[Supplementary Material]
CSDA: Learning Category-Scale Joint Feature for Domain Adaptive Object Detection

Changlong Gao[1*], Chengxu Liu[1,2*], Yujie Dun[1], Xueming Qian[1,2]
[1]Xi'an Jiaotong University
[2]Shaanxi Yulan Jiuzhou Intelligent Optoelectronic Technology Co., Ltd
{gaochanglong,chengxuliu}@stu.xjtu.edu.cn, {dunyj,qianxm}@mail.xjtu.edu.cn

In this supplementary material, we provide the results on ResNet50 [2] in Cityscapes [1] to Foggy Cityscapes [7] scenario in Sec. 1. We analyze the sensitivity of $\omega_{1,2}$ in Sec. 2. We analyze the scale mapping function $f_{sm}(\cdot)$ in Sec. 3. We visualize the feature distribution of each category and scale in Sec. 4. We analyze the limitations of CSDA in Sec. 5. Finally, we provide more visualization results in Sec. 6.

1. Results on ResNet

As presented in Tab. 1, we report the results of Cityscapes→Foggy Cityscapes based on ResNet50 [2]. We can find that our CSDA based on ResNet50 outperforms other state-of-the-art methods. Specifically, our method achieves the best performance in 3 (person, car, and truck) out of 8 categories. This demonstrates the superior performance of our CSDA on different backbones.

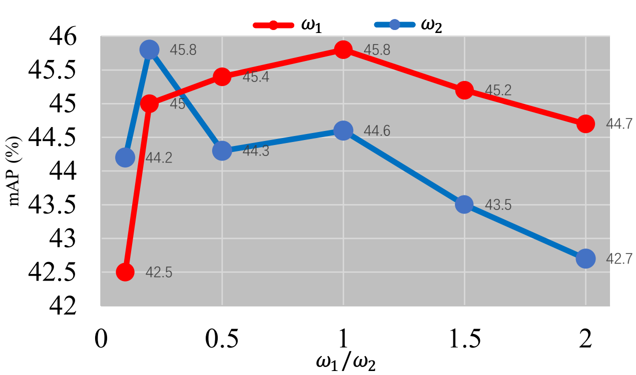

Figure 1. Comparsion results on Cityscapes → Foggy Cityscapes (%) of different ω_1 and ω_2. For the blue line, we set ω_1 = 1.0 and change ω_2. For the red line, we set ω_2 = 0.2 and change ω_1.

*Equal contribution.

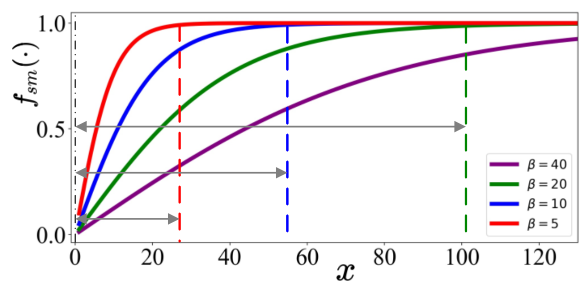

Figure 2. Function $f_{sm}(\cdot)$ with different β when $\alpha = 1$. Different β maps the object scales to different 0 to 1 space, which can represent the difference in scales.

2. Parameter Sensitivity Analysis

As shown in Fig. 1, we analysis the sensitivity of ω_1 and ω_2, where ω_1 works for the SGFF and ω_2 works for SAFE in our CSDA. We adopt a control variate method to study the effect of ω_1 and ω_2.

We first set $\omega_1 = 1.0$ and change the values of ω_2, finding that $\omega_2 = 0.2$ achieves the best results. By fixing ω_2, increasing ω_1 promotes some further improvements. However, the performance suffers a slight drop when $\omega_1 > 1.0$. These results demonstrate that suitable ω_1 can promote the feature alignment within the same scale and suitable ω_2 can promote the interaction learning among objects with different scales. ω_1 and ω_2, which are too small, will lead to the inadequate alignment and interaction of features. ω_1 and ω_2, which are too large, will lead to the degradation of the detection branch. We empirically set ω_1 to 1.0 and ω_2 to 0.2 in our model.

