mathematics

MDPI

Article

DE-MKD: Decoupled Multi-Teacher Knowledge Distillation Based on Entropy

Xin Cheng [1], Zhiqiang Zhang [2], Wei Weng [3], Wenxin Yu [2] and Jinjia Zhou [1,*]

[1] Graduate School of Science and Engineering, Hosei University, Tokyo 184-8584, Japan; xin.cheng.5x@stu.hosei.ac.jp
[2] School of Science and Technology, Southwest University of Science and Technology, Mianyang 621010, China; zzq.zhangzhiqiang2018@gmail.com (Z.Z.); yuwenxin@swust.edu.cn (W.Y.)
[3] Institute of Liberal Arts and Science, Kanazawa University, Kanazawa City 920-1192, Japan; weng@staff.kanazawa-u.ac.jp
* Correspondence: zhou@hosei.ac.jp

Abstract: The complexity of deep neural network models (DNNs) severely limits their application on devices with limited computing and storage resources. Knowledge distillation (KD) is an attractive model compression technology that can effectively alleviate this problem. Multi-teacher knowledge distillation (MKD) aims to leverage the valuable and diverse knowledge distilled by multiple teacher networks to improve the performance of the student network. Existing approaches typically rely on simple methods such as averaging the prediction logits or using sub-optimal weighting strategies to fuse distilled knowledge from multiple teachers. However, employing these techniques cannot fully reflect the importance of teachers and may even mislead student's learning. To address this issue, we propose a novel Decoupled Multi-Teacher Knowledge Distillation based on Entropy (DE-MKD). DE-MKD decouples the vanilla knowledge distillation loss and assigns adaptive weights to each teacher to reflect its importance based on the entropy of their predictions. Furthermore, we extend the proposed approach to distill the intermediate features from multiple powerful but cumbersome teachers to improve the performance of the lightweight student network. Extensive experiments on the publicly available CIFAR-100 image classification benchmark dataset with various teacher-student network pairs demonstrated the effectiveness and flexibility of our approach. For instance, the VGG8 | ShuffleNetV2 model trained by DE-MKD reached 75.25% | 78.86% top-one accuracy when choosing VGG13 | WRN40-2 as the teacher, setting new performance records. In addition, surprisingly, the distilled student model outperformed the teacher in both teacher-student network pairs.

Keywords: multi-teacher knowledge distillation; image classification; entropy; deep learning

MSC: 68T07

Citation: Cheng, X.; Zhang, Z.; Weng, W.; Yu, W.; Zhou, J. DE-MKD: Decoupled Multi-Teacher Knowledge Distillation Based on Entropy. *Mathematics* **2024**, *12*, 1672. https://doi.org/10.3390/math12111672

Academic Editors: Jüri Majak and Michael Voskoglou

Received: 15 April 2024
Revised: 15 May 2024
Accepted: 24 May 2024
Published: 27 May 2024

1. Introduction

Over the past decade, deep neural networks (DNNs) [1] have achieved remarkable success in various vision tasks, including image classification [2–4], object detection [5,6], and semantic segmentation [7,8]. However, high-performing models usually necessitate substantial computational resources and storage. This requirement often precludes their deployment on edge devices with limited computational and storage resources. One promising approach to tackle this issue is knowledge distillation (KD) [9], which leverages "dark knowledge" from a powerful yet cumbersome teacher network to guide a lightweight student network. KD aims to force the student network to mimic the teacher's prediction while utilizing only a small number of parameters.

According to the type of knowledge [10], KD can be categorized into three categories: response-based KD, feature-based KD, and relationship-based KD. Vanilla KD [9] guides the student network's training by encouraging it to mimic the teacher's prediction. To

enhance student performance further, researchers have begun scrutinizing the middle layer features of the teacher network as knowledge to guide student network training. FitNets [11] align the student's middle layer features with those of the teacher. Attention-transfer-based knowledge distillation (AT) [12] employs the attention map of features for knowledge transfer.

