Style Projected Clustering for Domain Generalized Semantic Segmentation Supplementary Material

Wei Huang[1*] Chang Chen[2†] Yong Li[2] Jiacheng Li[1]
Cheng Li[2] Fenglong Song[2] Youliang Yan[2] Zhiwei Xiong[1]
[1]University of Science and Technology of China [2]Huawei Noah's Ark Lab
{weih527,jclee}@mail.ustc.edu.cn, zwxiong@ustc.edu.cn,
{chenchang25,liyong156,licheng89,songfenglong,yanyouliang}@huawei.com

This supplementary document is organized as follows:
Sec. 1 provides further implementation details.
Sec. 2 provides additional ablation studies.
Sec. 3 provides additional generalization analysis.
Sec. 4 provides more qualitative comparisons on multiple unseen datasets, *i.e.*, Cityscapes [3], BDD100K [14] and Mapillary [7].

1. Further Implementation Details

Training. We train all models on four NVIDIA Tesla V100 GPUs, where the batch size per domain on each GPU is 4. In addition, we adopt the automold road augmentation library [6] to enhance the representation ability of style projection during the training phase, which enriches the style of images by the simulation of various urban scenarios, including diverse brightness, weather, motion blur and so on. Specially, for the single-source setting, we set multiply sets of style and semantic bases to improve the representation ability of our method. In other words, the number of bases M no longer depends on the number of source domains. During the training phase, the style and semantic information of current image is used to update the nearest style and semantic bases by Eq. 7 and Eq. 13 (in the main text).

Hyper-parameter setting. Following [2, 4], we adopt 19 classes that are compatible with all datasets as our prediction goal, *i.e.*, $C = 19$. The momentum coefficient α for style and semantic bases in Eq. 7 and Eq. 13 is set to 0.9 and 0.999, respectively, to achieve the best generalization performance. We empirically set the temperature parameter τ in Eq. 12 to 0.1, and set the weight β and γ of loss terms in Eq. 14 to 1.0 and 0.1, respectively, to balance the value of each loss term. In addition, we set the number of style and semantic bases to 10 for the single-source setting.

Network architecture. Following existing DGSS methods [2, 4, 5], we conduct main experiments by adopting DeepLabV3+ [1] with the ResNet-50 backbone. The output stride of DeepLabV3+ is set to 16 and 8 for ResNet-50 and ResNet-101, respectively. We replace the final segmentation classifier with an MLP projection head which consists of two standard convolution operations to generate 256-dimensional deep features. In addition, we remove the auxiliary per-pixel cross-entropy loss proposed in PSPNet [15] to avoid using the learnable classifier and fully demonstrate the effectiveness of semantic clustering.

Distance	C	B	M	Avg.-$\mathcal{T}$
$\mu/\sqrt{\sigma^2+\epsilon}$ [13]	45.87	42.18	47.48	45.18
KL-Divergence [9]	**47.13**	42.78	47.37	45.76
W-distance [12]	46.36	**43.21**	**48.23**	**45.93**

Table 1. Ablation results for different distance measures in Eq. 3 in the main text.

