Dual-level Interaction for Domain Adaptive Semantic Segmentation

Supplementary Material

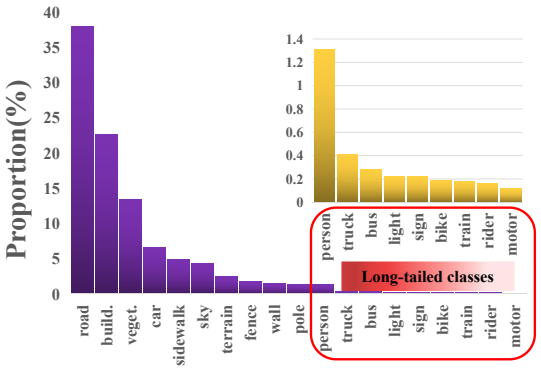

Figure 1: Category distribution of GTA5 [3]

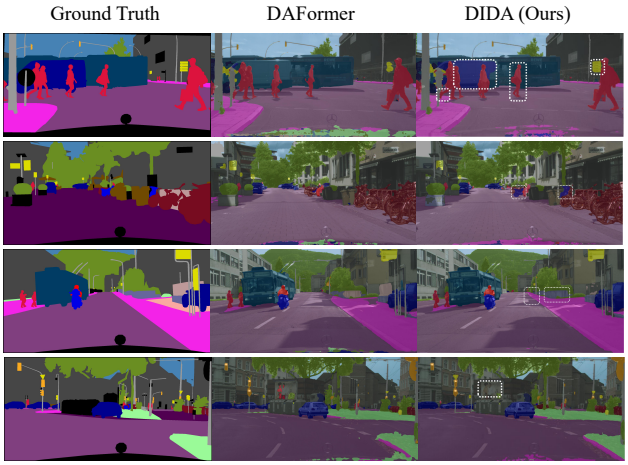

Figure 2: Visual comparisons with semantic-only self-training method DAFormer [2].

1. Additional Experiments

1.1. Class distribution analysis

We examine the distribution of all categories in the source GTA5 [3] dataset, and find the imbalanced long-tail distribution as displayed in Figure 1. In our further observation, the cause are mainly sorted into two reasons: on the one hand, some entities in small size possess very limited pixels, *e.g.*, "light" and "sign", on the other hand, some classes such as "train", "bike", "rider" and "motor" only appear in very few image samples. This also happens under real-world circumstances.

1.2. More ablation results

Temperature. The temperature tp in $q_{i'}^{A} = \frac{\exp\left((e_t^A)^T e_{i'}/tp\right)}{\sum_{k=1}^{K}\exp\left((e_t^A)^T e_k/tp\right)}$ and $q_{i'}^{\alpha} = \frac{\exp\left((e_t^\alpha)^T e_{i'}/tp\right)}{\sum_{k=1}^{K}\exp((e_t^\alpha)^T e_k/tp)}$, controls the sharpness of the instance distribution. (Noted $tp = 0$ is equivalent to the argmax operation). As shown in Table 1, the top result is obtained at $tp = 0.1$ and slightly drops when $tp = 0.05$.

Table 1: Results of temperature tp

tp	0.05	0.1	0.5	1.0
mIoU	70.5	**71.0**	70.3	70.0

Parameter Sensitivity of $\mathcal{L}_{ins}$. We experiment over different weights of $\mathcal{L}_{ins}$, and find that the weight of 1.0 secures the best mIoU result (shown in Table 2). This displays an equal contribution of the instance-level information as the semantic-level.

Table 2: Results of loss weight parameter λ

λ	0.05	0.1	0.5	1.0	2.0
mIoU	69.3	69.8	70.2	**71.0**	70.5

Runtime and Memory Consumption. DIDA can be trained on a single RTX 3090 GPU within 17 hours (0.65 it/s) while consuming about 20 GB GPU memory (24 GB total) during training. In comparison, our baseline model DAFormer [2] consumes 14 GB GPU memory (shown in Table 3). Noted that the extra 6 GB storage is the instance feature bank and all tensors for calculation combined. With our strictly controlled bank size K, we successfully introduce instance-level discrimination and improve adaptation performance by a notable margin.