However, vanilla KD methods rely on a single pre-trained teacher network. More recently, considering human cognitive learning processes, investigators have explored the idea that students can benefit from multiple teachers. This trigger has led to the conception of multi-teacher distillation (MKD), which aims to leverage the valuable and diverse knowledge presented by multiple teacher networks to improve the performance of the student network. Many MKD approaches have demonstrated that students can benefit from multiple teachers [13–18]. However, many MKD methods usually fail to reflect the importance of each teacher using the same or fixed weight [13–15]. This leads to unreasonable integration among knowledge from multiple teachers, and the student cannot fully utilize the integrated knowledge. Some methods [16–18] use various strategies to try to improve the problem of unreasonable knowledge integration. However, these methods suffer from certain deficiencies that prevent them from fully leveraging knowledge integration, which consequently leads to limited performance enhancements.

DKD [19] reveals that the vanilla KD loss is coupled, leading to insufficient exploitation of knowledge transferred by the teacher. Inspired by this, we propose a novel multi-teacher KD method called DE-MKD , which leverages the abundant distilled knowledge from multiple teachers to enhance students' performance. Specifically, we assign sample-aware teacher importance weights based on the entropy of teachers' predictions. The importance weight assigned to the teacher decreases as the information entropy increases. Recognizing the significance of features in representation learning, we further extend our method to incorporate intermediate features of teacher models for knowledge distillation. We demonstrated the effectiveness of our approach on the CIFAR-100 dataset, a widely recognized benchmark. Obviously, our work goals mainly focus on two points: **(1) how to integrate knowledge from multiple sources reasonably** and **(2) how to use the integrated knowledge fully**. Our main contributions are summarized as follows:

- We propose a novel method for multi-teacher knowledge distillation, namely, DE-MKD, which decouples the original KD loss function and assigns sample-aware weights to the teachers based on entropy;
- In order to further improve the performance of the student network, we also use the teacher's intermediate layer features as transmitted knowledge;
- Extensive experiments on the image classification dataset CIFAR-100 validated the effectiveness and flexibility of our approach.

The rest of this paper is organized as follows. Section 2 introduces the related work on knowledge distillation and multi-teacher knowledge distillation. Section 3 demonstrates the formulation and detail of the proposed DE-MKD. Section 4 shows the experimental results on the publicly available image classification benchmark dataset CIFAR-100 across various teacher-student network pairs. Finally, a conclusion is brought forth in Section 5.

2. Related Work

Knowledge Distillation. Knowledge distillation (KD) [20,21] has attracted substantial attention as a prospective method for model compression, leveraging supervisory signals from complex teacher networks to assist in training lightweight student models. Vanilla KD [9] only transfers the soft label of a teacher to a student. FitNet [11] pioneers the concept of letting the student network imitate the intermediary layer features of the teacher. Attention-transfer-based knowledge distillation (AT) proposes to align the attention maps of teacher and student features, resulting in improved student performance. Contrastive representation distillation (CRD) [22] utilizes contrastive learning strategies to enhance the distillation effectiveness. Simple knowledge distillation (SimKD) [23] directly employs the discriminative classifier from the pre-trained teacher model for student inference, training

a student encoder via feature alignment with a single $\mathcal{L}_2$ loss. Decoupled knowledge distillation (DKD) [19] decouples the original KD loss function into target class loss and the non-target class one. Ref. [24] proposed an approach for adaptively selecting the distillation layer to enhance the distillation process. They emphasized the importance of selecting layers that adapt to the training samples and placed greater emphasis on the position of distillation. Ref. [25] proposed a self-supervised knowledge distillation (KD) method for complementary label learning. This method utilizes self-supervised learning as an auxiliary tool to enhance the quality of knowledge. However, previous distillation methods relied on a single pre-trained network, while our approach differentiates by attempting to extract knowledge from multiple teachers.

