Focusing on Significant Guidance: Preliminary Knowledge Guided Distillation[⋆]

Qizhi Cao[0009−0007−3831−9358], Kaibing Zhang[0000−0002−6983−9854], Dinghua Xue[2222−−3333−4444−5555], and Zhouqiang Zhang[2222−−3333−4444−5555]

Xi'an Polytechnic University, Xi'an 710048, China

Abstract. Feature-based knowledge distillation has been recognized a remarkably effective way to transfer informative knowledge from a complicated teacher model to a simple student model. However, for most knowledge distillation methods, the teacher model merely regards the feature knowledge as a supervisory information but neglects its guidance to the student model, leading to large gap between the feature knowledge of the teacher and that of the student. To overcome this weakness, we propose a novel preliminary knowledge guided distillation that incorporates the layer-level features from the teacher as prior knowledge to guide the student to generate the guided features. The guided features can narrow the difference between the teacher knowledge and the student knowledge. Furthermore, to enhance the quality of teacher features, a Multi-Level Feature Fusion module is employed to integrate the rich context of the teacher features across different levels, which benefit to more comprehensive exploration of teacher features. We validate the superiority of our approach by performing experiments on three different tasks, i.e., image classification on CIFAR-100 and Tiny ImageNet datasets, object detection and instance segmentation on MS-COCO dataset, respectively, indicating more competitive performance than other typical approaches. The code will be available at https://github.com/CaoQiZhi/PKGD

Keywords: Knowledge Distillation · Prior Guidance · Feature Fusion.

1 Introduction

Neural networks have achieved tremendous success in computer vision, e.g., Image Classification [1,2], Object Detection [3,4], and Instance Segmentation [5,6]. To achieve superior performance, larger and deeper neural network models are universally employed with a large number of parameters, which imposes increasing burdens on hardware. Facing to the issue, many model lightweighting methods have been proposed, e.g., network pruning [7], parameter quantization [8].

[⋆] Supported by the Key Research and Development Program of Shaanxi Province under Grant 2024GX-ZDCYL-02-07 and the National Natural Science Foundation of China under Grant 61971339.

Besides, knowledge distillation [9] is also a popular and effective method, whose core idea is to train a smaller neural network under the supervision of a well-performing neural network to achieve excellent performance.

In knowledge distillation, feature-based knowledge distillation [10, 11] is always a class of methods with great concern. Taking object detection as an example, it involves two downstream tasks: bounding box regression and object classification, both of which share the same input features. This also leads us to the mind that high-quality feature extraction is crucial for the performance of neural network models. Hence, feature knowledge should be one of the primary focuses in knowledge distillation.

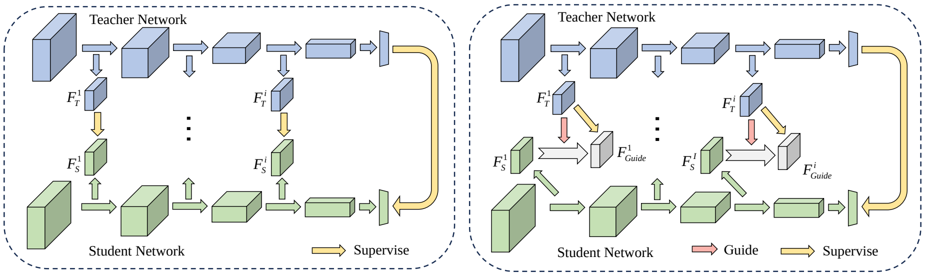

Fig. 1. Comparison between common distillation and our approach. The image in left hand represents common distillation and the right one represents our method.

Recently, the Fitnet [12] proposes to migrate relevant features between the teacher model and the student model, and AT [13] further transmits attention maps of corresponding layers. Consequently, Zhang et al. [14], achieve selective transfer of foreground and background features through attention guidance, while FGD [15] divides features into background and foreground and distributes different weights. Whereas, these methods only distill feature knowledge among features at the same level, ignoring the beneficial information contained in cross-level features, which seriously limits the improvement of distillation efficiency. More importantly, previous methods only take the feature knowledge of the teacher as supervision to assist the student. However, just like the learning process of students, the teachers should not only supervise the answers of the students, but also appropriately guide the students in problem-solving way.

