WildNet: Learning Domain Generalized Semantic Segmentation from the Wild (Supplementary Material)

Suhyeon Lee Hongje Seong Seongwon Lee Euntai Kim*
School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
{hyeon93, hjseong, won4113, etkim}@yonsei.ac.kr

A. More Analysis

In this section, we further analyze our method with additional qualitative results. We also provide semantic segmentation results on five different datasets, which consist of four unseen domain datasets and one seen domain dataset.

A.1. Content Extension Learning

Fig. 1 illustrates extended the wild contents from the source (*i.e.*, GTAV [8]) to the wild (*i.e.*, ImageNet [4]). The eight contents are extended from a centered image in GTAV to the eight ImageNet images, and each color represents the semantic label of the content in GTAV. After network training, we used our final ResNet-50 [5] with DeepLabV3+ [1] model to visualize the pixels in the wild image extended from each pixel in the source image. Although the source content was extended to the wild content closest to the stylized source content in the feature space without using any wild label, the source content was extended to the wild content with the same semantic information as itself, as shown in Fig. 1a. These content extensions increase the intra-class content variability in the latent embedding space and alleviate overfitting to the source contents.

There are various semantic classes in the wild dataset that are not considered in the source dataset, and we will refer to them as wild-only classes in this supplementary material. The source content is sometimes extended to wild-only class content, such as the thin pole-class pixel being extended to the thin bird's leg pixel in Fig. 1b. The proposed content extension learning is a pixel-wise approach. Therefore, if two pixels have similar features, content extension to other classes with similar shapes is observed. This is not limited to human-annotated class labels and encourages the network to learn generalized features by reducing the distance between contents with similar semantic information in the feature space. This may provide clues to generalization performance improvements for unseen contents. In Fig. 1c, it was observed that some road pixels were extended to the waterside ground and underwater ground pixels. It is expected that content extension to these wild-only classes will guide the network to correctly predict wet road and puddle pixels as road classes in rainy scenes. With content extension learning, WildNet makes reliable predictions in various environments, such as wet vegetation in the fifth row of Fig. 7 and light-reflected road in the first row of Fig. 3.

A.2. Wild-Stylized Features

Fig. 2 shows the importance of learning task-specific information from wild-stylized features. Given the source image and ground truth label (see the first and sixth columns in Fig. 2), we diversify source data by stylizing the source feature using the style of the wild feature from the given wild image (see the second column in Fig. 2). To maintain the spatial information of the source feature, we apply adaptive instance normalization [6] with channel-wise mean and standard deviation for the source and wild features. To visualize that the wild-stylized source feature contains the spatial information of the source feature and the style of the wild feature, we reconstructed the image from the wild-stylized feature using the U-Net [9] structure following the process of RobustNet [2] reconstructing the input image from the whitened feature. After training the baseline model and our model on the semantic segmentation task, we freeze the weights of the pre-trained model and add a decoder to learn the image reconstruction. The reconstructed images from the wild-stylized source features show that both the baseline model and our model transform the style while successfully maintaining the spatial information of the source features (see the third column in Fig. 2). Nevertheless, the baseline model fails to make accurate predictions from wild-stylized source features, as opposed to making accurate predictions from the original source features (see the fifth and fourth columns in Fig. 2).

To address this issue, we train our WildNet with the proposed style extension learning and semantic consistency regularization methods. The style extension learning enables our model to naturally adapt to various styles by learning task-specific information from the wild-stylized

*Corresponding author.

features. Moreover, the semantic consistency regularization regularizes the model, enabling the capture of consistent semantic information from the wild-stylized and original source features. As a result, our model captures generalized semantic information from features of various styles and makes correct predictions on wild-stylized source features (see the fifth column in Fig. 2).

