Style Blind Domain Generalized Semantic Segmentation via Covariance Alignment and Semantic Consistence Contrastive Learning

Supplementary Material

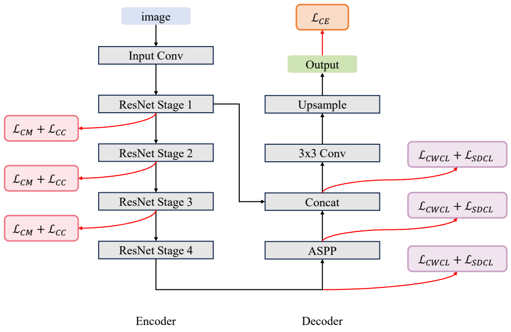
image
Input Conv
ResNet Stage 1
ResNet Stage 2
ResNet Stage 3
ResNet Stage 4
$\mathcal{L}_{CM} + \mathcal{L}_{CC}$
$\mathcal{L}_{CM} + \mathcal{L}_{CC}$
$\mathcal{L}_{CM} + \mathcal{L}_{CC}$
$\mathcal{L}_{CE}$
Output
Upsample
3x3 Conv
Concat
ASPP
$\mathcal{L}_{CWCL} + \mathcal{L}_{SDCL}$
$\mathcal{L}_{CWCL} + \mathcal{L}_{SDCL}$
$\mathcal{L}_{CWCL} + \mathcal{L}_{SDCL}$
Encoder
Decoder

Figure 1. Implementation details of proposed loss functions in the DeepLabV3+ architecture.

A. Implementation Details of BlindNet

As shown in Fig. 1, we apply our covariance alignment losses to the encoder features and the semantic consistency contrastive learning to the decoder features.

B. More Results

In this section, we show the detailed quantitative compassion results (Section B.1) and additional qualitative results (Section B.2) of our study.

B.1. Quantitative Results

Table 1 reports a comparison of pixel accuracy and IoU for each semantic class between DGSS methods. Our model significantly outperforms others in overall pixel accuracy, indicating its robust performance. In IoU for each semantic class, our model particularly excels in roads, sidewalks, sky, people, riders, and cars, which are commonly present in photos. However, the table also indicates a degraded performance in classes such as traffic signs, traffic lights, and trains, which are less frequently encountered in the source domain (GTAV). Our future work will aim to address this issue and improve performance across all classes.

B.2. Qualitative Results

Figs. 2 (G→C), 3 (G→B), and 4 (G→M) present qualitative comparisons between our model and others, including baseline [**?**], RobustNet [**?**], WildNet [**?**], SiamDoGe [**?**], and SPC [**?**]. The results clearly illustrate our model's consistent superiority, particularly in the segmentation of sidewalks, roads, buildings, terrain, and cars. The result demonstrates the robustness and effectiveness of our model in handling DGSS.

ω_1 ($\mathcal{L}_{CM}$)	ω_2 ($\mathcal{L}_{CC}$)	ω_3 ($\mathcal{L}_{CWCL}$)	ω_4 ($\mathcal{L}_{SDCL}$)	C	B	M	S
0.2	0.2	0.3	0.3	**45.72**	**41.32**	47.08	**31.39**
0.1	0.2	0.3	0.3	45.06	39.37	45.14	31.09
0.3	0.2	0.3	0.3	43.04	38.75	44.69	29.58
0.2	**0.1**	0.3	0.3	44.15	39.15	46.00	30.62
0.2	**0.3**	0.3	0.3	44.78	40.01	46.56	30.74
0.2	0.2	**0.2**	0.3	43.42	39.24	45.55	30.40
0.2	0.2	**0.4**	0.3	45.52	39.88	45.73	30.20
0.2	0.2	0.3	**0.2**	44.58	40.42	**47.35**	30.72
0.2	0.2	0.3	**0.4**	45.26	40.16	46.91	30.49

Table 2. Sensitivity to weighting parameters of each loss function

C. More Ablation Studies

In this section, we conduct more ablation studies on our model. In Section C.1, we show a qualitative analysis of the proposed loss functions, and in Section C.2, we experiment on the weight of the proposed loss functions.