To alleviate the shortcomings, we propose a novel feature-based knowledge distillation framework called Preliminary Knowledge Guided Distillation. In our approach, teacher models no longer directly supervise student models' features, but guide the student model to generate guided features through a Dual Dimension-Aware Guidance module before feature distillation (seeing Fig. 1). In the process, teacher models provide preliminary knowledge to guide student models before they learn complex knowledge, thereby reducing the difficulty of learning and effectively improving learning efficiency. Additionally, we introduce a Multi-Level Feature Fusion module to extract multi-level features from teacher models by bidirectional progressive fusion, which more fully exploiting the information in teacher features. In summary, our main contributions are threefold:

1) We propose a novel and effective distillation framework with Dual Dimension-Aware Guidance module, addressing the issue of missing prior guidance in the training process of the student model.
2) We introduce a Multi-Level Feature Fusion module to fully exploit the feature knowledge of the teacher, which covers the short of the insufficient utilization for cross-layer features of the teacher.
3) We validate our method on Cifar-100, ImageNet, and MS-COCO datasets, achieving competitive performance, and extend it from image classification to object detection and instance segmentation tasks.

2 Related Works

2.1 Feature-based Knowledge Distillation

In previous feature-based knowledge distillation, AT [13] transfers feature attention maps from corresponding layers, while FSP [16] generates a FSP matrix from layer features and utilizes it to supervise the student model. These methods only consider the feature transfer between the same levels, but rich information is also contained in the features from different levels. Whereupon, Chen et al. [17], propose knowledge review, which utilizes the teacher's shallow features to guide the student's deep features. However, it obviously ignores the useful infromation in the teacher's deep-layer features. To leverage these knowledge, Cao et al. [18], propose knowledge preview based on knowledge review. Knowledge preview progressively integrates shallow-layer knowledge into deep layers to obtain highly informative feature representations. Building upon the idea, we employ a Multi-level Feature Fusion module in our method, which effectively solves the current problem on insufficient use of teacher features.

2.2 Knowledge Guidance

'Supervision' and 'Guidance' are distinct concepts. From a physiological perspective, 'Supervision' can be understood as the teacher's correction and revision for the student's answers, with a greater emphasis on the outcome. On the other hand, "guidance" is more akin to the teacher's instruction during the student's completion of assignments, focusing more on the process.

Utilizing the ground truth to supervise the final output of the models is the core idea of current neural network training. Later, model guidance has also been proven effective in enhancing model performance. Wu et al. [19], propose a Semantic-Aware Embedding Module that utilizes the feature knowledge of segmentation networks to guide low-light enhancement networks. Nie et al. [20], distribute model branches to generate guidance to assist in training of the diagnostic classification model. However, in knowledge distillation, the knowledge of the teacher is treated purely as the supervisory signal, while the guidance role of the teacher is always ignored. Therefore, we propose a Dual Dimension-Aware Guidance module to compensate for the absence of teacher guidance during student training.

3 The Proposed Method

The overall framework of our approach is depicted in Fig. 2. We will explain it from three parts: Multi-Level Feature Fusion, Dual Dimension-Aware Guidance, and Optimization Function. The details will be described in subsequent sections.

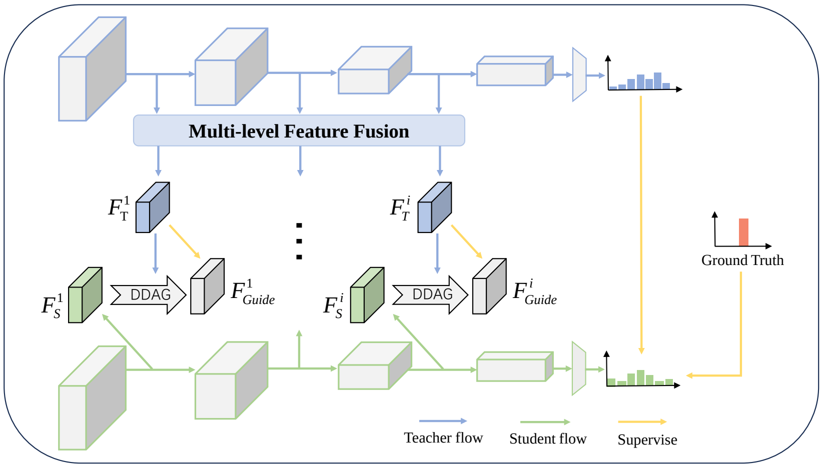

Fig. 2. The overall framework. Blue arrows represent the flow of teacher features; green arrows represent the flow of student features; yellow arrows represent supervision and the 'DDAG' represents the Dual Dimension-Aware Guidance module.

