Mamba as a Bridge: Where Vision Foundation Models Meet Vision Language Models for Domain-Generalized Semantic Segmentation

Supplementary Material

1. Evaluate on Additional VFMs

Besides DINOv2 in the main text, we additionally evaluate VFMs, BEiT2 [11] and iBOT [15]. Both of them are of the *Large* size. EVA02-CLIP is utilized as the VLM. As shown in Tab. 1, they also improve the performance of solely using VLM.

Table 1. Ablation studies on more VFMs under the G→{C, B, M} setting. EVA02-CLIP is utilized as the VLM by default. BEiT2 [11] and iBOT [15] are evaluated as VFMs, respectively. Both are of *Large* types.

	C	B	M	Avg.
VLM-only	68.26	60.02	70.18	66.15
+ BEiT2-L	69.60	60.19	70.39	66.73
+ iBOT-L	69.37	60.76	70.53	66.89

2. Evaluate on SYNTHIA Benchmarks

We compare the performance of the proposed MFuser with existing state-of-the-art DGSS methods under the Synthia→{C, B, M} (as shown in Tab. 2), G→Synthia and C→Synthia (as shown in Tab. 3) settings. MFuser achieves the best performance on all settings.

3. Evaluate on ACDC Benchmarks

We compare the performance of the proposed MFuser with existing state-of-the-art DGSS methods under the clear-to-adverse-weather setting. Models are trained on Cityscapes and tested on ACDC which is composed of four domains, namely *fog*, *night*, *rain* and *snow*. As shown in Tab. 4, we consistently outperform the existing methods by a large margin. Particularly, we surpass SET on *rain* by 3.79 mIoU.

Table 2. Performance comparison (mIoU in %) under the synthetic-to-real setting (S→{C, B, M}). Note that we implement DINOv2 [6] as the VFM and EVA02-CLIP [3] as the VLM. Our method is marked in gray. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Backbone	synthetic-to-real			
		S→C	S→B	S→M	Avg.
SAN-SAW [10]	RN101	40.87	35.98	37.26	38.04
TLDR [4]	RN101	42.60	35.46	37.46	38.51
IBAFormer [12]	MiT-B5	50.92	44.66	50.58	48.72
Rein [13]	DINOv2-L	48.59	44.42	48.64	47.22
SET [14]	DINOv2-L	49.65	45.45	49.45	48.18
MFuser	EVA02-L	**54.17**	**46.67**	**53.22**	**51.35**

Table 3. Performance comparison (mIoU in %) under G→S and C→S. Note that we implement DINOv2 [6] as the VFM and EVA02-CLIP [3] as the VLM. Our method is marked in gray. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Backbone	G→Synthia	C→Synthia
Rein [13]	DINOv2-L	48.86	48.56
SET [14]	DINOv2-L	50.01	49.61
tqdm [7]	EVA02-L	53.32	50.62
MFuser	EVA02-L	**54.04**	**54.13**

Table 4. Performance comparison (mIoU in %) on Cityscapes→ACDC. Note that we implement DINOv2 [6] as the VFM and EVA02-CLIP [3] as the VLM. Our method is marked in gray. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Backbone	clear-to-adverse-weather				
		→Fog	→Night	→Rain	→Snow	Avg.
IBN [8]	RN50	63.80	21.20	50.40	49.60	46.25
IW [9]	RN50	62.40	21.80	52.40	47.60	46.05
ISW [2]	RN50	64.30	24.30	56.00	49.80	48.60
ISSA [5]	MiT-B5	67.50	33.20	55.90	53.20	52.45
CMFormer [1]	Swin-L	77.80	33.70	67.60	64.30	60.85
Rein [13]	DINOv2-L	79.48	55.92	72.45	70.57	69.61
SET [14]	DINOv2-L	80.06	57.29	74.80	73.69	71.46
tqdm [7]	EVA02-L	81.28	54.80	72.92	72.41	70.35
MFuser	EVA02-L	**82.33**	**57.94**	**78.59**	**74.93**	**73.45**

4. Ablation on the Number of MVFusers

We evaluate the effect of the number of MVFusers utilized for feature fusion. To do so, MVFuser is inserted after every N blocks. As shown in Tab. 5, more MVFusers generally improve performance.