A.3. Qualitative Results

In Figs. 3 to 7, we present semantic segmentation results on four unseen domain validation sets (*i.e.*, Cityscapes [3], BDD100K [11], Mapillary [7], and SYNTHIA [10]) and a seen domain validation set (*i.e.*, GTAV [8]). We used ResNet-50 as the backbone network and trained on GTAV train set. To show the efficacy of the proposed method, we additionally present the results of the baseline and RobustNet [2]. As shown in Figs. 3 to 6, the baseline model works poorly on the unseen datasets, and RobustNet also often fails. In contrast, WildNet can accurately segment the road and sidewalk (*e.g.*, the top row in Fig. 3, the second row in Fig. 4, and the second row in Fig. 6) and correctly classify instances (*e.g.*, terrain in the fifth row of Fig. 3, riders and bicycles in the top row of Fig. 5, and a car in the fifth row of Fig. 6). Furthermore, as shown in Fig. 7, our WildNet performed well on the seen dataset even in some challenging cases, such as night-time (the top row in the figure), rainy (the fifth row), and backlight (the third row).

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*, pages 801–818, 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*, pages 11580–11590, 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*, pages 3213–3223, 2016. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[6] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 1

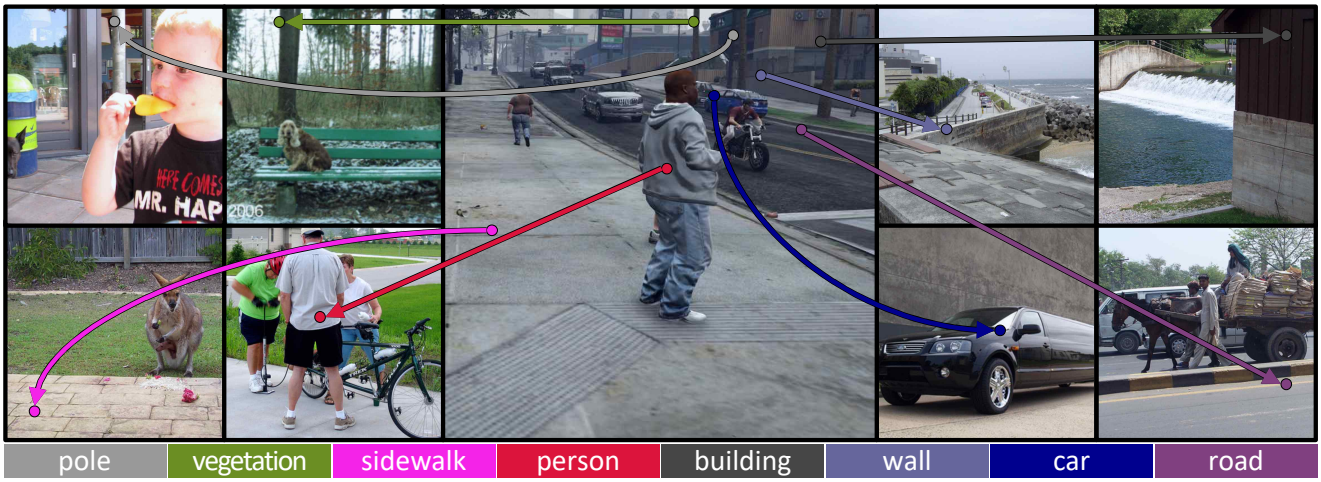
[7] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, pages 4990–4999, 2017. 2

[8] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, pages 102–118. Springer, 2016. 1, 2

