Deng LI, Peng LI, Aming WU, Yahong HAN, 2025. Prototype-guided cross-task knowledge distillation. *Frontiers of Information Technology & Electronic Engineering*, 26(6):912-929. https://doi.org/10.1631/FITEE.2400383

Prototype-guided cross-task knowledge distillation

Front Inform Technol Electron Eng

Key words: Knowledge distillation; Cross-task; Prototype learning

Corresponding author: Yahong HAN
E-mail: yahong@tju.edu.cn
ORCID: https://orcid.org/0000-0003-2768-1398

Motivation

- Due to the high computational complexity and massive storage requirements, large-scale pre-trained models are difficult to deploy in real-world scenarios with limited resources. Existing knowledge distillation methods typically require the teacher and student models to operate within the same label space, which significantly limits their applicability in multi-task or cross-task settings. To address the challenge posed by inconsistent label spaces, this paper proposes a prototype-guided cross-task knowledge distillation (ProC-KD) approach. The goal is to transfer the intrinsic local-level features learned by the teacher model to various downstream tasks, thereby improving the generalization ability and practicality of the student model.

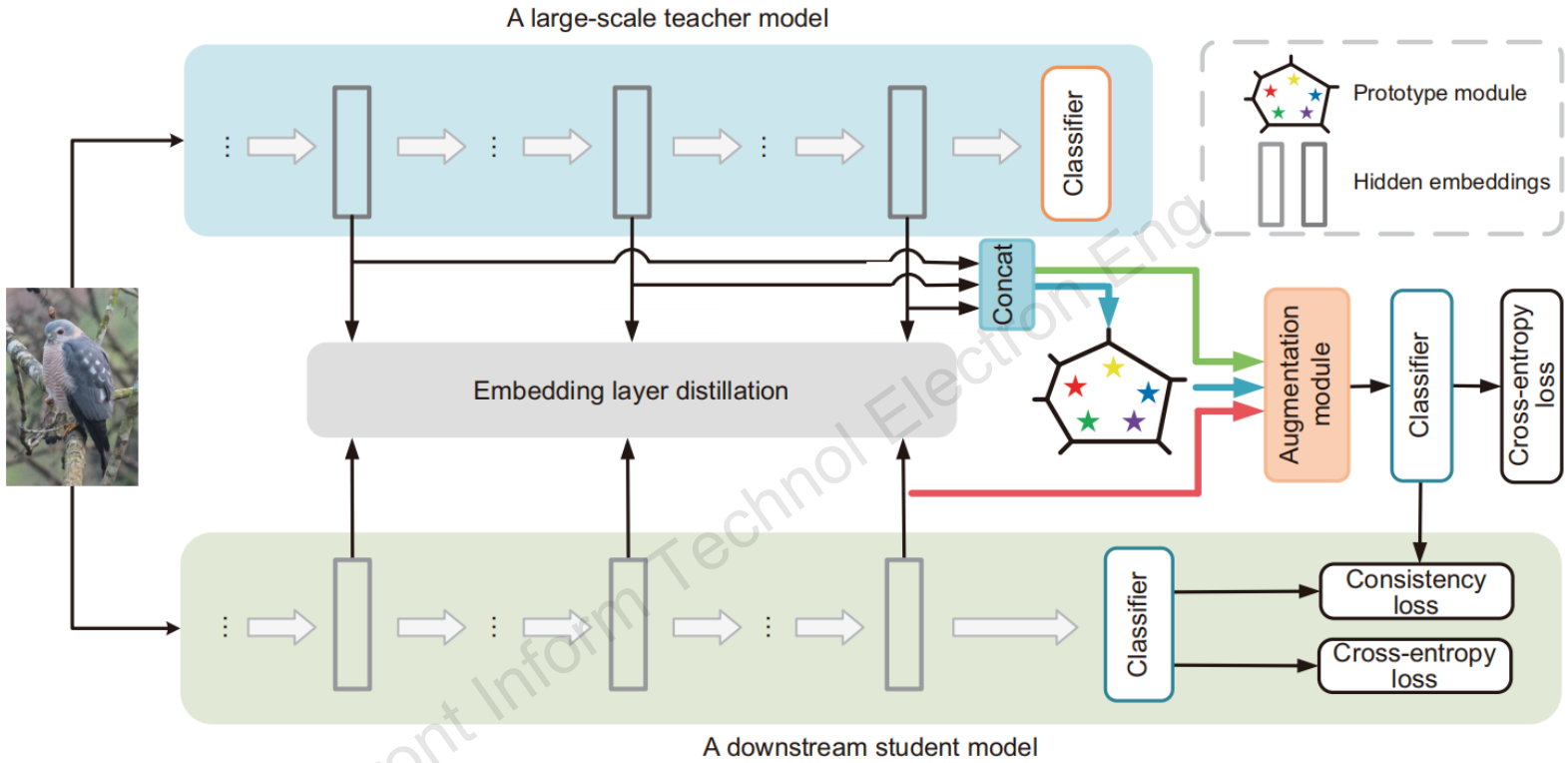
Front Inform Technol Electron Eng

Main idea

- A novel prototype-guided KD approach is proposed to migrate the intrinsic knowledge from a large-scale model to different small cross-task models without fine-tuning on the downstream task and improve the student model generalization ability.
- A prototype learning module is proposed to extract invariant object features from a large-scale teacher model. The learned prototypes capture generalized knowledge that is transferable across tasks with different label spaces.
- A feature augmentation module is designed to enhance the student model by integrating the learned prototypes. This module selectively strengthens features relevant to the prototypes and suppresses irrelevant ones, improving the generalization performance of the student model in cross-task scenarios.

Front Inform Technol Electron Eng

Method

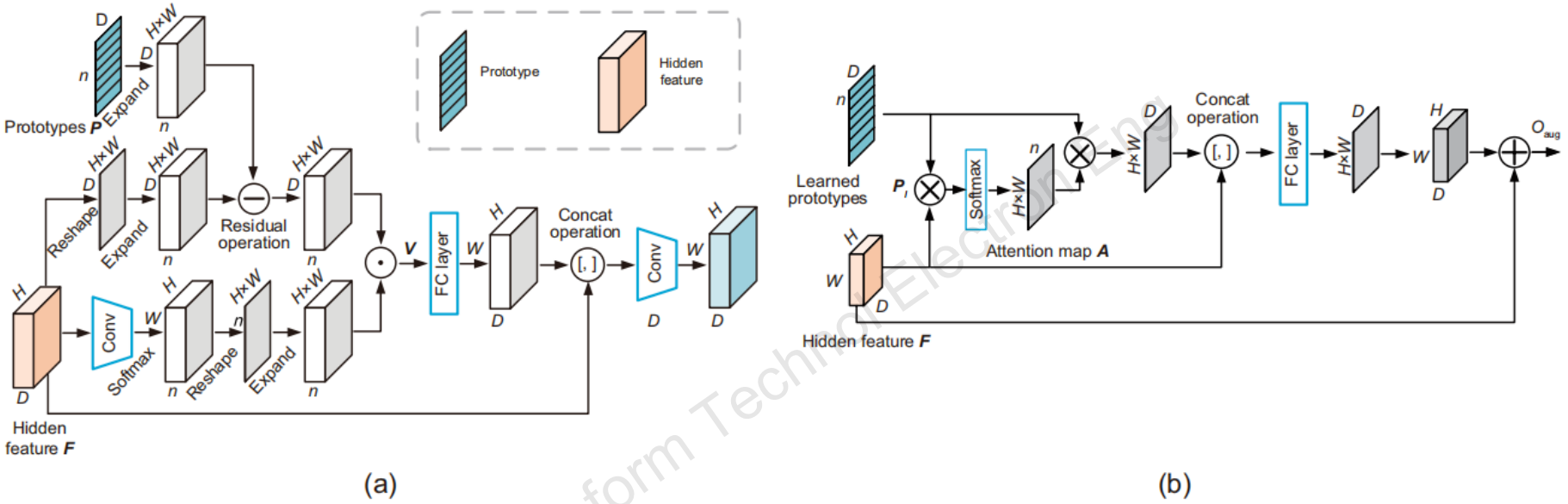


Front Inform Technol Electron Eng

Fig. 2 Illustration of our proposed ProC-KD framework. ProC-KD includes an embedding layer distillation module, a prototype-based representation 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. References to color refer to the online version of this figure