3.1 Multi-Level Feature Fusion

Unlike knowledge review and preview distillation [18], we utilize an attention-based fusion module in the multi-level feature fusion process to progressively fuse the intermediate features of the teacher model in both forward and backward directions. We also refine the attention-based fusion module to better preserve the original feature information, minimizing the miss of original information during multiple fusion processes. The specific process is as follows:

Attention Based Fusion Assuming F^a and F^b to be features from adjacent layers, as illustrated in Fig. 3. we firstly align features by convolution and interpolation. This process can be represented as:

$$F_{trans} = G(F^b), \tag{1}$$

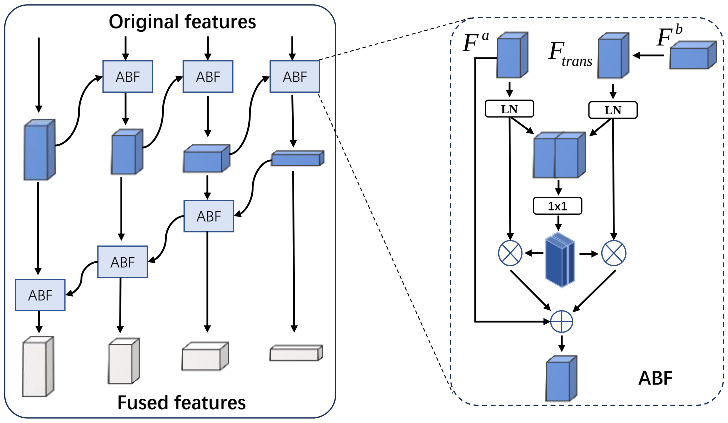

Fig. 3. Multi-Level Feature Fusion. The details of attention-based fusion module are depicted in the dashed box.

we use $G(\cdot)$ to represent the convolution and interpolation operations, F^b for features from the upper layer or the lower layer. The transformed feature representation F_{trans} maintains consistency in dimensions with the features F^a. Then, layer normalization is applied to the features, which can be represented as follows:

$$F_{LN} = LN(F), \tag{2}$$

where $LN(\cdot)$ represents the layer normalization process. After that, a 1×1 convolution is used to obtain the attention maps corresponding to features, represented as:

$$F_{Atten} = Conv(F_{LN}), \tag{3}$$

where F_{Atten} represents the attention map and $Conv(\cdot)$ denotes the 1×1 convolution operation. Then, the features and their corresponding attention maps are multiplied element-wise and summed to obtain the fused features. Finally, the fusion module can be as follows:

$$F_{ABF}(F_a, F_b) = F^a_{Atten} \odot F^a_{LN} + F^b_{Atten} \odot F^b_{LN} + F^a, \tag{4}$$

where F^a_{Atten} and F^b_{Atten} respectly represent the attention maps corresponding to the features F_a and F_b, while F^a_{LN} and F^b_{LN} represent the features after layer normalization. The '$\odot$' represents the element-wise multiplication.

Compared to the ABF module in RP KD [18], we introduce residual structure to maintain the original information of corresponding layer feature and add layer normalization to better avoid the overdue miss of feature information in multiple fusion processes.

Bidirectional Progressive Passing During the feature passing, we primarily obtain the first-order fused features by fusing the original features of the teacher model. Let's define $F_i \circ F_{i+1} = F_{ABF}(F_i, F_{i+1})$, the forward progressive fusion can be represented as:

$$\dot{F}_i = F_i \circ F_{i-1} \cdots \circ F_1, \tag{5}$$

where F_i represents the i-th layer original feature and $\dot{F}_i$ represents the first-order fused feature at i-th layer. The first-order fused features contain information from shallower layers than the i-th layer. Similarly, the backward progressive fusion can be represented as:

$$\ddot{F}_i = \dot{F}_i \circ \dot{F}_{i+1} \cdots \circ \dot{F}_n, \tag{6}$$

where $\ddot{F}_i$ represents the second-order fused feature. The second-order fused features contain information from each layer.

The Multi-Level Feature Fusion module is shown in Fig. 3. In contrast to the approach RP KD [18] that knowledge review and preview are performed on student features, we employ progressive forward and backward propagation of teacher features to obtain more informative second-order fusion features. Then, the second-order fusion features are utilized as prior knowledge to guide student features in subsequent processing.

