Open-Vocabulary Domain Generalization in Urban-Scene Segmentation (Supplementary Material)

Dong Zhao[1], Qi Zang[2] ✉, Nan Pu[2], Wenjing Li[2], Nicu Sebe[1], Zhun Zhong[2] ✉
[1] Department of Information Engineering and Computer Science, University of Trento, Italy
[2] School of Computer Science and Information Engineering, Hefei University of Technology, China

A. More Dataset Details

As shown in Table 1, our OVDG-SS benchmark spans a progressively expanding set of semantic categories, beginning with the 7 basic driving classes from Cityscapes and GTA, including *road, sidewalk, building, vegetation, sky, person, car*.

The 19-class datasets (ACDC-19 [21], BDD-19 [28], Mapi-19 [15]) extend this space by incorporating additional urban background categories *wall, fence, pole, traffic light, traffic sign, terrain* and dynamic traffic participants *rider, truck, bus, train, motorcycle, bicycle*. These datasets differ in environmental conditions, covering adverse weather, varied illumination, and globally diverse street scenes.

The 41-class synthetic datasets (ACDC-41, BDD-41) further expand the vocabulary with a wide range of open-world objects introduced through our diffusion-based inpainting pipeline. In addition to the 19 urban classes, these datasets include the full set of inpainted objects across multiple semantic groups: animals (*bird, cat, cow, deer, dog, elephant, horse, sheep, zebra*), man-made items (*bag, ball, barrel, bench, bottle, cart, chair, hat, table, toy, umbrella*), machines and devices (*drone, robot*), and other everyday objects. These inpainted categories enrich the open-vocabulary evaluation setting by introducing diverse unseen visual concepts that do not appear in the original driving datasets.

Mapillary-30 includes an extended set of fine-grained street-scene categories, covering structural elements *bridge, tunnel*, natural entities *water, snow, sand*, traffic-related elements *traffic sign, traffic light, signboard, railway*, vehicles *boat*, and everyday objects *chair, trash can*, in addition to the core driving classes.

ROADWork-10 [6] focuses on construction-related semantics and introduces the unseen categories *barrier, police vehicle, work vehicle, police officer, worker, cone, arrow board, TTC sign*.

This structured progression from basic driving categories to diverse inpainted open-world objects enables comprehensive evaluation of OVDG-SS models under increasingly complex and heterogeneous semantic spaces.

B. More Implementation Details

B.1. Text prompt templates.

In our implementation, we continue to generate text embeddings by composing natural-language sentences from class names, but we extend the original single-template design with a set of domain-aware prompts. Instead of relying solely on "A photo of a {class}", we experiment with ten prompt templates that describe variations in environment, lighting, weather, and scene context. These templates include sentences such as "A photo of {class} in different environments", "An image of {class} under various conditions", and "A picture of {class} in rainy or foggy weather". This experimental set of 10 domain-oriented prompts provides more diverse textual cues and encourages the text encoder to better capture cross-domain characteristics. A more systematic exploration of prompt learning is left for future work.

B.2. Inference at High Resolutions

To perform inference at high resolutions in urban-scene images (e.g., 2K) with ViT-B/16 and ViT-L/14, we adopt a sliding-window strategy that differs from the patch inference approach in [2]. Our implementation follows a window-based scanning scheme parameterized by `SW_KERNEL`, `SW_OVERLAP`, and `SW_OUT_RES`. Specifically, for ViT-L/14, we use a kernel size of 448 and an overlap ratio of 0.333, while the input image is first resized to an effective output resolution of 448×896. Given the resized image, we generate a grid of overlapping windows of size 448×448. The stride is computed as $(1 - \texttt{SW\_OVERLAP})$ times the kernel size, and when the spatial dimensions cannot be evenly covered, the window origin is backtracked to ensure full coverage of the image. Each window is independently normalized, fed through the dense CLIP encoder, and then passed into the segmentation head to produce local probability maps. These window-level predictions are then aggregated by summing the logits over all overlapping