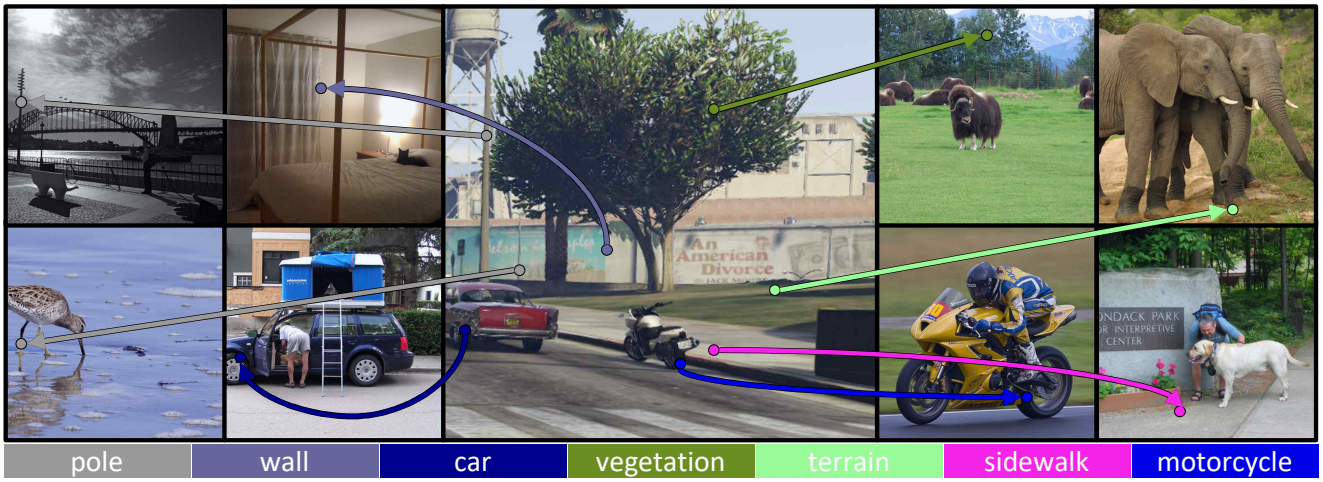
[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 1, 4

[10] 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*, pages 3234–3243, 2016. 2

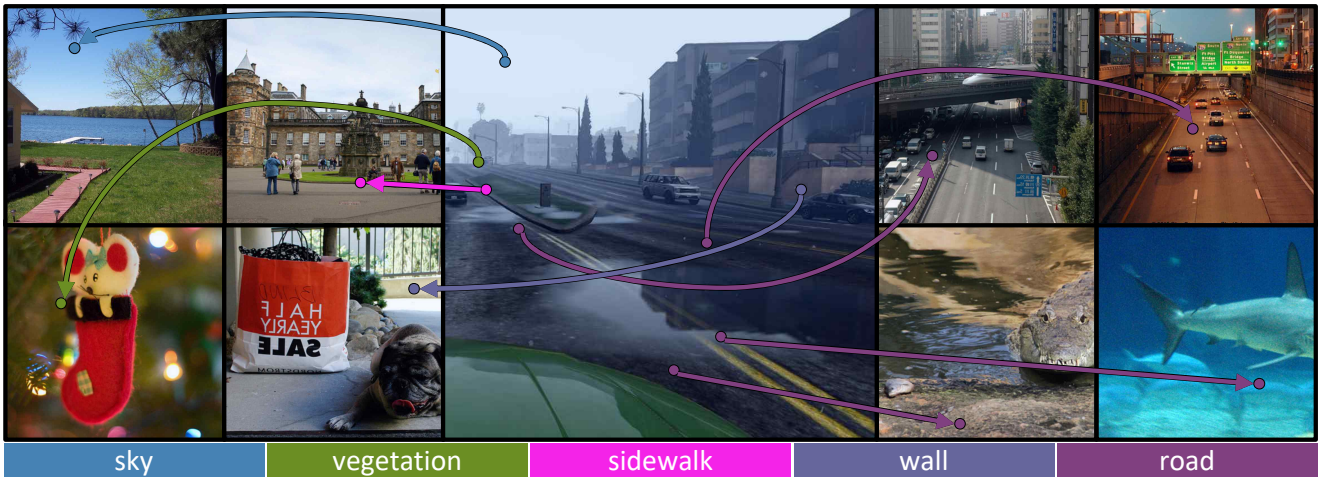
[11] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pages 2636–2645, 2020. 2



(a)



(b)



(c)

Figure 1. Visualization of extended wild contents. The center image is sampled from the GTAV dataset while the remaining eight images are sampled from ImageNet. The contents are extended from the centered GTAV image to the eight ImageNet images, and each color represents the semantic label of the content in GTAV.