- **Prototype learning module**: extracts generalized, invariant object representations (prototypes) from the teacher's hidden features.
- **Feature augmentation module**: enhances the student's features using the learned prototypes via attention mechanisms.

Method



Front Inform Technol Electron Eng

Fig. 3 Schematic of the prototype-based representation learning module (a) and the feature augmentation module (b). 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

1. Prototype learning module:

$$\boldsymbol{V}_i = \sum_{j=1}^{WH} \frac{\mathrm{e}^{\boldsymbol{L}_{ji}}}{\sum_{i=1}^{n} \mathrm{e}^{\boldsymbol{L}_{ji}}} \left(\boldsymbol{F}_j - \boldsymbol{p}_i\right)$$

$$\boldsymbol{O}_{\mathrm{pro}} = f\left(\mathrm{concat}\left[\boldsymbol{F}, \boldsymbol{V}_{\mathrm{r}}\boldsymbol{W}_{\mathrm{p}} + \boldsymbol{b}_{\mathrm{p}}\right]\right)$$

2. Feature augmentation module:

$$\boldsymbol{A} = \mathrm{softmax}\left(\boldsymbol{F}_{\mathrm{e}}\boldsymbol{P}_{\mathrm{l}}^{\mathrm{T}}\right)$$

$$\boldsymbol{O}_{\mathrm{aug}} = \mathrm{ReLU}(\Phi\left(\mathrm{concat}\left[\boldsymbol{F}_{\mathrm{e}}, \boldsymbol{A}\boldsymbol{P}_{\mathrm{l}}\right]\right) + \boldsymbol{F}_{\mathrm{r}})$$

Results

1. Cross-task KD

Table 1 Mean accuracy of three cross-task image classification knowledge distillation tasks

Teacher	Method	Number of parameters	Mean accuracy (%)				
			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 JH 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 JH 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**

Cross-domain image classification task is performed on the Office–Home dataset. LT-CIFAR indicates long-tailed CIFAR-100. M means million. The value in the bracket in the first column means the number of parameters of the teacher network. The best results are in bold, and the second-best results are underlined