3.2 Dual Dimension-Aware Guidance

In our method, we no longer simply regard the teacher's feature knowledge as supervision to assist in training the student. Instead, we take teacher features as prior knowledge to guide the student and generate guidance features, which is obtained by the Dual Dimension-Aware Guidance (seeing Fig. 4). The details are as follow:

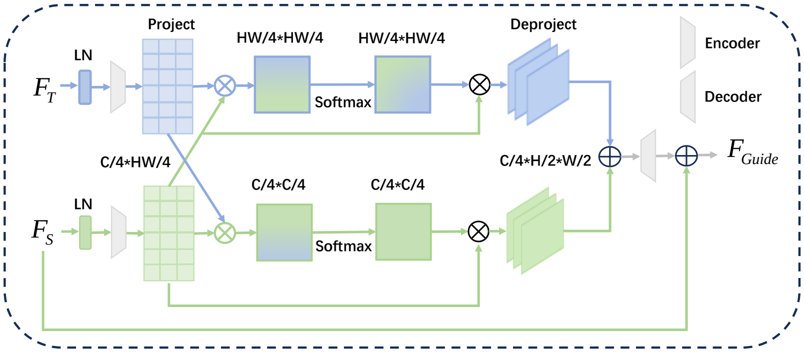

Fig. 4. The details of Dual Dimension-Aware Guidance module. The F_T represents the teacher features and the F_S represents the student features.

First of all, we perform layer normalization on the input features. Then, the normalized features are encoded to shrink the number of channels and dimensions, reducing the parameters and computational complexity and saving storage space in subsequent operations. Subsequently, the encoded features are projected into two-dimension, which can be represented as:

$$F^{Pro} = Project(Encoder(LN(F))), \tag{7}$$

where $Encoder(\cdot)$ represents the feature encoding, and $Project(\cdot)$ represents the feature projecting. Then, the transformed features are used to obtain spatial and channel-wise guideance maps through matrix multiplication and $Softmax$ function. This process can be represented as follows:

$$Map\_c = Softmax(F_T^{Pro} \times (F_S^{Pro})^T), \tag{8}$$

$$Map\_s = Softmax((F_T^{Pro})^T \times F_S^{Pro}), \tag{9}$$

where Map_c and Map_s respectively represent the generated channel and spatial attention guidance maps. Compared to common attention maps, the guidance maps incorporate student features and prior information from the teacher features, enabling better utilization of teacher knowledge to guide the student. The guidance features in spatial and channel can be obtained in line with the attention guidance maps, which can be represented as:

$$F_C^G = Map\_c \times F_S^{Pro}, \tag{10}$$

$$F_S^G = F_S^{Pro} \times Map\_s, \tag{11}$$

where F_C^G and F_S^G represent channel-guided features and spatial-guided features respectively. Then, they are projected into three-dimension and pointwise added. The process is illustrated as follows:

$$F^{De} = Deproject(F_C^G) + Deproject(F_S^G), \tag{12}$$

where $Deproject(\cdot)$ represents the feature inverse mapping operation, F^{De} represents the sum of the inverse mapped spatial and channel-guided features. After that, the features are fed into the decoder and the decoded features are then added to the original student features to obtain the guided features F_{Guide}, which can be described as follows:

$$F_{Guide} = Decoder(F^{De}) + F_S. \tag{13}$$

Finally, we utilize teacher features to supervise the guided features, and take HCL to obtain the feature knowledge distillation loss, represented as follows:

$$L_{feat} = HCL(F_T, F_{Guide}), \tag{14}$$

where $HCL(\cdot)$ represents the hierarchical context loss, which is the same as the $HCL(\cdot)$ function used in Review KD [17].

3.3 Optimization Function

In addition to features knowledge, we found, as demonstrated in method [21], that prior feature guidance is highly compatible with vanilla KD [22] targeting the logits knowledge, and we further confirmed it in subsequent ablation experiments. Therefore, the overall loss function is as follows:

$$Loss_{All} = \alpha Loss_{feat} + Loss_{KD}, \tag{15}$$

where $Loss_{KD}$ represents the vanillat KD loss and α is a coefficient for balancing the different loss functions.

4 Experiment

We conduct experiments on different tasks to validate our approach, including image classification on the Cifar-100 [23] and Tiny ImageNet [24] datasets, as well as object detection and instance segmentation on the MS-COCO [25] dataset. Additionally, we performe ablation experiments for further validation and analysis, and analyzed the advantages of guided features by feature visualization comparison.

Datesets

1) The CIFAR-100 dataset consists of 100 classes, with each class containing 600 images of size 32 × 32 pixels. Among these, 500 images are allocated for training and 100 images for testing. In total, there are 50,000 training images and 10,000 test images.
2) The Tiny ImageNet contains 100000 images of 200 classes downsized to 64×64 pixels. Each class has 500 training images, 50 validation images and 50 test images.
3) The MS-COCO dataset is a large-scale dataset used for computer vision tasks such as object detection, semantic segmentation. The dataset consists of 80 object categories, with the training set comprising over 118k images and the test set containing 5k images.

