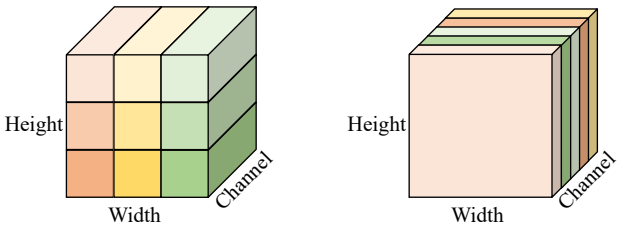
Supplementary Material for Holistic Weighted Distillation for Semantic Segmentation



(a) Spatial-wise Knowledge (b) Channel-wise Knowledge

Fig. 1. Schematic diagram of two categories of knowledge. A cube represents the transferred knowledge unit.

TABLE I
HYPER-PARAMETERS SETTINGS IN THE EXPERIMENTS.

Dataset	Student	μ	β
ADE20K	DeepLab-Res18 DeepLab-MBV2 PSPNet-Res18	0	0.1
Cityscapes	DeepLab-Res18	1	
COCO-Stuff	DeepLab-Res18	0.1	

I. Why Network-based Weight Calculation Fails?

The loss of HWD is

$$L_{\text{HWD}} = \frac{\tau^2}{C}\sum_{c=1}^{C} w_c \sum_{i=1}^{HW} p_{c,i}^{T} \log \frac{p_{c,i}^{T}}{p_{c,i}^{S}}. \tag{1}$$

If we take it as the optimization goal, then for the weight network, it only needs to assign the whole weight to channel with the smallest KL loss, which is obviously inappropriate. Other optimization objectives, such as hoping that the weights obtained by the weight network can help to minimize the cross-entropy loss with the label, will introduce approaches such as meta-learning, which will increase training complexity and overhead.

II. Experiment Environment

Ubuntu 18.04 LTS, Python 3.6.13, PyTorch 1.9.0, CUDA 11.4. Models are trained over 4 * NVIDIA GeForce RTX 2080Ti. We observed that the results are stable with multiple runs (the std of 3 runs of CWD obtained using ADE20K and DeepLab-Res18 is 0.101), therefore, to reduce training overhead, we train only once for each method and do not specify a seed, as in CIRKD [1].

III. Implementation Details

The training images are randomly flipped and scaled in the range of 0.5 to 2. After that, images are randomly cropped into 512×512 pixels. For the evaluation images, we do not process them. Following the classical setting for segmentation distillation [2]–[4], we set the batch size to 8 and the initial learning rate to 0.01. We used SGD with momentum 0.9 and weight decay 1e-4 for 40000 iterations, and the learning rate is decayed by $(1 - \frac{iter}{max\_iter})$. Since CIRKD uses a different setting, our results can be different from its original paper [1]. However, we have tried both settings and found similar conclusion (orders of method superiority). We have provided our code, which is built on CIRKD's code, to ensure reproducibility. Hyper-parameters are provided in Table I.

IV. Pre-trained Models

Readers can download the pre-trained teacher models and ImageNet pre-trained student models from https://github.com/winycg/CIRKD provided by CIRKD [1]. Since ADE20K and COCO-Stuff teacher models are not released when we submitted our paper, we trained these models by ourselves. Larger iterations (e.g., 80K, 160K) should be used when training the teacher.

V. Complete Results

The complete results, where additional results for IFVD [3] and DIST [5] is added, are shown in Table II.

References

[1] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang, "Cross-image relational knowledge distillation for semantic segmentation," in *CVPR*, 2022, pp. 12319–12328. 1

[2] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen, "Channel-wise knowledge distillation for dense prediction," in *ICCV*, 2021, pp. 5311–5320. 1

[3] Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu, "Intra-class feature variation distillation for semantic segmentation," in *ECCV*. Springer, 2020, pp. 346–362. 1

[4] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen, "Structured knowledge distillation for dense prediction," *TPAMI*, 2020. 1

[5] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu, "Knowledge distillation from a stronger teacher," *arXiv preprint arXiv:2205.10536*, 2022. 1

TABLE II
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART DISTILLATION METHODS OVER VARIOUS STUDENT MODELS ON DIFFERENT DATASETS. **BOLD VALUE** DENOTES THE BEST RESULT.

Dataset	Method	mIoU (%)
ADE20K	T: DeepLab-Res101	42.46
	S: DeepLab-Res18	29.86
	+SKD (TPAMI 2020)	31.05
	+IFVD (ECCV 2020)	29.87
	+CWD (ICCV 2021)	32.90
	+CIRKD (CVPR 2022)	30.47
	+DIST (arXiv 2022)	31.68
	+MGD (ECCV 2022)	30.32
	+HWD (ours)	**33.85**
	S: DeepLab-MBV2	27.16
	+SKD	26.96
	+IFVD	26.03
	+CWD	29.19
	+CIRKD	27.39
	+DIST	28.69
	+MGD	27.39
	+HWD (ours)	**31.07**
	S: PSPNet-Res18	31.64
	+SKD	31.30
	+IFVD	31.58
	+CWD	33.43
	+CIRKD	31.36
	+DIST	32.34
	+MGD	32.10
	+HWD (ours)	**34.47**
Cityscapes	T: DeepLab-Res101	78.07
	S: DeepLab-Res18	71.78
	+SKD	72.79
	+IFVD	72.20
	+CWD	73.57
	+CIRKD	72.23
	+DIST	71.98
	+MGD	71.79
	+HWD (ours)	**74.72**
COCO-Stuff	T: DeepLab-Res101	35.31
	S: DeepLab-Res18	26.05
	+SKD	25.94
	+IFVD	25.14
	+CWD	26.17
	+CIRKD	25.85
	+DIST	26.51
	+MGD	26.44
	+HWD (ours)	**26.76**