Supplementary Material for Multi-Source Domain Adaptation with Collaborative Learning for Semantic Segmentation

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

Abstract

In this supplementary material, we report more experimental results that adapting GTA5 [3] + Synscapes [7] to IDD [5] and Mapillary [2] respectively. First, we give a description of the datasets. And then, report the performance comparison between the reproduced AdaptSeg [4], Advent [6] and our proposed method. Moreover, we also evaluate the performance of collaborative learning between source on different target datasets (IDD and Mapillary). The results further validate the effectiveness of our proposed method.

1. More Experiments

1.1. Datasets

Mapillary and IDD are another two widely used benchmarks for autonomous driven scene. They are have more images sampled from more various scenes. Tab. 1 shows the statistics comparison of different datasets.

Mapillary Vistas dataset (**M**) is a large-scale diverse street-level image dataset that containing 25,000 high resolution images with densely pixel-level annotated into 66 object categories. It is designed and compiled to cover diversity, richness of detail and geographic extent. The images are from all around the world, captured at various conditions regarding weather, season and daytime. Moreover, these images come from different imaging devices (mobile phones, tablets, action cameras, professional capturing rigs) and differently experienced photographers. To evaluation our proposed method, we train models with the common 19 categories with Cityscapes [1] training labels.

IDD (India Driving Dataset) [5] (**I**) consists of 20,000 images, which are obtained from a front facing camera attached to a car and finely annotated with 34 classes collected from 182 drive sequences on Indian roads. Most of images are 1080p resolution with some are 720p. Their label set is expanded in comparison to Cityscapes [1], to account for new classes. We train all the models based on the common 19 classes with Cityscapes for adaptation setting. Note that, IDD has another 10k version and here we use thus 20k version one for evalutaion of our proposed method.

Table 1. The comparison of different datasets for semantic segmentation in autonomous driving.

Dataset	Num. of Images	Num. of Scenes	Cats. (Train/All)	Avg. Resolution
Cityscapes [1]	5K	50	19/30	2048×1024
Mapillary [2]	25K	–	19/66	≥1920×1080
IDD [5]	20k	180	19/34	1678×968

Table 2. The domain generalization ability comparison of Collaborative Learning Between Sources (Co-Learning-Src) with baseline and domain generalization method.

GTA5+Synscapes		
Method	Target	mIoU
Data Combination	I	47.06
MLDG+TN [8]		47.42
Co-Learning-Srcs		47.80
Data Combination	M	46.64
MLDG+TN [8]		47.11
Co-Learning-Srcs		47.16

1.2. Results

Tab. 2 shows the performance comparison of proposed collaborative learning between sources trained on the original images which is not translated with baseline that simple combination and domain generalization method MLDG [8]. From the results, we can see that our proposed collaborative learning can achieve better or comparable performance compared with the state-of-the-art domain generalization method. For example, we achieve 47.80% and 47.16% on the IDD and Mapillary dataset, respectively. Both of them are better or comparable to the MLDG.

Tab. 3 shows the comparison of i): the reproduce of AdaptSeg [4] and Advent [6] that adapting from GTA5,

Table 3. The quantitative results that adapting from GTA5 + Synscapes to IDD and Mapillary respectively. Here, Our-M* means the performance of model $\mathcal{M}_{S_*}$, and Ours-Ensemble means the results that ensemble of all outputs of models $\mathcal{M}_{S_*}$. † means training our proposed approach with stage-wise.