Multi-Teacher Knowledge Distillation. Multi-teacher knowledge distillation (MKD) comes from the idea that the wisdom of the crowd is greater than the wisdom of the smartest individual. MKD aims to leverage the valuable and diverse knowledge presented by multiple teacher networks to improve the performance of the student network. Various MKD methods have been proposed. These approaches [13–15] treat each teacher equally and assign equal weight to each teacher. Ref. [26] introduced a progressive training approach for knowledge fusion. Initially, it enables the student model to mimic the feature representations of multiple teachers. Subsequently, it hierarchically learns the weight parameter information from these teachers and ultimately trains a compact model capable of multitasking. RLKD [27] uses reinforcement learning to filter out unsuitable teachers and then averages the predicted logits of the remaining teachers. However, these approaches fail to capture the importance of diverse teachers. To fully integrate multi-teacher knowledge, AMTML-KD [16] is a proposed adaptive multi-level knowledge distillation technique. This method is distinguishable by assigning weights to individual teachers. Nonetheless, it does not alleviate the student's challenge in learning from the teachers; it merely enhances the integration of knowledge provided by multiple teachers. EBKD [28] assigns weights to teachers based on their predicted logits' information entropy. AEKD [17] explores the diversity of multiple teachers from the gradient space. Ref. [29] suggested a collaborative teaching KD (CTKD) method. One teacher imparts high-accuracy knowledge, while another furnishes intermediate spatial attention knowledge. CA-MKD [18] distinguishes the teacher's importance based on the teacher prediction confidence, which is calculated by the cross-entropy of the prediction logits with the ground-truth labels. Due to the coupled knowledge distillation loss, these methods are insufficient in utilizing integrated knowledge.

3. Method

In this section, we introduce our proposed Decoupled Multi-Teacher Knowledge Distillation based on Entropy (DE-MKD). We describe the method in detail, mainly by presenting the components of its loss function, which consists of decoupled logit distillation loss, intermediate feature distillation loss, and classification loss with the ground-truth label. Figure 1 is the illustration of our proposed method.

Notations. We define $D = \{x_i, y_i\}_i^N$ as the labeled training set, where N denotes the number of samples, and K denotes the number of teachers. Let F be the feature output of the second network block, where F is a tensor with dimensions $h \times w \times c$. We represent the logits' output as $z = [z_1, ..., z_C]$, where C is the number of categories. The final prediction of the model is obtained using a softmax function $\sigma(z^c) = \frac{exp(z^c/\tau)}{\sum_j exp(z^j/\tau)}$ with a temperature parameter τ.

