Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Prototype-guided cross-task knowledge distillation*

Deng LI[1], Peng LI[2], Aming WU[3], Yahong HAN[†‡1]

[1]*College of Intelligence and Computing, Tianjin University, Tianjin 300350, China*

[2]*Songshan Laboratory, Zhengzhou 450000, China*

[3]*School of Electronic Engineering, Xidian University, Xi'an 710401, China*

[†]E-mail: yahong@tju.edu.cn

Received May 12, 2024; Revision accepted Sept. 18, 2024; Crosschecked

Abstract: Recently, large-scale pretrained models have revealed their benefits in various tasks. However, due to the enormous computation complexity and storage demands, it is challenging to apply large-scale models to real scenarios. Existing knowledge distillation methods mainly require the teacher model and the student model to share the same label space, which restricts its application in the real scenario. To alleviate the constraint of different label spaces, we propose a prototype-guided cross-task knowledge distillation (ProC-KD) method to migrate the intrinsic local-level object knowledge of the teacher network to various task scenarios. First, to better learn the generalized knowledge in cross-task scenarios, we present a prototype learning module to learn the invariant intrinsic local representation of objects from the teacher network. Secondly, for diverse downstream tasks, a task-adaptive feature augmentation module is proposed to enhance the student network features with the learned generalization prototype representations and guide the learning of the student network to improve its generalization ability. The experimental results on various visual tasks demonstrate the effectiveness of our approach for cross-task knowledge distillation scenarios.

unedited

Key words: Knowledge distillation; Cross-task; Prototype learning
https://doi.org/10.1631/FITEE.2400383 **CLC number:** TP391.4

1 Introduction

Recently, the transformer network (Vaswani et al., 2017) has achieved great advances in some visual tasks, for example, image classification (Dosovitskiy et al., 2021; Liu et al., 2021; Touvron et al., 2021), object detection (Carion et al., 2020; Zhu et al., 2021; Zhou et al., 2023), image segmentation (Ye et al., 2019; Jain et al., 2023), and visual language joint learning (Chen et al., 2020; Li et al., 2020; Fu et al., 2023).

Based on the self-attention mechanism transformer networks can process complete input sequences and possess the advantage of parallelization. Therefore, these networks are usually used to obtain the pretrained model from large-scale datasets (Deng et al., 2009). Currently, fine-tuning is the common strategy for the utilization of pretrained models in cross-task learning scenarios. After learning the generalized feature representation from large-scale datasets, perform fine-tuning on the downstream task with the small dataset to boost the performance of the downstream task model. However, applying these large-scale models to practical application scenarios with limited resources (*e.g.*, mobile devices) has become a big challenge due to their enormous computation complexity and huge storage needs.

To solve the above model application issue, some model compression and acceleration technologies have been proposed, for example, parameter pruning (Molchanov et al., 2017; Zhu and Gupta, 2018), model quantization (Wu et al., 2016), and knowledge distillation (KD) (Hinton et al., 2015).

‡ Corresponding author
* Project supported by the National Natural Science Foundation of China (Nos. 62376186 and 61932009)
© Zhejiang University Press 2024

Particularly, KD is a valid approach for model compression, which transfers the knowledge from a large deep neural network into a small one (Hinton et al., 2015).

Different from other model compression approaches, KD can decreases the number of parameters and boost the performance of small models on downstream tasks, irrespective of the architectural differences between the teacher network and the student network. It has been successful in a variety of tasks, such as computer vision (Hinton et al., 2015; Romero et al., 2015; Yim et al., 2017; MÃijller et al., 2019; Park et al., 2019; Gou et al., 2021), natural language processing (Sanh et al., 2019; Sun et al., 2019; Jiao et al., 2020), and speech recognition (Chebotar and Waters, 2016; Kurata and Saon, 2020; Yoon et al., 2021).

However, these KD approaches mainly require the teacher network and the student network to perform the same task, for example, the teacher network and the student network share the same label space, which limits their application in the real scenario, such as downstream tasks in different label spaces (as shown in Fig. 1a). By transferring the knowledge from the teacher network to downstream tasks with differently labeled spaces, the cross-task KD method expands the application of the teacher model to a variety of downstream tasks. The existing same-task KD method works mainly to transfer the final prediction logits or the hidden layer knowledge, which are global-level knowledge alignments and cannot be applied to cross-task KD directly. An earlier work on cross-task KD (Ye et al., 2020) aligns the high-order comparison relationship between models in a local manner; however, this method lags in the representation power of the invariant intrinsic object and is a two-stage distillation method.

Under the scenario of cross-task KD, the intrinsic object characteristics can give beneficial guidance to train the student network, for example, the shape features of the legs of a cattle and a horse are similar while cattle belong to the teacher network dataset and the horse belongs to the student network. Considering the complexity and variability of real-scenario tasks and the generalization capability of large-scale pretrained models, we propose a prototype-guided cross-task KD (ProC-KD) approach to migrate local intrinsic knowledge of a large-scale teacher network to various task scenarios (as shown in Fig. 1b). Our method obtains the downstream-task small model with a one-stage training process that does not require fine-tuning. Specifically, our proposed approach consists of two integrated modules: the prototype-based representation learning module; and the feature augmentation module. The prototype learning module is carefully designed to capture intrinsic feature information from the hidden layer of the large teacher network. Next, we feed the learned prototype representation into the feature augmentation module. The feature augmentation module enriches the student model feature that is more related to the prototype representation, while suppressing unrelated features. To give guidance on the training of the student model with the learned generalized prototype representation, a consistency loss is designed to obtain the maximum agreement between the prototype augmented output and the student network output.

In the experiments, we first verify the efficacy of our approach on various cross-task KD tasks. Then, we evaluate our approach on standard KD tasks. The experimental results in the two scenarios show the effectiveness and versatility of our method.

Our contributions are concluded as follows:

(1) We propose a prototype-guided KD approach to migrate the intrinsic knowledge from a large-scale model to different cross-task small models without fine-tuning on the downstream task and improve the student model generalization ability.

(2) We propose a prototype learning module and a feature augmentation module to learn the invariant intrinsic knowledge from the large-scale teacher network and enhance the student network feature with the attention mechanism, respectively.

(3) We verify our method on both cross-task and same-task KD on various visual tasks. The experimental results show the validity and generality of our approach.

2 Related work

2.1 KD approach

KD compresses the mode by transferring knowledge from a larger network to a smaller one. It can be classified into three types, namely, response-based distillation (Hinton et al., 2015; MÃijller et al., 2019), feature-based distillation (Romero et al., 2015), and

unedited

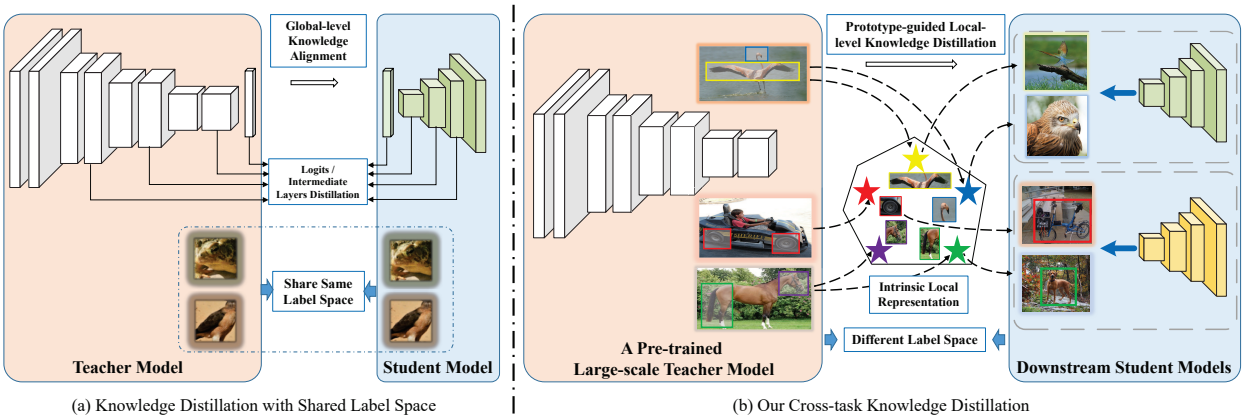


Fig. 1 Comparison between the same-task knowledge distillation and the proposed cross-task knowledge distillation methods: (a) the same-task knowledge distillation method in which the large teacher network and the small student network share the same label space; (b) the proposed prototype-guided cross-task knowledge distillation method in which the large teacher network and the small student network have different label spaces.

relation-based distillation (Yim et al., 2017; Park et al., 2019).

The response-based KD mimics the output layer neural response directly. Hinton et al. (2015) and Ba and Caruana (2014) proposed the mimicking of the knowledge by learning the probability distribution via soft labels. However, their methods need to obtain the class probability distribution. An effective method is to distill the feature knowledge or the relationship knowledge from the large teacher model. The goal of feature-based KD is to align the feature representation of the student network with the teacher network. FitNets (Romero et al., 2015) initially introduce intermediate representations learning, in which hints are defined as the outputs of the middle layer of the teacher network and are used to boost the learning of the student network. Inspired by Romero et al. (2015), various of feature-based KD approaches (Zagoruyko and Komodakis, 2017; Passalis and Tefas, 2018; Chen et al., 2021) have been proposed. The relation-based KD approach probes the relationships among different intermediate layer features (Yim et al., 2017) or data in the dataset (Park et al., 2019).