Methods	Source	Target	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
DT			80.5	7.8	51.1	17.8	6.4	23.4	4.0	22.4	77.5	9.2	90.4	41.4	37.3	68.6	32.0	27.9	0.0	55.7	18.6	35.37
AdaptSeg [4]	S		92.5	19.4	58.1	23.2	8.9	20.4	5.0	25.7	77.2	9.5	93.9	49.6	42.7	72.0	37.1	30.6	0.0	59.6	20.0	39.23
Advent [6]			93.2	19.5	59.1	21.9	8.4	23.9	5.6	24.8	79.1	9.4	94.7	48.2	40.2	71.4	37.1	29.7	0.0	58.9	21.3	39.28
DT			90.2	27.9	56.3	23.4	20.4	27.8	4.9	26.0	74.4	29.6	87.8	46.4	39.1	65.1	47.3	36.6	0.0	49.1	26.9	41.01
AdaptSeg [4]	G		92.8	21.4	64.7	25.0	23.3	26.9	6.0	40.7	76.7	30.5	92.5	45.7	34.0	70.9	50.5	37.5	0.0	47.6	26.2	42.78
Advent [6]			93.0	25.1	66.2	31.9	22.3	29.1	10.0	38.1	73.7	26.4	93.2	49.4	43.2	72.1	52.5	40.0	0.0	50.7	26.6	44.40
DT			92.2	19.1	66.0	32.1	19.4	29.4	9.5	45.1	80.3	35.7	94.8	59.4	40.5	76.4	49.3	46.6	0.0	59.9	38.4	47.06
AdaptSeg [4]		IDD	92.0	18.9	66.2	23.9	17.6	30.6	5.8	45.8	81.7	30.1	94.4	57.3	47.5	75.2	51.5	53.6	0.0	58.9	35.4	46.65
Advent [6]			93.9	28.8	68.2	32.1	20.0	32.1	8.8	44.9	77.1	23.1	95.0	58.8	47.1	74.3	57.4	49.4	0.0	61.0	32.8	47.61
Ours-M1			95.4	38.5	70.0	36.7	21.2	25.0	14.2	43.9	78.6	28.5	94.8	58.9	45.0	70.8	56.1	48.3	0.0	63.4	38.8	48.86
Ours-M2	S+G		95.1	35.2	71.2	39.0	19.3	27.2	11.5	48.1	77.8	26.3	95.3	57.6	39.2	69.7	52.2	46.1	0.0	60.0	34.0	47.63
Ours-Ensemble			95.8	41.8	72.9	39.5	21.5	26.4	18.2	44.5	78.1	28.1	95.5	62.2	43.0	70.6	58.9	49.5	0.0	63.5	38.9	49.94
Ours-M1†			95.6	39.6	71.5	38.4	19.9	30.1	12.8	47.8	78.3	31.5	95.3	55.6	47.5	74.6	48.9	54.9	0.0	64.5	39.9	49.83
Ours-M2†			95.3	37.5	71.5	36.4	21.1	31.2	13.1	44.6	79.4	33.0	95.2	55.4	46.9	73.4	51.6	44.8	0.0	64.8	41.5	49.30
Ours-Ensemble†			95.8	39.9	73.1	38.8	21.0	31.0	14.1	43.8	78.2	32.2	95.5	58.2	47.2	74.2	52.6	50.7	0.0	65.8	41.4	50.19
DT			70.4	23.6	63.6	14.8	12.0	25.8	30.7	32.7	75.2	41.2	89.4	36.2	22.0	73.0	19.5	17.2	0.2	27.7	31.1	37.18
AdaptSeg [4]	S		85.9	24.2	73.2	17.7	27.4	26.4	33.0	39.0	75.4	44.6	94.3	34.7	27.8	77.4	25.8	16.5	1.2	29.9	31.2	41.35
Advent [6]			86.2	23.9	74.6	17.8	26.8	29.5	35.9	39.8	79.4	43.6	96.2	37.3	27.5	78.4	26.3	16.1	1.4	29.1	29.1	42.04
DT			82.2	28.6	74.2	23.4	27.2	35.3	36.4	18.6	73.8	29.2	89.6	58.9	39.2	74.5	35.0	17.2	12.5	31.3	27.8	42.89
AdaptSeg [4]	G		86.5	31.6	78.2	24.6	30.0	36.1	35.8	31.6	73.4	33.2	93.7	59.2	44.5	78.6	41.2	39.3	14.8	36.5	32.3	47.44
Advent [6]			86.6	28.3	77.9	24.7	30.6	36.1	36.0	32.5	75.8	34.9	94.4	58.8	44.1	79.9	41.3	42.3	15.7	35.6	32.6	47.79
DT			77.7	30.9	75.2	27.0	27.5	33.4	37.2	37.3	76.9	43.1	93.3	55.8	38.0	72.5	38.4	40.2	2.8	36.9	42.3	46.64
AdaptSeg [4]		Mapillary	84.2	33.4	78.0	27.9	34.0	38.0	41.6	39.4	78.6	34.5	92.7	46.9	41.6	81.9	38.3	39.0	3.6	41.5	40.5	48.19
Advent [6]			87.2	36.2	78.0	27.1	31.2	38.4	40.8	40.2	80.8	44.2	96.0	47.1	43.5	82.3	39.0	39.3	5.0	42.0	40.3	49.40
Ours-M1			88.2	32.5	81.0	29.1	37.5	39.9	41.7	39.6	80.4	44.6	95.8	58.7	40.2	83.1	48.1	40.7	2.3	40.1	43.2	50.89
Ours-M2	S+G		87.8	31.6	81.0	30.0	37.8	34.8	38.3	41.3	78.1	39.1	95.1	60.1	49.5	82.2	42.7	39.0	19.2	45.9	48.0	51.67
Ours-Ensemble			88.5	34.3	81.9	31.9	41.1	39.0	40.1	41.5	79.7	45.0	95.7	62.7	51.1	83.3	49.9	45.9	8.5	46.4	47.5	53.37
Ours-M1†			87.5	40.1	80.9	31.0	37.4	40.0	42.5	40.6	79.6	42.4	95.2	55.5	46.5	84.5	45.1	40.3	16.5	41.6	39.1	51.92
Ours-M2†			88.6	36.5	81.4	29.7	38.2	41.3	43.0	43.4	80.2	45.8	95.6	58.3	43.8	84.5	42.5	42.0	10.1	46.2	43.9	52.37
Ours-Ensemble†			88.4	40.1	81.9	32.4	39.8	41.4	42.2	42.7	80.1	46.4	95.6	58.2	48.5	84.7	46.6	45.5	11.7	46.9	42.4	53.44

Synscapes and combination of GTA5 and Synscapes to IDD and Mapillary, and ii): Direct Transfer from GTA5, Synscapes and GTA5+Synscapes to IDD and Mapillary, and iii): each model and ensemble of our proposed method that adapting from GTA5 + Synscapes to IDD and Mapillary. Note that, the network architecture and hyperparameters for different losses are same as the setting to Cityscapes.

From Tab. 3, we can see that our proposed method achieve the best performance no matter what the target dataset, *i.e.*, achieving 50.19% and 53.44% on IDD and Mapillary respectively. Moreover, directly adopting UDA methods on combined sources data sometimes could not achieve better performance than direct transfer. For example, AdaptSeg only achieves 46.65% when IDD as target domain which is lower the performance of directly transfer based on combined data. All these results further validate the effectiveness of our proposed method.

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] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *International Conference on Computer Vision (ICCV)*, 2017. 1

[3] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen

Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016. 1

[4] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 1, 2

[5] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1743–1751. IEEE, 2019. 1

[6] 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, 2

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

[8] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Generalizable semantic segmentation via model-agnostic learning and target-specific normalization. *arXiv preprint arXiv:2003.12296*, 2020. 1