4.1 Image Classification

Implementation Details In our experiments, a standard data augmentation scheme is used, including random cropping, flipping, and padding. For the CIFAR-100 dataset, we set the initial learning rate to 0.02 for MobileNet [26] and ShuffleNet [27, 28], while other models are set to 0.1. Additionally, the learning rate decay by 0.1 times every 30 epochs after 150 epochs. All models are trained for 240 epochs with a batch size of 128. For the TinyImageNet dataset, all models are trained for 300 epochs with a batch size of 128. We set the initial learning rate to 0.1, and it is decayed by 0.2 times at epochs 90, 150, 210, and 270.

Table 1. Top-1 Accuracy Rate (%) of Various Methods on both Peer-Architecture.

Knowledge Type	Method	T: ResNet56	ResNet110	ResNet32x4	WRN40-2	WRN40-2	Vgg13
		73.08	74.42	79.42	76.48	76.48	74.65
		S: ResNet20	ResNet32	ResNet8x4	WRN16-2	WRN40-1	Vgg8
		69.14	71.14	72.32	73.42	71.98	70.63
Logits	Vanilla KD [22]	71.08	70.84	73.08	73.25	74.77	73.54
	Decouple KD [31]	71.94	71.49	74.11	76.32	76.24	74.81
Features	AT [13]	71.57	73.11	73.36	75.23	72.77	73.07
	FT [29]	71.52	73.37	73.64	75.10	73.04	73.14
	CRD [11]	71.62	73.57	75.44	75.51	74.14	74.11
	Review KD [17]	71.89	73.89	75.53	76.01	75.09	74.57
	SAKD [10]	71.79	73.42	75.60	75.85	74.67	74.29
	ICKD-C [31]	71.75	73.72	74.80	75.58	74.32	73.88
	RP KD [18]	72.21	73.99	75.74	76.05	75.11	74.59
	Ours	**72.49**	**74.26**	**77.21**	**76.27**	**75.27**	**75.18**

Results and analysis We compare our method with representative feature-based distillation methods, including AT [13], FT [29], CRD [11], Review KD [17], SAKD [10], ICKD-C [30], and RP KD [18] as well as typical logits-based distillation methods, Vanilla KD [22] and Decouple KD [31]. We evaluate the model performance under different methods by accuracy and the results on Cifar-100 are presented in Table 1. It can be observed that our method consistently outperforms the baseline by more than 2% on various teacher-student pairs. Particularly, the improvement in accuracy reaches to 4.89% for the ResNet32x4-ResNet8x4 pair and 4.55% for the Vgg13-Vgg8 pair. Moreover, our method exhibits significant superiority over other both feature-based and logits-based knowledge distillation methods in the comparison.

To further validate the effectiveness of our method, we conduct more challenging experiments on heterogeneous teacher-student pairs, as shown in Table 2. It is evident that our method continues to exhibit strong performance, surpassing other methods and achieving improvements of over 5% in accuracy compared to the baseline model. Noteworthy is the accuracy of MobileNetV2 exceeding 70% under our method and the accuracy of the student model even surpasses that of the teacher model in the WRN40-2-ShuffleNetV1 structure.

Additionally, we validate our method on the more abundant Tiny ImageNet dataset. We conduct experiments on both homogeneous (ResNet34-ResNet18) and heterogeneous (ResNet50-MobileNetV1) teacher-student pairs, and the results are illustrated in Fig. 5. We found that compared to logits-based distillation methods, feature-based distillation methods usually require longer training times but tend to yield better results. Although our method incurs higher computational costs, it consistently outperforms other methods in terms of model performance improvement.

4.2 Extended Tasks

In addition to image classification, we also apply our method to object detection and instance segmentation tasks. We follow the experimental setup described in

Table 2. Top-1 Accuracy Rate (%) of Various Methods on both Cross-Architecture.

Knowledge Type	Method	T: ResNet32x4	ResNet32x4	WRN40-2	Vgg13	ResNet50
		79.42	79.42	76.48	74.65	79.34
		S: ShuffleNetV1	ShuffleNetV2	ShuffleNetV1	MobileNetV2	MobileNetV2
		70.47	71.75	70.47	64.62	64.62
Logits	Vanilla KD [22]	72.73	74.11	74.36	74.69	67.41
	Decouple KD [31]	76.39	76.94	76.49	69.75	70.32
Features	AT [13]	73.62	73.57	73.80	60.80	59.72
	FT [29]	72.58	72.73	72.21	62.03	61.84
	CRD [11]	75.14	75.60	75.67	69.71	69.11
	Review KD [17]	77.04	77.50	77.14	69.91	69.82
	SAKD [10]	75.33	75.82	75.92	69.88	69.75
	ICKD-C [31]	74.69	74.56	74.63	67.89	67.80
	RP KD [18]	77.09	77.01	77.23	69.42	69.17
	Ours	**77.49**	**77.26**	**77.58**	**70.27**	**70.18**

