BPKD: Boundary Privileged Knowledge Distillation For Semantic Segmentation

Liyang Liu [1] **Zihan Wang** [2 3] **Minh Hieu Phan** [1] **Bowen Zhang** [1] **Jinchao Ge** [1] **Yifan Liu** [1*]

[1] The University of Adelaide, Australia [2] The University of Queensland, Australia
[3] CSIRO's Data61, Australia

The paper "Boundary Privileged Knowledge Distillation For Semantic Segmentation" introduces a novel method, BPKD, that decouples knowledge distillation into edge and body components, a first in the field. This decoupling enables the student model to focus on spatially sensitive features for edge prediction and semantically rich features for body segmentation. The method achieves state-of-the-art performance across multiple network architectures on three benchmark datasets, including a 13% Trimap improvement over existing methods. While it incurs a slightly higher training cost, the method is efficient for real-time applications post-training. The work has been implemented using the MMseg framework, promising full code and training log release for public scrutiny. The paper proposes a novel approach to solving "conceptual knowledge leaking" in student networks.

In the supplementary materials, we augment the core arguments of our main paper with additional analyses, expanded experimental details, and enhanced visualizations. Initially, we conduct a performance comparison using reduced schedules across three unique datasets, specifically chosen for ablation studies. These comparative results are encapsulated in Table 2. To facilitate reproducibility, we provide an in-depth guide to our implementation, along with PyTorch-compatible code for our specialized Edge Loss. Additional analytical insights into class-trimap IoU are available in Figure 1 and Table 3. Lastly, due to space constraints in the main manuscript, we include both visual and quantitative experiments on three distinct datasets within this supplementary section.

1. Performance On Reduced Schedules

Table 1. Performance Comparison with state-of-the-art distillation methods on ADE20K dataset, the student backbone is not pre-trained on ImageNet.

Methods	mIoU	mAcc(%)
T:PSPNet-R101 [11]	44.39	54.74
S:PSPnet-R18 [11]	17.11	22.99
SKDS [5]	20.79	27.74
IFVD [3]	20.75	27.6
CIRKD [10]	22.90	30.68
CWD [7]	24.79	31.44
BPKD(Ours)	**27.46**	**36.10**

In order to demonstrate the efficacy of leveraging soft-labels from teachers to accelerate convergence speed, we conducted a comparison experiment with reduced schedules, 40k iteration training, and a 512 × 512 resolution. The results of this experiment, as well as comparisons with other state-of-the-art algorithms, are reported in Table 2. To ensure a fair comparison, our proposed knowledge distillation framework was applied to different teachers. Our students were able to learn knowledge from the teacher network, resulting in significant performance gains and achieving state-of-the-art results on multiple datasets. As demonstrated in Table 2, on ADE20K our method outperformed strong baseline channel-wise distillation by 2.34%, 1.46%, 1.66%, 1.57%, and 1.59%, respectively. Furthermore, our methods were able to enhance the lightweight student network without increasing computational capacity, improving performance by 6.45%, 4.13%, 6.08%, and 8.49% compared to raw students. On Cityscape our method outperformed strong baseline channel-wise distillation by 2.22%, 1.35%, 4.40%, 1.52%, respectively. Furthermore, our methods were able to enhance the lightweight student network without increasing computational capacity, improving performance by 10.07%, 3.78%, 8.64%, and 1.75% compared to raw students. On Pascal Context our method outperformed strong baseline channel-wise distillation by 2.34%, 1.46%, 1.66%, 1.57%, and 1.59%, respectively. Furthermore, our methods were able to enhance the lightweight student network without increasing computational capacity, improving performance by 10.07%, 3.78%, 6.08%, and 8.49% compared to raw students. These numerical results suggest that

[*] Corresponding author. Contact Email: akide.liu@adelaide.edu.au.