Different from existing works, we explore the scenario of learning the intrinsic local-level features and reusing the knowledge of large-scale models for different downstream tasks in a cross-task manner.

2.2 Prototype learning

Prototype learning aims to learn a set of prototypes from the source set, enabling this set of prototypes to retain the maximum amount of information contained in the target set, while ensuring that all elements within the prototype set have minimal overlapping information. Deng et al. (2021) introduced a variational prototype learning (VPL) method, which represents each category as a distribution in the latent space to align samples with prototypes. CPR (Hur et al., 2023) links each prototype with the corresponding known class features while pushing the reciprocal components away from these prototypes, thereby learning class-wise prototypes in the potential unknown feature space. Some methods (Snell et al., 2017; Liu et al., 2020; Li et al., 2021) extract more representative prototypes by aggregating similar feature vectors for few-shot learning. Wei et al. (2023) proposed an online prototype updating method with adaptive momentum for aligning prototypes in online continual learning tasks. Here, we propose an implicit prototype learning approach to learn invariant essential representations of objects from the generalized features of a teacher model, thereby enhancing the performance of cross-task KD.

3 Main approach

Our approach is intended to distill the knowledge in the large-scale network to different down-

unedited

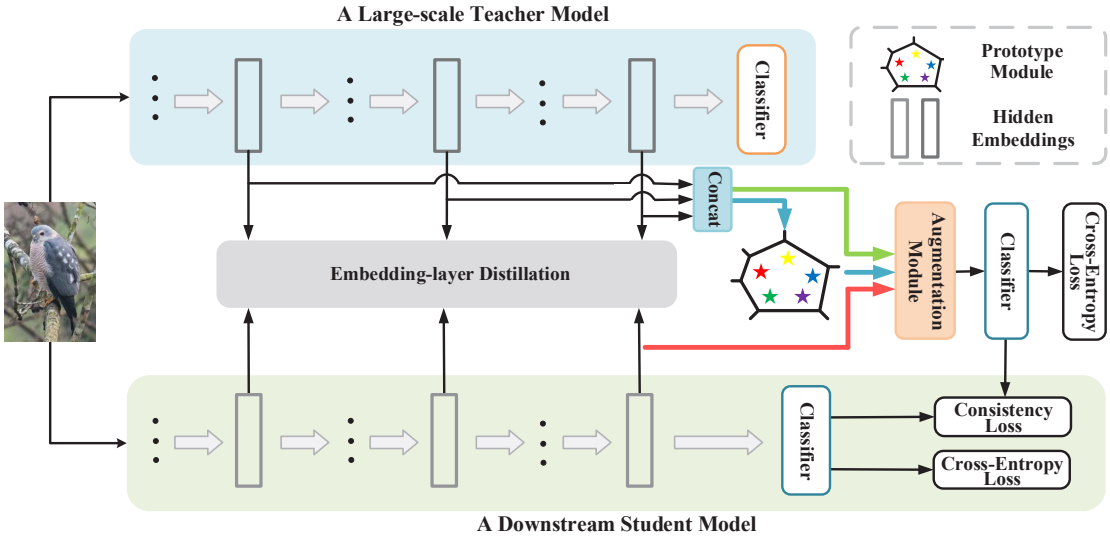


Fig. 2 Illustration of our proposed prototype-guided cross-task knowledge distillation (ProC-KD) framework. ProC-KD includes an embedding layer distillation module, a prototype learning module, and a feature augmentation module. The blue arrow in the framework represents generalized representation learning based on prototypes. The green and red arrows indicate that the prototypes are used to enhance the features extracted from the teacher network and the student network, respectively.

unedited

stream small networks. The label space of the teacher network is different from that of the student network, which is called a cross-task KD. Existing same-task KD methods mainly directly mimic the final prediction or the middle layers of the teacher model, thereby transferring the global features, and are task-specific. The local intrinsic representations can greatly benefit the cross-task student model in understanding the novel dataset of the downstream task. As shown in Fig. 2, to enhance the generalization performance of the downstream model, a ProC-KD method is proposed to migrate the invariant intrinsic knowledge from the large-scale teacher network to the small student network. The prototype learning module and the feature augmentation module are the key modules of our model.

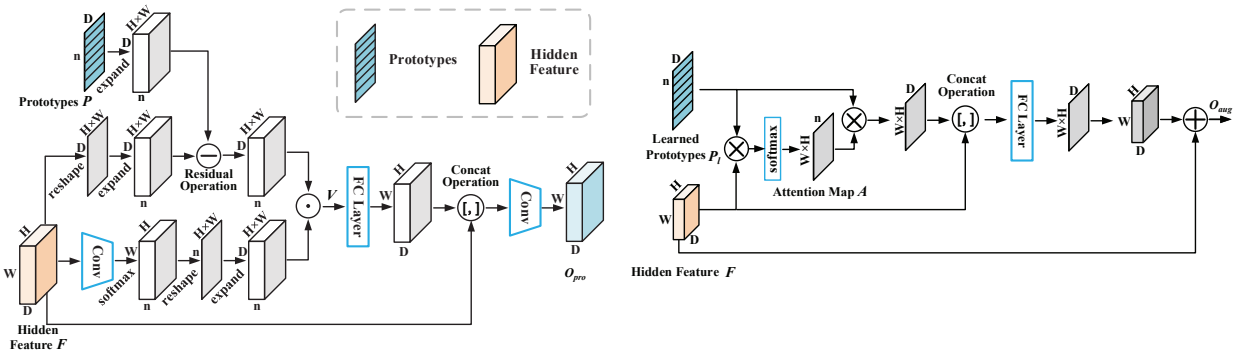
3.1 Prototype-based representation learning

Compared to previous methods that directly mimic the final predicted logits or the intermediate layers, our approach is to design a module to learn the intrinsic representation. Recent studies (Snell et al., 2017; Liu et al., 2020) have demonstrated that constructing prototype learning in models can help to solve novel dataset problems. The category-specific information can be captured by prototype learning. Thus, we propose a prototype-guided module for the teacher-student distillation architectures to learn the generalized representations with the guidance of prototypes.

Fig. 3a shows an illustration of the prototype learning module. The forward process is first to align the prototypes with the input feature, then reconstruct the prototype-related feature with the attention mechanism, and finally, aggregate the reconstructed attention features with the input features. The whole process can be divided into three subprocesses, which are *alignment*, *attention*, and *aggregation*.

Specifically, before feeding forward the two-dimensional (2D) hidden layer feature extracted by the transformer-based model, we reshape it to $F \in \mathbb{R}^{D\times H\times W}$, where D, H, and W indicate the feature dimension, height, and width, respectively. We define the prototypes as P ($p_i \in \mathbb{R}^D, i = 1, 2, ..., n$). In the *Alignment* subprocess, both the defined prototype tensor matrix and the input feature tensor matrix are expanded to $n \times D \times (W \times H)$, and we align them with residual operation $(F - P)$. In the *Attention* subprocess, we calculate the feature descriptors V based on the attention maps, which can be formulated as follows:

$$V_i = \sum_{j=1}^{WH} \frac{e^{L_{ji}}}{\sum_{i=1}^{n} e^{L_{ji}}} \left(F_j - p_i\right), \tag{1}$$



(a) Prototype Learning Module

(b) Feature Augmentation Module

Fig. 3 The schematic diagrams of (a) prototype learning module and (b) feature augmentation module. 'Conv' indicates convolution operation, and FC layer is the fully connected layer. $\ominus$, $\odot$, $\otimes$, $\oplus$, and [,] indicate the residual operation, element-wise multiplication, matrix multiplication, element-wise addition, and concatenation operation, respectively.

where, L_{ji} denotes the ji-th feature value within the feature map of dimensions $n \times D \times (W \times H)$, which is derived from the convolution operation on the hidden feature F.

In the *Aggregation* subprocess, we first concatenate the feature descriptors V and the input features and then transform the result with a nonlinear transformation block f. It can be presented as follows:

$$O_{pro} = f\left(concat\left[F,\ \ V_r W_p + b_p\right]\right), \tag{2}$$

where V_r is the reshaped feature descriptors of V, W_p and b_p indicate the weight and bias of the fully connected layer, respectively. Moreover, $concat[,]$ indicates the concatenation operation. The shape of the output O_{pro} and the input F are the same.

Through the above processes, the generalized representation of the prototypes could be learned from the input feature of the large-scale teacher model. Upon such generalized prototypes, we seek to enhance the student features with these prototypes.

3.2 Feature augmentation with prototypes

To strengthen the generalization ability of student networks of different downstream tasks, the learned generalized prototypes are used to enhance the feature in the feature augmentation module, as shown in the right part of Fig. 2. The main idea is to enrich the feature that is more related to the prototype representation while suppressing the unrelated feature.

Fig. 3b shows the illustration of the feature augmentation module. The forward process is first to pay attention to the feature correlated with the prototype representation, and then enhance the input feature with the prototype-related feature. The whole feature augmentation process can also be broken into two subprocesses, namely *Attention* and *Augmentation*.