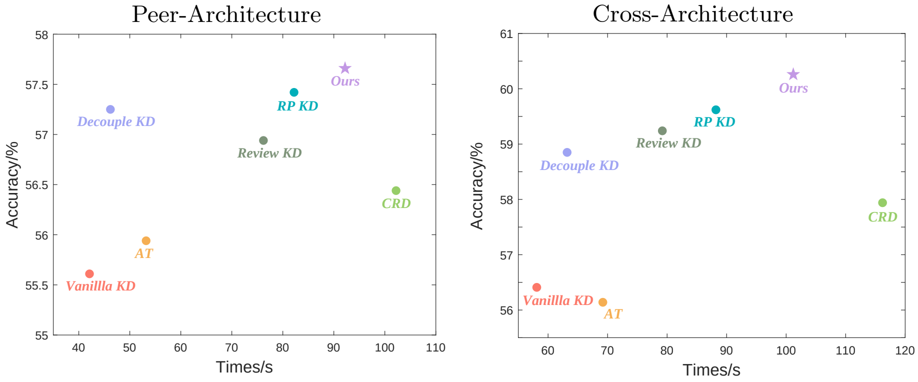

Fig. 5. Accuracy *vs.* Times on Tiny ImageNet.The teacher-student pair in the left image is ResNet34-ResNet18; That in the right image is ResNet50-MobileNetV1.

Review KD [17], and all experiments are conducted based on the Detectron2 [32] platform.

Object Detection We choose typical one-stage detector RetinaNet [3] and two-stage detector Faster RCNN [33] as the base models and distille feature knowledge from the backbone parts of the models. We compare our method with FitNet [12], FGFI [34], Review KD [17], Vanilla KD [22], and Decouple KD [31] on MS-COCO dataset and the results are shown in Table 3. It can be observed that our method outperforms other methods significantly in AP_{50}. Although the performance of ResNet101-RetinaNet in AP_{75} is slightly inferior, it still maintains the lead in mAP.

Instance Segmentation In instance segmentation task, we select Mask R-CNN [35] as the base model for distilling feature knowledge between different backbone networks. As far as we know, Review KD [17] is the first method to

Table 3. The results of different methods on Object Detection.

Base Model	Faster RCNN [33]						RetinaNet [3]		
Architecture	ResNet101-ResNet50			ResNet50-MoboileNetV2			ResNet101-ResNet50		
Evaluation Metric	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
Teacher	42.04	62.48	45.88	40.22	61.02	43.81	40.40	60.25	43.19
Student	37.93	58.84	41.05	29.47	48.87	30.90	36.15	56.03	38.73
Vanilla KD [22]	38.35	59.41	41.71	30.13	50.28	31.35	36.76	56.60	39.40
FitNet [12]	38.76	59.62	41.80	30.20	49.80	31.69	36.30	55.95	38.95
FGFI [34]	39.44	60.27	43.04	31.16	50.68	32.92	37.29	57.13	40.04
Review KD [17]	40.36	60.97	**44.08**	33.71	53.15	36.13	38.48	58.22	**41.46**
Decouple KD [31]	39.25	60.90	42.73	32.34	**53.77**	34.01	-	-	-
Ours	**40.62**	**61.43**	44.01	**33.92**	53.56	**36.77**	**38.57**	**58.96**	39.91

apply knowledge distillation to instance segmentation. Therefore, we compare our method with it in homogeneous and heterogeneous pairs, and evaluate model performance by mAP, AP_{50}, and AP_{75}. The results, as shown in Table 4, indicate that our method can improve the mAP of the baseline model by 2%, surpassing the performance of Review KD in all aspects.

Table 4. The comparison of Review KD and our method on Instance Segmentation.

Architecture	ResNet101-ResNet18			ResNet101-ResNet50			ResNet50-MoboileNetV2		
Evaluation Metric	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
Teacher	38.63	60.45	41.28	38.63	60.45	41.28	37.17	58.60	39.88
Student	31.25	51.07	33.10	35.24	56.32	37.49	28.37	47.19	29.95
Review KD [17]	33.62	53.91	35.96	36.98	58.13	39.60	31.56	50.70	33.44
Ours	33.86	54.43	36.44	37.33	58.86	39.77	31.70	50.96	33.91

4.3 Ablation Study

We conduct ablation experiments on the Cifar-100 dataset for image classification to observe the effects of different modules in our method by cross-combining them. We set ResNet56 as the teacher network and ResNet20 as the student network. The results are demonstrated in Table 5.