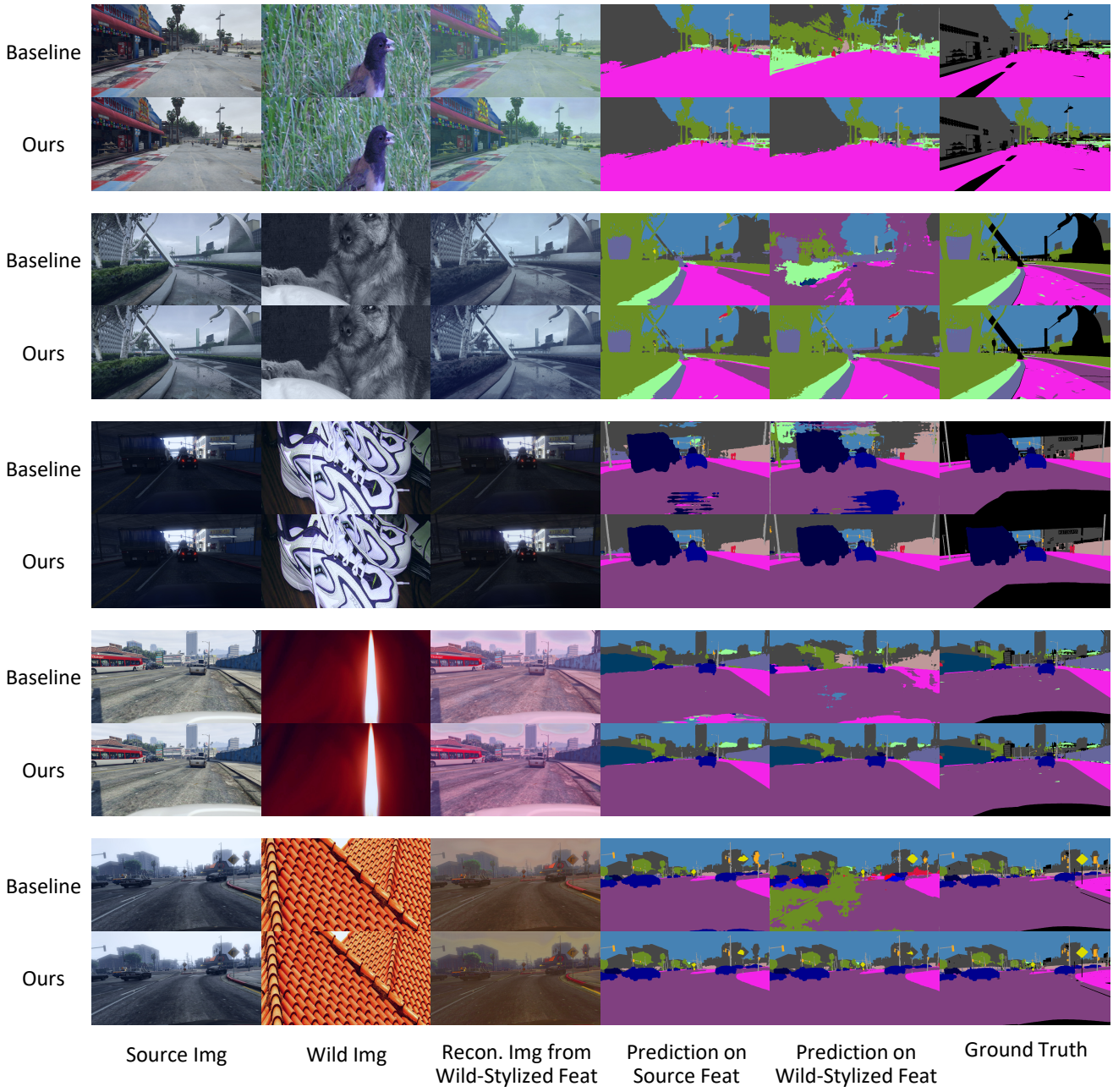


Figure 2. Given the source image and ground truth label, we stylize the source feature using the style of the wild feature from the given wild image. To visualize the wild-stylized source feature, we reconstructed an image from the wild-stylized feature using U-Net [9]. The reconstructed image from wild-stylized source feature includes spatial information of the source image and style information of the wild image. The baseline model fails to make correct predictions from wild-stylized features, as opposed to accurate predictions from source features. In contrast, the proposed WildNet makes accurate predictions on wild-stylized features by applying style extension learning and semantic consistency regularization in the training process.

