Learning to Adapt Structured Output Space for Semantic Segmentation

Yi-Hsuan Tsai[1*] Wei-Chih Hung[2*] Samuel Schulter[1] Kihyuk Sohn[1]
Ming-Hsuan Yang[2] Manmohan Chandraker[1,3]
[1]NEC Laboratories America [2]University of California, Merced [3]University of California, San Diego

1. Hyperparameter Analysis

In Table 1, we present hyperparameter analysis on the weights of (5) in the main manuscript for the two-level adversarial domain adaptation in the output space. We note that λ^1_{seg} and λ^1_{adv} correspond to the proposed adaptation module applied on *conv5*, while λ^2_{seg} and λ^2_{adv} correspond to the one applied on *conv4*. First, our single-level adaptation model performs the best with $\lambda^1_{adv} = 0.001$, and hence we use the same value of λ^1_{adv} for the following experiments. Second, to evaluate the impact of adding an auxiliary loss for segmentation in the lower level, we find that a smaller weight of $\lambda^2_{seg} = 0.1$ achieves the best result. While fixing the weights used in the previous steps, we further analyze the impact of the second adversarial loss λ^2_{adv} for our multi-level adaptation module and find that $\lambda^2_{adv} = 0.0002$ performs the best. As a result, we use $\lambda^1_{adv} = 0.001$, $\lambda^2_{seg} = 0.1$, $\lambda^2_{adv} = 0.0002$ as our final weights adopted in the main manuscript.

Table 1. Parameter analysis.

Parameter	λ^1_{seg}	λ^2_{seg}	λ^1_{adv}	λ^2_{adv}	Mean IoU
Baseline (ResNet)	1	0	0	0	36.6
Single-level	1	0	0.0005	0	40.2
	1	0	0.001	0	41.4
	1	0	0.002	0	40.4
	1	0	0.004	0	40.1
Auxiliary loss	1	0	0.001	0	41.4
	1	0.1	0.001	0	41.8
	1	0.2	0.001	0	41.5
	1	0.4	0.001	0	39.4
Multi-level	1	0.1	0.001	0.0001	40.9
	1	0.1	0.001	0.0002	42.4
	1	0.1	0.001	0.0005	42.1
	1	0.1	0.001	0.001	41.1

2. Detailed Hyperparameters for Optimization

In Table 2, we present detailed hyperparameters for optimizing our proposed method for both the generator and the discriminator, including baseline model, feature space adaptation and single/multi-level adaptation in the output space. In addition, we show the image size used for training and testing in different datasets. Note that we always maintain the aspect ratio of the image (without cropping) and down-sample it to the size as in Table 3.

*Both authors contribute equally to this work.

Table 2. Detailed hyperparameters for optimization.

Method	Baseline (ResNet)	Feature	Single-level	Multi-level
Training Iterations	250k	250k	250k	250k
Learning Rate Decay Policy	Poly	Poly	Poly	Poly
Decay Power	0.9	0.9	0.9	0.9
Learning Rate	2.5e-4	2.5e-4	2.5e-4	2.5e-4
Batch Size	1	1	1	1
Optimizer	SGD	SGD	SGD	SGD
Weight Decay	5e-4	5e-4	5e-4	5e-4
Momentum	0.9	0.9	0.9	0.9
Learning Rate ($\mathbf{D}$)	-	1e-4	1e-4	1e-4
Batch Size ($\mathbf{D}$)	-	2	2	2
Optimizer ($\mathbf{D}$)	-	Adam	Adam	Adam
Weight Decay ($\mathbf{D}$)	-	5e-4	5e-4	5e-4
Momentum ($\mathbf{D}$)	-	0.9	0.9	0.9
λ_{seg}^1	1.0	1.0	1.0	1.0
λ_{seg}^2	-	-	-	0.1
λ_{adv}^1	-	0.001	0.001	0.001
λ_{adv}^2	-	-	-	0.0002

Table 3. Image size for training and testing.