Table 3: Bank size K and corresponding memory consumption

K	–	100	150	200	250	300
Memory (GB)	14.1	16.5	17.8	18.6	20.4	23.6

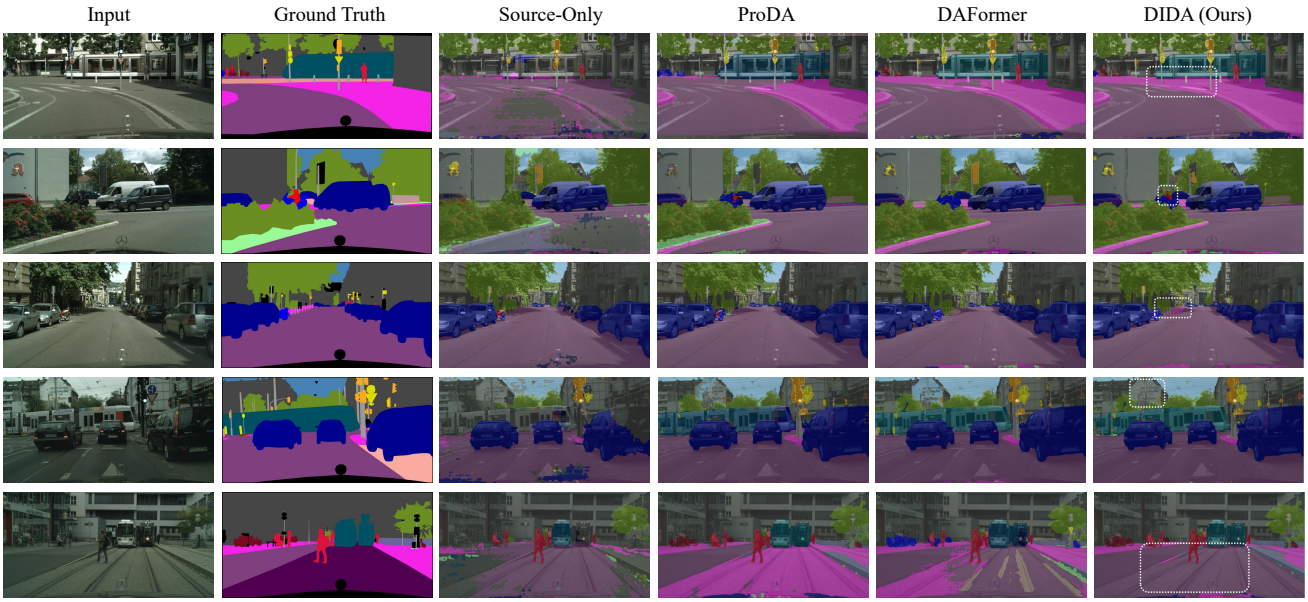

Figure 3: Qualitative results on GTA5 [3] → Cityscapes [1] adaptation. Comparison among (From left to right) Input target domain image, Ground Truth, Source-Only model, two semantic-level self-training models (ProDA [4], DAFormer [2]) and our DIDA method.

1.3. Visualization and Analysis

Qualitative comparison with semantic-only self-training method. As demonstrated in Figure 2, we visualize the segmentation results of the previous state-of-the-art model DAFormer [2], which used to be the most effective self-training method, and our DIDA. As we mentioned before, the semantic-level only self-training (*i.e.*, pseudo-labeling) method creates noisy semantic pseudo-labels leading to the wrong optimizing direction. This is particularly obtrusive when there are analogous and overlapping entities appearing in the same crop of input image. For example, in the case of the **1st row** in Figure 2, several "people" are walking in front of the "car" and the "bus" (these two entities are similar-looking), creating an overlapping situation. The baseline DAFormer may find it strenuous to clearly distinguish between the entity boundaries and end up with unsatisfactory performance. Similar problems also occur in three other rows of Figure 2: the **2nd row**, DAFormer misclassifies "motor" into "bike"; the **3rd row**, DAFormer struggles with "road"/"sidewalk" and "building"/"fence", which share similar textures; and the **4th row**, DAFormer recognizes person figures appeared in the billboard but are actually meant to be sorted as "building" as a whole. Instead, our proposed DIDA efficiently adjusts noisy semantic-level pseudo-labels using the discrimination and consistency regularization from the instance-level, resulting in a more explicit and accurate segmentation.

More visualization comparisons with ProDA [4] and DAFormer [2]. In Figure 3 we provide more visualized comparisons of baseline methods [4, 2]. As displayed in Figure 3, the major performance gain also comes from a better recognition of class "sidewalk" (1st row), "rider" (2nd row), "pole" and "vegetation" (3rd row), "sky" (4th row), and "road" (5th row). These results indicate that our DIDA outperforms existing baselines stably, especially on long-tailed and overlapping entities.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[2] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022.

[3] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016.

[4] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021.