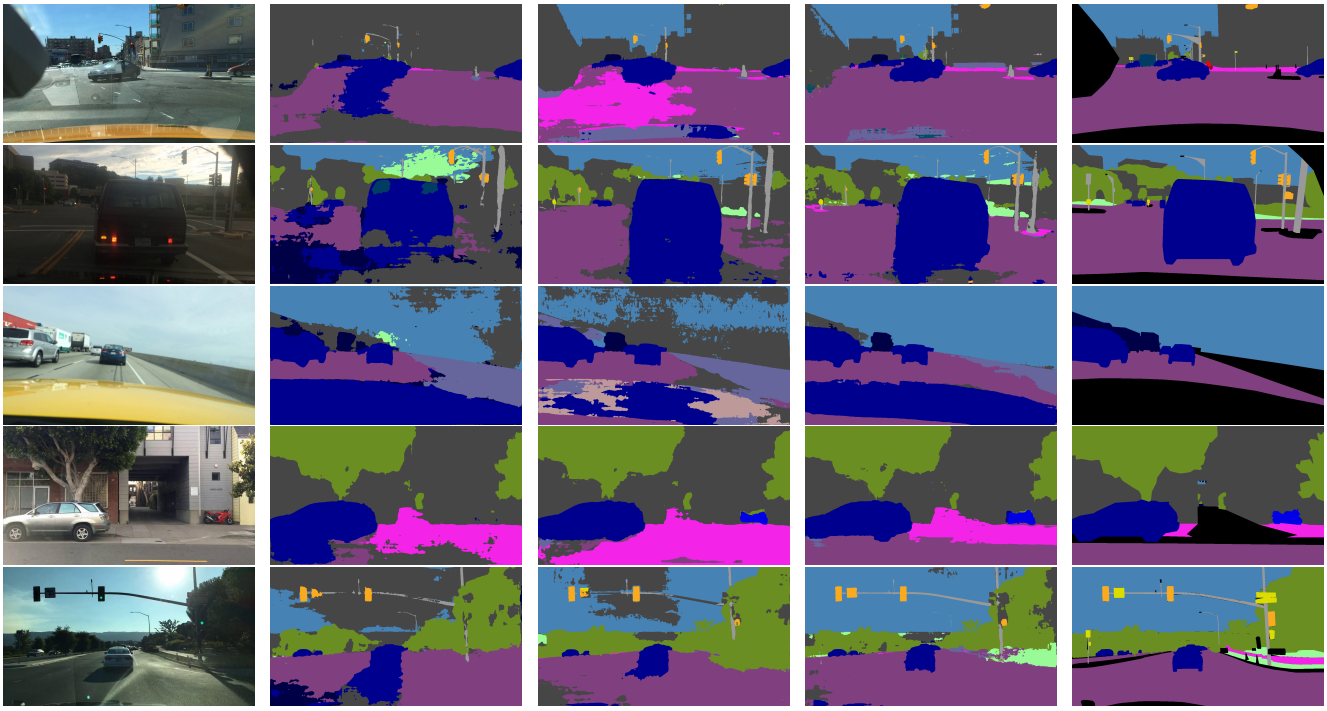


Figure 3. Semantic segmentation results on unseen domain images in BDD100K with the models trained on GTAV.