Dataset	Cityscapes [2]	GTA5 [3]	SYNTHIA [4]	Cross-City [1]
Training	512×1024	720×1280	760×1280	512×1024
Testing	512×1024	-	-	512×1024

3. Qualitative Comparisons

We present additional results for qualitative comparisons under various settings, including *GTA5* $\rightarrow$ *Cityscapes* (Figure 1 and 2), *SYNTHIA* $\rightarrow$ *Cityscapes* (Figure 3 and 4), and *Cityscapes* $\rightarrow$ *Cross-City* (Figure 5 and 6). In each setting, we show results of the baseline model, feature adaptation and our multi-level adversarial adaptation in the output space. We observe that the proposed domain adaptation method often yields qualitatively better segmentation outputs in the target domain, as compared to feature-level adaptation.

References

[1] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. F. Wang, and M. Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *ICCV*, 2017. 2

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2

[3] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 2

[4] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 2

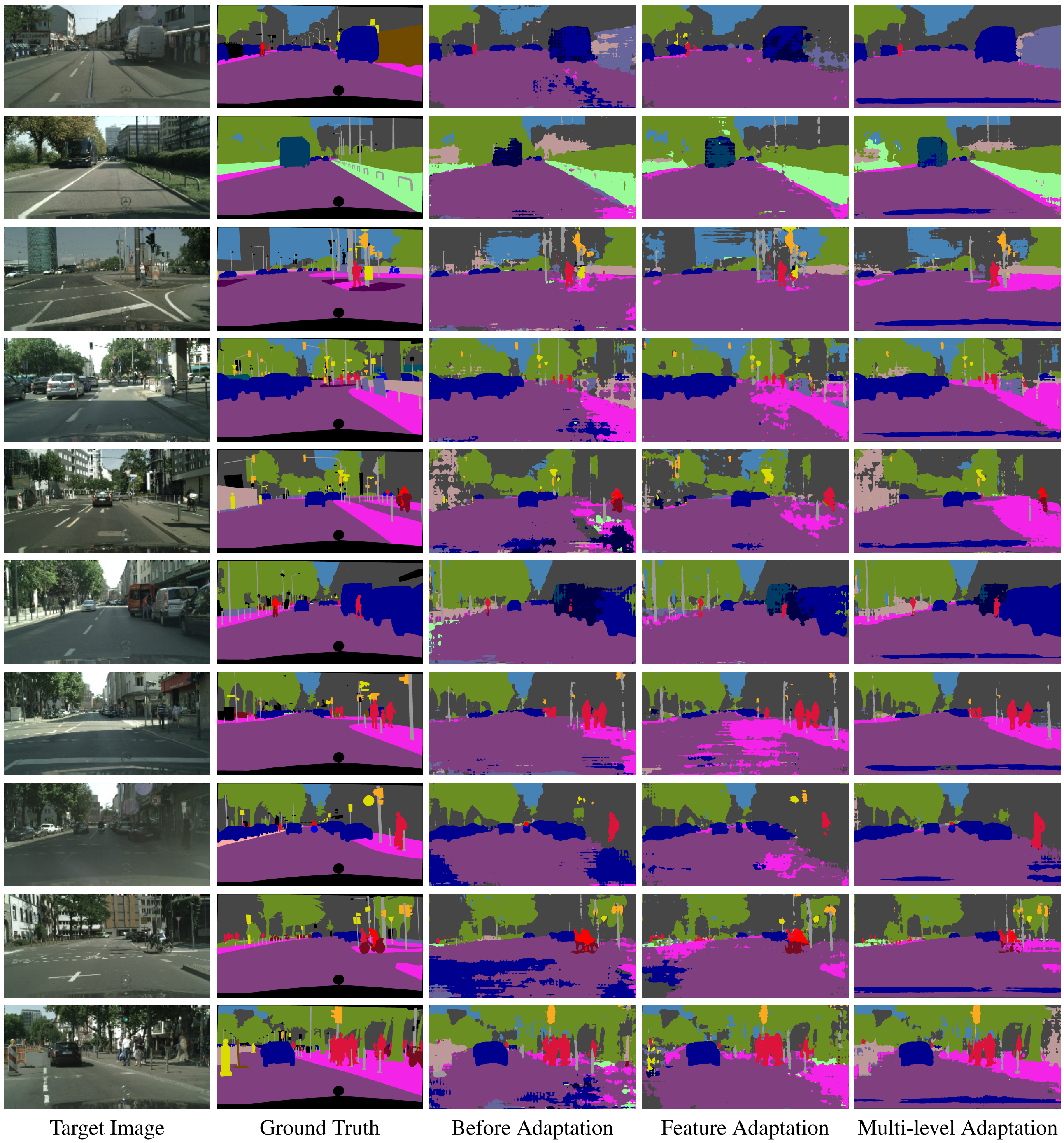
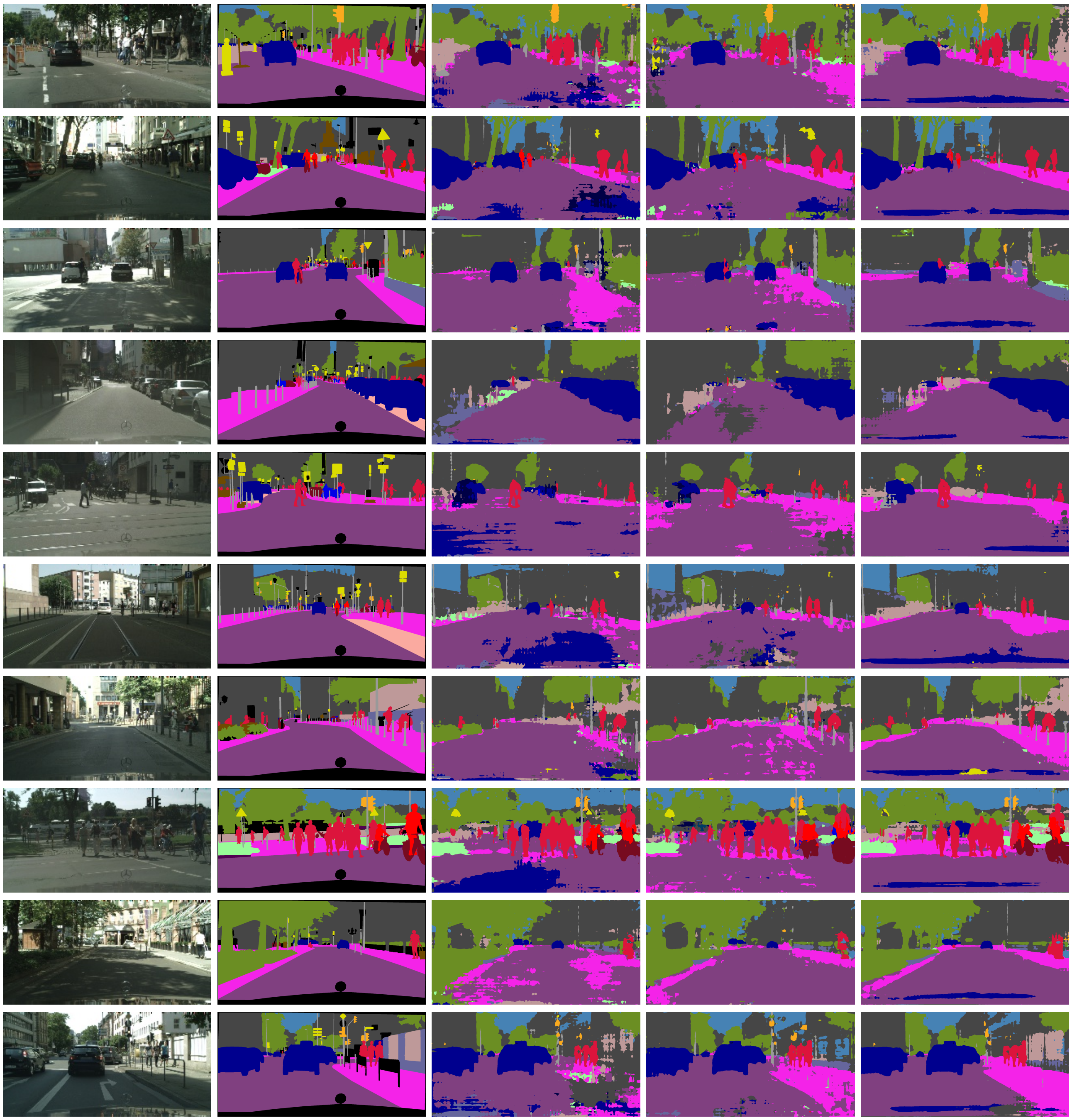


Figure 1. Example results of adapted segmentation for the GTA5-to-Cityscapes dataset. For each target image, we show results before adaptation, with feature adaptation and our adapted segmentations in the output space.