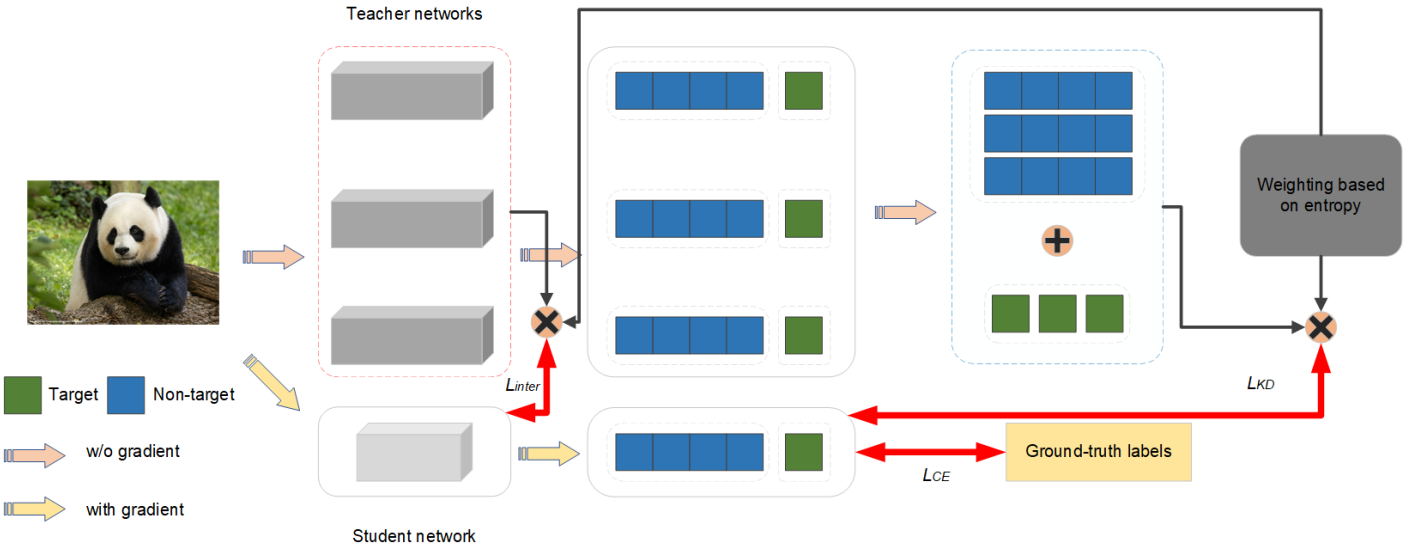

Figure 1. An illustration of the proposed multi-teacher KD method. The loss function consists of three parts: the loss of intermediate layer features, the loss of decoupled prediction, and the label loss. The red dotted box represents feature knowledge distillation. In our proposed method, we mainly use the features of the teacher's second network block as transferred knowledge. The blue dotted box denotes multi-teacher decoupled knowledge distillation.

3.1. The Loss of Decoupled Logit

To further investigate the workings of knowledge distillation (KD), decoupled knowledge distillation (DKD) [19] introduces a modified distillation loss function, which comprises a weighted sum of two components. The first component transfers knowledge concerning the "difficulty" of training samples and reflects the similarity of the target class prediction distribution of the student and teacher models, which is named target class knowledge distillation (TCKD). The second component, non-target class knowledge distillation (NCKD), is associated with the non-target classes and captures the similarity between the teacher and student predictions for these classes.

In vanilla KD, the loss function typically utilizes KL divergence. However, with the redefinition, the distillation loss function is transformed into the following form. Due to limited pages, you can refer to (DKD) [19] for specific derivation details.

$$\mathcal{L}_{KD} = TCKD + (1 - p_t^T)NCKD. \tag{1}$$

Here, p_t^T represents the prediction probability of the target class t of the teacher model. Notably, the weight of NCKD is coupled with p_t^T in this formulation.

To mitigate the suppressive impact on NCKD and enhance its significance in the loss function, DKD [19] reformulates the vanilla KD loss function as follows:

$$\mathcal{L}_{KD} = aTCKD + bNCKD. \tag{2}$$

The importance of TCKD and NCKD can be balanced by adjusting the weights a and b.

In multi-teacher knowledge distillation (MKD), distilled knowledge comes from multiple teachers. Intuitively, all knowledge should have a different weight corresponding to a different teacher. To be able to integrate all knowledge and make full use of it perfectly, we assign different weights to each teacher prediction logit based on the entropy of the teacher prediction. The greater the entropy, the lower the teacher's prediction confidence, the smaller the weight assigned, and vice versa. The weight is calculated as follows:

$$w_i = 1 - \frac{H(p^{T_t})}{\sum_{i=1}^{K} H(p^{T_i})}, \tag{3}$$

where w_i denotes the teacher, i denotes the importance weight, and $H(\cdot)$ represents the entropy of the teacher prediction.