Concretely, for the learned prototypes $P_l \in \mathbb{R}^{n \times D}$ and the hidden features $F_h \in \mathbb{R}^{t \times D}$ in the *Attention* subprocess, we first encode the prototypes and input features, respectively. Attention map A is obtained by calculating the softmax of the cross-product between learned prototypes P_l and encoded feature F_e, which can be expressed as follows:

$$A = softmax\left(F_e P_l^{\mathsf{T}}\right), \tag{3}$$

and then the attention feature is obtained by calculating the cross-product between the attention map and the prototypes.

In the *Augmentation* subprocess, we concatenate the attention feature with the encoded input features and then apply a fully connected layer to transform the shape of the result. The original input hidden layer feature is enhanced with the prototype-related feature through element-wise sum operation. This whole process can be written as follows:

$$O_{aug} = ReLU(\Phi\left(concat\left[F_e,\ \ AP_l\right]\right) + F_r), \tag{4}$$

where O_{aug} is the output, Φ indicates the function of the fully connected layer, F_r indicates the reshaped feature of F_h, W_h denotes the linear transform weight, and $concat[,]$ indicates the concatena-

unedited

tion operation. Finally, the enhanced feature is input to the classifier shared with the student model to predict the categories.

3.3 Cross-task KD

In this paper, the KD of student models and teacher models in different label spaces is defined as cross-task KD. We proposed a prototype learning module and a feature augmentation module to guide the student network training in cross-task distillation scenarios and improve its generalization ability. Here, we also design some loss functions to constrain the training of cross-task KD. Following Jiao et al. (2020), we also distill the knowledge of hidden state features from the large-scale teacher model for the embedding-layer KD. Assuming that we are distilling the knowledge from an m-layers teacher model to the n-layers small student model, we need to select n out of m layers from the large teacher network. The loss function for hidden layer KD can be expressed as follows:

$$L_{emb} = \sum_{j=1}^{n} MSE(F_i^S W_h,\ F_i^T), \tag{5}$$

where $F^S \in \mathbb{R}^{l\times d'}$ is the student hidden feature and $F^T \in \mathbb{R}^{l\times d}$ is the teacher hidden feature; d and d' denote the hidden embedding sizes of the teacher model and the student model, respectively; $W_h \in \mathbb{R}^{d'\times d}$ is a learnable transformation weight matrix, which transforms the hidden layer features of the student model into the same dimensions as the features of the teacher model; $MSE(\cdot)$ indicates the mean squared error function.

We also defined a consistency loss L_{con} and a classification loss L_{procls} for the prototype learning. The consistency loss is obtained by calculating the KullbackâĂŞLeibler (KL) divergence between the logits y_{con} from the prototype augmentation module and the predicted logits y_{stu} from the student model, $L_{con} = \mathcal{H}(y_{con}, y_{stu}))$. The classification loss is obtained with the softmax cross-entropy loss between y_{con} and the labels y. Thus the loss function of the prototype learning module can be defined as follows:

$$L_{pro} = L_{con} + L_{procls}, \tag{6}$$

In addition to the embedding layer feature distillation loss function and the prototype learning loss function, we define the student model loss function as L_{stu}. The joint training loss function for our cross-task KD can be expressed as follows:

$$L_{total} = \lambda_{emb} L_{emb} + \lambda_{pro} L_{pro} + \lambda_{stu} L_{stu}, \tag{7}$$

where λ_{emb}, λ_{pro}, and λ_{stu} are the weights of the embedding layer feature distillation loss, the prototype learning loss, and the student model loss, respectively.

4 Experiments

To evaluate the general effectiveness of our method, we conducted experiments on both cross-task KD and standard same-task KD settings. The experiments were carried out on image classification and object detection for each setting. In this paper, the KD scheme is set as offline distillation, which means the weights of the teacher model are fixed.

4.1 Cross-Task KD

4.1.1 Image classification

We carried out our experiments on the transformer-based model in three downstream tasks, including standard image classification, long-tailed image classification, and cross-domain image classification. Here, the teacher networks are trained on ImageNet-1K (Deng et al., 2009).

Datasets. CIFAR-100 (Rebuffi et al., 2017) consists of 50K and 10K images for training and validation, respectively. It contains 100 categories and 600 images per category. Following Cao et al. (2019) and Cui et al. (2019), we create the long-tailed CIFAR-100 by reducing the number of training samples for each category but with the verification set unchanged. We also define the imbalance ratio β to represent the ratio of sample sizes between the most- and the least-frequent category; this can be formulated as $\beta = N_{max}/N_{min}$. The number of samples decays exponentially between classes. The imbalance ratio is set as 10 in our experiment. Office-Home (Venkateswara et al., 2017) has been built to test domain adaptation methods for image classification. It contains four different domains and a total of 15,500 images; the four domains are named Clip Art (Cl), Artistic images (Ar), Real-World images (Rw), and Product images (Pr). Each domain in this dataset contains 65 categories, and the images

unedited

are from office or home scenarios. We set the Real-World images (Rw) to be the training set and other domains to be the test sets in our experiments.

Implementation details. The well-known ViT (Dosovitskiy et al., 2021) and Swin Transformer (Liu et al., 2021) models are used for our experiments. For the ViT, we set the teacher network as a 12-layer ViT-B and the student network as six-, four-, and two-layer small ViT models. The indexes of hidden layers selected for distillation in the teacher model are [2, 4, 6, 8, 10, 12], [3, 6, 9, 12], and [6, 12]. Both the attention map knowledge and hidden layer feature knowledge are distilled. We set the hyperparameter λ_{pro} and λ_{stu} in Eq. (7) to be 1.0 and set λ_{emb} to be 0.3. In the training phase, the embedding features are selected to concatenate and then input to the prototype learning module and the feature augmentation module. We set the number of prototypes as 72 and the optimizer as AdamW with a learning rate of 5e-4 and a weight decay of 0.05. The size of the input image is 224×224 and 32 images per batch for each GPU.

For the Swin Transformer, the teacher model is a 24-layer Swin-L, and the student models are smaller Swin Transformer models of 12 and four layers. We set the training optimizer as AdamW with a learning rate of 5e-4 and a weight decay of 0.05. The training batch has 64 images.

The experiments are run on 8× Nvidia Tesla V100 GPUs with 32GB VRAM. We use the NVIDIA collective communications library (NCCL) for multi-node parallel training. We also reduce multi-GPU communication overheads with gradient accumulation.

Results and analysis. The experimental results on the standard, long-tailed, and cross-domain image classification tasks are shown in Table 1. We reimplement Relation Knowledge Distillation (RKD) in our experimental setting. Following TinyBERT (Jiao et al., 2020), FBKD is a feature-based KD method for the Transformer-based model, which distills the knowledge from the embedding layers and attention maps to the student model. We also reimplement the hidden layer KD method in the PKD (Miles and Mikolajczyk, 2024) approach for comparison.

Image classification. Compared with the baseline method FBKD, for the ViT, our ProC-KD improves classification by 1.3% (87.46%âĂŞ86.16%), 0.67% (84.41%âĂŞ83.75%), and 0.71% (74.21%âĂŞ73.50%) on the six-, four-, and two-layer small ViT models, respectively. For the Swin Transformer, our ProC-KD improves classification by 0.58% (84.21%âĂŞ83.63%) and 3.05% (76.13%âĂŞ73.08%) on 12- and four-layer small Swin Transformer models, respectively. This demonstrates that our ProC-KD method can promote the prototypes to learn the generalized representation and improve the generalized ability of the student model in cross-task image classification KD scenarios.

Long-tailed image classification. Table 1 shows that our ProC-KD improves the performance upon the baseline approach FBKD by 5.63% (78.32%âĂŞ72.69%), 3.63% (73.46%âĂŞ69.83%) and 0.44% (58.87%âĂŞ58.43%) on six-, four-, and two-layer small ViT models, respectively, on the long-tailed CIFAR-100. For the Swin Transformer, our method improves the performance by 10.4% (68.23%âĂŞ57.83%) on the 12-layer student Swin Transformer model and reaches the second-best level on the four-layer student Swin Transformer model. This demonstrates that our ProC-KD can also boost the generalization performance in long-tailed image classification tasks.

Cross-domain image classification. The cross-domain image classification experiment is conducted on the Office-Home dataset. As can be seen, compared with the FBKD baseline, ProC-KD improves the performance by 1.3% (76.83%âĂŞ75.51%) on domain shift Rw → Pr and by 0.9% (40.96%âĂŞ40.02 %) on the hardest domain shift Rw → Cl for the six-layer ViT student model. Our method also achieves the best results on Swin Transformer compared with other KD methods. This demonstrates that distilling the knowledge from the large-scale network can boost the performance of the small network in the cross-domain scenario, and our ProC-KD can further improve the generalization ability.

