Supplemental Material for "Semi-supervised Domain Adaptation based on Dual-level Domain Mixing for Semantic Segmentation"

Shuaijun Chen[1†], Xu Jia[2†‡] Jianzhong He[1,3], Yongjie Shi[1,4], Jianzhuang Liu[1]
[1]Noah's Ark Lab, Huawei Technologies. [2]Dalian University of Technology.
[3]Data Storage and Intelligent Vision Technical Research Dept, Huawei Cloud.
[4]Key Lab of Machine Perception, Peking University
{chenshuaijun,jianzhong.he,shiyongjie2,liu.jianzhuang}@huawei.com,xjia@dlut.edu.cn

Abstract

In the supplemental material, we further conduct more experiments to demonstrate the effectiveness of proposed framework. We first show the results of our method and state-of-the-art unsupervised domain adaptation (UDA) and semi-supervised learning (SSL) methods on another synthetic-to-real benchmark. Secondly, more ablation studies are reported.

1. More Experiments

1.1. Datasets

Synscapes [8] is another photorealistic synthetic dataset for street scene parsing, which contains 25,000 RGB images with the resolution of 1440×720. Synscapes is designed to be similar in structure and content to the real-world Cityscapes dataset [1], and it includes all 19 training classes for semantic segmentation in Cityscapes. To further verify the effectiveness of our method, we use the entire synthetic dataset as another source domain and consider 19 common categories to train our models on Synscapes to Cityscapes benchmark.

1.2. Implementation Details

Architecture. As the description in the main paper, we also utilize the DeepLabV2 with ResNet101 as the segmentation model. In detail, following [6], we also adopt the multi-level adaptation architecture, which contains two additional ASPP modules on the last two convolutional layers, for fair comparison.

Training Details. During training, all the models are trained 250,000 iterations and early stopped at 120,000 iterations. Iterative rounds R is set to 3 on Synscapes to Cityscapes.

Table 1. Semantic segmentation performance comparison with the state-of-the-art UDA, SSL and SSDA methods on Synscapes→Cityscapes. 19-class mIoU (%) score are reported on Cityscapes validation set across 0, 100, 200, 500, 1000, 2975 numbers of labeled target images. "*" denotes our reimplementation on corresponding numbers of labeled Cityscapes images. Synscapes images are not introduced for implementing SSL methods. Best results are **highlighted**.

Type	Methods	Labeled target images					
		0	100	200	500	1000	2975
UDA	AdaptSeg* [6]	51.3	-	-	-	-	-
	Advent* [7]	**51.6**	-	-	-	-	-
Supervised	DeeplabV2	-	41.9	47.7	55.5	58.6	65.3
SSL	CutMix* [3]	-	50.8	54.8	61.7	64.8	-
	DST-CBC* [2]	-	48.7	54.1	60.6	63.2	-
SSDA	Baseline	-	57.3	58.1	61.5	63.9	67.4
	MME* [5]	-	56.6	57.1	60.6	63.1	67.9
	Ours	-	**62.0**	**62.5**	**65.1**	**68.2**	**71.0**

Table 2. The results of students trained on single-teacher and multi-teacher knowledge distillation method. "SL" and "RL" denote the teacher model trained on sample-level mixed data and region-level mixed data, respectively. E means ensemble operation of two domain-mixed teachers. All the results are obtained at first round on GTA5→Cityscapes.

Model		100	200	500	1000	2975
SL	$\mathcal{M}_{SL}^{1}$	53.9	54.4	58.4	61.7	65.8
	$\mathcal{M}_{S}^{1}$	55.9	56.2	61.5	64.5	68.1
RL	$\mathcal{M}_{RL}^{1}$	53.5	56.6	61.7	65.4	68.2
	$\mathcal{M}_{S}^{1}$	54.8	57.1	61.9	65.3	69.6
SL & RL	$E(\mathcal{M}_{SL}^{1}, \mathcal{M}_{RL}^{1})$	56.2	57.5	62.3	65.8	69.1
	$\mathcal{M}_{S}^{1}$	57.1	58.3	62.6	65.5	69.8

1.3. Results on Synscapes to Cityscapes

We show the results of our methods and several state-of-the-art methods on Synscapes to Cityscapes in Table 1.