Table 5. Ablation studies on the number of MVFusers under the G→{C, B, M} setting. Note that we implement DINOv2 [6] as the VFM and EVA02-CLIP [3] as the VLM.

N	C	B	M	Avg.
8	69.20	61.85	69.24	66.76
4	68.02	61.69	69.96	66.56
2	70.49	62.71	70.78	67.99
1	**70.19**	**63.13**	**71.28**	**68.20**

5. More Qualitative Results

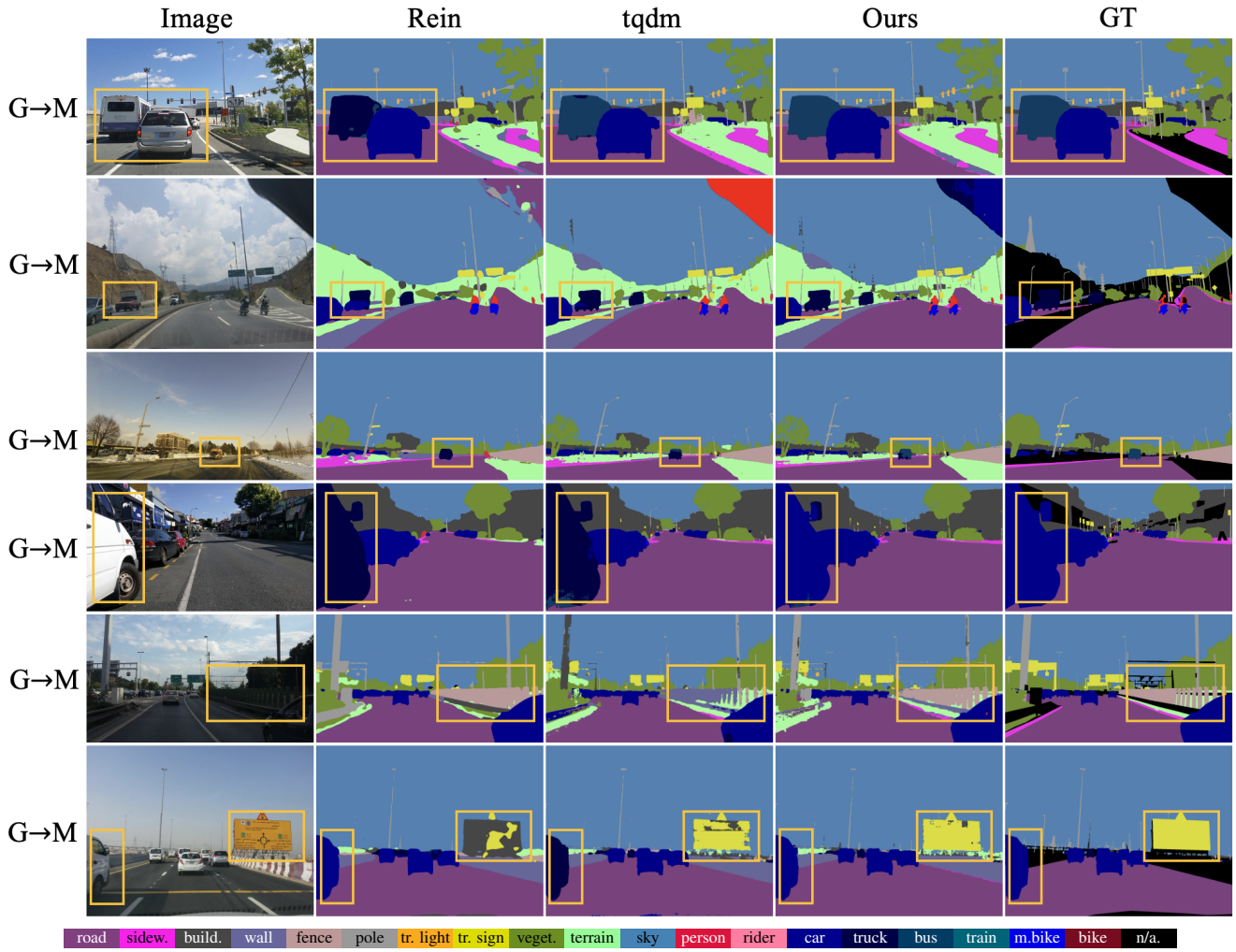

Figure 1. Qualitative results on unseen target domains under the G→M setting. MFuser is compared with Rein [13] and tqdm [7].