3. Analysis of Scale Mapping Function

As shown in Fig. 2, when the β increases, the function $f_{sm}(\cdot)$ has an expanding range of perception for scale (red, blue, and green dotted line), which helps SAFE perceive the objects of various scales better. However, the distinguishing

Method	Backbone	person	rider	car	truck	bus	train	mbike	bicycle	mAP
GPA [10]		32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5
EPM [3]		39.9	38.1	57.3	28.7	50.7	37.2	30.2	34.2	39.5
DIDN [5]		38.3	44.4	51.8	28.7	**53.3**	34.7	32.4	40.4	40.5
DSS [9]	ResNet50	42.9	**51.2**	53.6	33.6	49.2	18.9	**36.2**	**41.8**	40.9
SDA [6]		38.8	45.9	57.2	29.9	50.2	**51.9**	31.9	40.9	43.3
SIGMA [4]		44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2
CSDA(Ours)		**44.3**	44.0	**61.9**	**34.0**	52.7	51.0	35.8	38.6	**45.3**

Table 1. Experimental results (%) on Cityscapes→Foggy Cityscapes.

ability of the function $f_{sm}(\cdot)$ for objects at different scales starts to decrease when $\beta > 20$, which leads to a drop in performance. Empirically, we set β to 20 in the final model.

4. Visualization of Feature Distribution

In Fig. 3, we utilize category-wise t-SNE [8] to visualize the feature distribution of source and target domain on Cityscapes→Foggy Cityscapes. Compared with the baseline model, the features of the same category are better aligned, and the features of different categories are separated more clearly.

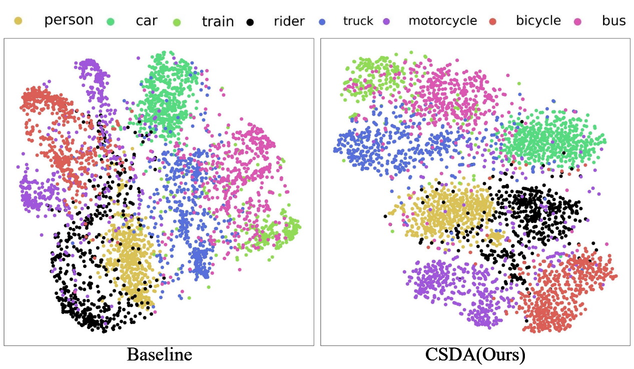

Figure 3. Feature comparison via category-wise t-SNE between the baseline model and our CSDA. For each category, we randomly sample object features (marked as circles) inside bounding boxes in the source domain and target domain.

In Fig. 4, the features (CSDA) of different scales are separated more clearly compared with the baseline model.

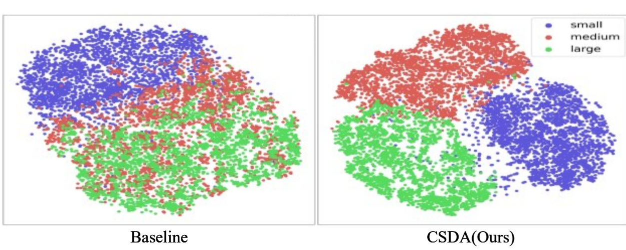

Figure 4. Feature comparison via scale-wise t-SNE between the baseline model and our CSDA. For each scale, we randomly sample object features (marked as circles) inside bounding boxes in the source domain and target domain.

5. Limitation Analysis

Our CSDA is good at aligning the features with the small aspect ratio. However, for objects with large aspect ratios, there are still some limitations (*e.g.*, bus in Fig. 5). It is because the aspect ratio of objects has a significant effect on features, especially for large objects. It makes our CSDA ineffective in aligning the features of the large scale while large aspect ratio objects.

Figure 5. Illustration of some failure examples. The bus with large aspect ratios cannot be detected well.

6. More Visualization Results

In this section, we provide more adaptation visualization results to further verify the effectiveness of our CSDA. The results on Cityscapes→Foggy Cityscapes, Sim10k→Cityscapes, and KITTI→Cityscapes are shown in Fig. 6, Fig. 7, and Fig. 8, respectively.