dataset	From	Size	Resolution	Classes
CS-7	Cityscapes Train [3]	2,975	2048×1024	Rd, Sw, Bldg, Veg, Sky, Per, Car
GTA-7	GTA Train [20]	24999	2048×1024	
ACDC-19	ACDC Train [21]	1,600	1920×1080	Rd, Sw, Bldg, Wal, Fen, Pol, TL, TS
BDD-19	BDD100K Validation [28]	1000	1280×720	Veg, Ter, Sky, Per, Rid, Car, Trk, Bus, Trn, Mot, Bic
Mapi-19	Mapillary Validation [28]	2000	2k-4k	
ACDC-41	SD 2.1 + ACDC Train	1000	1920×1080	Rd, Sw, Bldg, Wal, Fen, Pol, TL, TS, Veg, Ter, Sky, Per, Rid, Car, Trk, Bus, Trn, Mot, Bic, Bag, Bal, Bar, Bnc, Brd, Btl, Crt, Cat, Chr, Cow, Der, Dog, Drn, Elp, Hat, Hrs, Rbt, Shp, Tbl, Toy, Umb, Zbr
BDD-41	SD 2.1+ BDD100K Validation	1000	1280×720	
Mapi-30	Mapillary Training+Validation [28]	3943	2k-4k	Rd, Sw, Bldg, Wal, Brg, Tun, TS, TL, Pol, Fen, Sky, Veg, Ter, Wtr, Snw, Snd, Per, Rid, Car, Trk, Bus, Trn, Bic, Mot, Anm, Sbd, Rwy, Bot, Chr, Tc
RW-10	ROADWork Training [6]	2098	1280×720	Rd, Sw, Bar, PV, WV, PO, Wkr, Con, ABd, TTC

Table 1. Summary of datasets used in the OVDG-SS benchmark.

Algorithm 1 Spatial Aggregation in S^2-Corr

Require: Correlation embeddings $\mathbf{E} \in \mathbb{R}^{C\times T\times H\times W}$, appearance guidance $\mathbf{G}$, Chunk size L, decay prior γ.
Ensure: Refined spatial embeddings $\widetilde{\mathbf{E}}$.

1: **Modulation before aggregation:** compute modulation from $\mathbf{G}$ and update each location by Eq. (6) $\widehat{\mathbf{E}}_{:,h,w} = \mathbf{E}_{:,h,w} \odot (1+\gamma_{h,w}) + \beta_{h,w}$.
2: **Row-wise chunking:** for each row, split the W tokens into chunks of length L and initialize a row state $\mathbf{s}_h^{(0)} = \mathbf{0}$.
3: **Chunk-wise 1D state-space update:** for each chunk and token $\mathbf{x}_t$, compute gates α_t, β_t and the decayed update coefficient $A_t = \sigma(w)\alpha_t + (1-\sigma(w))\gamma$. update the state and output token as $\mathbf{s}_t = A_t \odot \mathbf{s}_{t-1} + \beta_t \odot \mathbf{x}_t, \quad \mathbf{y}_t = W_t\mathbf{s}_t + U_t\mathbf{x}_t$, and store the final state of the chunk as $\mathbf{s}_{h,k}^{\text{end}}$.
4: **Cross-row state propagation:** for each row $h > 1$, initialize the new row state by $\mathbf{s}_h^{(0)} \leftarrow \eta_{\text{cross}}\, \mathbf{s}_{h-1,k}^{\text{end}}$, which induces snake-shaped spatial dependencies across rows.
5: **Reconstruction:** reassemble the tokens $\mathbf{y}_t$ back to the (H, W) layout to obtain $\widetilde{\mathbf{E}}$.

regions and normalizing by the accumulated weights, yielding a full-resolution prediction.

B.3. Algorithm Details

We provide the detailed algorithmic procedure of the proposed spatial aggregation in S^2-Corr module in Algorithm 1.

VLM	ACDC-19	BDD-19	Mapi-19	ACDC-41	BDD-41	Mapi-30	RW-10	Ave.
EVA-02-CLIP-L/14 [5]	54.3	53.1	60	62	61.7	47.4	41.9	44.7
SigLIP-ViT-L/16 [23]	52.7	54.1	59.2	60.5	60.4	**48.3**	**42.5**	54.0
CLIP-ViT-L/14 [19]	53.7	52.8	59.0	61.6	60.9	46.7	40.9	53.7
SelfCLIP-ViT-L/14 [25]	**56.3**	**53.6**	**60.9**	**63.8**	**63.5**	48.1	41.9	**59.6**

Table 2. Comparison of different Vision-Language Models (VLMs) under the OVDG-SS setting across multiple datasets. The best results are underlined.