Front Inform Technol Electron Eng

Results

1. Cross-task KD

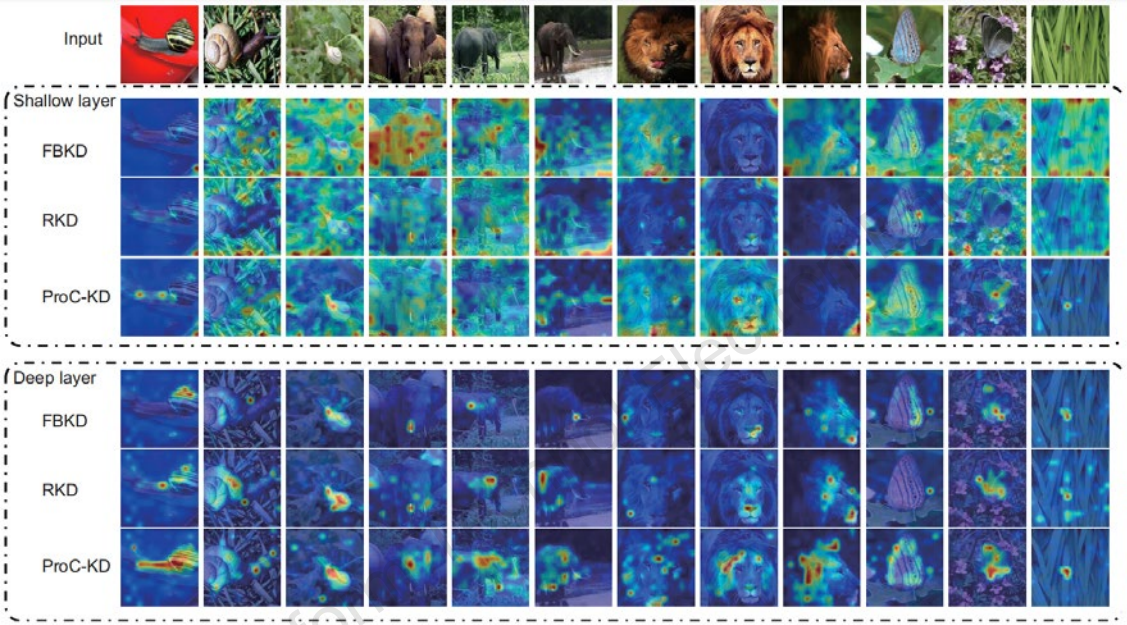


Fig. 4 Comparison of attention maps by using the Transformer interpretability method (Chefer et al., 2021). Here, the second and the last layers of the ViT are selected as the shallow layer and the deep layer, respectively

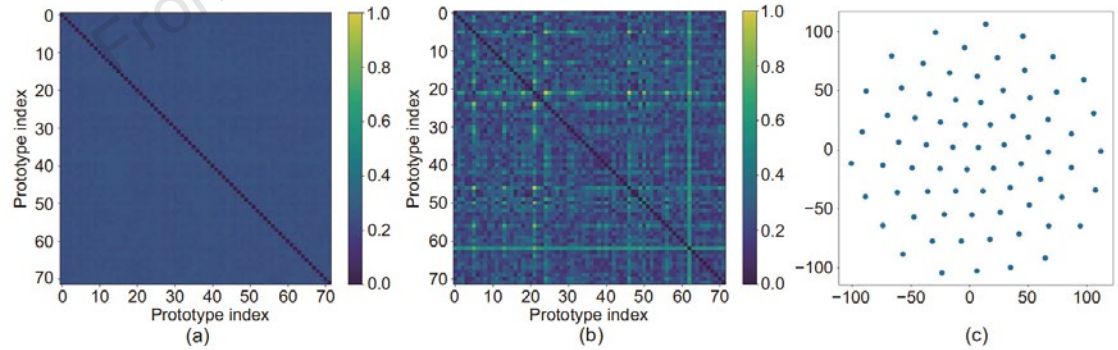


Fig. 5 Visualization of prototypes on the CIFAR-100 dataset of the cross-task knowledge distillation task: (a) distance matrix of the initial prototypes; (b) distance matrix of the learned prototypes; (c) t-SNE of the learned prototypes

Results

1. Cross-task KD

Table 2 Detection results of cross-task knowledge distillation of object detection on the Cityscapes and Foggy Cityscapes datasets

Method	Detection accuracy on Cityscapes (%)								
	Bicycle	Bus	Car	Motorcycle	Person	Rider	Train	Truck	mAP
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., 2022b)	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**
Method	Detection accuracy on Foggy Cityscapes (%)								
	Bicycle	Bus	Car	Motorcycle	Person	Rider	Train	Truck	mAP
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., 2022b)	47.3	49.4	66.3	32.3	47.8	28.5	31.3	**41.0**	43.0
SKD (Zhang LF and Ma, 2023)	45.4	53.7	69.3	41.4	52.1	51.2	42.6	36.0	49.0
ProC-KD (Ours)	**51.5**	**57.7**	**73.1**	**44.2**	**53.8**	**57.9**	**51.2**	40.3	**53.7**

Teacher networks are trained on the COCO and the weight is frozen during distillation training. The best results are in bold, and the second-best results are underlined

Table 3 Detection results of cross-task knowledge distillation on domain adaptive object detection of Daytime-sunny→Night-rainy and Daytime-sunny→Dusk-rainy