Target Image | Ground Truth | Before Adaptation | Feature Adaptation | Multi-level Adaptation

Figure 2. Example results of adapted segmentation for the GTA5-to-Cityscapes dataset. For each target image, we show results before adaptation, with feature adaptation and our adapted segmentations in the output space.

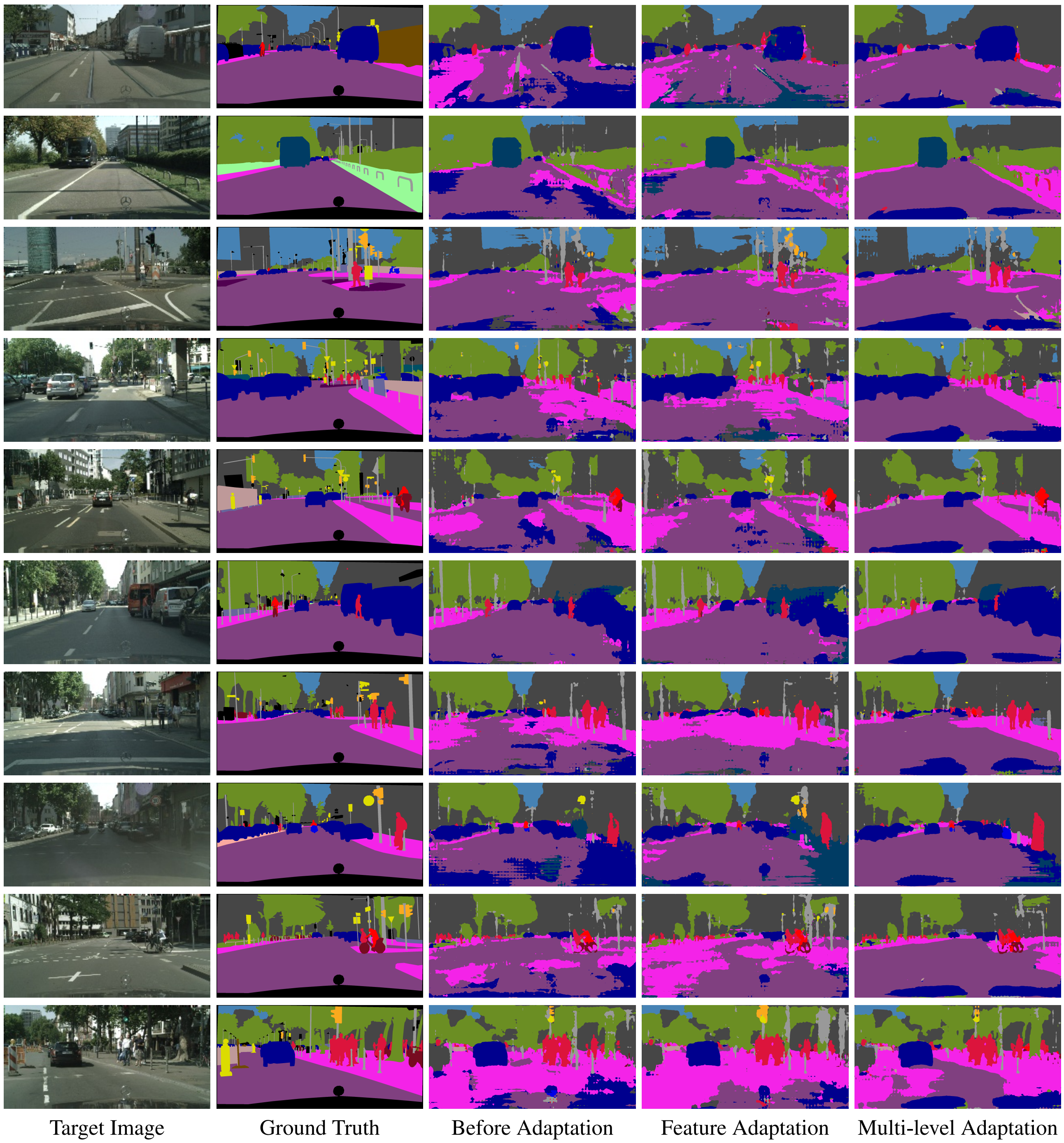


Figure 3. Example results of adapted segmentation for the SYNTHIA-to-Cityscapes dataset. For each target image, we show results before adaptation, with feature adaptation and our adapted segmentations in the output space.