Table 2. Performance comparison of different distillation methods with state-of-the-art techniques in **Reduced schedules**. We set a reduced training setting to reducethe computational cost, including crop size reduced to 512×512, and training schedule to 40k iterations. We test these methods on various segmentation networks for both student and teacher models, using datasets including Cityscapes [1], ADE20K [12], and Pascal Context [2]. The FLOPs are obtained on 512×512 resolutions. Our BPKD outperforms all previous methods in large margins across multiple datasets and network architectures.

Methods	FLOPs(G)	Parameters(M)	ADE20K		Cityscapes		Pascal Context 59	
			40k 512*512		40k 512*512		40k 512*512	
			mIoU(%)	mAcc(%)	mIoU	mAcc(%)	mIoU	mAcc(%)
T: PSPNet-R101	256.89	68.07	44.39	54.75	79.74	86.56	52.47	63.15
S:PSPnet-R18	54.53	12.82	29.42	38.48	68.99	75.19	43.07	53.79
SKDS	54.53	12.82	31.80	42.25	69.33	75.37	43.93	54.01
IFVD	54.53	12.82	32.15	42.53	71.08	77.46	44.75	54.99
CIRKD	54.53	12.82	32.25	43.02	72.23	78.79	44.83	55.3
CWD	54.53	12.82	33.53	41.71	74.29	80.95	45.92	55.50
BPKD(Ours)	54.53	12.82	**35.87**	**45.42**	**75.94**	**82.62**	**47.16**	**57.61**
T: HRNetV2P-W48	95.64	65.95	42.02	53.52	80.65	87.39	51.12	61.39
S:HRNetV2P-W18S	10.49	3.97	28.69	37.86	73.77	82.89	40.82	51.70
SKDS	10.49	3.97	30.49	40.19	74.75	83.23	42.91	53.63
IFVD	10.49	3.97	30.57	40.42	75.33	83.83	43.12	54.03
CIRKD	10.49	3.97	31.34	41.45	74.63	83.72	43.45	54.10
CWD	10.49	3.97	31.36	40.24	75.54	84.08	45.50	56.01
BPKD(Ours)	10.49	3.97	**32.82**	**43.49**	**76.56**	**85.34**	**46.12**	**57.63**
T:DeeplabV3P-R101	255.67	62.68	45.47	56.41	80.98	88.7	53.2	64.04
S:DeeplabV3P+MV2	69.6	15.35	22.38	31.71	70.49	80.11	37.16	49.1
SKDS	69.6	15.35	24.65	35.07	70.81	79.31	39.18	51.13
IFVD	69.6	15.35	24.53	35.13	71.82	80.88	38.8	50.79
CIRKD	69.6	15.35	25.21	36.17	72.39	81.84	39.99	52.66
CWD	69.6	15.35	26.89	35.79	73.35	82.41	42.52	53.24
BPKD(Ours)	69.6	15.35	**28.46**	**41.45**	**76.58**	**84.14**	**44.32**	**56.04**
T:ISANet-R101	228.21	56.8	43.8	54.39	80.61	88.29	52.94	63.52
S: ISANet-R18	54.33	12.46	27.68	36.92	71.45	78.65	41.08	50.62
SKDS	54.33	12.46	28.70	38.51	70.65	77.53	42.87	52.89
IFVD	54.33	12.46	29.66	38.80	70.30	77.79	43.19	53.46
CIRKD	54.33	12.46	29.79	40.48	72.00	79.32	43.49	53.89
CWD	54.33	12.46	34.58	43.04	71.61	80.02	44.63	55.01
BPKD(Ours)	54.33	12.46	**36.17**	**45.26**	**72.72**	**81.50**	**45.50**	**56.55**

our method is not dependent on a specific model structure, and that it produces significant performance gains with the pure student network, even without ImageNet Pre-train shows in table 1. To further demonstrate the effectiveness of our method, qualitative segmentation results are visualized in Figure 5.