C. More Results

C.1. Different VLM Pretrain

As shown in Table 2, among the evaluated VLMs, SelfCLIP-ViT-L/14 and SigLIP-ViT-L/16 achieve consistently stronger performance across the OVDG-SS datasets. Both models are post-distilled variants of CLIP that explicitly enhance the spatial alignment between image and text embeddings, making them inherently more suitable for pixel-level tasks such as semantic segmentation. In contrast, EVA-02-CLIP-L/14 and CLIP-ViT-L/14 do not include such spatial-consistency optimization. Although EVA-02 occasionally attains competitive scores due to its strong backbone pretraining, its performance varies more significantly across domains. Overall, these results indicate that VLMs with improved spatial alignment capabilities (SelfCLIP and SigLIP) provide more stable and robust open-vocabulary generalization, especially when transferring to unseen categories and diverse real-world conditions.

C.2. Impact of Hyperparameters

Chunk size L. Fig. 1(a) reports the sensitivity of S^2-Corr to the chunk size L. The results show that performance remains remarkably stable across a wide range of L, indicating that the proposed chunked state-space design is inherently robust to this hyperparameter. Using a moderate chunk size (e.g., L=12–16) provides the best balance, while further increasing L introduces longer-range dependencies

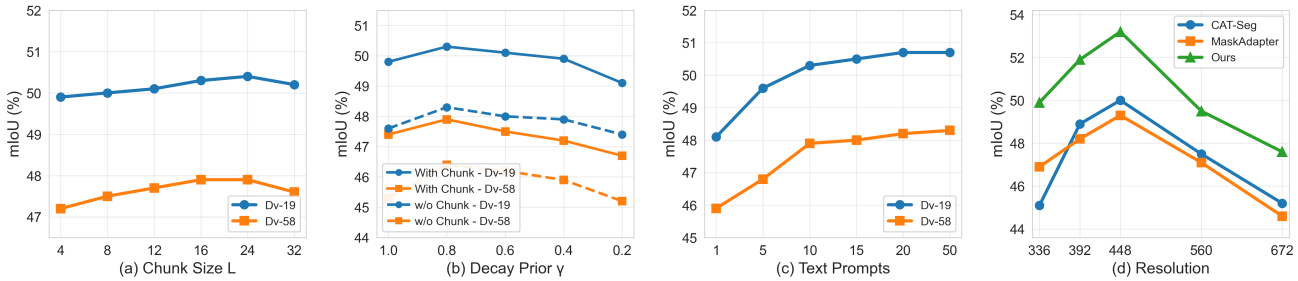

Figure 1. Impact of hyperparameters: chunk size, prior decay, text prompt templates, and training resolution

that do not yield additional gains and may slightly degrade performance due to accumulated noise. Overall, the method shows weak sensitivity to L, highlighting the effectiveness of chunk-wise aggregation.

Decay prior γ. Fig. 1(b) summarizes the effect of the decay prior γ with and without the chunk mechanism. With chunking enabled, performance peaks around γ=0.8, while both overly large and overly small values lead to slight drops. This indicates that a moderate decay prior provides the right balance between retaining useful long-range information and suppressing noise accumulation. Without chunking, the model becomes more sensitive to γ, showing consistently lower performance and a larger performance gap between γ=1 and γ=0.2. These results demonstrate that chunking stabilizes the state-space recurrence and makes the decay prior significantly more robust.

Number of text prompt templates. We further examine the effect of varying the number of text prompt templates. As shown in Fig. 1(c), increasing the prompt pool from a single template to a more diverse set consistently boosts performance, confirming that multi-domain textual descriptions help the VLM produce more robust class embeddings for OVDG-SS. However, the performance gain saturates when the number exceeds 10–15, while the computational cost continues to grow. We therefore adopt 10 templates as a practical trade-off between accuracy and efficiency.

Training Resolution. We also investigate the impact of training resolution when fine-tuning EVA02-based ViT-L/14 models. As shown in Fig. 1(d), performance is highly sensitive to the choice of resolution: resolutions that deviate too much from the original pre-training size lead to significant degradation for all methods (CAT-Seg, MaskAdapter, and ours). Very small resolutions underfit the spatial details of downstream data (typically around 512×1024), whereas excessively large resolutions break the alignment with the ViT-L/14 pre-training scale, resulting in severe mismatch and loss of geometric consistency. Our method consistently achieves the best performance across settings, and peaks near the pre-training–aligned resolution (around 448), confirming the importance of maintaining compatibility with the VFM's original image scale during fine-tuning.