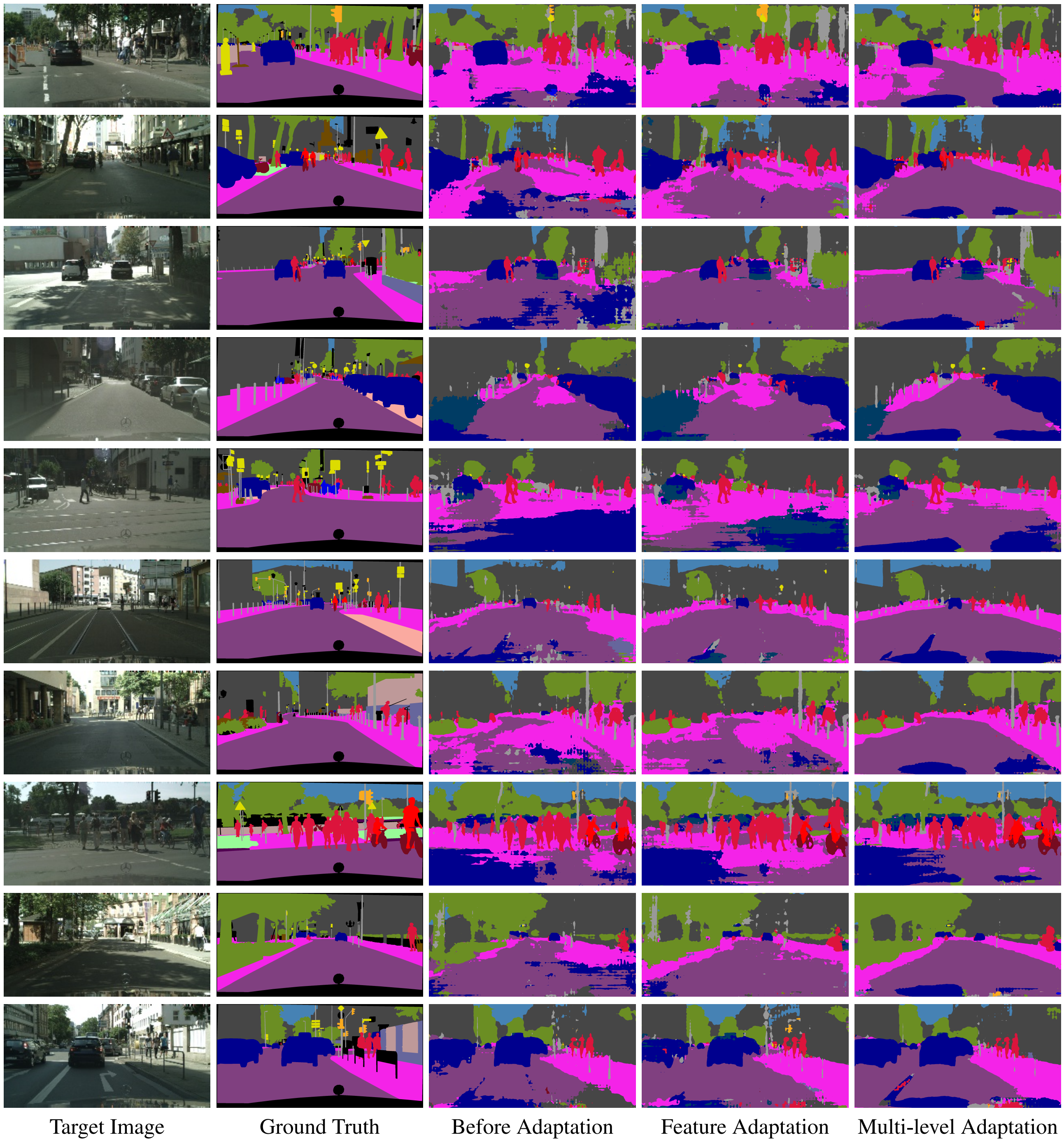


Figure 4. Example results of adapted segmentation for the SYNTHIA-to-Cityscapes dataset. For each target image, we show results before adaptation, with feature adaptation and our adapted segmentations in the output space.