2. Implementation Details of Edge Loss

This section presents the implementation details of the Edge Loss in PyTorch, aimed at facilitating reproducibility. Our implementation consists of well-annotated components that are incorporated into our distillation system. To begin with, the Pre Mask Filter operation is applied to the logits obtained from both teachers and students. Subsequently, a dimensional rearrangement is performed to optimize the process. The KL divergence serves as the core component to estimate the distance between the student and teacher probability distributions, which establishes an embryonic reference for calculating the loss. The Post Mask Filter takes input from the unreduced criterion and aggregates the distance on the channel dimension. It is followed by an additional spatial expansion that repeats the spatial information based on the given channels. Finally, the inner weights vector is applied to corresponding categories to enhance the learning capacity for hard edge samples. The EDM loss terms are then finalized by overall weights adaption and average reduction. Furthermore, the Edge Detection Module is another crucial sub-component that employs multiple-level edge masks by providing input and ground truth labels.

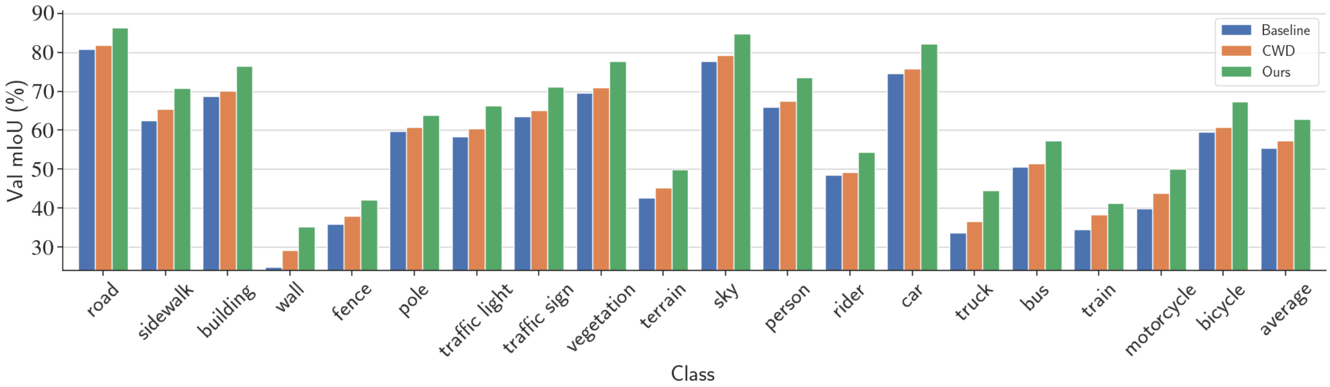

Figure 1. Illustration of our proposed **Boundary Privilege Knowledge Distillation** schemes in terms of class **Trimap IoU** metrics with PSPnet + Resnet18 network architecture over the Cityscapes validation set. It can be seen from the figure that our method has different degrees of improvement for all categories, meanwhile , we have a significant improvement for categories that are difficult to distinguish by boundaries.

Class	road	sidewalk	building	wall	fence	pole	light	sign	vege.	terrain
Baseline	80.88	62.44	68.72	24.82	35.91	59.73	58.39	63.58	69.64	42.58
CWD	81.84	65.36	70.14	29.22	37.98	60.72	60.43	65.12	70.91	45.28
BPKD(Ours)	86.29	70.73	76.45	35.13	42.12	63.96	66.25	71.09	77.65	49.87
Class	sky	person	rider	car	truck	bus	train	moto.	bicycle	average
Baseline	77.72	65.94	48.45	74.55	33.63	50.58	34.58	39.83	59.56	55.34
CWD	79.2	67.58	49.11	75.76	36.53	51.37	38.29	43.82	60.75	57.34
BPKD(Ours)	84.87	73.56	54.4	82.28	44.49	57.27	41.27	50.14	67.37	62.91

Table 3. Illustration of our proposed **Boundary Privilege Knowledge Distillation** schemes in terms of class **Trimap IoU** metrics with PSPnet + Resnet18 network architecture over the Cityscapes validation set.