Table 5. Ablation Study on Integrating Different Modules.

Architecture	ResNet56(73.08%)-ResNet20(69.14%)			
Modules	MLFF	DDAG	VanillaKD	Accuracy(%)
Association Schemes			√	71.08
	√			72.18
		√		71.94
	√	√		72.32
	√	√	√	**72.49**

In the table, MLFF stands for Multi-Level Feature Fusion module and DDAG represents Dual Dimension-Aware Guidance module. It can be observed that each module can significantly improve the performance of the baseline model when used alone. Furthermore, when all of them are combined, the effect is the best, with an improvement of 3.35% in accuracy.

4.4 Visualization

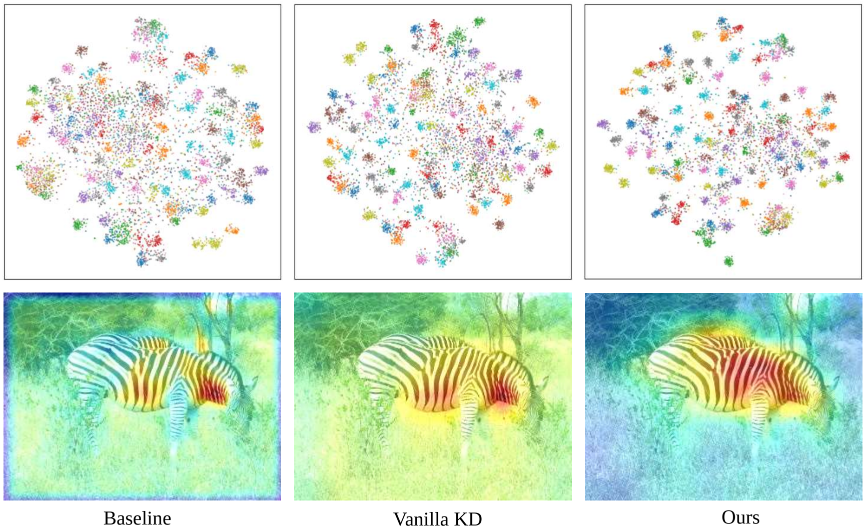

Fig. 6. Visualization comparisons of different methods. The first row depicts the visualization results of feature t-SNE and the second row is that of feature attention.

To better illustrate the enhancement effects of our method on features, we conduct the t-SNE and feature attention visualization comparison between our method and the traditional Vanilla KD with setting ResNet8x4 as the student and ResNet32x4 as the teacher, as shown in Fig. 6.

As observed from the first row, our method enhances the feature clustering capability of the baseline model. Furthermore, compared to Vanilla KD, our method also demonstrates better clustering effect on classification features. Similarly, the feature attention visualization in the second row presents consistent conclusions. The feature attention obtained by our method is evidently more accurate and dense compared to that of the baseline model and the feature attention obtained through Vanilla KD. All of the above indicates that our method has better capturing ability for distinct features.

5 Conclusion

In this paper, we propose a novel knowledge distillation method called Preliminary Knowledge Guided Distillation. Unlike previous approaches, in addition to the supervisory role of the teacher model, we exploit the guiding role of the teacher model in the training process of the student model. In our method, a Dual Dimension-Aware Guidance module is introduced to use teacher features to guide student features. Moreover, we propose a Multi-Level Feature Fusion module to utilize the beneficial information in teacher cross-layer features. Extended experiments based on Cifar-100, Tiny ImageNet, and MS-COCO are conducted on different tasks including Image Classification, Object Detection, and Instance Segmentation tasks, where our method demonstrates superior performance.