Visualization analysis. Fig. 4 shows the attention map visualizations of baseline FBKD and our ProC-KD. As can be seen, compared with the FBKD baseline, the attention map of our method ProC-KD focuses more on objects in both the shallow and the deep layers. It indicates that our ProC-KD method can promote student network learning in both the shallow layers and the deep layers. Fig. 5 shows the visualization of the 72 prototypes of the

unedited

Table 1 The mean accuracy (%) of three cross-task image classification knowledge distillation tasks

Teacher (parameters)	Method	Parameters	Standard	Long-tailed	Cross-domain		
			CIFAR-100	LT-CIFAR	Rw→Ar	Rw→Cl	Rw→Pr
ViT-B (86M)	Student model	43M	78.84	55.83	20.85	16.91	34.85
	RKD (Park et al., 2019)	43M	87.13	76.52	60.61	39.04	76.19
	ABLoss (Heo et al., 2019b)	43M	81.11	76.73	57.31	40.64	73.50
	OFD (Heo et al., 2019a)	43M	82.83	76.84	60.19	40.49	75.19
	FBKD (Jiao et al., 2020)	43M	86.16	72.69	60.28	40.02	75.51
	PKD (Miles and Mikolajczyk, 2024)	43M	86.37	75.81	59.17	40.16	75.22
	ProC-KD (Ours)	43M	**87.46**	**78.32**	**61.41**	**40.96**	**76.83**
ViT-B (86M)	Student model	29M	73.48	49.05	18.62	15.30	31.94
	FBKD (Jiao et al., 2020)	29M	83.75	69.83	19.53	16.17	35.41
	OFD (Park et al., 2019)	29M	77.78	72.20	**43.15**	**36.91**	63.87
	PKD (Miles and Mikolajczyk, 2024)	29M	79.69	71.84	36.55	27.93	57.85
	ProC-KD (Ours)	29M	**84.41**	**73.46**	42.69	32.58	**64.65**
ViT-B (86M)	Student model	15M	68.65	45.98	17.51	15.21	29.35
	FBKD (Jiao et al., 2020)	15M	73.50	58.43	19.37	16.20	32.28
	OFD (Park et al., 2019)	15M	60.83	53.15	24.80	**21.40**	**46.83**
	PKD (Miles and Mikolajczyk, 2024)	15M	73.08	58.72	24.85	20.55	44.38
	ProC-KD (Ours)	15M	**74.21**	**58.87**	**25.67**	21.17	45.51
Swin-L (197M)	Student model	110M	78.90	41.48	26.87	20.51	43.32
	RKD (Park et al., 2019)	110M	83.99	58.94	27.71	21.73	46.32
	ABLoss (Heo et al., 2019b)	110M	83.26	67.08	36.88	23.01	52.64
	OFD (Heo et al., 2019a)	110M	80.99	47.91	40.49	26.81	59.20
	FBKD (Jiao et al., 2020)	110M	83.63	57.83	41.29	29.03	63.40
	AttentionProbe (Wang et al., 2022)	110M	80.78	66.34	38.03	26.87	58.05
	ProC-KD (Ours)	110M	**84.21**	**68.23**	**42.16**	**30.24**	**64.27**
Swin-L (197M)	Student model	44M	74.89	55.82	25.01	19.95	40.14
	ABLoss (Heo et al., 2019b)	44M	75.23	58.03	34.33	20.84	51.43
	OFD (Heo et al., 2019a)	44M	75.21	52.00	28.75	19.49	44.25
	FBKD (Jiao et al., 2020)	44M	73.08	45.94	26.28	19.16	40.96
	AttentionProbe (Wang et al., 2022)	44M	75.28	**56.71**	32.98	24.97	49.47
	ProC-KD (Ours)	44M	**76.13**	56.50	**34.45**	**25.53**	**55.32**

The cross-domain image classification task is performed on the Office-Home dataset. The abbreviation LT-CIFAR indicates long-tailed CIFAR-100.

unedited

CIFAR-100 dataset on the six-layer ViT model. Fig. 5a is the distance matrix of the initial prototypes, Fig. 5b is the distance matrix of the learned prototypes, and Fig. 5c is the t-distributed stochastic neighbor embedding (t-SNE) visualization of the learned prototypes. The visualization results show that the distances between the initial prototypes are relatively small, concentrating around 0.3, and are tightly packed in the feature space. In contrast, the distance of the prototype after learning is relatively divergent, and the training process enables the prototype to be better distributed in the feature space to capture the diversity and complexity of the data.

4.1.2 Object detection

In addition to the image classification tasks, we also evaluate our approach on standard object detection and domain-adaptive object detection tasks. We only take the source domain as the training domain and then test on the target domain directly for domain-adaptive object detection. Here, we train the teacher models on the COCO (Lin et al., 2014) dataset, and the weights of the teacher model are fixed.

Datasets. The Cityscapes (Cordts et al., 2016) is an urban street dataset with eight categories of objects. It contains about 3K and 500 images for training and validation, respectively. It is often used for object detection and image segmentation tasks. FoggyCityscapes (Sakaridis et al., 2018) is a dataset generated by synthesizing different degrees of fog on Cityscapes (Cordts et al., 2016). Thus, the quantity of images in the training and validation set is the same with Cityscapes. Daytime-sunny, Dusk-rainy, and Night-rainy (Wu et al., 2021) are three street-scene datasets under different weather environments collected from the BDD-100k dataset. In our experiments, about 28K images from Daytime-sunny are selected as the training set, and 2.5K and 3.5K images from Night-rainy and Dusk-rainy are selected as the test set for the two scenarios, respectively.

Implementation details. For the cross-task KD experiment of standard object detection and cross-domain object detection, we set the teacher

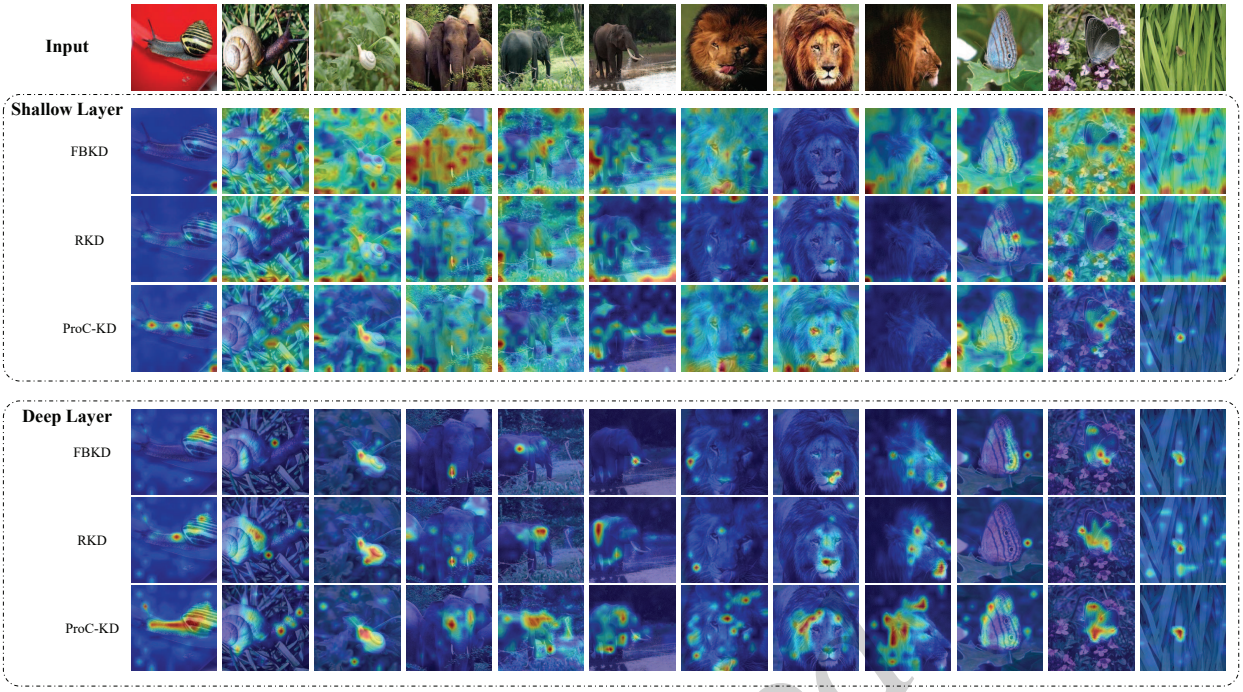


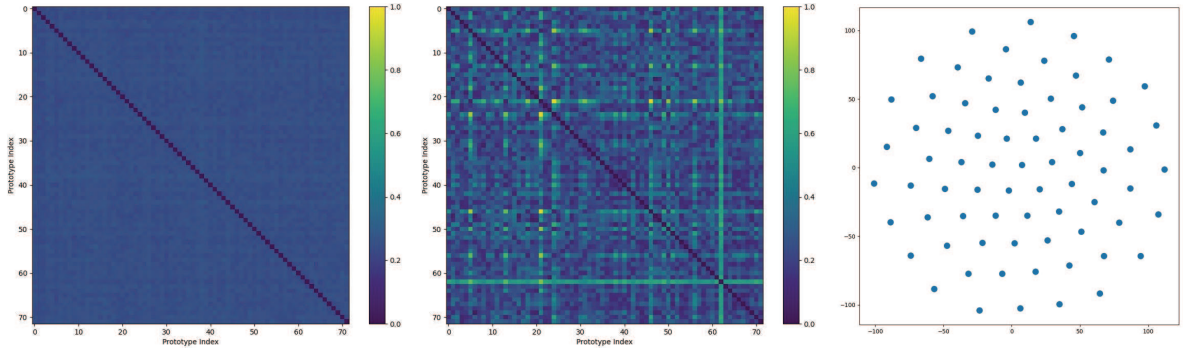
Fig. 4 The comparison of attention maps by using the transformer interpretability method (Chefer et al., 2021). Here, the second and the last layers of the ViT are selected as the shallow layer and the deep layer, respectively.

unedited

network as the Cascade Mask-RCNN with the backbone of 24-layer Swin-Base (Liu et al., 2021) model trained on COCO (Lin et al., 2014), and we set the student network as the Cascade Mask-RCNN with the backbone of 12-layer Swin-Tiny. The fourth layer feature of the feature pyramid Network (FPN) in the teacher network is selected to input to the prototype learning module and feature augmentation module for generalized representation learning. The coefficients of λ_{pro} and λ_{emb} are both set to '1.0'. The optimizer used for training is the SGD optimizer with a learning rate of 0.01 and a parameter decay of 1e-4. We set the training batch size to be 2/GPU in the cross-task object detection KD.