We utilize dilation and erosion to retrieve the edges, with the hyperparameter, $kernel\_size$, controlling the edge width. To address the computational pressure arising from a large kernel, the $compute\_iters$ is introduced for GPU memory optimization. Additionally, the edge detection module employs average pooling as a downsampling policy, considering progressively decreasing importance from internal boundaries to outlines.

3. Categorical trimap Performance

This section presents an evaluation of the categorical Edge IoU using the PSPnet encoder and Resnet18 backbone architecture on the Cityscape validation set. Figure 1 and Table 3 illustrate the results. Our proposed methods demonstrated significant improvement in most classes based on edge factors compared to the raw student model and strong baseline channel-wise distillation method. We also observed explicit improvement for categories that are difficult to distinguish by boundaries. For instance, we achieved a 10.31% improvement for the wall category and a 6.69% improvement for the bus category.

4. Instance segmentation

We conducted experiments using SOLOv2 on the COCO dataset to demonstrate the general adaptability of our method. Specifically, we selected SOLOV2 [8] X-101(DCN) as the teacher and Light SOLOV2 [8] R-18 as the student. Table 4 presents the results, which show that our method improved the raw student by 6.5%, 9.5%, 7.5%, 6.9%, 8.8%, and 4.8% on the corresponding metrics. The results demonstrate that our distillation method can easily adapt to instance segmentation tasks and outperforms previous methods in the small-scale training setting. While instance segmentation and semantic segmentation tasks have distinct differences, they share similar properties in that they predict masks for target senses and given pixel-level annotations. During the experiments, we applied knowledge distillation methods multiple times on pyramid classification logits and masked representations, followed by calculating the average across all levels of sub-terms to obtain the distillation loss. From the numerical results, AP metrics explicitly increased for small and medium objects. However, there was no performance improvement for large objects, indicating that our method has room for improvement in instance segmentation tasks. As a future prospect, we aim to adjust the current method and design a specialized loss term for instance-level knowledge distillation.

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
T:SOLOv2-X101	41.7	63.2	45.1	18.0	45.0	61.6
S:SOLOv2-R18*	26.7	44.1	27.5	6.50	27.2	45.8
CWD	28.4	47.1	29.6	9.90	30.3	44.2
BPKD(Ours)	**33.2**	**53.6**	**35.0**	**13.4**	**36.0**	**50.6**

Table 4. The instance segmentation results presented in this report were obtained on the COCO [4] validation set using single-model results. All distillation methods and the student network baseline were trained using a 1x schedule with multiple-scale training disabled. The table below demonstrates that our distillation method can easily adapt to instance segmentation tasks and outperforms previous methods in a small-scale training setting.

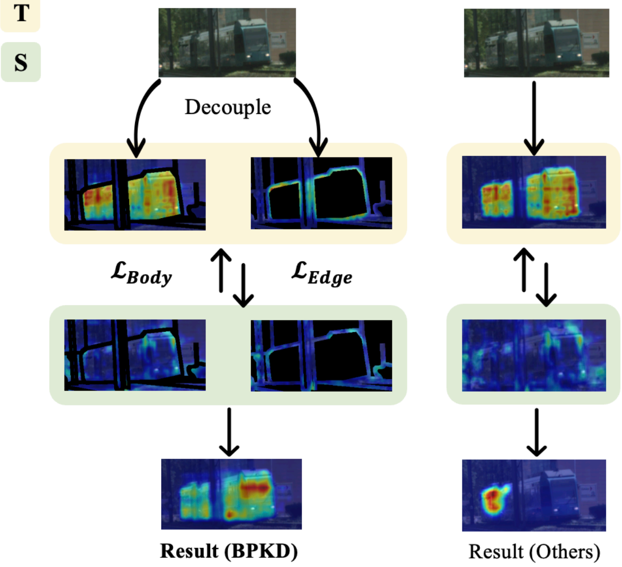

Figure 2. BPKD (left) decouples the edge and body information from the input image and creates two parallel distillation pipelines compared to pioneering works, such as CWD [7], SKDS [5], IFVD [9], CIRKD [10]. The proposed compositional loss forces the student to learn each part separately. The result shows the explicit performance gain.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 6