C.3. Impact of Scanning ways in S^2-Corr

Variant	Row	Col	mIoU Dv-19	mIoU Dv-58	Training Time (Min)
Row-only (ours)	✓	×	50.3	47.9	140
Col	×	✓	48.6	45.7	140
Row+Col	✓	✓	50.4	48.1	165

Table 3. Ablation on different scanning strategies used in the spatial aggregation module.

As shown in Table 3, row-only scanning achieves a better balance between accuracy and efficiency. Pure column-wise scanning is clearly inferior on both domain settings, suggesting that horizontal spatial dependencies are more informative for correlation aggregation. The row+column variant brings only marginal improvements over row-only while introducing a notable increase in training time. Therefore, we adopt the row-only design as the default configuration in S^2-Corr.

C.4. Training on Large Vocabulary

Table 4 shows that under the full 19-class GTA training vocabulary, our OVDG framework continues to demonstrate strong cross-domain generalization. Compared with proposal-based OVSS approaches such as MaskAdapter and MAFT+, correlation-refinement–based methods (e.g., Cat-Seg) show clear advantages on seen-class evaluation, confirming the effectiveness of refining text–image correlations rather than relying on region proposals alone. More importantly, when trained with the larger visible vocabulary, our OVDG method surpasses Cat-Seg on both seen-class benchmarks (CS-19, BDD-19, Mapi-19) and on unseen-class evaluations across Mapi-30, RW-10, ACDC-41, and BDD-41. Finally, although existing DG methods (Rein, tqdm, SoMA) can only evaluate seen classes, our OVDG achieves comparable performance on the same 19-class benchmarks while additionally providing unseen-class detection capability, which DG methods cannot offer.

C.5. Detailed OVDG-SS results

We provide additional experimental results using the ViT-L backbone in Table 6. As shown in the table, our method consistently delivers highly competitive perfor-

Training Data: GTA-19											
	Type	Pretrain	CS-19	BDD-19	Mapi-19	Ave.	Mapi-30	RW-10	ACDC-41	BDD-41	Ave.
Rein (CVPR'24) [24]	DG	EVA02-L	65.3	60.5	64.9	63.6	-	-	-	-	-
tqdm (ECCV'24) [16]	DG	EVA02-L	68.9	59.2	70.1	66.1	-	-	-	-	-
SoMA (CVPR'25) [29]	DG	EVA02-L	68.1	60.8	68.3	65.7	-	-	-	-	-
MaskAdapter (CVPR'25) [11]	OVSS	EVA02-L	61.5	55.5	65.7	60.9	48.5	41.0	63.3	61.8	53.7
MAFT+(ECCV24) [8]	OVSS	EVA02-L	61.2	55.0	65.0	61.2	48.2	40.1	63.9	62.4	53.7
Cat-Seg(CVPR'24) [2]	OVSS	EVA02-L	64.1	59.0	66.7	63.3	50.8	44.2	65.8	66.2	56.7
S^2-Corr (Ours)	OVDG	EVA02-L	65.2	61.8	67.6	64.9	51.9	44.8	66.2	66.5	57.4

Table 4. Extended results under a larger 19-class training vocabulary. The main paper reports OVDG performance using a compact 7-class training set, this table presents results obtained with the full 19-class vocabulary (GTA-19).