Method	Detection accuracy on Daytime-sunny→Night-rainy (%)							
	Bicycle	Bus	Car	Motorcycle	Person	Rider	Truck	mAP
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., 2022b)	32.6	10.6	42.5	1.4	21.4	9.9	27.8	20.9
SKD (Zhang LF 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**	**28.2**
Method	Detection accuracy on Daytime-sunny→Dusk-rainy (%)							
	Bicycle	Bus	Car	Motorcycle	Person	Rider	Truck	mAP
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., 2022b)	45.1	26.8	72.0	10.8	39.9	22.5	49.1	38.0
SKD (Zhang LF 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	**45.3**

Teacher models are all trained on the COCO dataset, and the weight is frozen during distillation training. The best results are in bold, and the second-best results are underlined

Front Inform Technol Electron Eng

Results

1. Cross-task KD



Fig. 6 Qualitative results on Cityscapes. The first, second, and third rows represent the ground truth, 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

Results

1. Cross-task KD



Fig. 7 Qualitative results on Foggy Cityscapes. The first, second, and third rows represent the ground truth, 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 foggy scene, for example, the rider, bicycle, bus, and car

Results

1. Cross-task KD



Fig. 8 Qualitative results of domain-adaptive object detection on Daytime-sunny→Night-rainy. The first, second, and third rows represent the ground truth, 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 Night-rainy scene, for example, the person, bus, truck, and car

