.37</td>
<td>79.38</td>
</tr>
<tr>
<td>OtO (ours)</td>
<td>83.20</td>
<td>79.74</td>
</tr>
<tr>
<td>OtO∗ (ours)</td>
<td>83.47</td>
<td>81.17</td>
</tr>
</tbody>
</table>

Table 5. Comparison of accuracy (%) of the proposed method with those of other KD methods for Tiny ImageNet. The best and second-best results are shown in bold and underlined, respectively.

<table>
<thead>
<tr>
<th>Teacher Student</th>
<th>ResNet34</th>
<th>ResNet34</th>
</tr>
</thead>
<tbody>
<tr>
<td>KD [14]</td>
<td>67.90</td>
<td>68.80</td>
</tr>
<tr>
<td>FitNet [41]</td>
<td>64.44</td>
<td>66.18</td>
</tr>
<tr>
<td>AT [5]</td>
<td>66.54</td>
<td>67.16</td>
</tr>
<tr>
<td>SP [42]</td>
<td>67.06</td>
<td>68.36</td>
</tr>
<tr>
<td>VID [44]</td>
<td>66.92</td>
<td>68.08</td>
</tr>
<tr>
<td>RKD [45]</td>
<td>65.90</td>
<td>67.80</td>
</tr>
<tr>
<td>PKT [23]</td>
<td>65.76</td>
<td>68.18</td>
</tr>
<tr>
<td>CRD [26]</td>
<td>67.94</td>
<td>68.96</td>
</tr>
<tr>
<td>AFD [49]</td>
<td>68.56</td>
<td>69.58</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Teacher Student</th>
<th>ResNet34</th>
<th>ResNet34</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSKD (ours)</td>
<td>69.04</td>
<td>70.78</td>
</tr>
<tr>
<td>OtO (ours)</td>
<td>69.86</td>
<td>71.38</td>
</tr>
</tbody>
</table>

5. Analysis

In this section, we first explore the validity of each part of the proposed method using ablation experiments. Subsequently, we explore the validity of the ensemble in the proposed method.
5.1. Ablation Study

To verify the effectiveness of each part of our OtO method, several teacher–student architectures are selected to conduct an ablation study for CIFAR100. Overall, we explored the following: (i) a multi-exit architecture network (Mu) using only classification loss, i.e., setting $\alpha$ in Equation (5) to 0; (ii) applying a conventional KD method on a multi-exit architecture network (Mu+KD), i.e., only the final exit of the multiple-exit student network architecture receives guidance from the original teacher network; (iii) OtO and the use of ensemble learning in the above three training modalities. The experimental results are shown in Figure 4. As shown in the figure, the multi-exit architecture student network with only one last exit to receive guidance from the teacher network improves the accuracies of the single last exit and ensemble learning. This illustrates the effectiveness of KD applied to multi-exit architectures. Furthermore, our proposed method is substantially better than the original multi-exit architecture network and conventional KD method, which indicates that a multi-exit architecture teacher–student network for one-on-one instruction stimulates the learning ability of the student network and results in a greater amount of knowledge. As shown in Figure 4b, in several teacher–student architectures, the accuracy is eventually improved by the ensemble, which illustrates the practical relevance of the ensemble for combining multiple stages. In conclusion, our proposed method can achieve better results using the ensemble.
Figure 4. Ablation study of multiple teacher–student architectures for CIFAR100. The horizontal coordinate refers to the teacher–student architectures. (a) Accuracy of the final exit corresponding to different training methods. (b) Accuracy after the ensemble of multiple exits corresponding to different training methods.

5.2. Ensemble Effectiveness of the Proposed Method

Here we will discuss the effectiveness of ensemble learning in proposing methods. In our approach, we first construct the student network as a multi-exit architecture network and add early classifiers to different depths of the neural network to form a multi-exit network. The multi-exit architecture network jointly optimizes the loss of multiple exits, each of which is equivalent to a regularization term for the remaining exits, which prevents overfitting and improves accuracy. Additionally, the early exit can reduce the tendency of gradient disappearance [50]. The student model is divided into multiple stages sharing the first few layers of the network structure, which can use the same low-level features, isolate unique features at their respective exits, and then use individual branches for higher-level features to achieve better classification results. To ensure that the ensemble is effective, we must preserve as much branch diversity as possible. Firstly, we added multiple exits at different depths when constructing a multi-exit architecture student network, so that each exit uses a different depth of the shared backbone. Secondly, we used a simplified attention module at the diversion in order to make that exit acquire unique features, and finally, we constructed branching networks at different depths according to the depth of the shared backbone of the exits. This set of measures allows each branch network to acquire different features and therefore different classification results, and the experimental data are shown in Tables 2 and 3. As can be seen from the data in the table, the classification accuracy of each branch network varies, showing a gradual increase in general, and the accuracy can be further improved after the final ensemble. For instance, when VGG13 is the teacher network and VGG8 is the student network in Table 2, the final integrated accuracy is improved by 3.62% compared to the final exit accuracy.