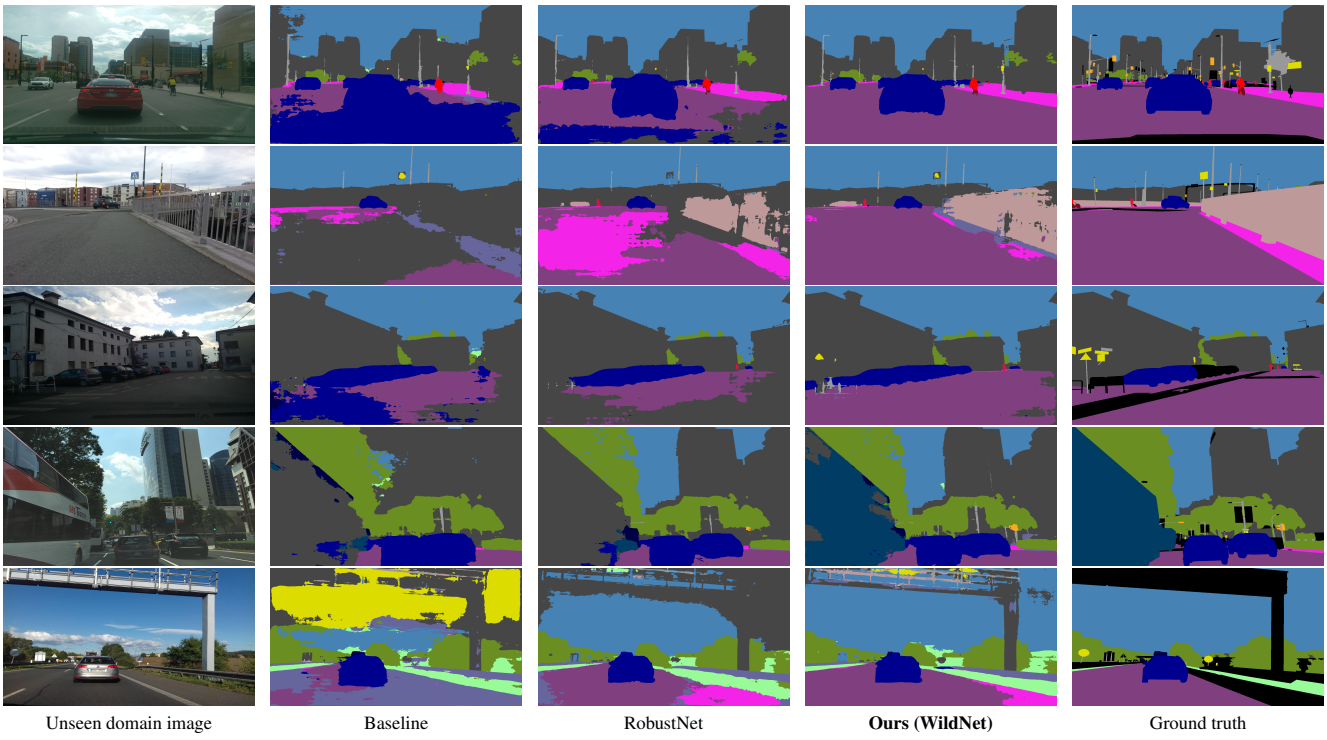


Figure 4. Semantic segmentation results on unseen domain images in Mapillary with the models trained on GTAV.

