1 Supplementary

1.1 Target-to-source Translation

Inspired by image-to-image translation networks [24, 14], existing domain adaptation methods [1, 17, 13] translate images from the source domain to the target domain (source-to-target) to reduce pixel-level domain discrepancy. This is achieved by an unsupervised image translation model $\mathcal{F}^{-1}$ such as CycleGAN [24] to learn a mapping $\mathcal{F}^{-1} : \mathcal{X}_s \rightarrow \mathcal{X}_t$. However, such strategy introduces inevitable bias to the translated images $\mathcal{F}^{-1}(\mathcal{X}_s)$, stemming from that $\mathcal{F}^{-1}(\mathcal{X}_s)$ and $\mathcal{X}_t$ cannot be guaranteed to follow the exactly identical distribution through the adversarial learning [8]. This problem can get even worse in the source-to-target translation, as $|\mathcal{X}_s| \gg |\mathcal{X}_t|$ in most of domain adaptation problems. For example, GTA5 [19] (i.e., $\mathcal{X}_s$) contains 24,966 images, while Cityscapes (i.e., $\mathcal{X}_t$) [7] has only 2,975 images. As a consequence, $\mathcal{F}^{-1}(\mathcal{X}_s)$ contains massive amounts of translation bias which will further induces negative effects when adapting domain knowledge between $\mathcal{F}^{-1}(\mathcal{X}_s)$ and $\mathcal{X}_t$. To alleviate this problem, for the first time, we translate images from the target domain to the source domain through the mapping $\mathcal{F} : \mathcal{X}_t \rightarrow \mathcal{X}_s$ (where $\mathcal{F}$ is the reverse function of $\mathcal{F}^{-1}$). $\mathcal{X}_s$ and $\mathcal{F}(\mathcal{X}_t)$ are then used for further domain knowledge transfer. Doing so significantly reduces the translation bias in the translated images $\mathcal{F}(\mathcal{X}_t)$ and is much more computationally efficient than the source-to-target translation. Another advantage is that the segmentation network can be trained using original source images $\mathcal{X}_s$ and their corresponding labels.

1.2 Network Architecture

The three multi-scale discriminators (i.e., D_1, D_2, and D_3) used in our reconstruction model follow the identical network architecture. Each of them is a 70×70 PatchGAN [10] containing five convolution layers with kernel number {64, 128, 256, 512, 1}. The kernel size for each layer is 4×4. The first three layers use stride 2, while the last two layers with stride 1. A leaky ReLU parameterized by 0.2 is applied to the first four layers. We also apply BatchNorm to each layer, except the first and last one.

1.3 Implementation Details

For multi-scale discriminators, Adam optimizer with initial learning rate 2×10^{-4} and momentum { 0.5, 0.999} are used in our study. The learning rate is linearly decayed to zero with step size 100.

1.4 mIoU Gap

We investigate our model's ability in narrowing the mIoU gap between the model (Oracle) that is trained in a fully-supervised matter. Compared to existing state-of-the-art methods, we significantly recover the performance loss based

Table 1. A performance comparison of our method with other state-of-the-art models on "GTA5 to Cityscapes". The performance is measured by the mIoU gap between each model and the fully-supervised model (Oracle). Two base architectures, i.e., VGG16 (V) and ResNet101 (R) are used in our study.

GTA5 → Cityscapes			
	Base	**mIoU Gap**	**Oracle**
Source only	R	-28.5	65.1
SIBAN [15]	R	-22.5	65.1
CLAN [16]	R	-21.9	65.1
DISE [2]	R	-19.7	65.1
IntraDA [18]	R	-18.8	65.1
BDL [13]	R	-16.6	65.1
CrCDA [9]	R	-16.5	65.1
SIM [22]	R	-15.9	65.1
Kim et al. [11]	R	-14.9	65.1
FDA-MBT [23]	R	**-14.65**	65.1
Ours	R	-15.6	65.1
Source only	V	-46.7	64.6
SIBAN [15]	V	-26.1	60.3
ASN [20]	V	-25.2	61.8
CyCADA [1]	V	-24.9	60.3
CLAN [16]	V	-23.7	60.3
CrDoCo [5]	V	-22.2	60.3
CrCDA [9]	V	-22.7	61.8
BDL [13]	V	-19.0	60.3
FDA-MBT [23]	V	-18.1	60.3
Kim et al. [11]	V	-18.0	60.3
SIM [22]	V	-17.9	60.3
Ours	V	**-16.8**	60.3

on the ResNet101 backbone on GTA5 to Cityscapes. Although we are inferior to [11, 23] which are published simultaneously with our work, we outperforms these two methods on VGG16-based backbone by a large margin (Table 1). Similar improvements can also be observed on the adaptation from SYNTHIA to Cityscapes as shown in Table 2.