C.1. Qualitative Results

We incrementally added each loss function ($\mathcal{L}_{CM}$, $\mathcal{L}_{CC}$, $\mathcal{L}_{CWCL}$, $\mathcal{L}_{SDCL}$) to the baseline model to validate the impact of loss. Fig. 5 presents the qualitative results of the ablation studies on the proposed loss functions.

For our qualitative ablation study, we added each loss function ($\mathcal{L}_{CM}$, $\mathcal{L}_{CC}$, $\mathcal{L}_{CWCL}$, $\mathcal{L}_{SDCL}$) to the baseline model, validating their contributions. The results are depicted in Fig. 5. The introduction of CML ($\mathcal{L}_{CM}$) enhances the capture of the details such as traffic lights, as illustrated in Fig. 5 row 1. Adding CCL ($\mathcal{L}_{CC}$) further strengthens content representation, leading to an improvement in overall accuracy. The CWCL ($\mathcal{L}_{CWCL}$) strengthens semantic understanding, allowing for better detection of smaller objects. However, this enhancement comes with a tradeoff, as it introduces some degree of confusion among similar classes (*e.g.* sidewalk and road). The application of SDCL ($\mathcal{L}_{SDCL}$) effectively disentangles misclassified features, leading to clearer class distinctions.

C.2. Hyper-parameter

We varied the weighting parameters for each loss function in (**??**), and conducted experiments by adjusting each loss weight by 0.1, using the model configuration that initially showed the best performance as our baseline, reported in Table 2. The CML ($\mathcal{L}_{CM}$), a key component for style blindness, shows that an overly strong influence can significantly degrade network performance. Conversely, the CCL ($\mathcal{L}_{CC}$) and the CWCL ($\mathcal{L}_{CWCL}$) exhibit improved performance with a slightly higher influence than a lower influence.

Methods	Pixel Accuracy	mIoU	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic light	Traffic sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle
			🟪	🟪	⬛	🟪	🟫	⬜	🟧	🟨	🟩	🟩	🟦	🟥	🟥	🟦	⬛	🟦	🟦	🟦	🟥
Baseline [?]	71.02	29.0	51.9	20.6	57.2	22.4	21.0	25.3	24.9	10.1	61.3	23.7	52.0	53.8	13.6	51.2	19.5	21.2	0.3	12.0	8.1
RobustNet [?]	77.18	37.3	58.9	27.7	63.2	22.8	23.1	26.4	30.6	20.7	85.1	39.2	69.8	62.4	15.9	76.7	23.2	22.3	3.9	18.4	18.6
SiamDoGe [?]	84.73	43.0	83.7	34.1	78.6	26.4	25.6	26.0	**42.4**	**28.6**	84.3	28.1	68.9	62.1	31.1	85.6	31.3	28.9	3.5	22.8	23.3
WildNet [?]	84.57	44.6	81.2	38.2	76.9	28.1	25.1	35.1	32.1	24.5	**85.4**	35.4	72.2	65.0	27.3	85.5	29.7	33.2	**12.6**	**32.8**	27.4
SPC [?]	86.65	44.1	86.9	37.8	81.2	28.9	26.9	**36.9**	35.1	25.2	83.7	36.2	78.5	63.9	30.4	84.1	24.8	28.1	12.1	19.3	17.9
DPCL [?]	82.22	44.7	75.6	32.8	73.2	26.1	23.5	34.1	42.3	28.2	85.2	38.5	**81.2**	63.8	25.0	76.6	31.7	33.9	5.7	27.6	45.0
Ours	**87.91**	**45.7**	**88.3**	**44.1**	**82.4**	**30.9**	**26.8**	35.4	33.4	20.3	85.0	**34.2**	78.5	**66.0**	**33.7**	**86.8**	**33.0**	**41.1**	1.4	25.3	22.1

Table 1. Quantitative results for pixel accuracy and each semantic class. The models are trained on GTAV and tested on Cityscapes using a ResNet50 backbone. The best and second best results are **bolded** and underlined, respectively