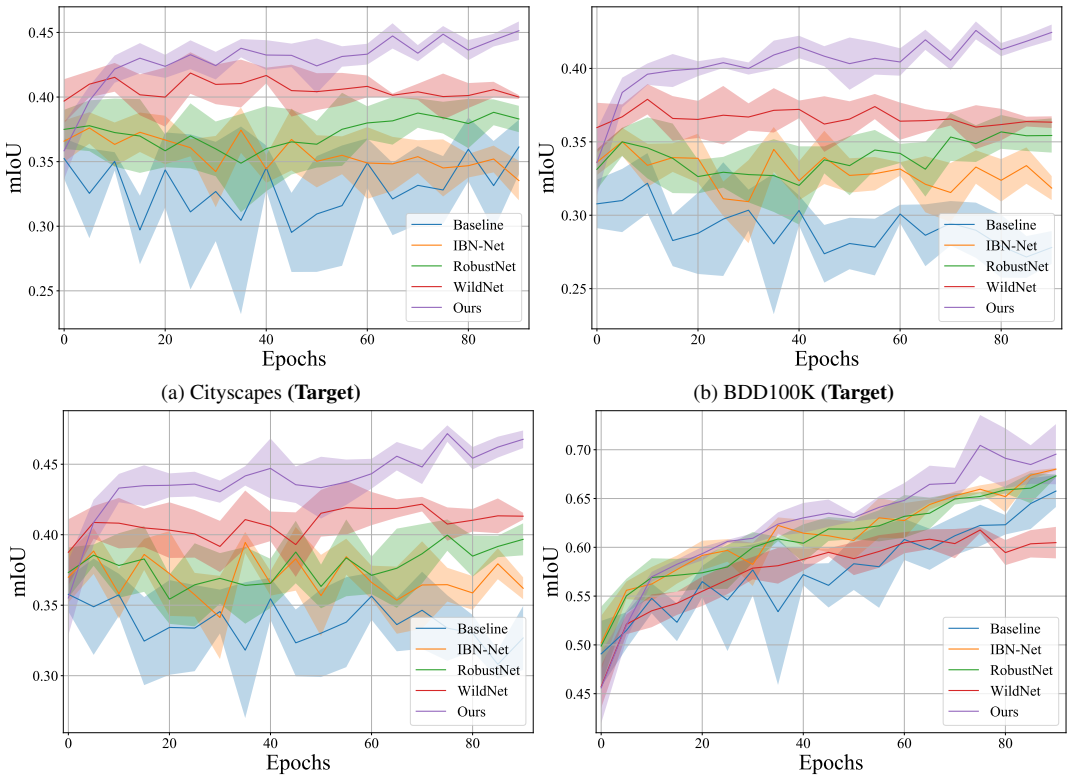
Layer1	Layer2	Layer3	C	B	M	Avg.-$\mathcal{T}$
			40.60	36.06	41.39	39.35
✔			45.76	41.67	46.36	44.60
✔	✔		46.36	**43.18**	**48.23**	**45.92**
✔	✔	✔	**46.42**	42.52	47.49	45.48

Table 2. Ablation results for our proposed style projection behind different layers of networks.

2. Additional Ablation Studies

Distance measure. We investigate the influence of different distance measures used to estimate the style similarity between the current image and style bases in Eq. 3 (in the main text). As listed in Table 1, we test three common distribution distance functions according to the value of mean μ and variance σ, *i.e.*, $\mu/\sqrt{\sigma^2+\epsilon}$ [13], KL-Divergence [9] and Wasserstein distance [12]. We can find that the Wasserstein distance performs better than other two distance measures. Remarkably, the performance differences between these three functions are not significant, which demon-

*This work was done during W. Huang's internship at Noah's Ark Lab.
†Corresponding author



(a) Cityscapes **(Target)** (b) BDD100K **(Target)**

(c) Mapillary **(Target)** (d) Synthia **(Source)**

Figure 1. **Source (G, S) → Target (C, B, M):** Comparison of validation performance with existing DGSS methods (*i.e.*, IBN-Net [8], RobustNet [2] and WildNet [5]) in different epochs, where all methods with the ResNet-50 backbone are trained 40K iterations on two synthetic (GTAV, Synthia) datasets.

Sty.-Pro.	Aug.	Sem.-Clu.	C	B	M	Avg.-$\mathcal{T}$
			36.03	28.15	32.61	32.26
✔			43.73	39.38	43.92	42.34
✔	✔		44.87	42.42	46.37	44.55
✔	✔	✔	**46.36**	**43.18**	**48.23**	**45.92**

Table 3. Ablation results for strong augmentation used in style projection. Sty.-Pro., Aug. and Sem.-Clu. indicate style projection, style augmentation and semantic clustering, respectively.

strates the robustness of our style projection for different distance measures.

Different layers. As listed in Table 2, we investigate the influence of applying our proposed style projection behind different layers of networks. We can find that there is the best generalization performance when style projection is applied behind both the first and second layers.

Style augmentation. As listed in Table 3, we conduct ablation experiments to demonstrate the effectiveness of automold road augmentation for style projection. Compared with the second and third lines, we can find that the style augmentation brings approximately 2% mIoU gains in average, which demonstrates the style augmentation successfully enhances the representation ability of style projection for unseen domains.

3. Additional Generalization Analysis

To demonstrate the superior generalization ability of our method, we conduct a set of contrast experiments with the same training iterations (*i.e.*, 40K). We show the validation curves of different DGSS methods on both source and target datasets in Fig. 1. We can find that the performances of all methods on the source dataset (Fig. 1d) are gradually increasing as the training goes on, while the performances of existing methods on target datasets (Fig. 1a-1c) are continuously declining or fluctuating. On the contrary, the performance of our method is also gradually increasing on the target datasets, which fully demonstrates that our method successfully avoids overfitting on the source dataset and consistently outperforms existing DGSS methods.