[†] Equal contribution

[‡] Part of this work was done while he was in Noah's Ark Lab

Table 3. The detailed results of student model during different rounds through vanilla self-training and our proposed progressive improving scheme on GTA5→Cityscapes.

Number	100				200				500				1000			
Rounds R	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Vanilla $\mathcal{M}_S^r$	57.1	56.7	55.1	53.3	58.3	57.9	57.6	56.5	62.5	60.8	59.3	58.4	65.5	62.8	61.6	60.3
Ours $\mathcal{M}_S^r$	57.1	59.8	61.0	61.2	58.3	60.2	60.3	60.5	62.5	63.7	64.1	64.3	65.5	66.0	66.6	66.0

From Table 1, our approach obtains superior results on all ratios of labeled data compared with UDA and SSL methods on Synscapes to Cityscapes. Due to the similarity of style and content between these two datasets, significant performance improvement can be obtained by our method. It is noteworthy that our method achieves 71.0% mIoU when using full data in target domain.

2. More Ablation Studies

2.1. Single-teacher VS. Multi-teacher

In our proposed framework, a good student can be obtained by distilling knowledge from multi domain-mixed teachers, *i.e.*, teachers trained on sample-level and region-level mixed data. Here, we compare the results of students via different knowledge distillation from one single teacher and multi teachers. We just run first round of our iterative framework on GTA5 [4] to Cityscapes, and the results are shown in Table 2. From Table 2, one best student model is achieved by our multi-teacher knowledge distillation framework with a large performance gain at 100 and 200 labeled images. Thus more accurate pseudo labels generated by student model can promote the next round training of teachers. However, at 500, 1000 and 2975 labled images, the multi-teacher knowledge distillation has the weak advantage compared with single region-level teacher. We argue that compared with full labeled data, such a lot of labeled images will provide enough information especially for region-level data mixing to train a better teacher network. The rest of unlabeled target images cannot provide extra information for further improving the student model.

2.2. Vanilla Self-training VS. Progressive Improving Scheme

Self-training is proposed to address the scarceness of labeled training data and successfully used in UDA and SSL tasks. Vanilla self-training aims to generate pseudo labels of unlabeled data by one model and leverage them to retrain this model. We instead use the pseudo labels to train two stronger teachers. To further demonstrate the advantage of progressive improving scheme between domain-mixed teachers and student, we conduct the vanilla self-training method on the student model obtained at first round on GTA5 to Cityscapes. In experiments, the portion of selected pseudo labels and the confidence threshold are kept same and set to 0.5 and 0.9 respectively. Table 3 shows the performance comparison between different self-training strategies. As the number of rounds increases, the performance of the student model obtained by vanilla self-training deceases. We explain that the initial student cannot be further improved through the pseudo labels generated by itself in our framework. The key of self-training is by generating pseudo labels of unlabeled data to further improve performance of model. However, the initial student model for self-training in our framework is obtained through the supervision of soft labels generated by ensemble of multi teachers on labeled and unlabeled target data, *i.e.*, this supervsion of pseudo- or soft- label mechanism has been used in the process of obtaining the student model. In addition, soft label has the more robustness ability than pseudo label because wrong pixels usually existing in pseudo label. Thus the vanilla self-training will lead to the performance drop through the pseudo labels generated by itself. However, in the progressive improving scheme, we instead use the pseudo labels for training two domain-mixed teachers. Due to accurately labeled ground truths in source domain images, the wrong pixels in pseudo labels has less impact after two kinds of data mixing methods.

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. 1

[2] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum. *arXiv preprint arXiv:2004.08514*, 2020. 1

[3] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 1

[4] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 2

[5] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8050–8058, 2019. 1

[6] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmenta-

tion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 1

[7] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2517–2526, 2019. 1

[8] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint arXiv:1810.08705*, 2018. 1