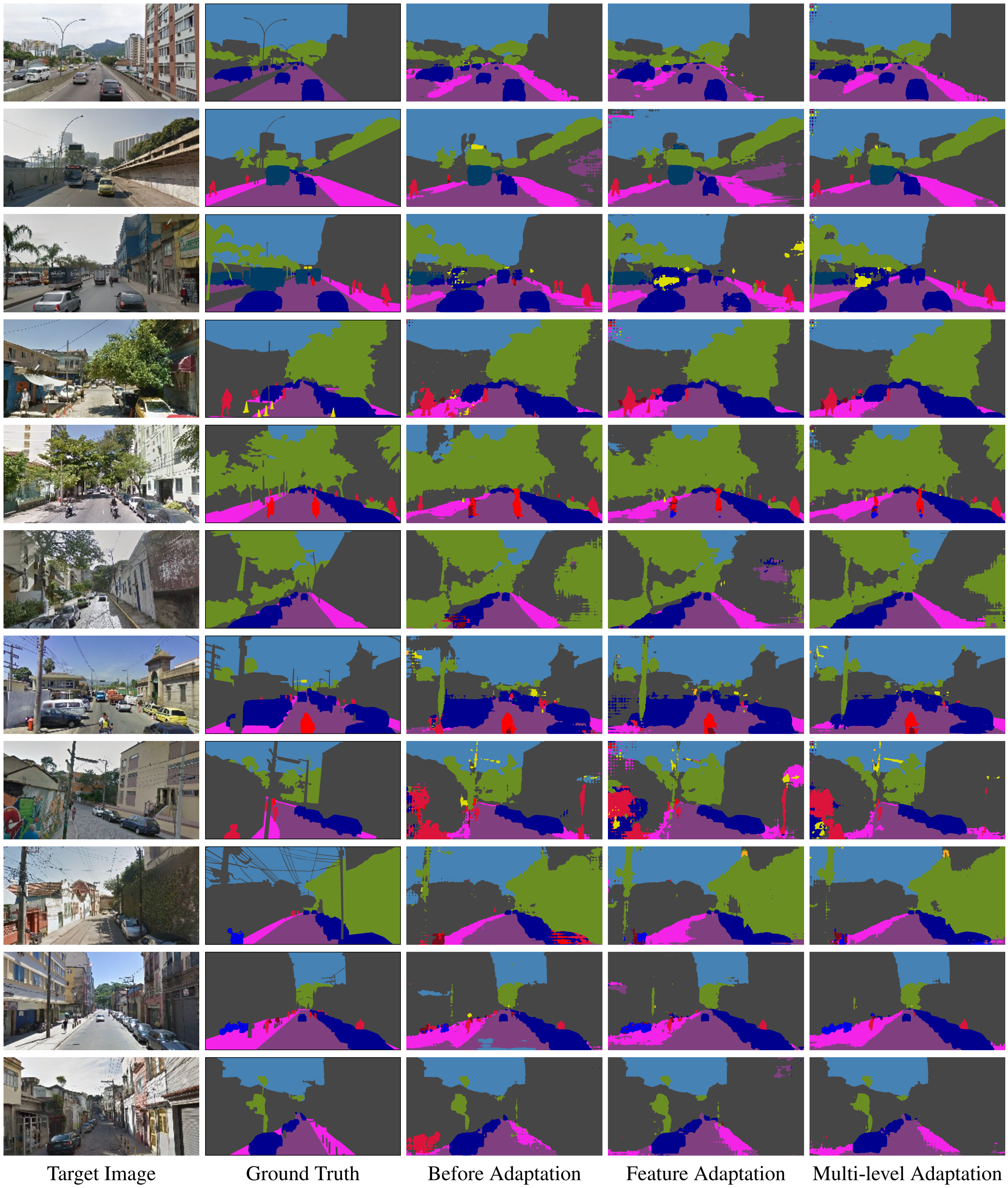


Figure 5. Example results of adapted segmentation for the Cityscapes-to-CrossCity dataset. For each target image, we show results before adaptation, with feature adaptation and our adapted segmentations in the output space.