4. More Qualitative Results

To qualitatively demonstrate the superior generalization of our proposed method, we further compare the visual segmentation results with existing state-of-the-art methods (*i.e.*, PintheMem [4] and WildNet [5]). All methods with the ResNet-50 backbone are trained on two synthetic

datasets (*i.e.*, GTAV [10], Synthia [11]), and tested on three real-world datasets (*i.e.*, Cityscapes [3], BDD100K [14] and Mapillary [7]).

Cityscapes. As shown in Fig. 2, we first provide qualitative comparisons on the Cityscapes dataset. Compared with synthetic (source) datasets, the brightness of images in the Cityscapes dataset is relatively dim. Due to the change of brightness, Baseline and other DGSS methods are weakened to predict some objects, such as road, sidewalk, person and so on. On the contrary, our method successfully predicts these objects, which demonstrates our method is well generalized to brightness changes.

BDD100K. As shown in Fig. 3, Fig. 4 and Fig. 5, we provide comprehensively qualitative comparisons on the BDD100K dataset. Compared with the Cityscapes dataset, the BDD100K dataset contains various urban scenarios which are acquired in adverse weather (snow and rain), special time (dusk and night), unseen structure (overpass and bridge) and so on. To demonstrate the wide effectiveness of our method, we qualitatively compare our method with other DGSS methods in various conditions, including diverse weather, illumination, reflection, dusk, night, shadow and unseen structure. We can find that our method shows superior robustness over existing DGSS methods in various real-world scenarios.

Mapillary. As shown in Fig. 6, we further provide qualitative comparisons on the Mapillary dataset. Similar to the BDD100K dataset, the Mapillary dataset contains various urban scenarios captured from different conditions. We also compare our method with other DGSS methods in different scenarios, including dusk, unseen structure and so on. Compared with these methods, our method predicts given objects more accurately, such as road, sky, traffic sign and so on, which demonstrates the superior generalization of our method for unseen scenarios again.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1

[2] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, 2021. 1, 2

[3] 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 *CVPR*, 2016. 1, 3

[4] Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, and Kwanghoon Sohn. Pin the memory: Learning to generalize semantic segmentation. In *CVPR*, 2022. 1, 2, 4, 5, 6, 7

[5] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *CVPR*, 2022. 1, 2, 4, 5, 6, 7

[6] Road Augmentation Library. https://github.com/UjjwalSaxena/Automold--Road-Augmentation-Library. 1

[7] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 1, 3

[8] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018. 2

[9] Fernando Pérez-Cruz. Kullback-leibler divergence estimation of continuous distributions. In *ISIT*, 2008. 1

[10] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 3

[11] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 3

[12] SS Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974. 1

[13] Ximei Wang, Ying Jin, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Transferable normalization: Towards improving transferability of deep neural networks. *NeurIPS*, 2019. 1