In our proposed method, the calculation of the loss of decoupled multi-teacher KD is as follows:

$$\mathcal{L}_{KD} = \sum_{i=1}^{K} w_i(aTCKD + bNCKD). \tag{4}$$

3.2. The Loss of Intermediate Features

Features are widely recognized to have a significant impact on representation learning. FitNets [11] have also been experimentally verified to improve the performance of the student model in knowledge distillation by using the features from the teacher's hidden layer as transferred knowledge. After the introduction of FitNets, a plethora of derivative methods has emerged. Feature distillation allows students to imitate the intermediate features of teachers. Almost all existing feature distillation methods use $\mathcal{L}_2$ distance or its subtle variants as the distance metric between teacher and student features. For simplicity, in our method, we directly use the $\mathcal{L}_2$ loss function.

$$\mathcal{L}_{inter} = \sum_{i=1}^{K} w_i||F_{T_k} - r(F_s)||_2^2, \tag{5}$$

K represents the number of teachers, and w_i denotes the teacher importance weight, which is the same size as before in the prediction loss. The function $r(\cdot)$ is employed to align the dimensions of student and teacher features. In addition, in order to reduce the amount of calculation, we only use the features of the second hidden layer as distilled knowledge.

3.3. The Overall Loss

We describe the distillation process of our method in Algorithm 1 to make it easier to understand. In addition to the aforementioned two losses, a regular cross-entropy loss between the ground-truth labels and student prediction is calculated,

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C} y^c log(\sigma(z_s^c)). \tag{6}$$

The overall loss function of the proposed DE-MKD is given as

$$\mathcal{L}_{overall} = \gamma\mathcal{L}_{CE} + \alpha\mathcal{L}_{KD} + \beta\mathcal{L}_{inter}, \tag{7}$$

Here, α, β, and γ are hyperparameters that balance the effects of each loss.

Algorithm 1 Our proposed DE-MKD on CIFAR-100

Input: Training dataset $\mathcal{D}$, teacher models $\{T_1, T_2, T_3\}$, student model S
Output: Distilled student model S'
Initialize student model S
Initialize optimizer **optimizer**
for each training epoch **do**
  for each mini-batch $(X, Y) \in \mathcal{D}$ **do**
    Compute teacher predictions: $Z_i = T_i(X), i \in \{1, 2, 3\}$
    Compute student predictions: $\hat{Y} = S(X)$
    Compute distillation loss: $\mathcal{L}_{\text{KD}}$
    Compute feature loss: $\mathcal{L}_{inter}$
    Compute classification loss: $\mathcal{L}_{\text{CE}} = \text{CrossEntropy}(Y, \hat{Y})$
    Compute total loss: $\mathcal{L}_{overall} = \gamma\mathcal{L}_{CE} + \alpha\mathcal{L}_{KD} + \beta\mathcal{L}_{inter}$
    Update student model parameters: $\theta_S = \theta_S - \text{optimizer}(\nabla_{\theta_S}\mathcal{L}_{\text{total}})$
  end for
end for
return Distilled student model S'

4. Experiments

4.1. Dataset and Details

Dataset. We conducted the experiments on CIFAR-100 [30], which is a popular publicly available image classification dataset containing 32 × 32 images in 100 categories. Training and validation sets are composed of 50k and 10k images, respectively.

Implementation Details. We adopted a stochastic gradient descent (SGD) optimizer with 0.9 Nesterov momentum for all teacher-student pairs. The total training epoch was set to 240, and the learning rate was divided by 10 at the 150th, 180th, and 210th epochs. The initial learning rate was set to 0.01 for MobileNet/ShuffleNet [4,31,32] series architectures and 0.05 for other architectures [2,33,34]. The mini-batch size was set to 64, and the weight decay was set to 5×10^{-4}. The temperature τ in the KD loss was set to 4, and the α, β, γ, a, and b were set to 1, 100, 1, 1, and 8, respectively.

