Domain Generalization via Balancing Training Difficulty and Model Capability (Supplementary Material)

Xueying Jiang, Jiaxing Huang, Sheng Jin, Shijian Lu*
S-lab, Nanyang Technological University
xueying003@e.ntu.edu.sg
{Jiaxing.Huang, Sheng.Jin, Shijian.Lu}@ntu.edu.sg

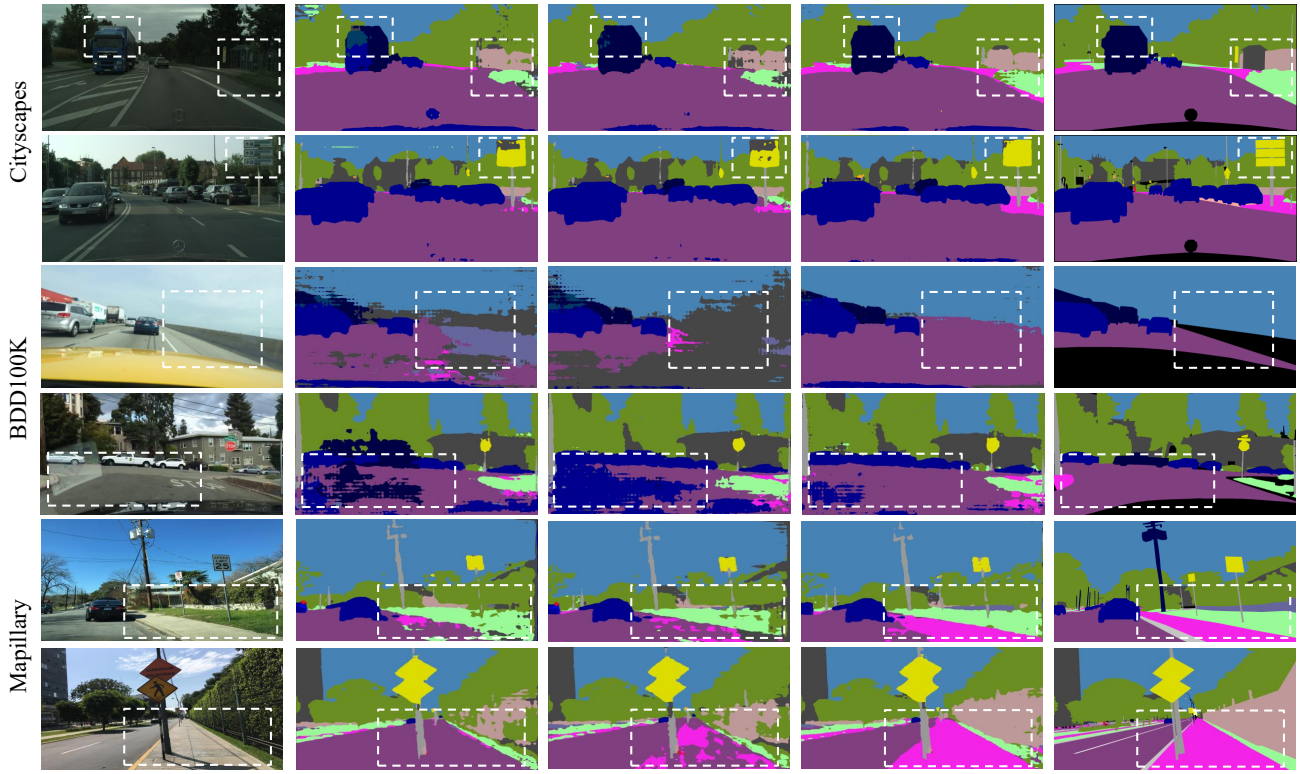

Figure 1: Qualitative results of semantic segmentation for GTAV to Cityscapes (Row 1, 2), BDD100K (Row 3, 4), and Mapillary (Row 5, 6). White boxes highlight regions with clear difference across the compared methods.

This supplementary material provides additional discussions and experimental details, including qualitative and quantitative results.