Results and analysis of Cityscapes and FoggyCityscapes. Table 2 shows the detection results on Cityscapes and FoggyCityscapes. Here, we reimplement the method of FBKD (Jiao et al., 2020), CWD (Shu et al., 2021), MGD (Yang et al., 2022a), and SKD (Zhang and Ma, 2023) in our experimental setting. FBKD is a KD method that follows TinyBERT (Jiao et al., 2020). It shows that our method boosts the performance under the cross-task KD scenario significantly. For Cityscapes, compared to the second best method CWD (Shu et al., 2021), our method separately improves the performance by 1.7%. For FoggyCityscapes, compared to the second best method CWD (Shu et al., 2021), our method boosts the mAP by 1.6%. The results demonstrate that the generalized prototype representation is helpful for the learning of the student network in the context of object detection.

Results and analysis of domain-adaptive object detection. We also evaluate our method on the domain-adaptive object detection scenario. Here, the student model is trained on the Daytime-sunny domain and tested on the Night-rainy domain and the Dusk-rainy domain. The teacher networks are all trained on the COCO dataset, and the weight is frozen during distillation training. As shown in Table 3, for Daytime-sunny → Night-rainy, compared with the CWD (Shu et al., 2021) baseline, our ProC-KD method boosts the mAP by 0.9%. For Daytime-sunny → Dusk-rainy, our ProC-KD method improves the mAP by 0.5%. The reason that the performance of our ProC-KD is lower than CWD on the object of the motorcycle may be that the quantity of the ground truth of the motorcycle is quite small. It only contains 49 ground truths of motorcycles in the Daytime-sunny → Night-rainy test-



(a) Distance Matrix of Initialize Prototype (b) Distance Matrix of Learned Prototype (c) T-SNE of Learned Prototype

Fig. 5 The visualization of prototypes on the CIFAR-100 dataset of the cross-task knowledge distillation task.

Table 2 Results (%) of cross-task knowledge distillation of object detection on the Cityscapes and the FoggyCityscapes datasets

Method	Bicycle	Bus	Car	Motorcycle	Person	Rider	Train	Truck	mAP
Cityscapes									
Student model	47.4	46.1	68.4	27.0	33.9	35.6	32.1	31.7	40.3
CWD (Shu et al., 2021)	57.9	62.1	77.5	**41.4**	51.1	49.4	51.5	**52.3**	55.4
MGD (Yang et al., 2022a)	52.6	55.2	73.8	31.0	45.7	47.0	42.4	38.5	48.3
ProC-KD (Ours)	**58.7**	**66.7**	**77.5**	39.2	**53.4**	**55.2**	**56.3**	50.0	**57.1**
FoggyCityscapes									
Student model	31.8	42.3	63.0	27.3	40.8	40.3	11.6	27.8	35.6
FBKD (Jiao et al., 2020)	49.0	53.9	68.6	40.6	52.0	54.1	35.7	37.5	48.9
CWD (Shu et al., 2021)	50.9	57.1	71.9	43.2	53.3	55.8	46.8	37.5	52.1
MGD (Yang et al., 2022a)	47.3	49.4	66.3	32.3	47.8	28.5	31.3	41.0	43.0
SKD (Zhang and Ma, 2023)	45.4	53.7	69.3	41.4	52.1	51.2	42.6	36.0	48.9
ProC-KD (Ours)	**51.5**	**57.7**	**73.1**	**44.2**	**53.8**	**57.9**	**51.2**	**40.3**	**53.7**

The teacher networks are trained on the COCO and the weights are frozen during distillation training.

unedited

Fig. 6 Qualitative results on Cityscapes. The first, the second, and the third rows represent the Ground Truth (GT), the results of the CWD method, and the results of our ProC-KD method, respectively. Compared with the CWD baseline, our ProC-KD method could detect objects more accurately, for example, the bus, bicycle, truck, and person.

ing set and only 110 ground truths in the Daytime-sunny→Dusk-rainy testing set.

Visualization analysis. The visualization of the detection results on the Cityscapes and FoggyCi-

Table 3 Results (%) of cross-task knowledge distillation on domain adaptive object detection of Daytime-sunny → Night-rainy and Daytime-sunny → Dusk-rainy

Method	Bicycle	Bus	Car	Motorcycle	Person	Rider	Truck	mAP
Daytime-sunny → Night-rainy								
Student Model	24.3	9.1	33.8	1.1	12.3	9.1	16.1	15.1
FBKD (Jiao et al., 2020)	35.7	17.0	47.1	**9.8**	22.7	13.9	31.7	25.4
CWD (Shu et al., 2021)	38.6	17.1	49.4	9.7	24.4	15.6	34.4	27.0
MGD (Yang et al., 2022a)	32.6	10.6	42.5	1.4	21.4	9.9	27.8	20.9
SKD (Zhang and Ma, 2023)	36.9	14.0	46.4	6.6	21.6	13.0	31.6	24.3
ProC-KD (Ours)	**40.9**	**18.3**	**49.4**	8.6	**26.1**	**18.2**	**35.7**	**27.9**
Daytime-sunny → Dusk-rainy								
Student Model	40.6	14.9	66.0	11.5	25.8	15.2	39.7	30.5
FBKD (Jiao et al., 2020)	48.2	33.0	73.1	21.5	42.2	28.7	53.7	42.9
CWD (Shu et al., 2021)	49.9	34.8	**73.9**	**24.0**	43.9	**32.0**	**54.7**	44.7
MGD (Yang et al., 2022a)	45.1	26.8	72.0	10.8	39.9	22.5	49.1	38.0
SKD (Zhang and Ma, 2023)	48.1	30.3	73.4	20.9	41.5	26.7	52.1	41.9
ProC-KD (Ours)	**52.6**	**36.6**	**73.3**	21.6	**46.5**	31.6	**54.6**	**45.2**

The teacher models are all trained on the COCO dataset, and the weight is frozen during distillation training.

Fig. 7 Qualitative results on FoggyCityscapes. The first, the second, and the third rows represent the Ground Truth (GT), the results of the CWD method, and the results of our ProC-KD method, respectively. Compared with the CWD (Shu et al., 2021) baseline, our ProC-KD method could detect objects more accurately in the foggy scene, for example, the rider, bicycle, bus, and car.

tyscapes datasets are shown in Figs. 6 and 7, respectively. Here, the first row is the ground truth, the second row shows the results of the baseline method CWD (Shu et al., 2021), and the third row shows the detection results of ProC-KD. We can see that our method ProC-KD could detect objects more precisely in normal and foggy scenes compared with the second-best method CWD. The visualization results of domain-adaptive object detection on Daytime-sunny → Night-rainy and Daytime-sunny → Dusk-rainy are shown in Figs. 8 and 9, respectively. Here, we take CWD (Shu et al., 2021) as the baseline. The first row is the ground truth, the second row contains the results of the baseline method CWD (Shu et al., 2021), and the third row shows the results of our ProC-KD. The visualization results show that our ProC-KD could localize and recognize objects in both night-rainy and dusk-rainy images accurately compared with the second-best method CWD.

4.2 Same-task KD

Our method is also verified in the standard KD setting in which the teacher network and the student network share the same label space.

Fig. 8 Qualitative results of domain-adaptive object detection on Daytime-sunny→Night-rainy. The first, the second, and the third rows represent the Ground Truth (GT), the results of the CWD method, and the results of our ProC-KD method, respectively. Compared with the CWD (Shu et al., 2021) baseline, our ProC-KD method could detect objects more accurately in the night rainy scene, for example, the person, bus, truck, and car.