We compared the proposed DE-MKD with five methods. They were AVER-KD [9], AVER-FitNet [11], EBKD [28], AEKD [17], and CA-MKD [18]. In this paper, we used top-1 accuracy to evaluate all methods and employed 3 teacher models by default in all experiments except for special declarations. All results are reported as the mean and standard deviation of 3 runs using different random seeds. AVER-KD [9] represents multi-teacher KD based on logit distillation, and teachers are assigned the same weight. AVER-FitNet [11] is a multi-teacher version of feature-based knowledge distillation, where the same teachers are treated equally. EBKD [28] refers to entropy-based multi-teacher knowledge distillation, which only considers logit distillation and does not consider feature distillation. AEKD [17] is also based on logit knowledge distillation. CA-MKD [18] considers logit and feature distillation and adaptively assigns different weights to each teacher with the help of ground-truth labels.

4.2. Main Results

Results on teacher–student pairs have similar architectures. Table 1 presents a comparison of top-one accuracy among various methods using the CIFAR-100 dataset. The teacher and student models share similar architectures, such as VGG13-VGG8 and ResNet32x4-ResNet8x4 pairs. Significantly, our proposed method, DE-MKD, outperformed all the compared methods in these two teacher–student pairs. Particularly, when compared to the second-best method (CA-MKD), DE-MKD demonstrated a remarkable improvement in accuracy of 1.09% absolute accuracy in the best-case scenario.

Table 1. The top-1 test accuracy (%) of diverse multi-teacher knowledge distillation methods on the CIFAR-100 dataset utilizing teacher-student pairs with similar architectures. The teachers possess identical architectures but exhibit distinct initializations. *Note that the bold indicates the best performance, and the underline represents the second best.*

Teacher	VGG13 74.89 ± 0.18	ResNet32x4 79.45 ± 0.19
Student	VGG8 70.70 ± 0.26	ResNet8x4 72.97 ± 0.22
AVER-KD [9]	74.08 ± 0.09	75.01 ± 0.41
AVER-FitNet [11]	73.99 ± 0.18	74.78 ± 0.04
AEKD [17]	73.90 ± 0.19	74.82 ± 0.10
EBKD [28]	73.89 ± 0.34	74.44 ± 0.33
CA-MKD [18]	<u>74.30 ± 0.24</u>	<u>75.66 ± 0.13</u>
DE-MKD	**75.25 ± 0.17**	**76.75 ± 0.13**

Results on teacher–student pairs have different architecture. We validated our method not only on teacher-student pairs with similar architectures but also on pairs with different architectures. The experimental results are shown in Table 2, where our method

is shown to have consistently outperformed the comparison methods. Notably, for the ResNet32x4-ShuffleNetV2 pair, our method achieved an absolute improvement of 1.06% over the second-best method.

Table 2. The top-1 test accuracy (%) of different multi-teacher knowledge distillation methodologies on CIFAR-100 with teacher–student pairs featuring varied architectures. Here, the teachers share identical architectures but undergo diverse initialization. *Note that the bold indicates the best performance, and the underline represents the second best.*

Teacher	WRN40-2 76.62 ± 0.17	ResNet56 73.19 ± 0.30	VGG13 74.89 ± 0.18	ResNet32x4 79.45 ± 0.19	ResNet32x4 79.45 ± 0.19
Student	ShuffleNetV2 73.07 ± 0.06	MobileNetV2 65.46 ± 0.10	MobileNetV2 65.46 ± 0.10	ShuffleNetV1 71.58 ± 0.30	VGG8 70.70 ± 0.26
AVER-KD [9]	76.98 ± 0.19	70.68 ± 0.11	68.89 ± 0.10	75.02 ± 0.25	73.51 ± 0.22
AVER-FitNet [11]	77.29 ± 0.14	70.63 ± 0.23	68.87 ± 0.06	74.75 ± 0.27	73.00 ± 0.16
AEKD [17]	77.02 ± 0.17	70.36 ± 0.19	69.07 ± 0.22	75.11 ± 0.19	73.21 ± 0.04
EBKD [28]	76.75 ± 0.13	69.89 ± 0.14	68.09 ± 0.26	74.95 ± 0.14	73.01 ± 0.01
CA-MKD [18]	<u>77.64 ± 0.19</u>	<u>71.19 ± 0.28</u>	<u>69.29 ± 0.09</u>	<u>76.37 ± 0.51</u>	<u>75.02 ± 0.12</u>
DE-MKD	**78.86 ± 0.15**	**71.48 ± 0.23**	**70.05 ± 0.16**	**77.43 ± 0.16**	**75.39 ± 0.20**