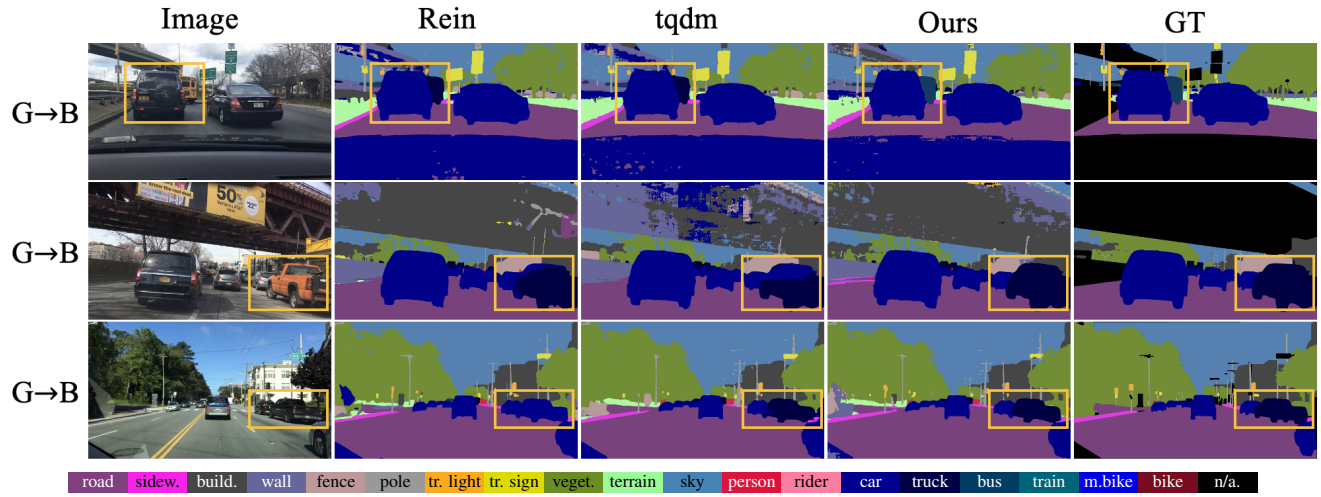

Figure 2. Qualitative results on unseen target domains under the G→B setting. MFuser is compared with Rein [13] and tqdm [7].

References

[1] Qi Bi, Shaodi You, and Theo Gevers. Learning content-enhanced mask transformer for domain generalized urban-scene segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 819–827, 2024. 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11580–11590, 2021. 1

[3] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. Image and Vision Computing, page 105171, 2024. 1

[4] Sunghwan Kim, Dae-hwan Kim, and Hoseong Kim. Texture learning domain randomization for domain generalized segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 677–687, 2023. 1

[5] Yumeng Li, Dan Zhang, Margret Keuper, and Anna Khoreva. Intra-source style augmentation for improved domain generalization. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 509–519, 2023. 1

[6] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023. 1

[7] Byeonghyun Pak, Byeongju Woo, Sunghwan Kim, Dae-hwan Kim, and Hoseong Kim. Textual query-driven mask transformer for domain generalized segmentation. In European Conference on Computer Vision, pages 37–54. Springer, 2025. 1, 2

[8] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In Proceedings of the european conference on computer vision (ECCV), pages 464–479, 2018. 1

[9] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1863–1871, 2019. 1

[10] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2594–2605, 2022. 1

[11] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint arXiv:2208.06366, 2022. 1

[12] Qiyu Sun, Huilin Chen, Meng Zheng, Ziyan Wu, Michael Felsberg, and Yang Tang. Ibaformer: Intra-batch attention transformer for domain generalized semantic segmentation. arXiv preprint arXiv:2309.06282, 2023. 1

[13] Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 28619–28630, 2024. 1, 2

[14] Jingjun Yi, Qi Bi, Hao Zheng, Haolan Zhan, Wei Ji, Yawen Huang, Yuexiang Li, and Yefeng Zheng. Learning spectral-decomposited tokens for domain generalized semantic segmentation. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 8159–8168, 2024. 1

[15] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832, 2021. 1