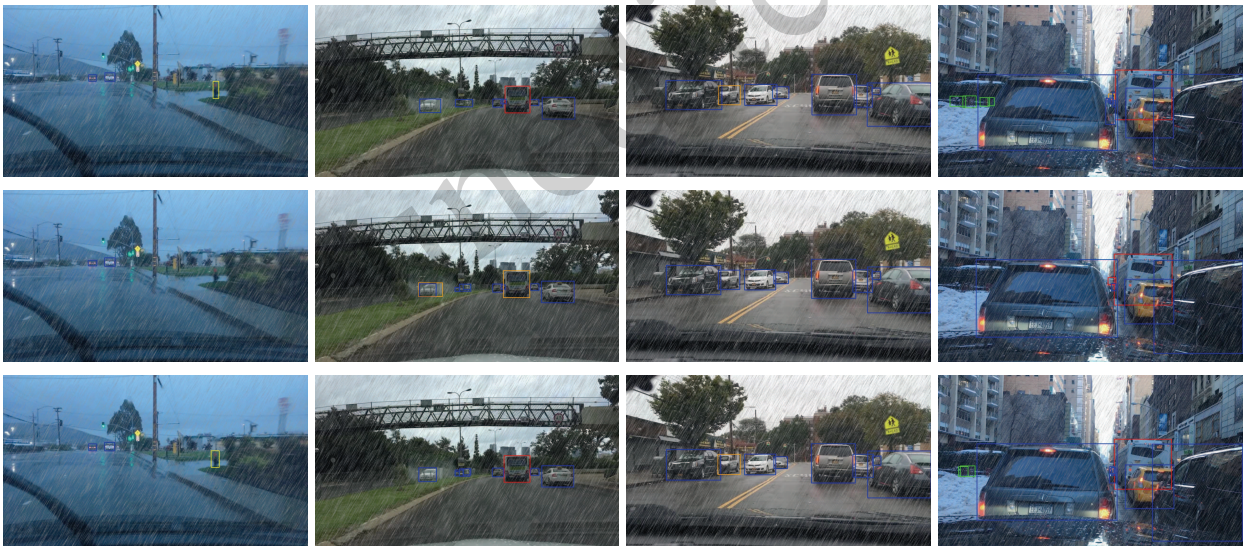


Fig. 9 Qualitative results of domain adaptive object detection on Daytime-sunny → Dusk-rainy. The first, the second, and the third rows represent the Ground Truth (GT), the results of the CWD method, and the results of our ProC-KD method, respectively. Compared with the CWD baseline, our ProC-KD method could detect objects more accurately in the dusk rainy scene, for example, the person, bus, car, truck, and bicycle.

4.2.1 Image classification

In this part, we verify our method on the image classification model of CNN structure. The experiments are conducted with the Wide-ResNet network. By changing the depth and the width of the student model, we can get different student models, thus verifying the adaptability of the method to networks of different scales. The dataset used in the experiment is CIFAR-100 (Rebuffi et al., 2017). Following with REFILLED (Ye et al., 2020), all teacher models are set as Wide-ResNet with a depth of 40 and width of '2'in these experiments.

The comparison results between our ProC-KD

Table 4 Results (%) of the standard image classification knowledge distillation scenario

Method/(depth, width)	(40, 1)	(16, 2)	(16, 1)
Student	68.97	70.15	65.44
KD (Hinton et al., 2015)	70.46	71.87	66.54
FitNet(Romero et al., 2015)	68.66	70.89	65.38
VID-I (Ahn et al., 2019)	71.51	73.31	66.32
RKD (Park et al., 2019)	72.18	72.56	65.22
REFILLED (Ye et al., 2020)	72.72	74.01	67.56
DKD (Zhao et al., 2022)	74.81	76.24	67.46
MASCKD (Gou et al., 2023)	-	-	67.26
BookKD (Zhu et al., 2023)	-	-	69.29
ProC-KD (Ours)	**75.14**	**76.43**	**69.36**

The teacher network and student network share the same label space of the CIFAR-100 dataset.

Table 5 Results (%) on the standard object detection knowledge distillation scenario

Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Teacher	44.3	62.7	48.4	25.4	48.4	58.1
Student (ResNet50)	38.4	59.0	42.0	21.5	42.1	50.3
Chen (Chen et al., 2017)	38.7	59.0	42.1	22.0	41.9	51.0
Wang (Wang et al., 2019)	39.1	59.8	42.8	22.2	42.9	51.1
Heo (Heo et al., 2019a)	38.9	60.1	42.6	21.8	42.7	50.7
CWD (Shu et al., 2021)	41.7	62.0	45.5	23.3	45.5	55.5
FGD (Yang et al., 2022b)	42.0	-	-	**23.8**	**46.4**	55.5
MGD (Yang et al., 2022a)	42.1	-	-	23.7	46.4	56.1
SKD (Zhang and Ma, 2023)	41.5	62.2	45.1	23.5	45.0	55.3
AKD (Zhang et al., 2023)	42.0	62.3	45.7	23.6	45.9	55.7
ProC-KD (Ours)	**42.1**	**62.7**	**46.0**	23.5	45.8	**57.1**

The teacher network and the student network share the same label space of the COCO dataset.

and other state-of-the-art (SOTA) distillation methods with different student models are shown in Table 4. Similar to REFILLED (Ye et al., 2020), we test the model after training convergence on the training set. Table 4 shows that our method achieves the best accuracy in three student models with different structures compared with other KD methods.

The visualization results of the randomly sampled 10 classes embedding features with t-SNE (Van Der Maaten and Weinberger, 2012) are shown in Fig. 11. It shows that for the embedding features of 10 categories sampled randomly, the embedding representation of our method is more discriminative.

4.2.2 Object detection

Our prototype-guided KD method is also applied to the standard object detection KD. The teacher model in the experiments is set as Cascade Mask-RCNN with ResNeXt101 backbone, and the student model is Faster-RCNN with the backbone of

Table 6 Quantitative results for diverse lightweight detectors on the COCO dataset

Student	Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-18	Student	34.6	55.0	37.1	19.3	36.9	45.9
	SKD	37.0	57.2	39.7	19.9	39.7	50.3
	ProC-KD	**37.5**	**57.4**	**40.5**	**20.0**	**40.9**	**50.7**
ResNet-50	Student	38.4	59.0	42.0	21.5	42.1	50.3
	SKD	41.5	62.2	45.1	23.5	45.0	55.3
	ProC-KD	**42.1**	**62.7**	**46.0**	**23.5**	**45.8**	**57.1**

The backbone of the teacher model is ResNeXt-101

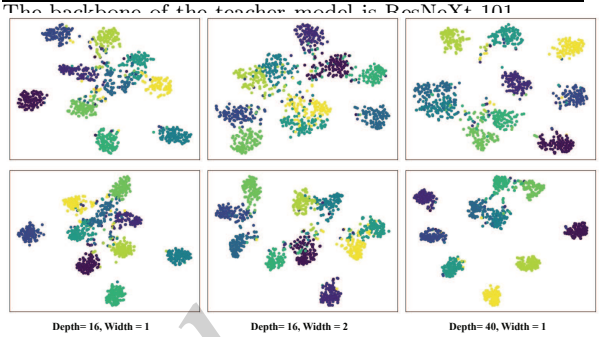


Fig. 11 The t-SNE of the baseline (upper) and our (bottom) methods over 10 classes randomly sampled from the CIFAR-100 dataset.

ResNet-18 and ResNet-50. Different from the cross-task KD experiments, here, the training dataset of the pretrained teacher network and the distillation process are both performed on the COCO dataset.

Table 5 shows the comparison results with other SOTA methods on object detection. Our method has performance comparable to that of other KD methods in different IoUs and with different object sizes. In particular, we achieve a 1.0% improvement over the second-best method on AP_l. Table 6 illustrates the results for various lightweight detectors on the COCO dataset, and it shows that our proposed method outperforms the SKD approach on the detector with the backbone of ResNet-18 and ResNet-50. These results demonstrate the robustness of our methodology across different compression ratio detectors.

Fig. 10 shows the error analysis PrecisionâĂŞRecall curve of all objects, large-sized objects, medium-sized objects, and small-sized objects under different conditions on the COCO dataset. The top row shows the detection results of the baseline method CWD (Shu et al., 2021), and the bottom row shows the detection results of our ProC-KD. We can see that our ProC-KD achieves better performance on different IoU thresholds for all different-sized objects. Compared with the baseline method, our method improves performance by 0.029