1. Discussions

Domain-level and sample-level difficulty. The difficulty degree defined in the proposed MoDify framework has two levels. The first level is the domain-level difficulty degree, which denotes the overall difficulty strength across all the training samples in the dataset. The second level is the more detailed sample-level difficulty degree, which is adjusted in an online manner during training.

Specifically, the domain-level difficulty degree provides an overall measure of the level of difficulty in the dataset, whereas the sample-level difficulty degree takes into account the specific characteristics of each sample and adjusts

* Corresponding author.

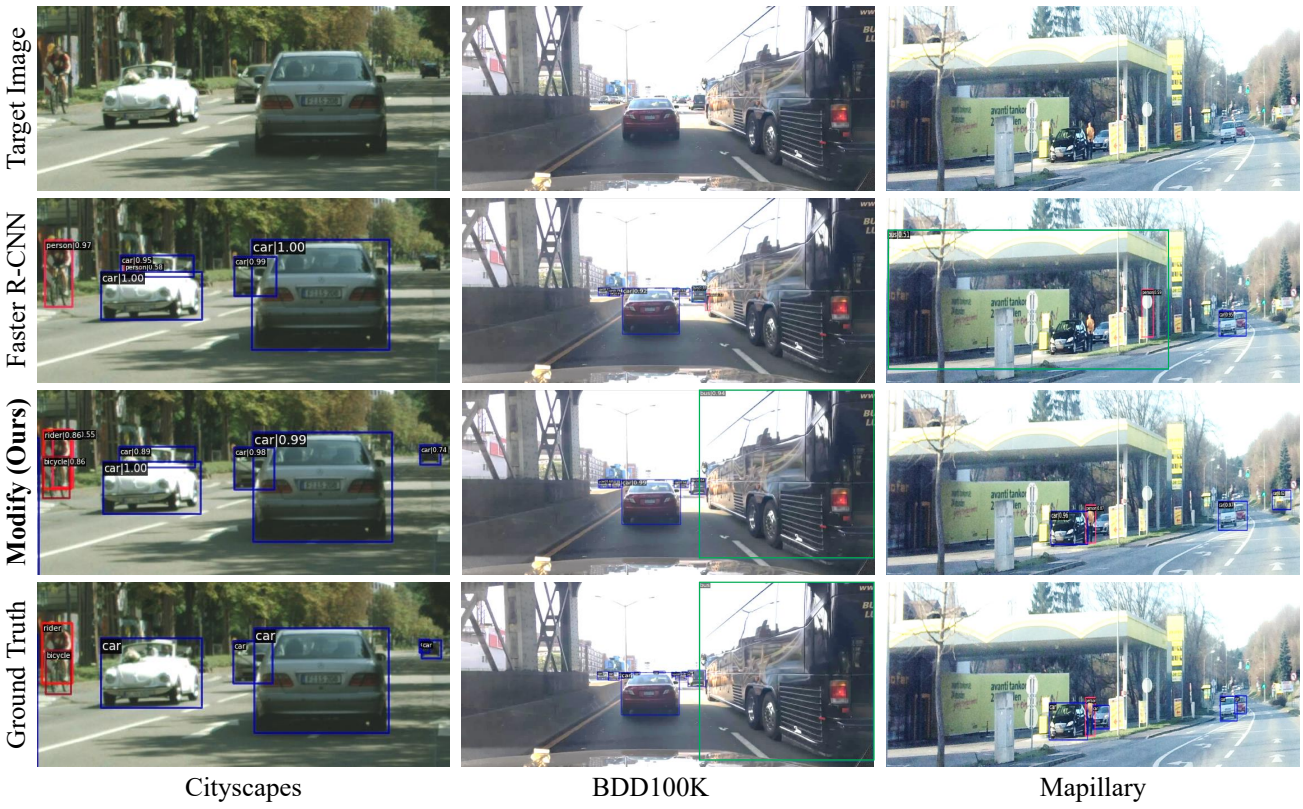

Figure 2: Qualitative results of the object detection task for the setting of training on SYNTHIA and validating on Cityscapes (Column 1), BDD100K (Column 2), and Mapillary (Column 3), respectively. Zoom in for a better view.