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2

[3] Yuenan Hou, Zheng Ma, Chunxiao Liu, Tak-Wai Hui, and Chen Change Loy. Inter-region affinity distillation for road marking segmentation. In *CVPR*, pages 12486–12495, 2020. 1

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[5] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. 1, 4

[6] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 7

[7] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021. 1, 4, 5, 6, 7

[8] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020. 3

[9] Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In *European Conference on Computer Vision*, pages 346–362. Springer, 2020. 4

[10] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12319–12328, 2022. 1, 4

[11] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1, 5, 6, 7

[12] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2, 5

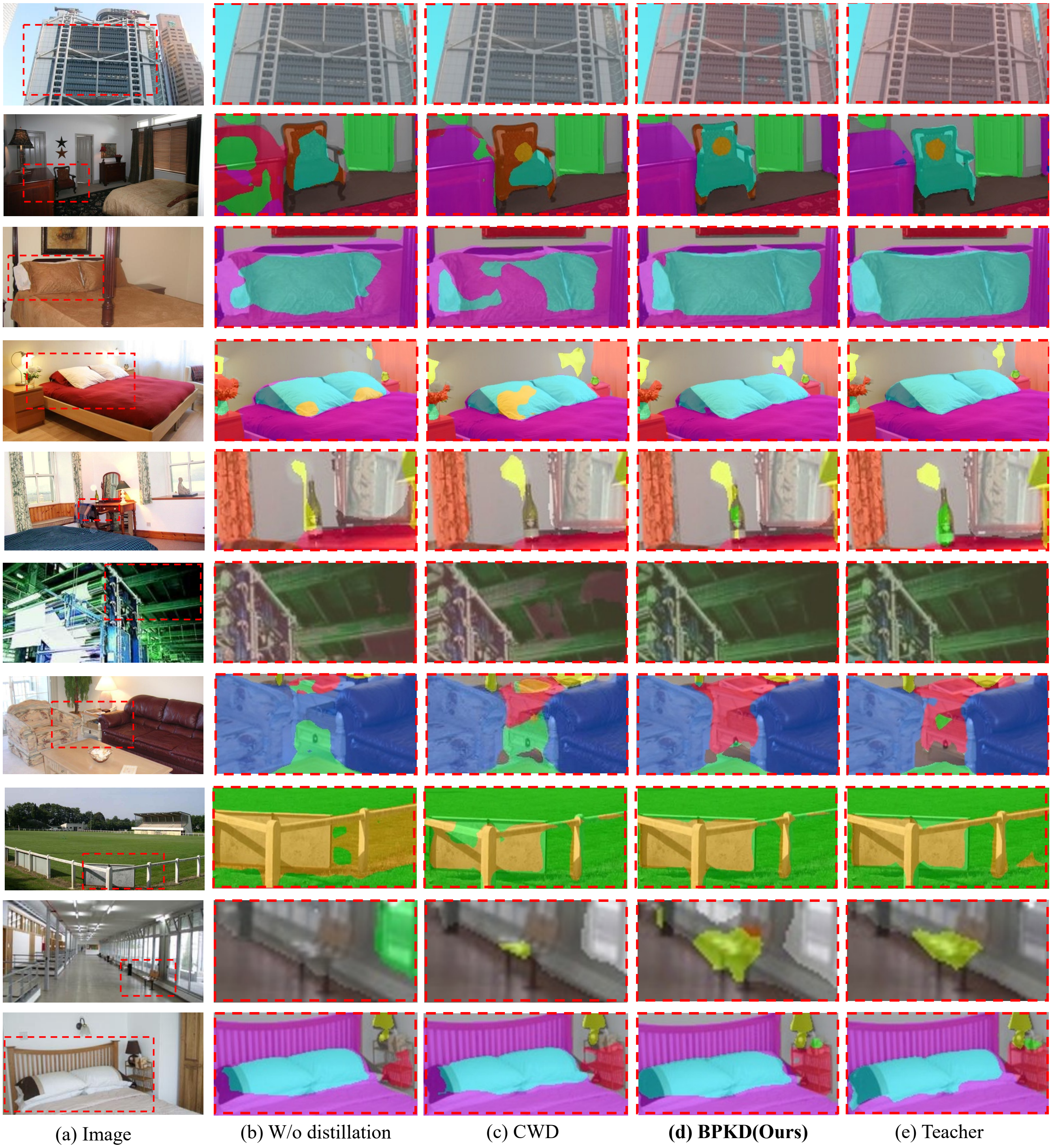

Figure 3. Qualitative results on the ADE20K [12] validation set produced by PSPNet [11] and ResNet18 network architecture: (a) Initial images, (b) w/o distillation scheme, (c) sate-of-the-art method channel-wise distillation [7], (d) **BPKD** our method, (e) teacher. This figure shows that our methods segment the small complex objects with explicit boundaries. Zoom in for a better view. In the second row, BPKD demonstrates superior segmentation results; for instance, the contour of the chair is more distinct and the overall accuracy is enhanced. In the third, fourth, and last rows, the contours of the pillows are clearly delineated, coming closer to the results generated by the Teacher. In the seventh row, the coffee table situated between the two chairs is rendered with greater clarity, thereby suggesting that our method achieves commendable performance in resolving the ambiguity among multiple proximate objects. These visual outcomes indicate that our approach not only improves semantic boundaries but also leverages prior knowledge of contours and shapes to produce outstanding segmentation results.

