A Appendix

A.1 More implementation details

We take the images from PASCAL VOC 2012 [11] [1], SBD [18] [2], and Cityscapes [8] [3]. The Cityscapes dataset is processed with these scripts [4].

A.2 More analysis on representation knowledge transmission

The representation knowledge transmission in our gentle teaching assistant is conducted merely on the feature extractor. In our main paper, we view the network backbone (*e.g.* ResNet-101 [19]) in the segmentation model as the feature extractor and the decoder (*e.g.* DeepLabv3++) as the mask predictor. Meanwhile, there are other variants of such a division, *i.e.*, taking fewer or more layers as the feature extractor and all the remaining layers as the decoder. Here, we present the experimental results when taking these divisions.

Table 10: Results on PASCAL VOC 2012, original training set, with different divisions on feature extractor and mask predictor in our method. We use ResNet-101 as the backbone and DeepLabv3+ as the decoder. We report the structure and parameter of the feature extractor and the mask predictor and denote the stem layer in ResNet-101 as layer0 for clarity. The experimental settings follow Table 4.

Method	Feature Extractor		Mask Predictor		mIoU
	Structure	Param (M)	Structure	Param (M)	
Ours	ResNet-101 + Decoder.feature layers	60.9	Decoder.classifier	3.6	70.67
	ResNet-101 (main paper)	42.7	Decoder (main paper)	21.8	**73.16**
	ResNet-101.layer0,1,2,3	27.7	Decoder + ResNet-101.layer4	36.8	68.41
	ResNet-101.layer0,1,2	1.5	Decoder + ResNet-101.layer3,4	63.0	66.23
	ResNet-101.layer0,1	0.3	Decoder + ResNet-101.layer2,3,4	64.2	62.11
	ResNet-101.layer0	0.1	Decoder + ResNet-101.layer1,2,3,4	64.4	60.88
Original EMA	ResNet-101 + Decoder	64.5	-	-	64.07
SupOnly	-	-	ResNet-101 + Decoder	64.5	54.92

We note that when taking the whole ResNet-101 and decoder as the mask predictor (the last row in Table 10), our method shrinks to the model trained only on supervised data (SupOnly). And when they both act as the feature extractor, our representation knowledge transmission boils down to the original EMA update in [42]. From Table 10, we can observe that compared to SupOnly, conducting representation knowledge transmission consistently brings about performance gains. And when taking suitable layers (the first four rows in 'Ours'), our method can achieve better performance than the original EMA. Among them, the most straightforward strategy (also the one in our main paper), which considers ResNet-101 as the feature extractor and decoder as the mask predictor, boasts the best performance.

These experimental results demonstrate that 1) utilizing unlabeled data is crucial to semi-supervised semantic segmentation, 2) transmitting all the knowledge learned from the pseudo labels will mislead the model prediction, 3) our method, which only conveys the representation knowledge in the feature extractor, can alleviate the negative influence of unreliable pseudo labels, making use of unlabeled data in a better manner.

[1] http://host.robots.ox.ac.uk/pascal/VOC/voc2012/

[2] http://home.bharathh.info/pubs/codes/SBD/download.html

[3] https://www.cityscapes-dataset.com/

[4] https://github.com/mcordts/cityscapesScripts