[14] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 1, 3

[15] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1

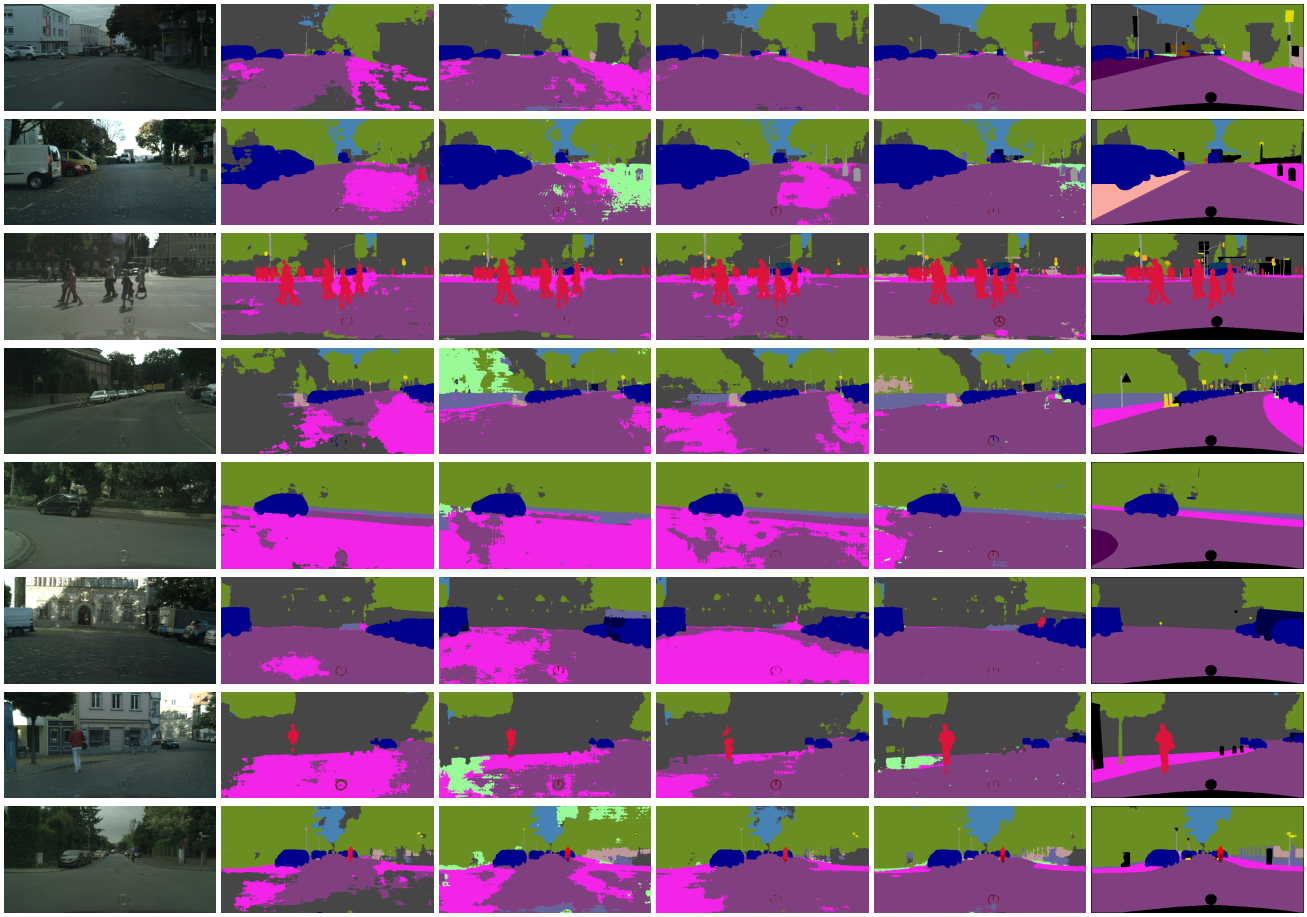


Figure 2. **Source (G+S) → Target (C):** Qualitative comparison on the Cityscapes dataset. All methods adopt DeepLabV3+ with the ResNet-50 backbone.