Results

1. Cross-task KD

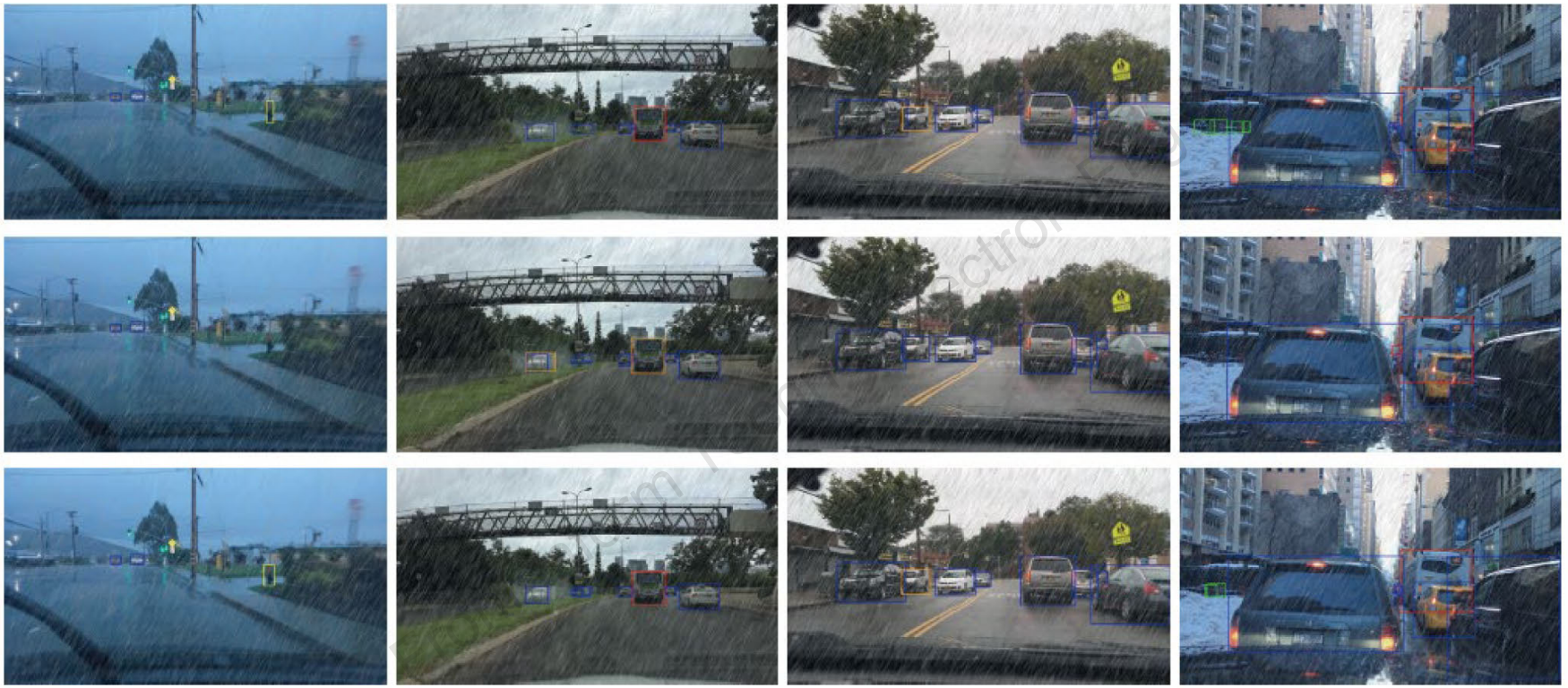


Fig. 9 Qualitative results of domain-adaptive object detection on Daytime-sunny→Dusk-rainy. The first, second, and third rows represent the ground truth, the results of the CWD method, and the results of our ProC-KD method, respectively. Compared with the CWD baseline, our ProC-KD method could detect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

Results

2. Same-task KD

Table 4 Classification results of the standard image classification knowledge distillation scenario

Method	Accuracy (%)		
	(40, 1)	(16, 2)	(16, 1)
Student	68.97	70.15	65.44
KD (Hinton et al., 2015)	70.46	71.87	66.54
FitNets (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 HJ 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 SL 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. The first and second values in the bracket are the depth and width, respectively. "–" means that the values of the compared methods are not provided in the literature. The best results are in bold, and the second-best results are underlined

Table 5 Detection 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 GB et al. (2017)'s	38.7	59.0	42.1	22.0	41.9	51.0
Wang T et al. (2019)'s	39.1	59.8	42.8	22.2	42.9	51.1
Heo et al. (2019a)'s	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., 2022a)	42.0	–	–	**23.8**	**46.4**	55.5
MGD (Yang et al., 2022b)	**42.1**	–	–	23.7	**46.4**	56.1
SKD (Zhang LF and Ma, 2023)	41.5	62.2	45.1	23.5	45.0	55.3
AKD (Zhang Y 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 student network share the same label space of the COCO dataset. "–" means that the values of the compared methods are not provided in the literature. AP_{50} and AP_{75} refer to the detection accuracy at the 0.50 and 0.75 IoU threshold, respectively. AP_S, AP_M, and AP_L refer to the detection accuracy of small-, medium-, and large-sized objects, respectively. The best results are in bold, and the second-best results are underlined