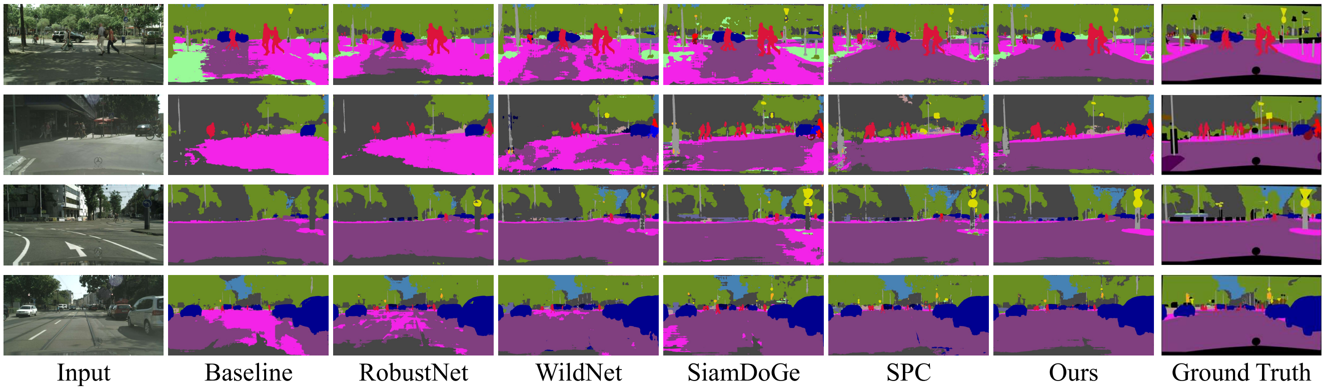

Figure 2. Qualitative comparison between DGSS methods trained on GTAV (G) and tested on unseen target domains of Cityscapes (C) using DeeplabV3+ with ResNet50 backbone.

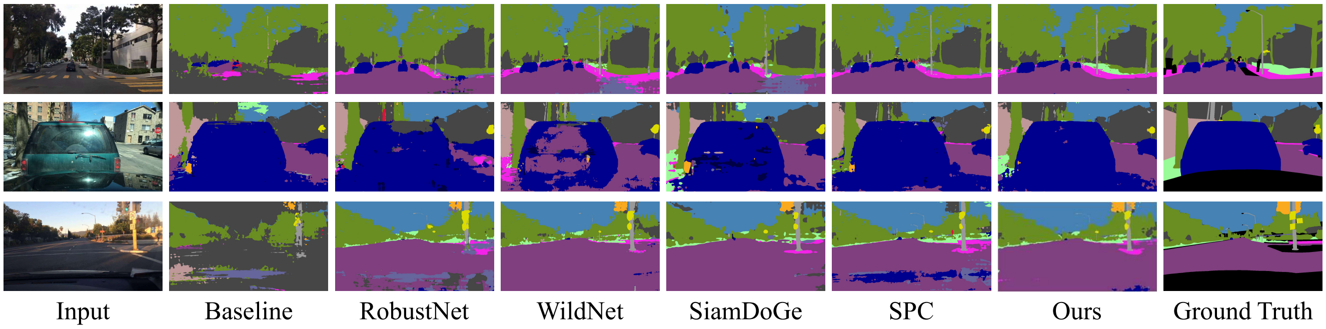

Figure 3. Qualitative comparison between DGSS methods trained on GTAV (G) and tested on unseen target domains of BDD100K (B) using DeeplabV3+ with ResNet50 backbone.