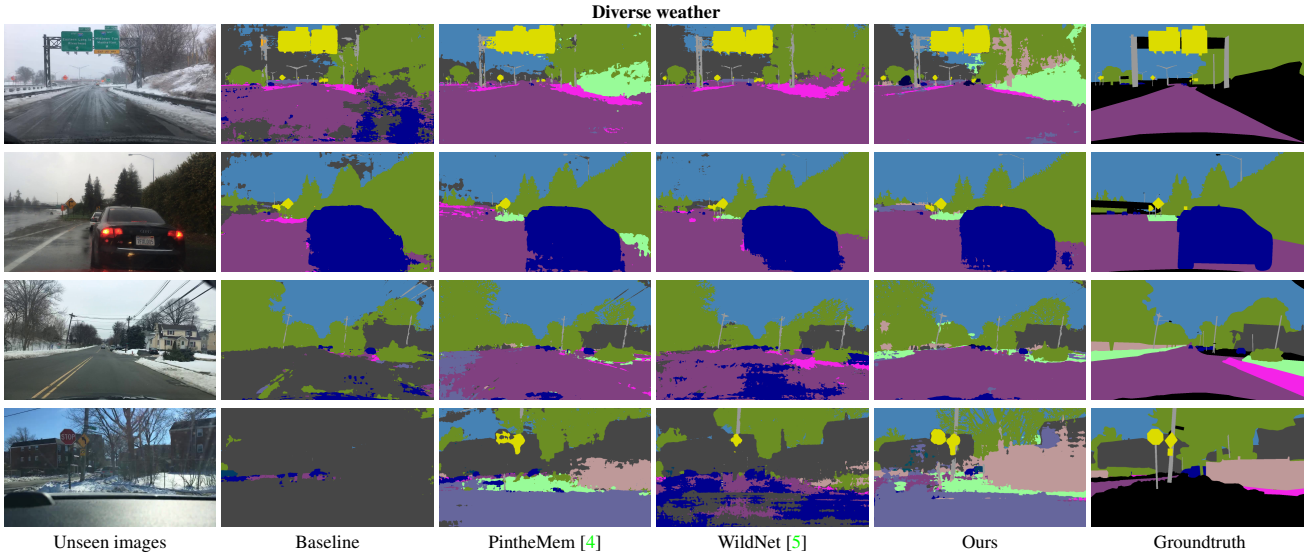


Figure 3. **Source (G+S) → Target (B):** Qualitative comparison on the BDD100K dataset. All methods adopt DeepLabV3+ with the ResNet-50 backbone.

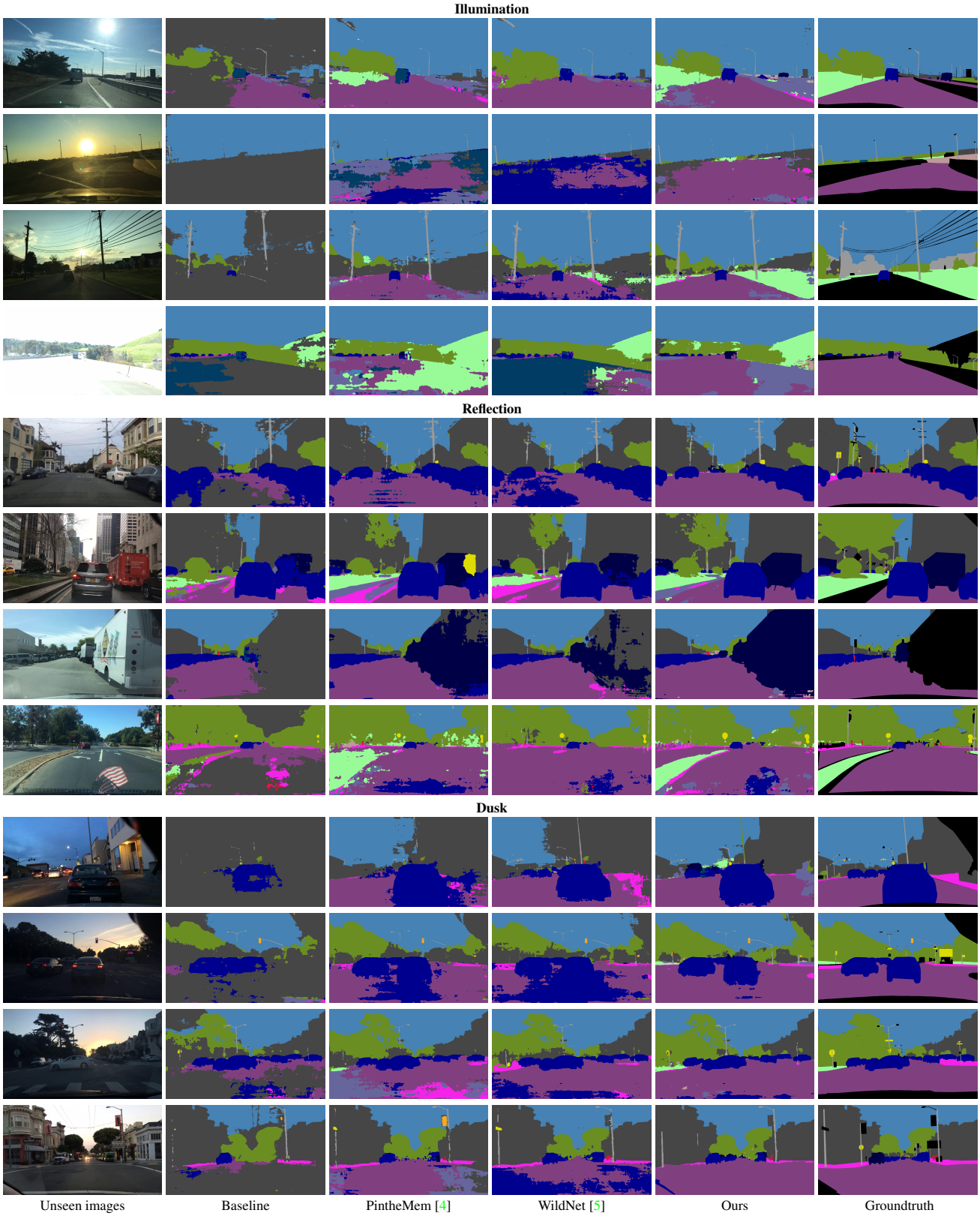


Figure 4. **Source (G+S) → Target (B):** Qualitative comparison on the BDD100K dataset. All methods adopt DeepLabV3+ with the ResNet-50 backbone.