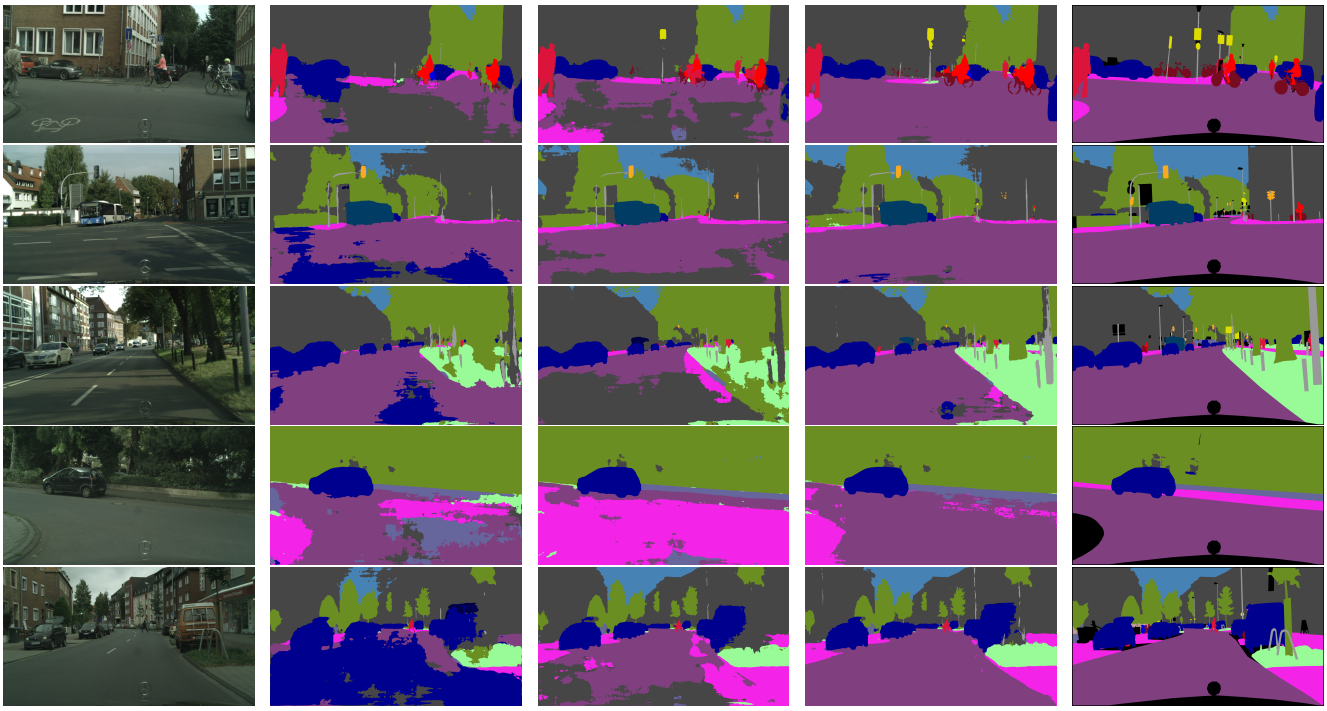


Figure 5. Semantic segmentation results on unseen domain images in Cityscapes with the models trained on GTAV.