1.5 Qualitative Comparison

GTA5→ Cityscapes As shown in Figure 1, we present the qualitative comparison in Cityscapes based on the VGG16 model from GTA5→Cityscapes. Our results reveal the effectiveness of target-to-source translation and joint distribution alignment in adapting cross-domain knowledge.

SYNTHIA → Cityscapes The qualitative comparison for ResNet101 and VGG16 model from SYNTHIA→Cityscapes are showcased in Figure 2 and Figure 3, respectively. Similarly, each component in our framework contributes to the overall performance improvement.

Table 2. A performance comparison of our method with other state-of-the-art models on "SYNTHIA to Cityscapes". The performance is measured by the mIoU gap between each model and the fully-supervised model (Oracle). Two base architectures, i.e., VGG16 (V) and ResNet101 (R) are used in our study.

SYNTHIA → Cityscapes			
	Base	**mIoU Gap**	**Oracle**
Source only	R	—	—
ASN [20]	R	-25.0	71.7
DISE [2]	R	-22.9	71.7
IntraDA [18]	R	-22.8	71.7
Kim et al. [11]	R	-22.4	71.7
DADA [21]	R	-21.9	71.7
CrCDA [9]	R	-21.7	71.7
BDL [13]	R	-20.3	71.7
SIM [22]	R	-19.6	71.7
FDA-MBT [23]	R	-19.2	71.7
Ours	R	**-18.6**	71.7
CrCDA [9]	V	-28.9	64.1
ROAD-Net [4]	V	-27.6	64.1
SPIGAN [12]	V	-22.7	59.5
GIO-Ada [3]	V	-26.8	64.1
TGCF-DA [6]	V	-25.6	64.1
BDL [13]	V	-20.5	59.5
FDA-MBT [23]	V	-19.0	59.5
Ours	V	**-18.4**	59.5

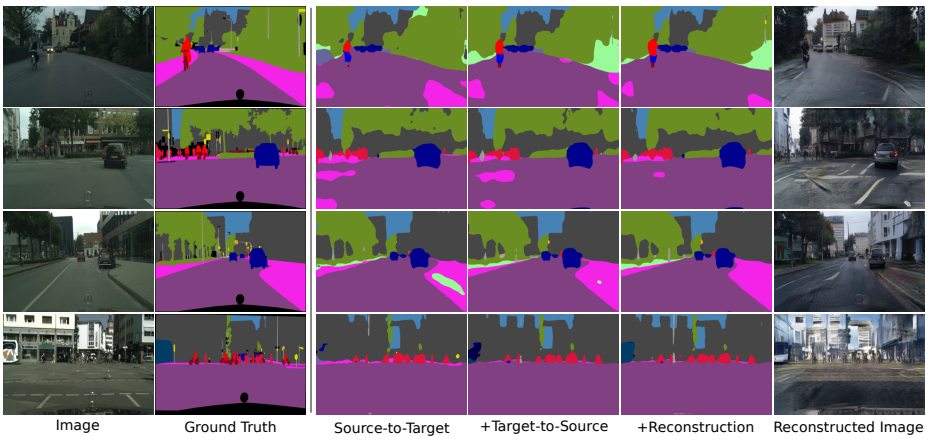

Fig. 1. Qualitative examples of semantic segmentation results in Cityscapes (GTA5→Cityscapes with VGG16). For each target-domain image (first column), its ground truth and the corresponding segmentation output from the baseline model (source-to-target) are given. The following are predictions of our method by incorporating target-to-source translation and reconstruction, together with the reconstructed image.