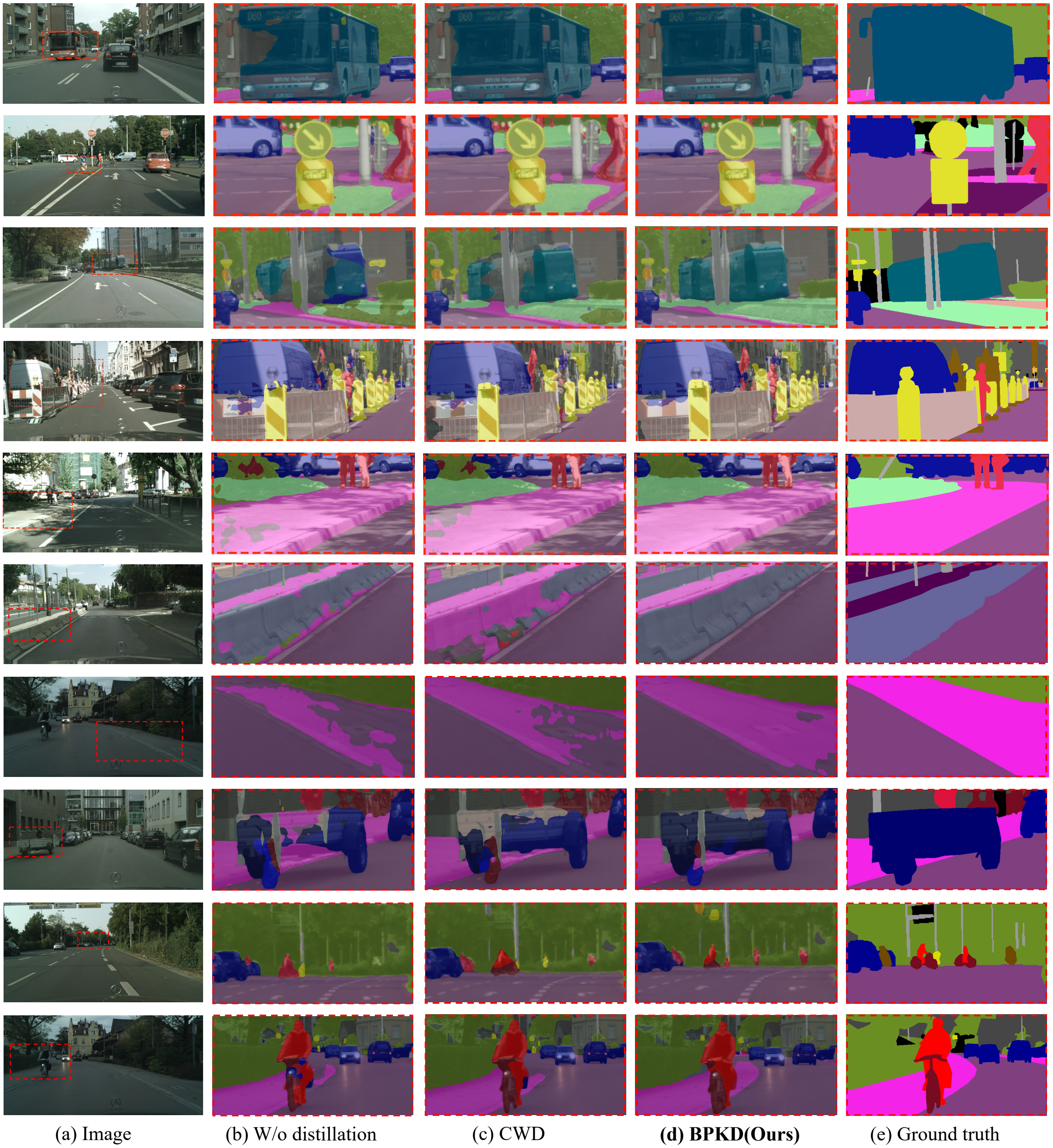

Figure 4. Qualitative results on the Cityscapes [1] validation set produced by PSPNet [11] and ResNet18 network architecture: (a) Initial images, (b) w/o distillation, (c) sate-of-the-art method channel-wise distillation [7], (d) **BPKD** our method, (e) Ground truth. This figure shows that our methods segment the small complex objects with explicit boundaries. Zoom in for a better view. In the first row, BPKD effectively addresses the issue of the front windshield of the bus being wrongly segmented into multiple classes. In the second row, the approach enhances the contours of the upper and lower sections of the road signs. The third row presents a particularly striking result in the segmentation of a train; through differentiated supervision and distillation of the edges and main body, the train is clearly segregated from obstructions. In the fourth row, distant pedestrians present a challenging case for segmentation and identification due to their far-off camera angle and significant occlusions; despite these complexities, BPKD still manages to produce stable edges around the human figures. In the remaining images, BPKD consistently shows improvements, achieving commendable segmentation outcomes on both pedestrian paths and roadways.