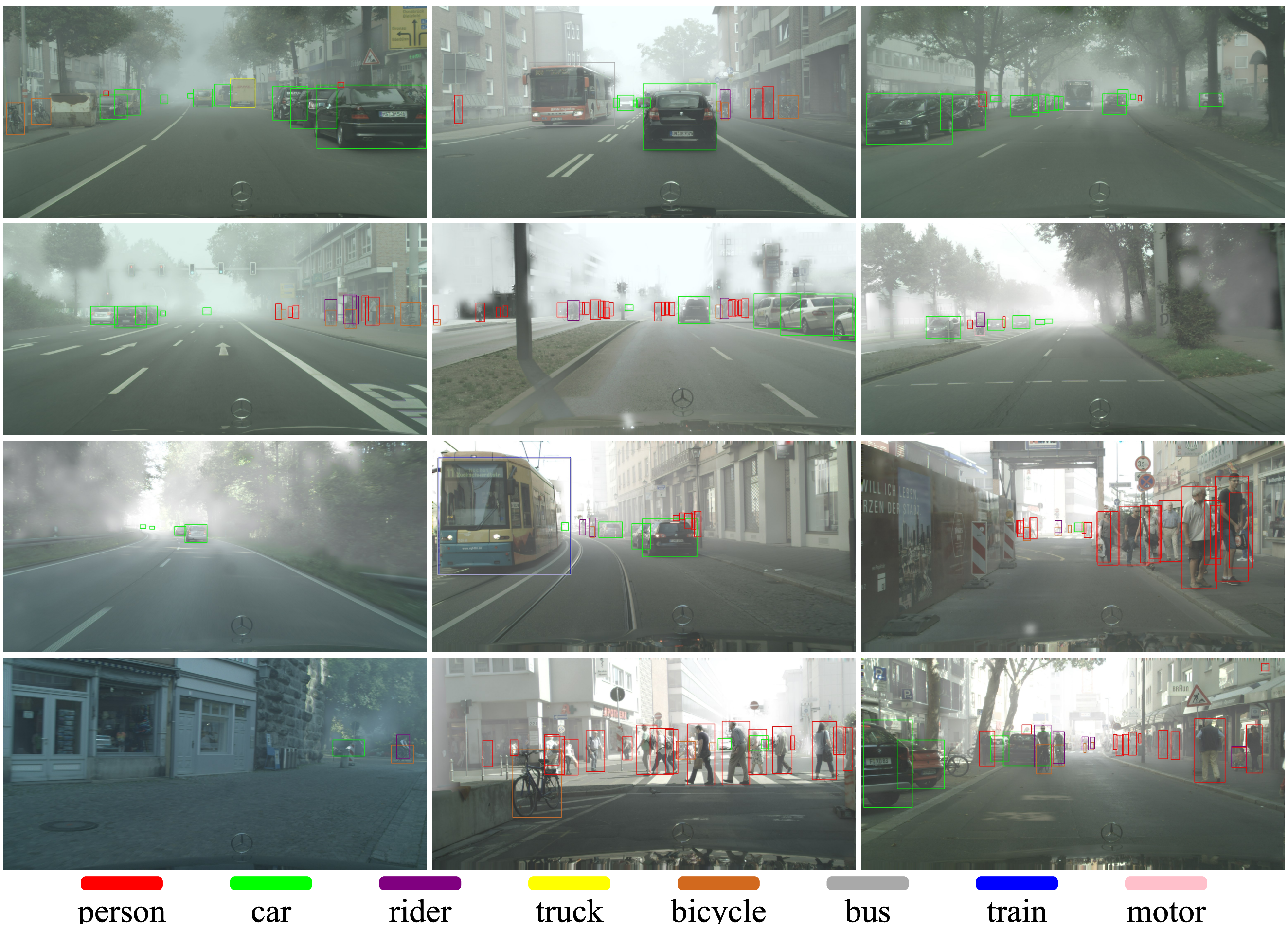

Figure 6. Normal to foggy adaptation visualization results, Cityscapes→Foggy Cityscapes.

Figure 7. Synthetic to real adaptation visualization results, Sim10k→Cityscapes.

Figure 8. Cross camera adaptation visualization results, KITTI→Cityscapes.

References

[1] 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. 1

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

[3] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *ECCV*, pages 733–748. Springer, 2020. 2

[4] Wuyang Li, Xinyu Liu, and Yixuan Yuan. SIGMA: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, pages 5291–5300, 2022. 2

[5] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *ICCV*, pages 8771–8780, 2021. 2

[6] Farzaneh Rezaeianaran, Rakshith Shetty, Rahaf Aljundi, Daniel Olmeda Reino, Shanshan Zhang, and Bernt Schiele. Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In *ICCV*, pages 9204–9213, 2021. 2

[7] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126:973–992, 2018. 1

[8] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 2

[9] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, YangYang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *CVPR*, pages 9603–9612, 2021. 2

[10] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, pages 12355–12364, 2020. 2