unedited

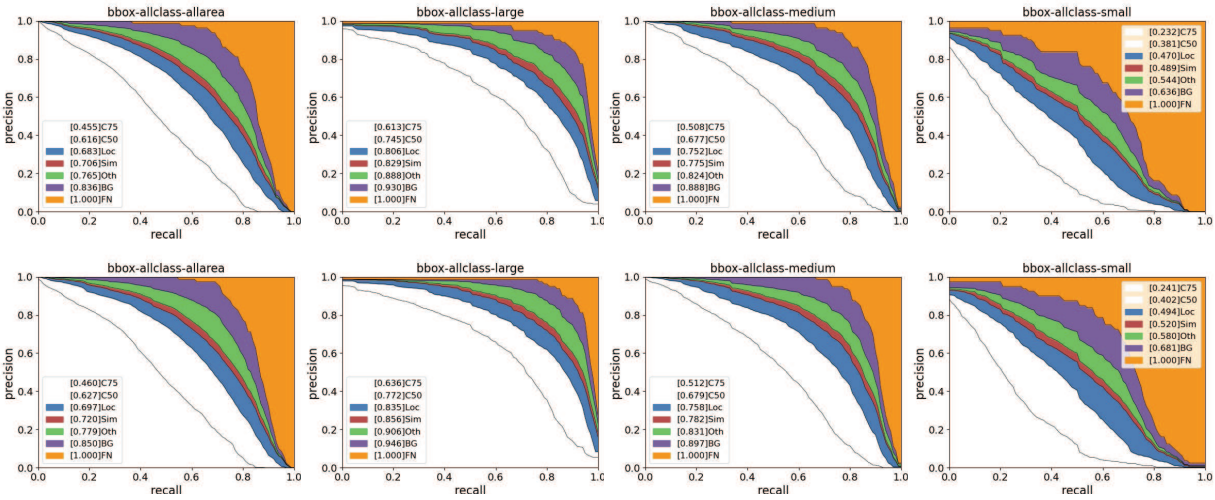


Fig. 10 The error analysis PrecisionâĂŞRecall curves of all objects, large-sized objects, medium-sized objects, and small-sized objects on the COCO dataset. The top row is from the baseline CWD, and the bottom row is from our ProC-KD. Here, C75 indicates the results at the 0.75 IoU threshold, C50 indicates the results at the 0.50 IoU threshold, Loc indicates the results after ignoring localization errors, Sim indicates the results after ignoring similar classes from the same supercategory false positives, Oth indicates the results after ignoring all category confusions, BG indicates the results after ignoring all false positives, and FN indicates the results after ignoring all false negatives.

unedited

and 0.024 on large and small objects, respectively, when ignoring the localization errors. This indicates that our method can provide more precise classification information. Our ProC-KD also outperforms by an average of 0.014 on all area objects after ignoring localization errors, ignoring similar classes from the same supercategory, ignoring all category confusions, and ignoring all false positives, which demonstrate a better location and recognition ability of our method.

4.3 Ablation study

We conduct the ablation study on the design elements, the number of prototypes, and the weight of the loss function in the proposed prototype-guided KD.

4.3.1 Design elements

We study the effects of the design elements on the FoggyCityscapes. The ablation results in Table 7 show the following: (a) compared with the baseline, a 3.5% (52.4%âĂŞ48.9%) mAP boost can be obtained with the prototype learning module, indicating that the prototype representation is beneficial to the learning of the student network; (b) the combination of the prototype learning module and feature augmentation module leads to a significant mAP improvement, which is 1.3% (53.7%âĂŞ52.4%). The reason may be that the feature augmentation module enriches the feature that is more related to the object.

4.3.2 Number of prototypes

We ablate the number of prototypes on the long-tailed image classification task with the ViT model. The 12-layer Base version of the ViT model is set as the teacher network, and a six-layer ViT model is set as the student network. We set the number of prototypes as 24, 48, 72, and 96. Here, we only change the number of prototypes and keep other network settings unchanged to evaluate the training of KD. Table 8 shows that the performance increases as the number of prototypes increases, and the best accuracy obtained is 78.32% when the number of prototypes is 72. It indicates that the small number of prototypes could not learn the generalized representation sufficiently. In other experiments, we set the number of prototypes in prototype-guided KD methods as 72.

4.3.3 Hyperparameters

We conduct the ablation study of hyperparameters in Eq. (7) on the FoggyCityscapes. As shown in

Table 7 Ablation study results (%) of design elements on FoggyCityscapes

Method	PL	FA	Bicycle	Bus	Car	Motorcycle	Person	Rider	Train	Truck	mAP
Baseline	✗	✗	49.0	53.9	68.6	40.6	52.0	54.1	35.7	37.5	48.9
Ours	✗	✓	50.5	55.3	72.3	**45.4**	53.4	56.5	47.7	38.1	52.4
Ours	✓	✓	**51.5**	**57.7**	**73.1**	44.2	**53.8**	**57.9**	**51.2**	**40.3**	**53.7**

'PL' is the prototype learning module and 'FA' is the feature augmentation module.

Table 8 Ablation study results (%) of the number of prototypes on the long-tailed CIFAR-100

Model/Number	24	48	72	96
ProC-KD	77.75	78.11	**78.32**	78.28

Table 9, we set the loss weights λ_{emb}, λ_{pro}, and λ_{stu} with different values and get the object detection results. The ablation study results of the loss weights also demonstrate the effectiveness of our prototype learning method in object detection KD. When λ_{emb} and λ_{pro} are disabled and λ_{stu} is set to '1.0', it represents the student model. When λ_{pro} is disabled and both λ_{emb} and λ_{stu} are set to '1.0', it represents the FBKD method. Our proposed method introduces λ_{pro}, and the results demonstrate a significant improvement of our method over the FBKD approach.

5 Conclusions

To solve the issue of applying a large-scale model to different downstream tasks, a prototype-guided cross-task KD method (ProC-KD is proposed, wherein the label spaces of the teacher network and the student network are inconsistent. Specifically, the prototype learning module is trained to capture the invariant intrinsic local-level representations of objects, leveraging the robust capabilities of the teacher network. Then, the learned prototypes are used to augment the student network features to improve the generalization ability of the student network. We conduct experiments on image classification and object detection tasks, and the quantitative and qualitative results demonstrate the effectiveness of our ProC-KD for cross-task KD.

Contributors

Deng Li and Ya-hong HAN conceived the method. Deng Li and A-ming WU designed and implemented the ProC-KD model. Deng LI and Peng LI processed the data. Deng Li drafted the paper. Peng LI, A-ming WU, and Ya-hong HAN helped organize the paper. Ya-hong HAN supervised all aspects of the project.

Table 9 Ablation study of loss function hyperparameter on FoggyCityscapes

Number	λ_{emb}	λ_{pro}	λ_{stu}	mAP(%)
Student	✗	✗	1	35.6
0	1	✗	1	48.9
1	0.3	1	1	49.5
2	1	0.3	1	52.5
3	1	1	0.3	51.2
4	0.5	1	1	50.7
5	1	0.5	1	52.5
6	1	1	0.5	52.0
7	0.8	1	1	52.1
8	1	0.8	1	52.9
9	1	1	0.8	52.8
10	1	1	1	**53.7**

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are openly available in mmrazor (https://github.com/open-mmlab/mmrazor) and mdistiller (https://github.com/megvii-research/mdistiller).

unedited

References

Ahn S, Hu SX, Damianou A, et al., 2019. Variational information distillation for knowledge transfer. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9163-9171. https://doi.org/10.1109/CVPR.2019.00938

Ba J, Caruana R, 2014. Do deep nets really need to be deep? Proc 27th Int Conf on Neural Information Processing Systems, p.2654-2662.

Cao KD, Wei CL, Gaidon A, et al., 2019. Learning imbalanced datasets with label-distribution-aware margin loss. Proc 33th Int Conf on Neural Information Processing Systems.

Carion N, Massa F, Synnaeve G, et al., 2020. End-to-end object detection with transformers. Proc 16th European Conf on Computer Vision, p.213-229. https://doi.org/10.1007/978-3-030-58452-8_13

Chebotar Y, Waters A, 2016. Distilling knowledge from ensembles of neural networks for speech recognition. Proc 17th Annual Conf of the Int Speech Communication Association, p.3439-3443.

Chefer H, Gur S, Wolf L, 2021. Transformer interpretability beyond attention visualization. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.782-791. https://doi.org/10.1109/CVPR46437.2021.00084

Chen DF, Mei JP, Zhang Y, et al., 2021. Cross-layer distillation with semantic calibration. Proc 35th AAAI Conf on Artificial Intelligence, p.7028-7036. https://doi.org/10.1609/aaai.v35i8.16865

Chen GB, Choi W, Yu X, et al., 2017. Learning efficient object detection models with knowledge distillation. Proc 31th Int Conf on Neural Information Processing Systems, p.742-751.

Chen YC, Li LJ, Yu LC, et al., 2020. Uniter: Universal image-text representation learning. Proc 16th European Conf on Computer Vision, p.104-120.
https://doi.org/10.1007/978-3-030-58577-8_7

Cordts M, Omran M, Ramos S, et al., 2016. The cityscapes dataset for semantic urban scene understanding. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3213-3223.
https://doi.org/10.1109/CVPR.2016.350

Cui Y, Jia M, Lin TY, et al., 2019. Class-balanced loss based on effective number of samples. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9268-9277.
https://doi.org/10.1109/CVPR.2019.00949

Deng J, Dong W, Socher R, et al., 2009. Imagenet: A large-scale hierarchical image database. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.248-255.
https://doi.org/10.1109/CVPR.2009.5206848

Deng JK, Guo J, Yang J, et al., 2021. Variational prototype learning for deep face recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.11906-11915.
https://doi.org/10.1109/CVPR46437.2021.01173

Dosovitskiy A, Beyer L, Kolesnikov A, et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Proc 9th Int Conf on Learning Representations*, .