Relevant experiments were conducted to illustrate the differences in the classification performance of each exit. Herein, we choose ResNet34 and ResNet18 as the teacher and student network, respectively, for the Tiny ImageNet dataset and explore the samples that
are correctly classified for each exit. The results reveal that the samples classified correctly in each exit are not identical, and there are cases in which samples classified correctly at the earlier exit are misclassified at the later exit. Figure 5 shows a set of samples that are correctly classified at the first exit only. Grad-CAM [51] can be used to visualize the importance of this set of samples at the first exit versus that at the last exit in the prediction process for each position of the image for the corresponding class. Notably, the first and last exits are concerned with the general structure of the input image and image details, respectively. The candle sample in Figure 5 is an example in which the first exit focuses only on the simple structure of the candle to arrive at the correct classification. In contrast, the final exit focuses on the image details. More specifically, the final exit only focuses on the local information of the candle, which results in the classifier mistaking the image for that of a jellyfish and finally misclassifying it. The different exits of the student network focus on different aspects of the samples, which results in different classifications. Furthermore, variabilities exist among the multiple exits of a multi-exit architecture student network and the validity of the ensemble is ensured.

<table>
<thead>
<tr>
<th>The original image</th>
<th>Exit1 (Correct)</th>
<th>Exit4 (Wrong)</th>
<th>The original image</th>
<th>Exit1 (Correct)</th>
<th>Exit4 (Wrong)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pole</td>
<td>Pole</td>
<td>Oboe</td>
<td>Barbershop</td>
<td>Barbershop</td>
<td>Oboe</td>
</tr>
<tr>
<td>Tailed frog</td>
<td>Tailed frog</td>
<td>Tabby</td>
<td>Candle</td>
<td>Candle</td>
<td>Jellyfish</td>
</tr>
</tbody>
</table>

**Figure 5.** A set of Tiny ImageNet samples with only the first exit classified correctly in the case of a teacher–student architecture of ResNet34–ResNet18, and heat maps drawn with Grad-CAM. Below the original image column is the true label for that sample, and below each sample in each exit column is the predicted category label for that exit.

5.3. Exploratory Experiments

In the multi-teacher KD method in the KD approach [6], the student network can receive guidance from multiple teachers simultaneously. We added a final exit guide of the teacher network to the OtO method, i.e., the early exit of the student network had two teachers: one at the corresponding exit and another at the final exit. This was equivalent to combining the MSKD and OtO methods. Herein, we used a simple average teacher output to instruct the student network, and the results of the experiment are shown in Figure 6.

Here, we chose two teacher–student architectures for our experiments and compare Mu+KD, MSKD, OtO, and the combination of multiple-stage KD and OtO (OtO+last). To visualize the differences in the methods, the framework diagram of the above methods is displayed in Figure 7. From the experimental results in Figure 6, the classification accuracy of MU+KD and MSKD is relatively low, OtO+last is observed to possess the best accuracy at the first exit; however, its performance does not exceed that of OtO at the later exits, and classification accuracy is still the best for OtO method after the final ensemble. Additionally, the accuracies of the early exits improve when guided by multiple teachers; however, when the learning goals of multiple exits tend to converge, the diversity among exits reduces and results in a poor ensemble. Therefore, to achieve the best final ensemble, a balance of
accuracy and variability must be maintained across multiple exits. Of the four methods compared, the OtO method has the best balance of accuracy and variability.

Figure 6. Comparison of the accuracies of different methods for multiple exits on CIFAR100.

Figure 7. The framework diagram of each method. The blue line indicates the supervision of the true label, the green line indicates the supervision of the teacher network, and the red line indicates the supervision of the early exit of the teacher network. (a–d) denote the four training methods mentioned in the context, respectively.

6. Conclusions

Inspired by the learning process of humans, this study divides the learning process of a DNN into multiple stages to stimulate the learning ability of students for improved learning. Herein, we propose an MSKD method in which the student network is constructed as a multi-exit architecture network that mimics the output of the teacher network at different
depths. Additionally, an ensemble is used to combine the results of the multiple stages of learning to form the final output. To match the learning capacities of the teacher and student networks, the teacher network also possesses a multi-exit architecture, thereby allowing the teacher branch network to provide one-to-one guidance to the student branch network. This results in the formation of the OtO method. The proposed OtO method is at least 4% more accurate than the benchmark method for a variety of teacher–student architectures, and can achieve a high level compared with other mainstream distillation methods. Numerous experimental results revealed that the proposed method can stimulate the learning potential of the student by changing their learning styles, which provides a new perspective on KD.

In this study, we only used logits as the knowledge being transferred. We can be considered to introduce the definition of knowledge from other studies or will be able to achieve a better distillation effect. Furthermore, we treated all the stages of learning equally. In the future, we may consider highlighting certain learning stages using learning adaptive weights.

Author Contributions: Conceptualization, C.X. and N.B.; methodology, C.X. and N.B.; software, G.L.; validation, W.G., T.L. and M.L.; formal analysis, W.G.; investigation, M.L.; resources, G.L.; data curation, T.L.; writing—original draft preparation, N.B.; writing—review and editing, C.X. and W.G.; visualization, N.B.; supervision, Y.Z.; project administration, Y.Z.; funding acquisition, C.X. and G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the China Chongqing Science and Technology Commission under Grants cstc2020jscx-msxmX0086, cstc2019jscx-zdztx0043, cstc2019jcyj-msxmX0442, and cstc2021jcyj-msxmX0605; China Chongqing Banan District Science and Technology Commission project under Grant 2020QX413; China Chongqing Municipal Education Commission under Grant KJQN202001137.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References