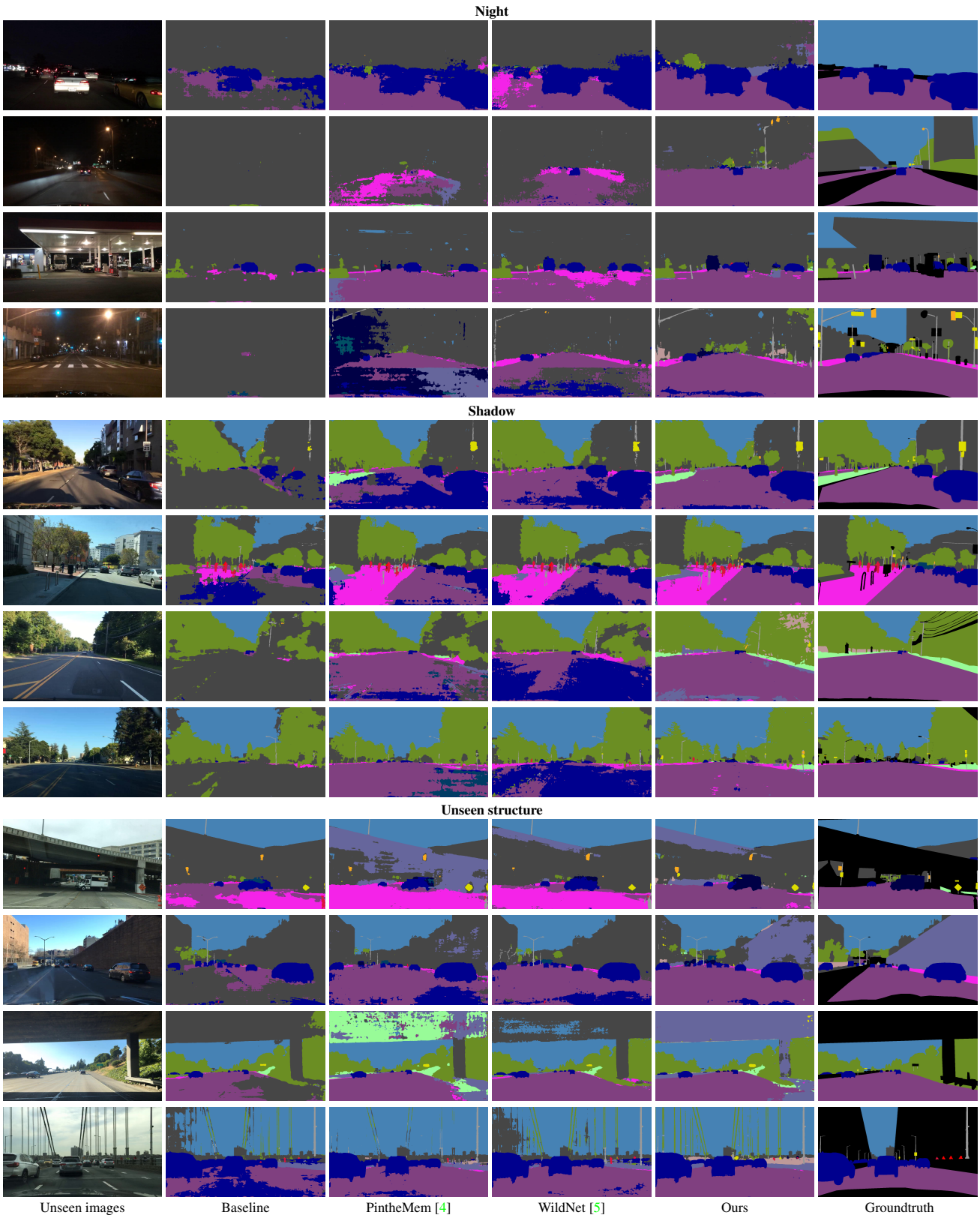


Figure 5. **Source (G+S) → Target (B):** Qualitative comparison on the BDD100K dataset. All methods adopt DeepLabV3+ with the ResNet-50 backbone.