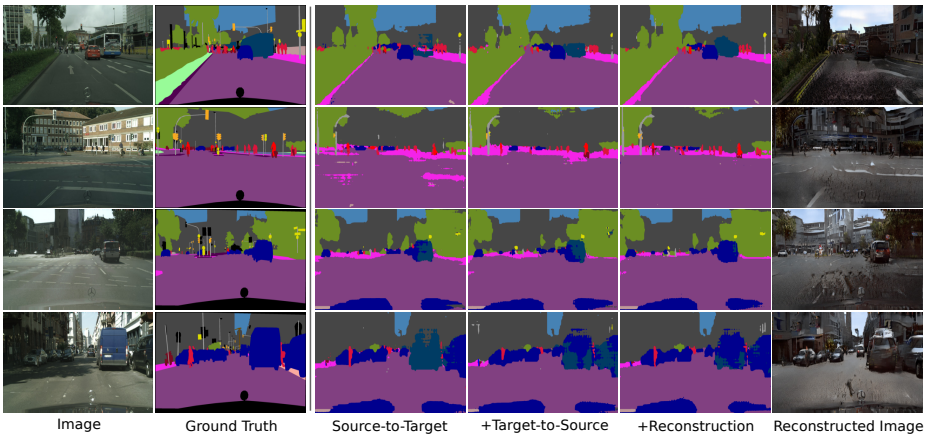

Fig. 2. Qualitative examples of semantic segmentation results in Cityscapes (SYNTHIA→Cityscapes with ResNet101). For each target-domain image (first column), its ground truth and the corresponding segmentation output from the baseline model (source-to-target) are given. The following are predictions of our method by incorporating target-to-source translation and reconstruction, together with the reconstructed image.

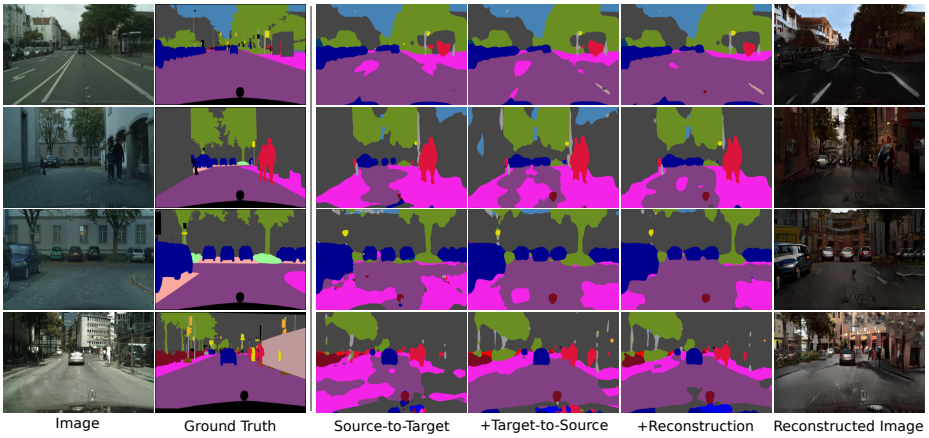

Fig. 3. Qualitative examples of semantic segmentation results in Cityscapes (SYNTHIA→Cityscapes with VGG16). For each target-domain image (first column), its ground truth and the corresponding segmentation output from the baseline model (source-to-target) are given. The following are predictions of our method by incorporating target-to-source translation and reconstruction, together with the reconstructed image.

References

1. Cycada: Cycle consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018.
2. Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1900–1909, 2019.
3. Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1841–1850, 2019.
4. Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7892–7901, 2018.
5. Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1791–1800, 2019.
6. Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2019.
7. 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 (CVPR)*, pages 3213–3223, 2016.
8. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems (NIPS)*, pages 2672–2680, 2014.
9. Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
10. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.
11. Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12975–12984, 2020.
12. Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. Spigan: Privileged adversarial learning from simulation. In *International Conference on Learning Representations (ICLR)*, 2019.
13. Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
14. Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems (NIPS)*, pages 700–708, 2017.

15. Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, October 2019.
16. Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2507–2516, 2019.
17. Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 13, 2018.
18. Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3764–3773, 2020.
19. Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 102–118, 2016.
20. 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 (CVPR)*, 2018.
21. Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2019.
22. Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12635–12644, 2020.
23. Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4085–4095, 2020.
24. Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017.