References

1. Zakariae, A., et al.: Riemannian Generalized Gaussian Distributions on the Space of SPD Matrices for Image Classification. IEEE Access **12**, 26096–26109 (2024)
2. Amna, A., et al.:Adaptive Feature Selection and Image Classification Using Manifold Learning Techniques. IEEE Access **12**, 40279–40289 (2024)
3. Tsung-Yi, L., et al.: Focal Loss for Dense Object Detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp.2999-3007. IEEE, Italy (2017)
4. Carion, N., et al.: End-to-End Object Detection with Transformers. ArXiv preprint ArXiv:2005.12872 (2020)
5. Shu, L., et al.: Path Aggregation Network for Instance Segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.8759-8768. IEEE, USA(2018)
6. Golnaz, G., et al.: Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.2917-2927. IEEE, USA (2021)
7. Yanming, C., et al.: FPC: Filter pruning via the contribution of output feature map for deep convolutional neural networks acceleration. Knowl. Based Syst. **238**, 107876 (2021)
8. Steve, D., et al.: VS-Quant: Per-vector Scaled Quantization for Accurate Low-Precision Neural Network Inference. ArXiv preprint ArXiv:2102.04503 (2021)
9. Gou, J., et al.: Knowledge Distillation: A Survey. International Journal of Computer Vision, **129**: 1789-1819 (2020)
10. Song, J., et al.: Spot-Adaptive Knowledge Distillation. IEEE Transactions on Image Processing, **31**: 3359-3370 (2022)
11. Tian, Y., et al.: Contrastive representation distillation. In: Proc. 8th Int. Conf. Learn. Representations, pp.26-30. OpenReview.net, Ethiopia (2020)
12. Adriana, R., et al.: FitNets: Hints for Thin Deep Nets. In: 3rd International Conference on Learning Representations (ICLR), USA (2015)
13. Sergey, Z., Nikos, K.: Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In: 5th International Conference on Learning Representations (ICLR), OpenReview.net, France (2017)
14. Linfeng, Z., Kaisheng, M.: Improve Object Detection with Feature-based Knowledge Distillation: Towards Accurate and Efficient Detectors. In: 9th International Conference on Learning Representations (ICLR), OpenReview.net, Austria (2021)

15. Zhendong, Y., et al.: Focal and Global Knowledge Distillation for Detectors. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.4633–4642. IEEE, USA(2022)
16. Yim, J., et al.: A gift from knowledge distillation: fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4133-4141. IEEE, USA (2017)
17. Chen, P., et al. Distilling Knowledge via Knowledge Review. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.5006-5015. IEEE, USA(2021)
18. Qizhi, C., et al.: Be an Excellent Student: Review, Preview, and Correction. IEEE Signal Process. Lett, **30**: 1722–1726 (2023)
19. Yuhui, Wu., et al.: Learning Semantic-Aware Knowledge Guidance for Low-Light Image Enhancement. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.1662–1671. IEEE, Canada (2023)
20. Weizhi, Nie., et al.: Deep reinforcement learning framework for thoracic diseases classification via prior knowledge guidance. Comput. Medical Imaging Graph, **108**: 102277 (2023)
21. Martin, Z., et al.: Better Teacher Better Student: Dynamic Prior Knowledge for Knowledge Distillation. In: The Eleventh International Conference on Learning Representations (ICLR), OpenReview.net, Rwanda (2023)
22. Hinton, G., et al.: Distilling the Knowledge in a Neural Network. Computer Science, **14**(7): 38-39 (2015)
23. Krizhevsky, A., et al.: Learning multiple layers of features from tiny images. In: Handbook of Systemic Autoimmune Diseases, pp.54-57. Elsevier, Amsterdam(2009)
24. Ya, Le., Xuan, Yang.,: Tiny imagenet visual recognition challenge. CS 231N, **7**(7): 3 (2015)
25. Lin, T., et al.: Microsoft coco: Common objects in context. In: 13th European Conference, pp.740-755. Springer International Publishing, Switzerland (2014)
26. Sandler, M., et al.: MobileNetV2: Inverted residuals and linear bottlenecks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp.4510-4520. IEEE, USA (2018)
27. Hluchyj, M., et al.: ShuffleNet: An application of generalized perfect shuffles to multihop lightwave networks. Lightw. Technol, **9**(10): 1386–1397 (1991)
28. Zhang, X., et al.: ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, pp.6848–6856. IEEE, USA (2018)
29. Kim, J., et al. Paraphrasing complex network: Network compression via factor transfer. In: Proc. Int. Conf. Neural Inf. Process. Syst., pp.2765-2774. Canada (2018)
30. Liu, L., et al.: Exploring inter-channel correlation for diversity-preserved knowledge distillation, In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp.8251-8260. IEEE, Canada (2021)
31. Zhao, B., et al.: Decoupled knowledge distillation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, pp.11943-11952. IEEE, USA (2022)
32. Abhishek, A., et al.: Detectron2 object detection & manipulating images using cartoonization. Int. J. Eng. Res. Technol.(IJERT), **10**: 1-5 (2021)
33. Shao, R., et al.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, **28** (2015)
34. Wang, T., et al.: Distilling object detectors with fine-grained feature imitation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp.4933-4942. IEEE, USA (2019)
35. Kai, He., et al.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp.2961–2969. IEEE, Italy (2017)