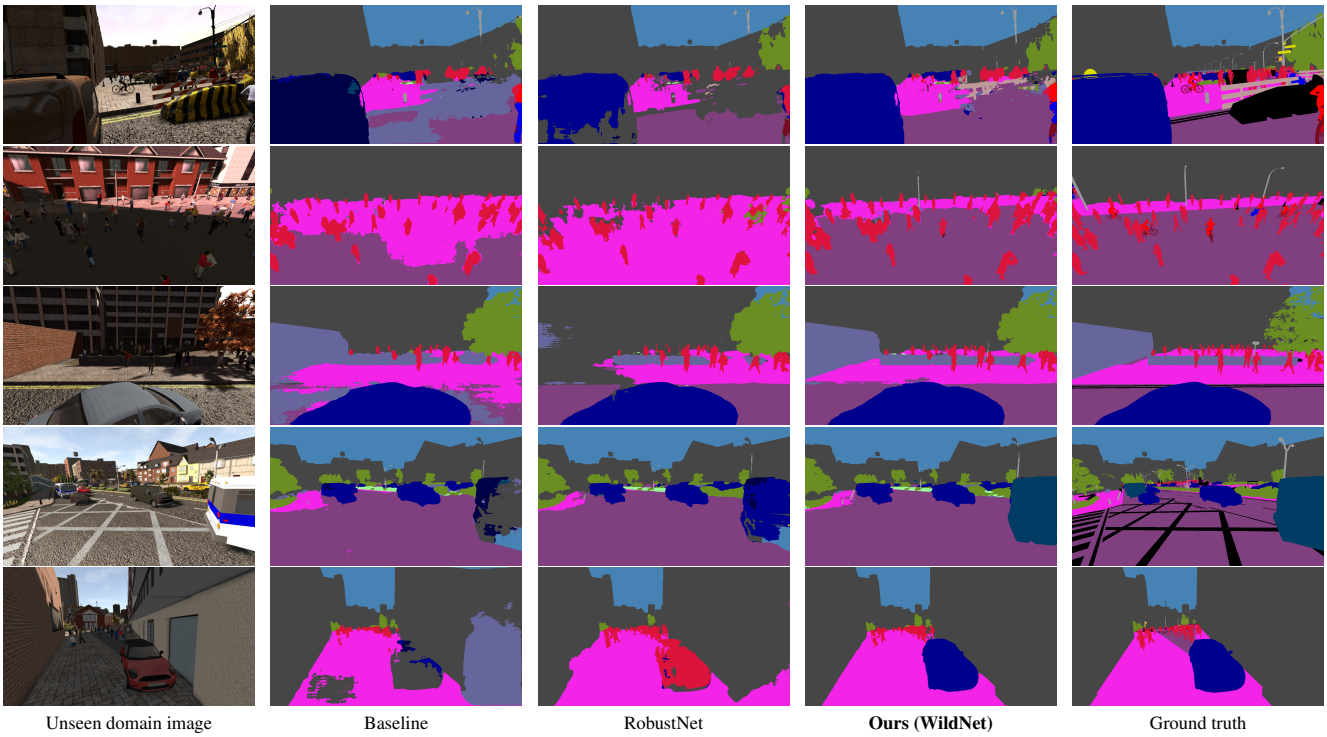


Figure 6. Semantic segmentation results on unseen domain images in SYNTHIA with the models trained on GTAV.

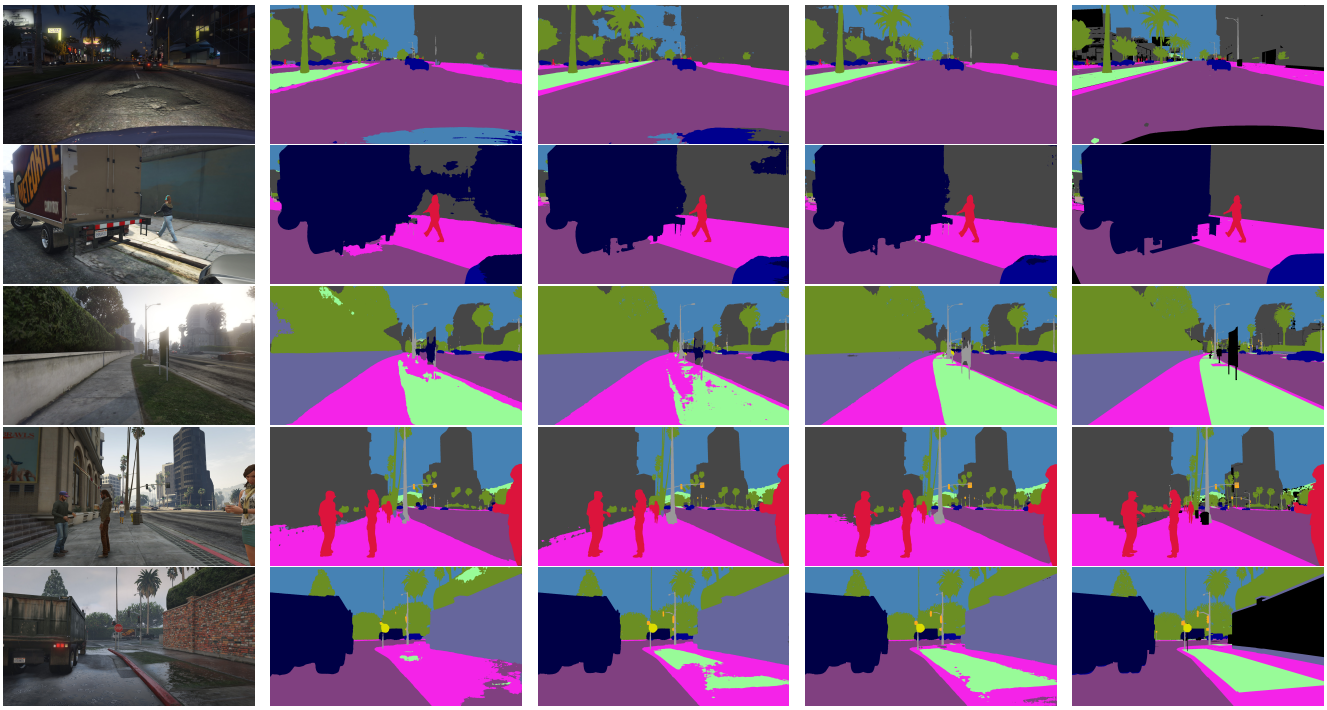


Figure 7. Semantic segmentation results on seen domain images in GTAV with the models trained on GTAV.