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

Backbone	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**

ResNeXt101 is the backbone of the teacher model. The best results are in bold

Front Inform Technol Electron Eng

Results

2. Same-task KD

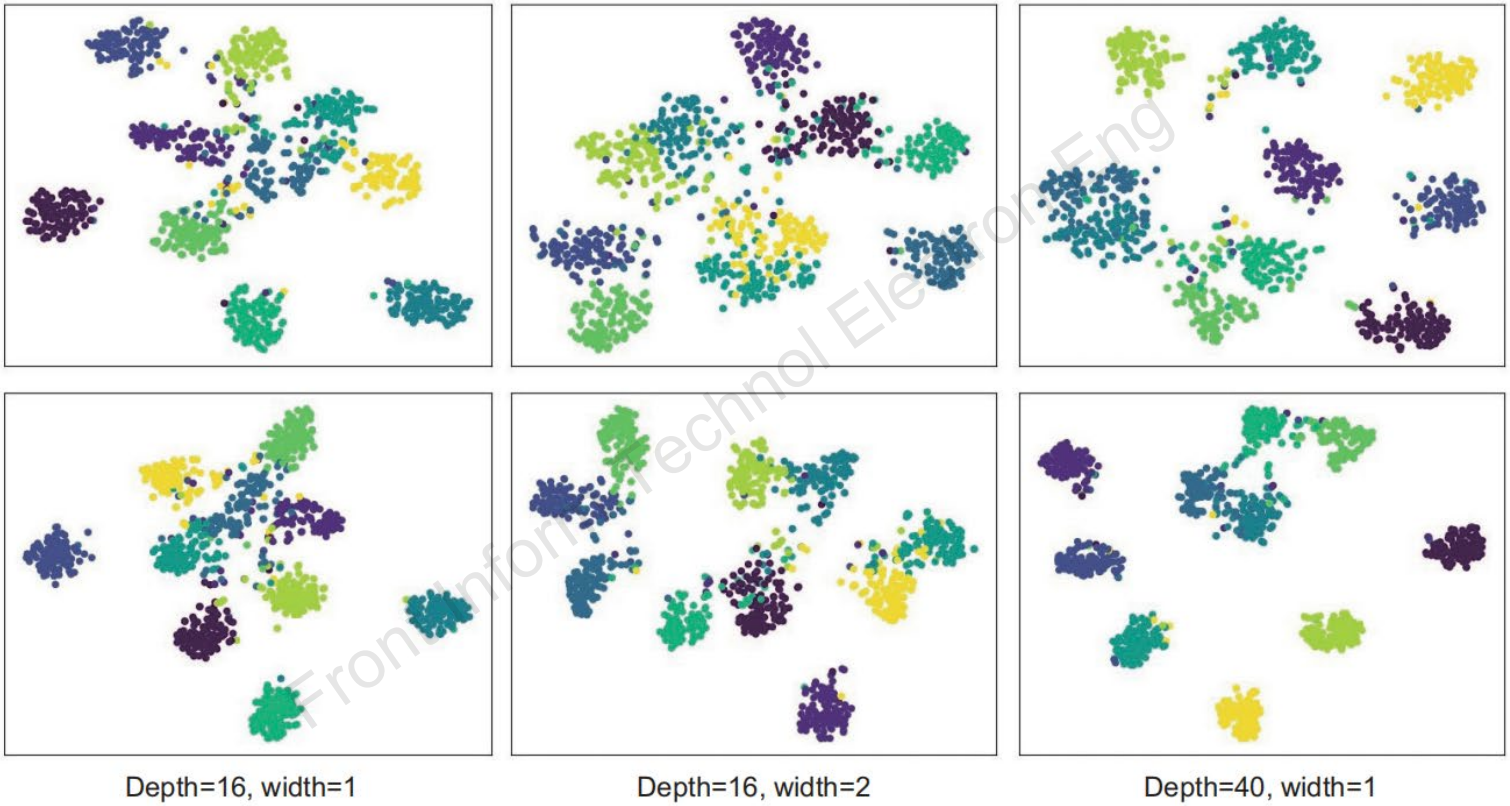


Fig. 10 t-SNE of the baseline (upper) and our (bottom) methods over 10 classes randomly sampled from the CIFAR-100 dataset