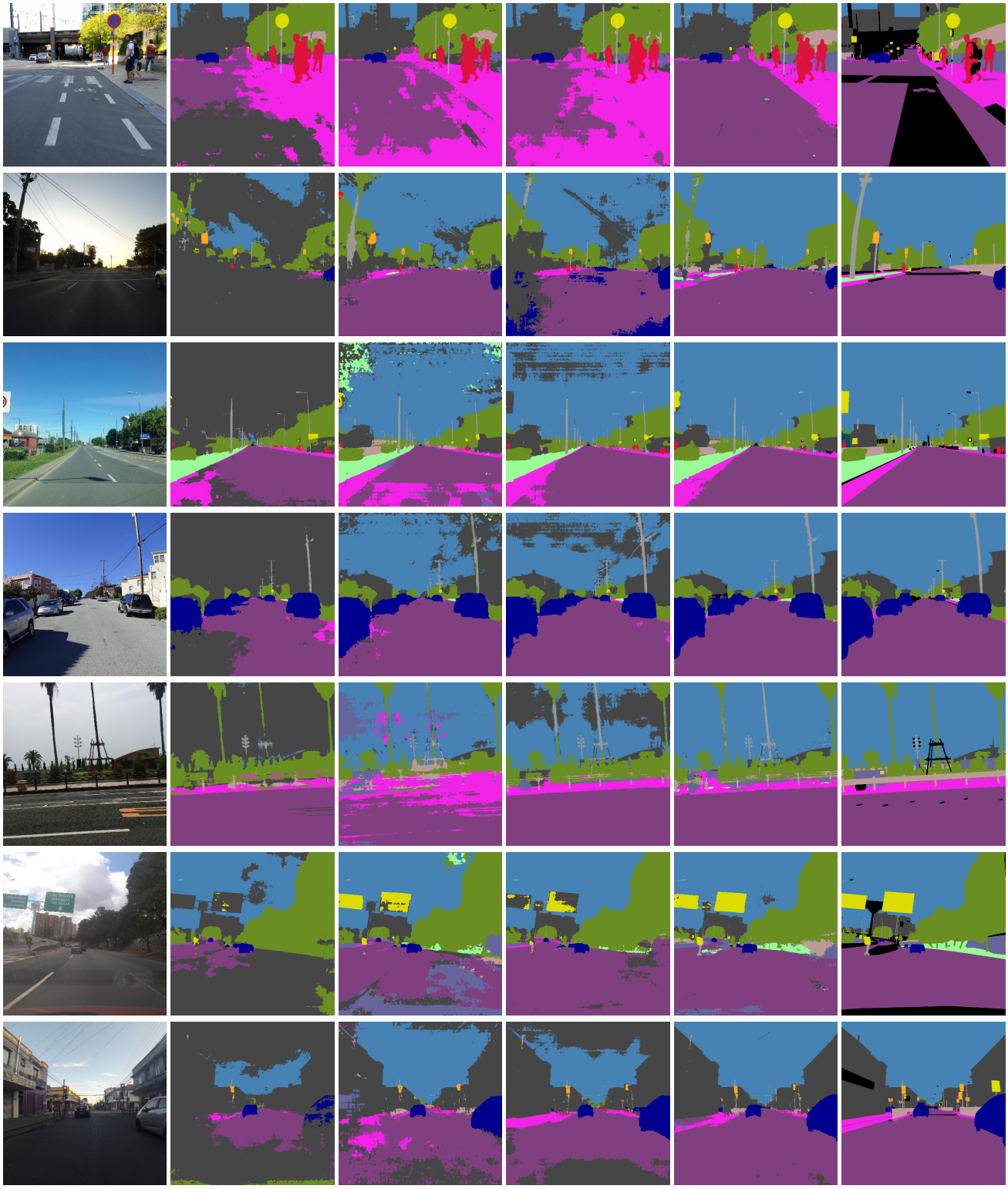


Figure 6. **Source (G+S) → Target (M):** Qualitative comparison on the Mapillary dataset. All methods adopt DeepLabV3+ with the ResNet-50 backbone.