Results on teachers with different architectures. The experiment conducted above-involved teacher networks in each teacher-student pair with identical architecture. In order to assess the flexibility of our approach, we employed disparate teacher networks across the teacher-student pairs. Specifically, we opted for ResNet8x4, ResNet20x4, and ResNet32x4 as the teacher combination networks, while VGG8 served as the student network. The comparative results of top-one accuracy are presented in Table 3, which further highlights the superiority of our method in relation to other compared methods.

Table 3. The top-1 test accuracy (%) of various multi-teacher knowledge distillation methodologies on CIFAR-100 where teachers possess diverse architectures. *Note that the bold indicates the best performance and the underline represents the second best.*

Teacher	ResNet8x4 72.69	ResNet20x4 78.28	ResNet32x4 79.31
Student		VGG8 70.70 ± 0.26	
AVER-KD [9]		74.53 ± 0.17	
AVER-FitNet [11]		74.38 ± 0.23	
AEKD [17]		74.75 ± 0.21	
EBKD [28]		74.27 ± 0.14	
CA-MKD [18]		<u>75.21 ± 0.16</u>	
DE-MKD		**75.56 ± 0.17**	

4.3. Ablation Study

The loss function of the proposed DE-MKD method mainly consists of three parts: the loss with the ground-truth label, the decoupling loss with the teacher's prediction, and the matching loss with the teacher's intermediate features. In order to explore the impact of each part on the performance of DE-MKD, we performed ablation experiments on the following four variants: (1) Variant A, which just uses $\mathcal{L}_{ce}$ and which refers to normal training from scratch; (2) Variant B, which represents the original multi-teacher knowledge distillation while treating teachers equally. $\mathcal{L}_{okd}$ refers to the KD loss function that is not decoupled; (3) Variant C, which represents decoupled multi-teacher knowledge distillation

but does not consider feature distillation; (4) Variant D, which represents our proposed method DE-MKD.

Table 4 presents the results of the ablation experiments, clearly showing the superior performance of our proposed new method compared to all other variants. Additionally, the improvement in accuracy for each variant demonstrates the effectiveness of each component in our method. Particularly, a comparison of Variant A with other variants demonstrates the effectiveness of knowledge distillation. The comparison between Variant B and Variant C shows the effectiveness of our proposed entropy-based decoupled multi-teacher distillation strategy. The comparison between Variant C and Variant D proves that the intermediate layer features have a positive effect on performance.

Table 4. Ablation study with ResNet32-VGG8 pair on CIFAR-100. *Note that the bold indicates the best performance.*

Variants	$\mathcal{L}_{ce}$	$\mathcal{L}_{okd}$	$\mathcal{L}_{dkd}$	$\mathcal{L}_{inter}$	Top-1
A	✔	✘	✘	✘	70.70 ± 0.26
B	✔	✔	✘	✘	73.51 ± 0.22
C	✔	✘	✔	✘	75.10 ± 0.18
D	✔	✘	✔	✔	$\mathbf{75.39 \pm 0.20}$

5. Conclusions

In this paper, we proposed a novel multi-teacher knowledge distillation method called DE-MKD. Our method integrates teacher-distilled knowledge by decoupling the loss function of traditional knowledge distillation and adaptively assigns teacher importance weights based on the entropy of teacher predictions. To further enhance the performance of student networks, we extended our method to intermediate features. We conducted extensive experiments using various teacher–student pairs to validate the effectiveness and flexibility of our method.