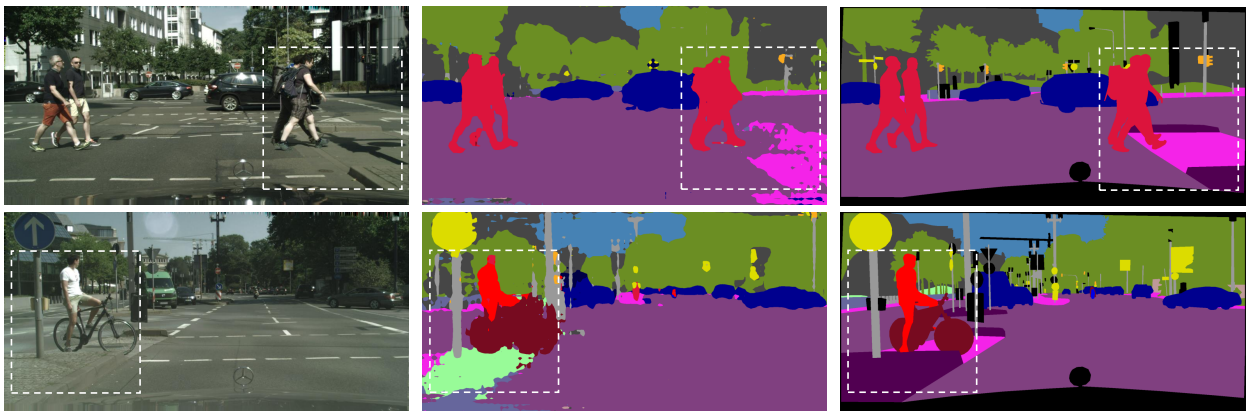

Figure 3: Qualitative results of failure cases. The proposed MoDify has relatively poor performance on certain region-level difficult samples. White boxes highlight regions with relatively poor prediction results. Results are obtained over the domain generalizable semantic segmentation task GTAV $\rightarrow$ Cityscapes, using ResNet-101 as the backbone.

the difficulty level accordingly during training.

In practice, we achieve the domain-level difficulty adjustment by adjusting each sample's difficulty during training in a sample-level manner.

2. Experimental Details

In this section, we provide additional details, including visualizations and quantitative experimental results.

Figure 4: Visualization of images using the proposed RGB Shuffle technique.

$[T_{easy}, T_{hard}]$	[0.05, 0.95]	[0.1, 0.9]	[0.2, 0.8]	[0.3, 0.7]
Performance	**48.8**	48.3	47.1	45.2

Table 1: Ablation study on the thresholds $[T_{easy}, T_{hard}]$ in MoDify-NO over the domain generalizable semantic segmentation task GTAV $\rightarrow$ Cityscapes, using ResNet-101 as the backbone.

λ	0.1	0.3	0.5	0.7	0.9
Performance	46.9	48.0	**48.8**	47.6	47.1

Table 2: Ablation study on the momentum coefficient λ over the domain generalizable semantic segmentation task from GTAV to Cityscapes, using ResNet-101 as the backbone.

2.1. Qualitative Results

Semantic Segmentation. We present additional qualitative illustrations on the semantic segmentation task in Fig. 1. Compared with other methods, our proposed MoDify method exhibits superior performance, as evidenced by its ability to predict more complete small objects (Row 1 and 2), as well as more accurate sidewalk and road segmentation predictions (Row 3 to 6).

Object Detection. Fig. 2 presents additional qualitative results on the object detection task. MoDify demonstrates superior performance in detecting small objects (Column 1 and 3) and accurately predicting big objects, while Faster-RCNN fails to find objects (Column 2, "bus" on the right side) or makes false-positive predictions (Column 3, "bus" on the left side).

Failure Cases. Qualitative results of failure cases are provided in Fig. 3. On certain region-level difficult samples, the proposed MoDify achieves relatively poor performance. We can observe that MoDify faces challenges in distinguishing between two pedestrians nearby (Row 1), and identifying a rider on a bicycle (Row 2). Future work could investigate a more focused approach through the implementation of a fine-grained region adaptive strategy, wherein data augmentation techniques are applied with varying levels of intensity to different regions of an image.