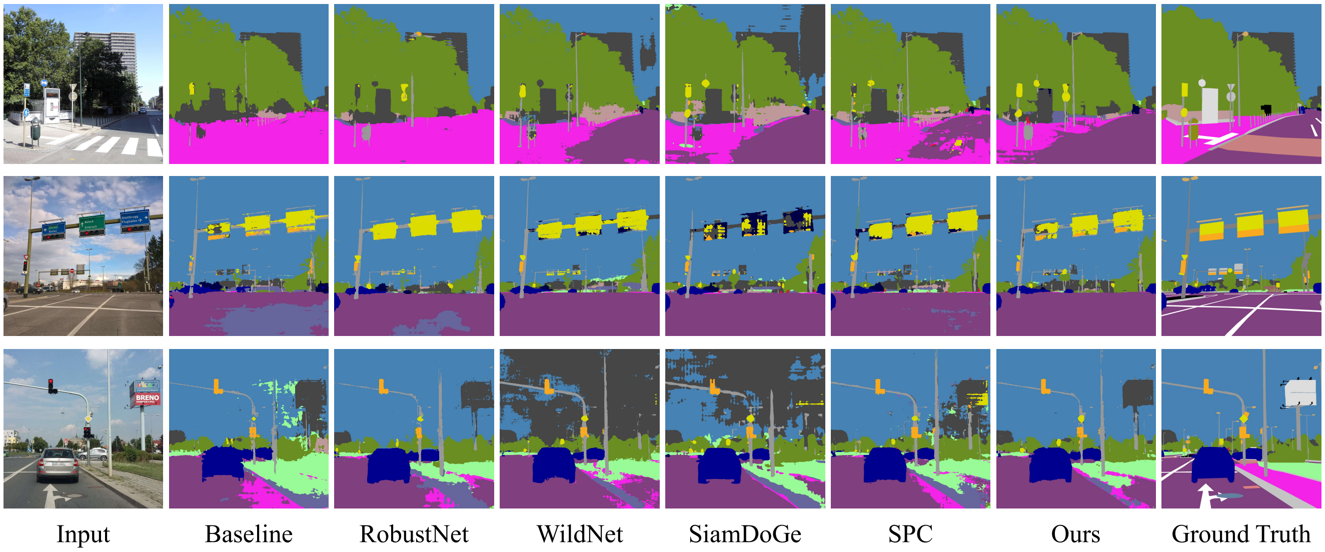

Figure 4. Qualitative comparison between DGSS methods trained on GTAV (G) and tested on unseen target domains of Mapillary (M) using DeeplabV3+ with ResNet50 backbone.

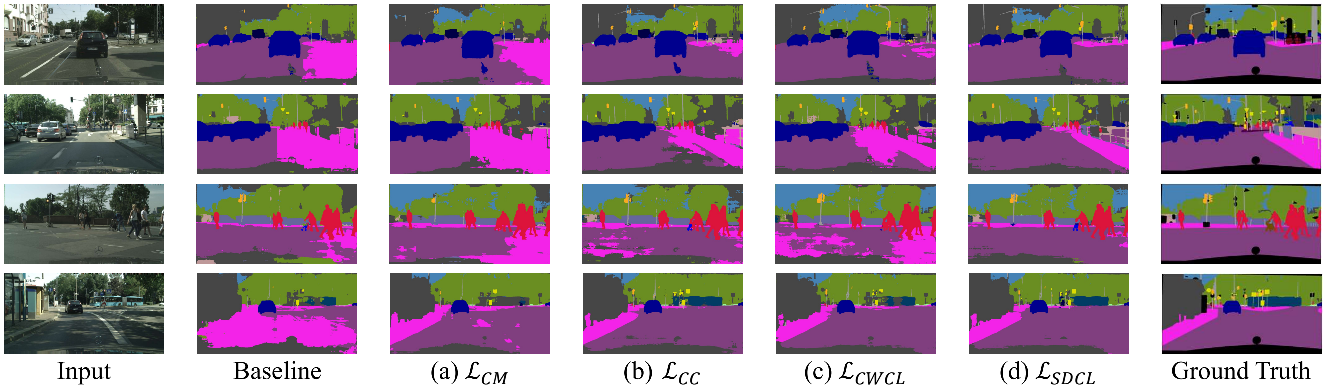

Figure 5. Qualitative comparison for ablation studies. The models are trained on GTAV (G) and tested on unseen target domains of Mapillary (M) using DeeplabV3+ with ResNet50 backbone. (a) $\mathcal{L}_{CM}$. (b) $\mathcal{L}_{CM} + \mathcal{L}_{CC}$. (c) $\mathcal{L}_{CM} + \mathcal{L}_{CC} + \mathcal{L}_{CWCL}$, (d) $\mathcal{L}_{CM} + \mathcal{L}_{CC} + \mathcal{L}_{CWCL} + \mathcal{L}_{SDCL}$