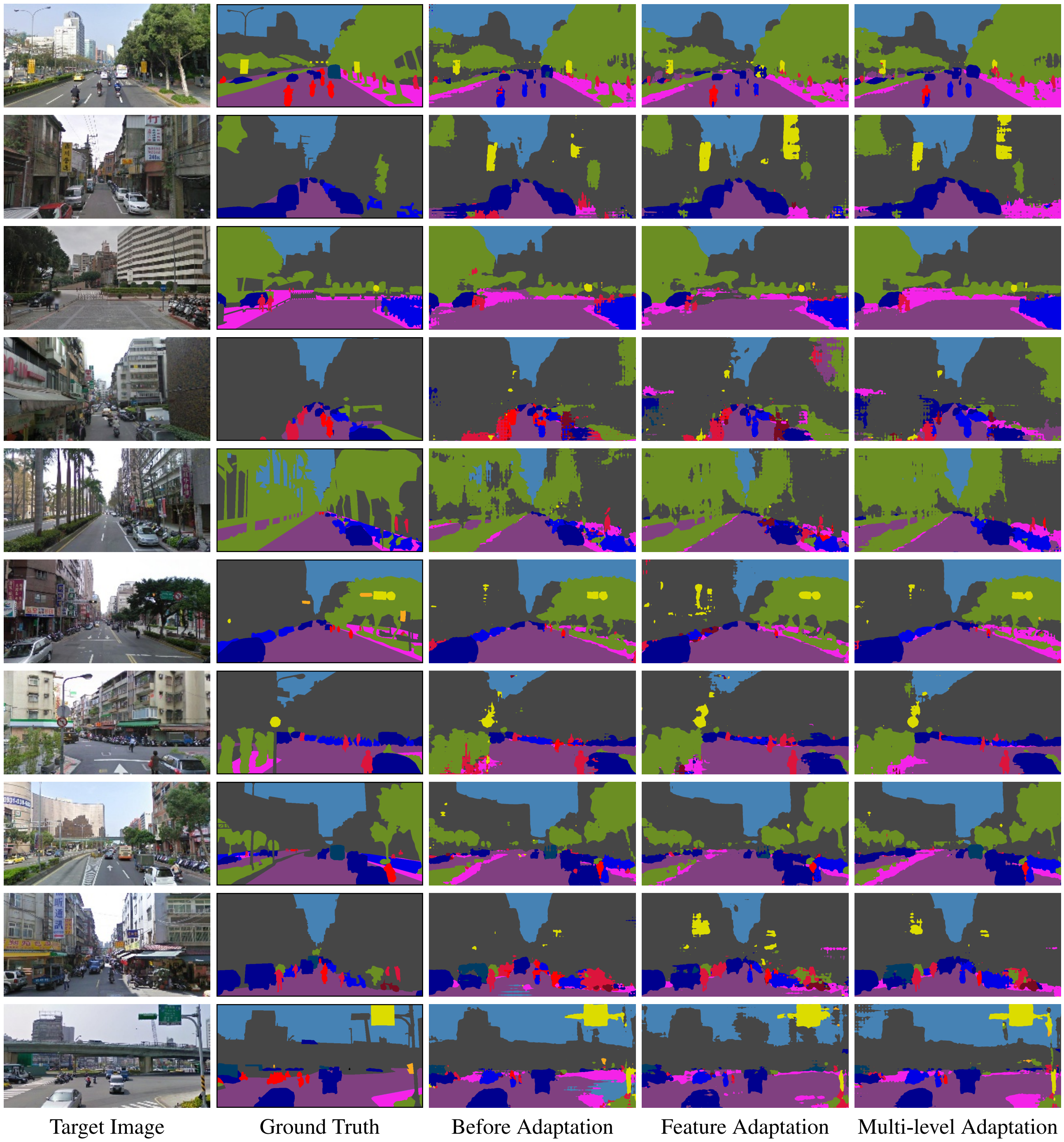


Figure 6. Example results of adapted segmentation for the Cityscapes-to-CrossCity dataset. For each target image, we show results before adaptation, with feature adaptation and our adapted segmentations in the output space.