RGB Shuffle. Visualization results of the proposed RGB Shuffle technique are provided in Fig. 4. By permuting the R, G, and B channels of an input image, RGB Shuffle alters the style in color while preserving the spatial layout, aiding the model in learning domain-invariant features.

2.2. Quantitative Results

The experimental settings used in the following are the same as the ones in the main text of the submission.

Momentum Coefficient of Loss Bank. In the proposed MoDify framework, the momentum coefficient λ is used in updating the Loss Bank to balance between losses of historical and current epochs. As shown in Tab. 2, MoDify achieves the best performance when λ equals 0.5.

Thresholds for filtering out samples in MoDify-NO. In the proposed MoDify-NO strategy, we utilize two thresholds T_{easy} and T_{hard} to filter out samples that are either too easy or too difficult. Tab. 1 shows the results of using different thresholds. We can observe that MoDify achieves the best performance when T_{easy} and T_{hard} equal 0.05 and 0.95, while performance drops with the range of two thresholds narrows, such as T_{easy} and T_{hard} equal to 0.3 and 0.7.

Data Augmentation	mIoU
Baseline	36.6
LAB-based Image Translation [1]	45.1
RGB Shuffle (ours)	**48.8**

Table 3: Comparison of different style-transfer data augmentation strategies over the domain generalizable semantic segmentation task GTAV → Cityscapes, using ResNet-101 as the backbone. For a fair comparison, only the data augmentation strategy of the compared methods is different.

Learning Strategy	mIoU
Baseline	36.6
Curriculum Learning [3]	42.3
MoDify (ours)	**48.8**

Table 4: Comparison of different learning strategies over the domain generalizable semantic segmentation task GTAV → Cityscapes, using ResNet-101 as the backbone. For a fair comparison, only the learning strategy of the compared methods is different.

Data Augmentation Strategies. Table 3 compares a representative style-transfer data augmentation with ours in domain generalization (DG). Most data augmentation strategies used in DG are complex and integrated with networks during training [4, 2], while the proposed RGB Shuffle is parameter-free and can be easily incorporated into existing networks. In this study, we compared the proposed RGB Shuffle technique with another straightforward technique, LAB-based image translation [1], as shown in Table 3. The baseline achieves 36.6% mIoU when no data augmentation is used. Using LAB-based image translation improves the performance to 45.1% mIoU, while RGB Shuffle achieves the best results with a 3.7% mIoU improvement over LAB-based image translation.

Learning Strategies. To show the difference between applying another learning strategy, curriculum learning, and the proposed MoDify, we conduct experiments to compare the two strategies, as shown in Tab. 4. We re-implement a curriculum learning method [3] on the same baseline for a fair comparison. As shown in Tab. 4, MoDify can outperform the re-implemented curriculum learning method and the baseline with 6.5% and 12.2% mIoU, respectively. This is because the curriculum learning approach concentrates primarily on pre-defined difficulty levels for sub-tasks during training. In contrast, MoDify dynamically augments and reserves difficulty-aware training samples according to the capability of contemporarily trained network models along the training process.

Loss bank length	100	1000	5000	10000	20000	24966
Performance	44.9	46.1	47.9	48.2	48.6	**48.8**

Table 5: Comparison of different lengths of loss bank over the domain generalizable semantic segmentation task GTAV → Cityscapes, using ResNet-101 as the backbone.

Length of loss bank. We constrain the length of loss bank by randomly selecting anchor samples and updating their values in the loss bank. Hence, MoDify doesn't store millions of variables while handling large training dataset. Tab. 5 shows the performance under different loss bank lengths on GTAV → Cityscapes.

References

[1] Jianzhong He, Xu Jia, Shuaijun Chen, and Jianzhuang Liu. Multi-source domain adaptation with collaborative learning for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11008–11017, 2021. 4

[2] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. WildNet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9946, 2022. 4

[3] Yang Zhang, Philip David, Hassan Foroosh, and Boqing Gong. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (8):1823–1841, 2019. 4

[4] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 535–552. Springer, 2022. 4