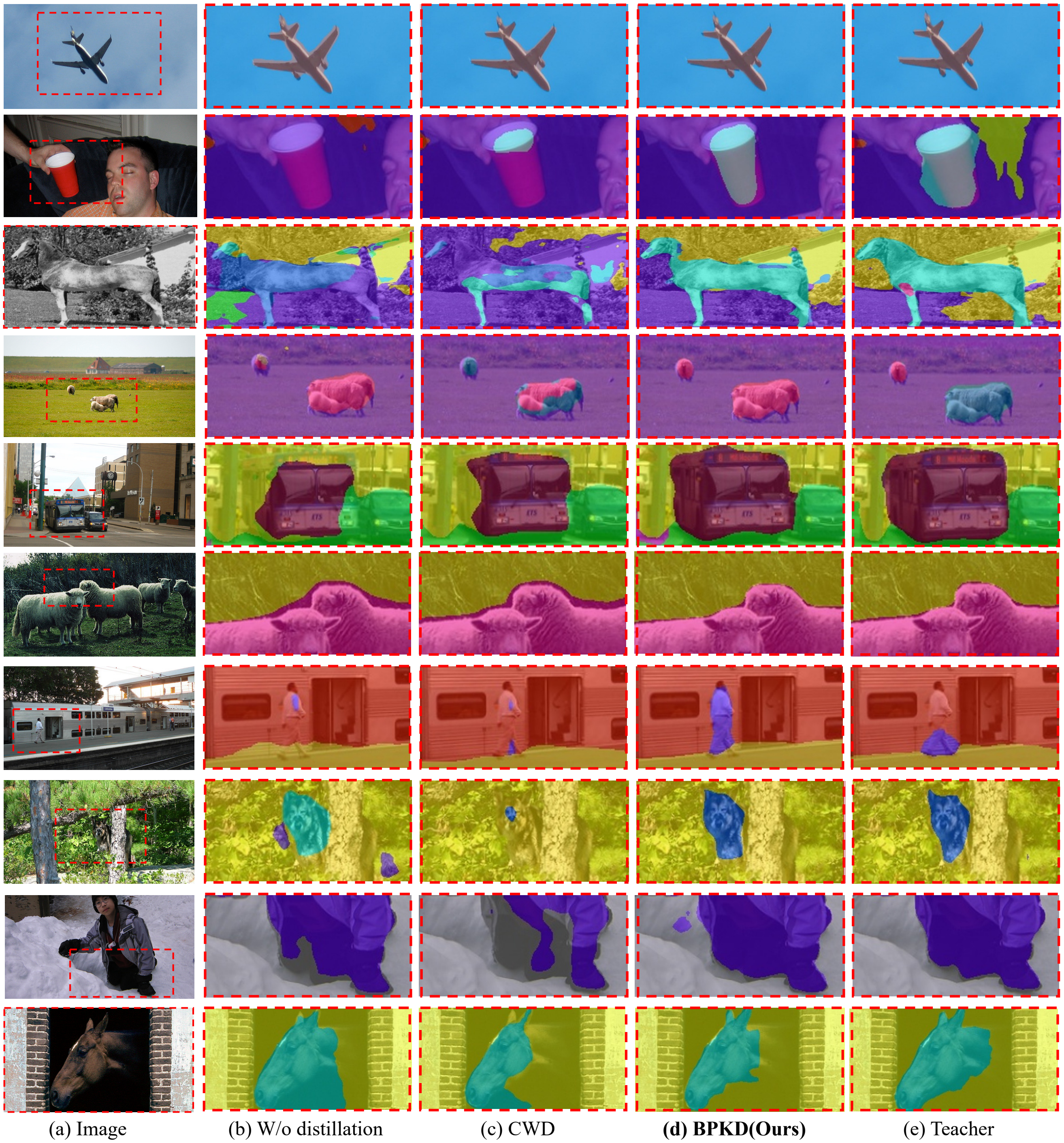

Figure 5. Qualitative results on the Pascal-Context [6] validation set produced by PSPNet [11] and ResNet18 network architecture: (a) Initial images, (b) w/o distillation scheme, (c) sate-of-the-art method channel-wise distillation [7] , (d) **BPKD** our method, (e) teacher. This figure shows that our methods segment the small complex objects with explicit boundaries. Zoom in for a better view. In the first row, BPKD exhibits more fine-grained edge segmentation compared to the non-distilled student network. In the second row, neither the student network nor CWD manages to achieve effective segmentation of the cup, owing to complex angles and environmental factors. Our approach, however, accomplishes accurate segmentation by leveraging prior knowledge of the cup's boundary and supervised learning on the main body of the cup. In the remaining examples, BPKD demonstrates more intricate and advanced segmentation results, thereby revealing that under the influence of edge loss, the entire network's convergence space benefits from shape and spatial priors across different classes. This, in turn, implicitly grants better main-body supervision and assistance, ultimately achieving state-of-the-art (SOTA) distillation results in segmentation.