Fu TJ, Li L, Gan Z, et al., 2023. An empirical study of end-to-end video-language transformers with masked visual modeling. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.22898-22909.
https://doi.org/10.1109/CVPR52729.2023.02193

Gou JP, Yu BS, Maybank SJ, et al., 2021. Knowledge distillation: A survey. *Int J Comput Vis*, 129(6):1789-1819.
https://doi.org/10.1007/s11263-021-01453-z

Gou JP, Sun LY, Yu BS, et al., 2023. Multilevel attention-based sample correlations for knowledge distillation. *IEEE Trans Ind Inf*, 19(5):7099-7109.
https://doi.org/10.1109/TII.2022.3209672

Heo B, Kim J, Yun S, et al., 2019a. A comprehensive overhaul of feature distillation. Proc IEEE/CVF Int Conf on Computer Vision, p.1921-1930.
https://doi.org/10.1109/ICCV.2019.00201

Heo B, Lee M, Yun S, et al., 2019b. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. Proc 33th AAAI Conf on Artificial Intelligence, 33(01):3779-3787.
https://doi.org/10.1609/aaai.v33i01.33013779

Hinton G, Vinyals O, Dean J, 2015. Distilling the knowledge in a neural network, .

Hur S, Shin I, Park K, et al., 2023. Learning classifiers of prototypes and reciprocal points for universal domain adaptation. Proceedings of the IEEE/CVF Winter Conf on Applications of Computer Vision, p.531-540.
https://doi.org/10.1109/WACV56688.2023.00060

Jain J, Li J, Chiu MT, et al., 2023. Oneformer: One transformer to rule universal image segmentation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2989-2998.
https://doi.org/10.1109/CVPR52729.2023.00292

Jiao XQ, Yin YC, Shang LF, et al., 2020. Tinybert: Distilling bert for natural language understanding. Proc Findings of the Association for Computational Linguistics, p.4163-4174.
https://doi.org/10.18653/v1/2020.findings-emnlp.372

Kurata G, Saon G, 2020. Knowledge distillation from offline to streaming rnn transducer for end-to-end speech recognition. Proc 21th Annual Conf of the Int Speech Communication Association, p.2117-2121.

Li G, Jampani V, Sevilla-Lara L, et al., 2021. Adaptive prototype learning and allocation for few-shot segmentation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8334-8343.
https://doi.org/10.1109/CVPR46437.2021.00823

Li L, Chen YC, Cheng Y, et al., 2020. Hero: Hierarchical encoder for video+language omni-representation pre-training. Proc Conf on Empirical Methods in Natural Language Processing, p.2046-2065.
https://doi.org/10.18653/v1/2020.emnlp-main.161

Lin TY, Maire M, Belongie S, et al., 2014. Microsoft coco: Common objects in context. Proc 13th European Conf on Computer Vision, p.740-755.
https://doi.org/10.1007/978-3-319-10602-1_48

Liu JI, Song L, Qin YQ, 2020. Prototype rectification for few-shot learning. Proc 16th European Conf on Computer Vision, p.741-756.

Liu Z, Lin Y, Cao Y, et al., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. Proc IEEE/CVF Int Conf on Computer Vision, p.10012-10022.
https://doi.org/10.1109/ICCV48922.2021.00986

Miles R, Mikolajczyk K, 2024. Understanding the role of the projector in knowledge distillation. Proc 38th AAAI Conf on Artificial Intelligence, 38(5):4233-4241.
https://doi.org/10.1609/aaai.v38i5.28219

MÃijller RR, Kornblith S, Hinton G, 2019. When does label smoothing help? Proc 33th Int Conf on Neural Information Processing Systems, p.4694-4703.

Molchanov P, Tyree S, Karras T, et al., 2017. Pruning convolutional neural networks for resource efficient inference.

Park W, Kim D, Lu Y, et al., 2019. Relational knowledge distillation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3967-3976.
https://doi.org/10.1109/CVPR.2019.00409

Passalis N, Tefas A, 2018. Learning deep representations with probabilistic knowledge transfer. Proc 16th European Conf on Computer Vision, p.283-299.
https://doi.org/10.1007/978-3-030-01252-6_17

Rebuffi SA, Bilen H, Vedaldi A, 2017. Learning multiple visual domains with residual adapters, :506-516.

Romero A, Ballas N, Kahou SE, et al., 2015. Fitnets: Hints for thin deep nets.

Sakaridis C, Dai D, Van Gool L, 2018. Semantic foggy scene understanding with synthetic data. *Int J Comput Vis*, 126(9):973-992.
https://doi.org/10.1007/s11263-018-1072-8

Unedited

Sanh V, Debut L, Chaumond J, et al., 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, .
https://doi.org/https://arxiv.org/abs/1910.01108

Shu CY, Liu YF, Gao JF, et al., 2021. Channel-wise knowledge distillation for dense prediction. Proc IEEE/CVF Int Conf on Computer Vision, p.5311-5320.
https://doi.org/10.1109/ICCV48922.2021.00526

Snell J, Swersky K, Zemel R, 2017. Prototypical networks for few-shot learning, :4080-4090.

Sun SQ, Cheng Y, Gan Z, et al., 2019. Patient knowledge distillation for bert model compression. Proc Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing, p.4322-4331.
https://doi.org/10.18653/v1/D19-1441

Touvron H, Cord M, Douze M, et al., 2021. Training data-efficient image transformers & distillation through attention. Proc 38th Int Conf on Machine Learning, p.10347-10357.

Van Der Maaten L, Weinberger K, 2012. Stochastic triplet embedding. Proc IEEE Int Workshop on Machine Learning for Signal Processing, p.1-6.
https://doi.org/10.1109/MLSP.2012.6349720

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need, :6000-6010.

Venkateswara H, Eusebio J, Chakraborty S, et al., 2017. Deep hashing network for unsupervised domain adaptation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5018-5027.
https://doi.org/10.1109/CVPR.2017.572

Wang JH, Cao MD, Shi SW, et al., 2022. Attention probe: Vision transformer distillation in the wild. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.2220-2224.

Wang T, Yuan L, Zhang X, et al., 2019. Distilling object detectors with fine-grained feature imitation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4933-4942.
https://doi.org/10.1109/CVPR.2019.00507

Wei YJ, Ye JX, Huang ZZ, et al., 2023. Online prototype learning for online continual learning. Proc IEEE/CVF Int Conf on Computer Vision, p.18764-18774.
https://doi.org/10.1109/ICCV51070.2023.01720

Wu AM, Liu R, Han YH, et al., 2021. Vector-decomposed disentanglement for domain-invariant object detection. Proc IEEE/CVF Int Conf on Computer Vision, p.9342-9351.
https://doi.org/10.1109/ICCV48922.2021.00921

Wu J, Leng C, Wang Y, et al., 2016. Quantized convolutional neural networks for mobile devices. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4820-4828.
https://doi.org/10.1109/CVPR.2016.521

Yang ZD, Li Z, Shao MQ, et al., 2022a. Masked generative distillation. Proc 17th European Conf on Computer Vision, p.53-69.
https://doi.org/10.1007/978-3-031-20083-0_4

Yang Z, Li Z, Jiang X, et al., 2022b. Focal and global knowledge distillation for detectors. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4643-4652.
https://doi.org/10.1109/CVPR52688.2022.00460

Ye HJ, Lu S, Zhan DC, 2020. Distilling cross-task knowledge via relationship matching. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.12396-12405.
https://doi.org/10.1109/CVPR42600.2020.01241

Ye LW, Rochan M, Liu Z, et al., 2019. Cross-modal self-attention network for referring image segmentation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10502-10511.
https://doi.org/10.1109/CVPR.2019.01075

Yim J, Joo D, Bae J, et al., 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7130-7138.
https://doi.org/10.1109/CVPR.2017.754

Yoon JW, Lee H, Kim HY, et al., 2021. Tutornet: Towards flexible knowledge distillation for end-to-end speech recognition. *IEEE/ACM Trans Audio, Speech, Lang Processing*, 29:1626-1638.
https://doi.org/10.1109/TASLP.2021.3071662

Zagoruyko S, Komodakis N, 2017. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer.

Zhang LF, Ma KS, 2023. Structured knowledge distillation for accurate and efficient object detection. *IEEE Trans Pattern Anal Mach Intell*, 45(12):15706-15724.
https://doi.org/10.1109/TPAMI.2023.3300470

Zhang Y, Chen WH, Lu YC, et al., 2023. Avatar knowledge distillation: self-ensemble teacher paradigm with uncertainty. Proc 31st ACM Int Conf on Multimedia, p.5272-5280.

Zhao BR, Cui Q, Song RJ, et al., 2022. Decoupled knowledge distillation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.11953-11962.
https://doi.org/10.1109/CVPR52688.2022.01165

Zhou C, Zhang Y, Chen J, et al., 2023. Octr: Octree-based transformer for 3d object detection. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5166-5175.
https://doi.org/10.1109/CVPR52729.2023.00500

Zhu MH, Gupta S, 2018. To prune, or not to prune: exploring the efficacy of pruning for model compression.

Zhu SL, Shang RH, Tang K, et al., 2023. Bookkd: A novel knowledge distillation for reducing distillation costs by decoupling knowledge generation and learning. *Knowl-Based Syst*, 279:110916.
https://doi.org/10.1016/j.knosys.2023.110916

Zhu XZ, Su WJ, Lu LW, et al., 2021. Deformable detr: Deformable transformers for end-to-end object detection.

unedited