Front Inform Technol Electron Eng

Results

2. Same-task KD

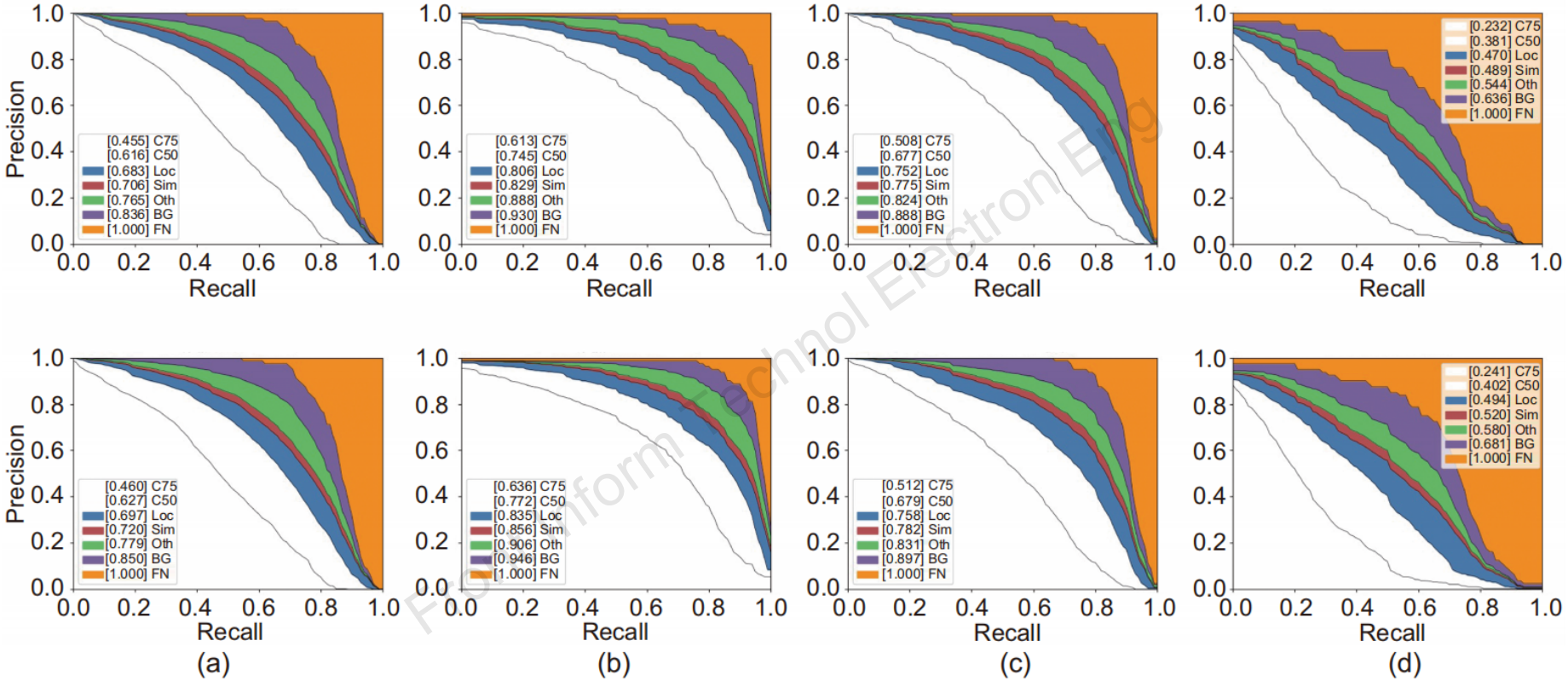


Fig. 11 Error analysis of precision–recall curves of all-area objects (a), large-sized objects (b), medium-sized objects (c), and small-sized objects (d) on the COCO dataset. The top row shows the results of the baseline CWD, and the bottom row shows the results of our ProC-KD. Here, C75 indicates the results at a 0.75 IoU threshold, C50 indicates the results at a 0.50 IoU threshold, Loc indicates the results after ignoring localization errors, Sim indicates the results obtained by ignoring false positives from similar classes within the same supercategory, 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

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-based representation learning module is trained to capture the invariant intrinsic local-level representations of objects, leveraging the robust capability 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.

Front Inform Technol Electron Eng



Deng LI received the MS degree from the University of Chinese Academy of Sciences, China, in 2018. Currently, he is pursuing his PhD degree with the College of Intelligence and Computing, Tianjin University. His research interests include computer vision, multimodal understanding, and transfer learning.



Peng LI, is a postgraduate Senior Engineer. His main research interests include digital government, network security, artificial intelligence, big data mining, and public security.



Aming WU received the PhD degree from Tianjin University, Tianjin, China, in 2021. He is currently an associate professor with the School of Electronic Engineering, Xidian University. He is the author or co-author of more than 20 scientific articles at top venues. His current research interests include computer vision, multimedia analysis, and machine learning.



Yahong HAN (Member, IEEE) received the PhD degree from Zhejiang University, Hangzhou, China, in 2012. He is currently a professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. From Nov. 2014 to Nov. 2015, he visited Prof. Bin Yu's group with UC Berkeley as a visiting scholar. His current research interests include multimedia analysis, computer vision, and machine learning.

Front Inform Technol Electron Eng