However, this method has some limitations. For instance, in our experiments, we only utilized the second layer features to simplify the calculations. In our future work, we plan to embrace more comprehensive feature knowledge. Additionally, self-distillation [35] would have a notable contribution. We believe that the integration of students' reflections into multi-teacher knowledge distillation could further enhance students' performance.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available because the associate code is still under development.

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

References

1. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]
2. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

3. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
4. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
5. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef]
6. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
7. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]
8. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
9. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
10. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge distillation: A survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]
11. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv* **2014**, arXiv:1412.6550.
12. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* **2016**, arXiv:1612.03928.
13. You, S.; Xu, C.; Xu, C.; Tao, D. Learning from multiple teacher networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Alifax, NS, Canada, 13–17 August 2017; pp. 1285–1294.
14. Fukuda, T.; Suzuki, M.; Kurata, G.; Thomas, S.; Cui, J.; Ramabhadran, B. Efficient Knowledge Distillation from an Ensemble of Teachers. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 3697–3701.
15. Wu, M.-C.; Chiu, C.-T.; Wu, K.-H. Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2202–2206.
16. Liu, Y.; Zhang, W.; Wang, J. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing* **2020**, *415*, 106–113. [CrossRef]
17. Du, S.; You, S.; Li, X.; Wu, J.; Wang, F.; Qian, C.; Zhang, C. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12345–12355.
18. Zhang, H.; Chen, D.; Wang, C. Confidence-aware multi-teacher knowledge distillation. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 23–27 May 2022; pp. 4498–4502.
19. 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.
20. Tang, J.; Liu, M.; Jiang, N.; Cai, H.; Yu, W.; Zhou, J. Data-free network pruning for model compression. In Proceedings of the 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Republic of Korea, 22–28 May 2021; pp. 1–5.
21. Tang, J.; Chen, S.; Niu, G.; Sugiyama, M.; Gong, C. Distribution Shift Matters for Knowledge Distillation with Webly Collected Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 17470–17480.
22. Tian, Y.; Krishnan, D.; Isola, P. Contrastive representation distillation. *arXiv* **2019**, arXiv:1910.10699.
23. Chen, D.; Mei, J.P.; Zhang, H.; Wang, C.; Feng, Y.; Chen, C. Knowledge distillation with the reused teacher classifier. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11933–11942.
24. Song, J.; Chen, Y.; Ye, J.; Song, M. Spot-adaptive knowledge distillation. *IEEE Trans. Image Process.* **2022**, *31*, 3359–3370. [CrossRef] [PubMed]
25. Liu, J.; Li, B.; Lei, M.; Shi, Y. Self-supervised knowledge distillation for complementary label learning. *Neural Netw.* **2022**, *155*, 318–327. [CrossRef] [PubMed]
26. Shen, C.; Wang, X.; Song, J.; Sun, L.; Song, M. Amalgamating knowledge towards comprehensive classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 29–31 January 2019; Volume 33, pp. 3068–3075.
27. Yuan, F.; Shou, L.; Pei, J.; Lin, W.; Gong, M.; Fu, Y.; Jiang, D. Reinforced multi-teacher selection for knowledge distillation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 14284–14291.
28. Kwon, K.; Na, H.; Lee, H.; Kim, N.S. Adaptive knowledge distillation based on entropy. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7409–7413.
29. Zhao, H.; Sun, X.; Dong, J.; Chen, C.; Dong, Z. Highlight every step: Knowledge distillation via collaborative teaching. *IEEE Trans. Cybern.* **2020**, *52*, 2070–2081. [CrossRef] [PubMed]
30. Krizhevsky, A.; Hinton, G. *Learning Multiple Layers of Features from Tiny Images*; Technical Report; University of Toronto: Toronto, ON, Canada, 2009.

31. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
32. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
33. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
34. Zagoruyko, S.; Komodakis, N. Wide residual networks. *arXiv* **2016**, arXiv:1605.07146.
35. 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 IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3713–3722.