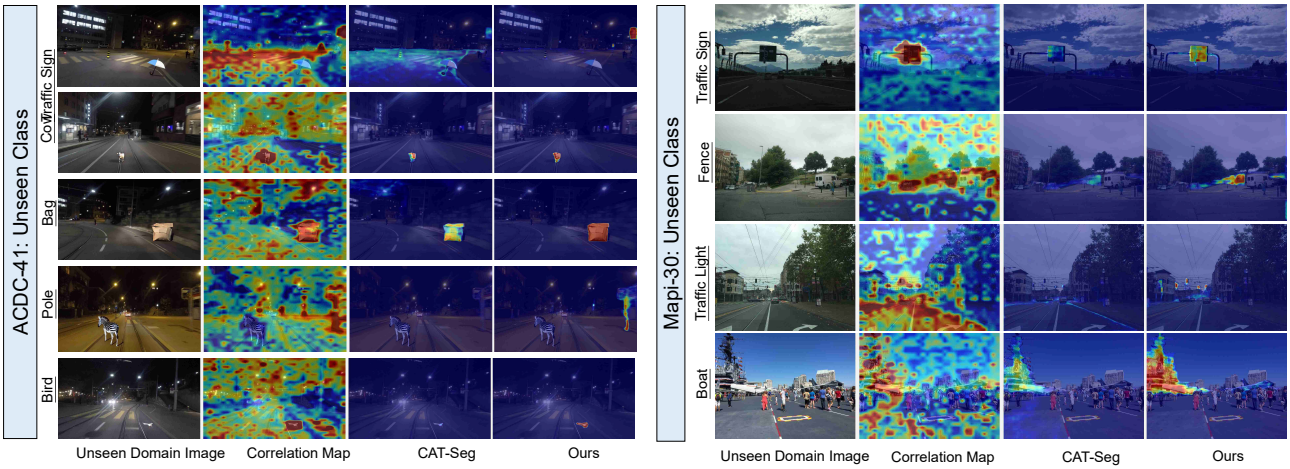

Figure 2. More comparison of initial text–image correlation aggregation on seen and unseen classes from unseen domains.

mance across all evaluated OVDG-SS settings. As shown in Tables 7, 8, 9, and 10, we report the per-class OVDG-SS results on CS-7 for all evaluation datasets. The results demonstrate that our method achieves consistently strong performance on both seen and unseen classes.

D. More Comparison

D.1. Correlation Refinement

We further provide additional comparisons of the initial text-image correlation aggregation on unseen classes from unseen domains in Fig. 2. Across multiple unseen categories and datasets, our method consistently filters domain-induced noise and produces more accurate spatial localization of unseen objects. These results reinforce the observations the main paper, demonstrating that our correlation refiner maintains robust semantic alignment even under significant domain shifts.

D.2. Compared with Clost-Set DG

Table 4 further compares our OVDG framework with recent DG approaches under the full 19-class training vocabulary. Rein [24] improves domain robustness by introducing carefully engineered adapters that regularize the VLM's vision encoder, yet its performance remains clearly lower than ours across all seen-class benchmarks. tqdm [16] modifies the VLM through a different strategy, constructing a fixed set of category queries and effectively collapsing the open vocabulary space into a closed-set formulation. Despite this constrained design tailored for DG, our method still achieves competitive or superior results on CS-19, BDD-19, and Mapi-19. These comparisons show that although DG methods explicitly optimize VLMs for closed-set generalization, our OVDG approach achieves comparable or competitive performance on seen classes while also providing open-vocabulary recognition, a capability that DG methods do not possess.

D.3. Compared with OOD Segmentation

Unlike conventional out-of-domain(OOD) segmentation in urban-scene understanding [4, 22], which focuses on detecting "anything outside a fixed set of ID classes" and typically represents all unknown content with a single anomaly score or a binary OOD mask, our open-vocabulary segmentation framework addresses a more general and more semantically meaningful problem setting. In many real-world road-driving scenarios, merely flagging "unknown" is insufficient: different unknown objects (e.g., animals, construction tools, debris, or uncommon vehicles) require distinct semantics for safe decision-making. Open-vocabulary

Methods	FS Static		FS Lost&Found		SMIYC-Anomaly		SMIYC-Obstacle		RoadAnomaly	
	IoU	mean F1	IoU	mean F1	IoU	mean F1	IoU	mean F1	IoU	mean F1
Synboost [4]	32.8	25.7	18.4	10.9	42.0	39.5	14.0	9.0	27.2	29.3
PEBAL [22]	26.9	13.3	6.4	2.6	42.4	35.1	6.7	1.1	33.8	23.9
RPL+CoroCL [14]	36.5	13.2	15.8	3.9	68.8	31.6	28.7	5.7	51.0	24.6
S2M [30]	70.0	70.2	30.5	35.3	77.5	60.4	67.6	65.0	58.5	61.7
Ours	**74.6**	**76.7**	**36.3**	**42.2**	**82.9**	**69.7**	**76.2**	**73.8**	**66.5**	**66.9**

Table 5. Comparison of open-set semantic segmentation on five anomaly benchmarks. Existing OOD segmentation methods (SynBoost, PEBAL, RPL+CoroCL, S2M) rely on COCO-Stuff segmentation supervision or additional anomaly masks for training. In contrast, our open-vocabulary segmentation model performs OOD detection purely through a text-extended vocabulary (Cityscapes 19 ID classes + 150 curated COCO-Stuff OOD concepts vocabulary) without using any COCO-Stuff pixel annotations.

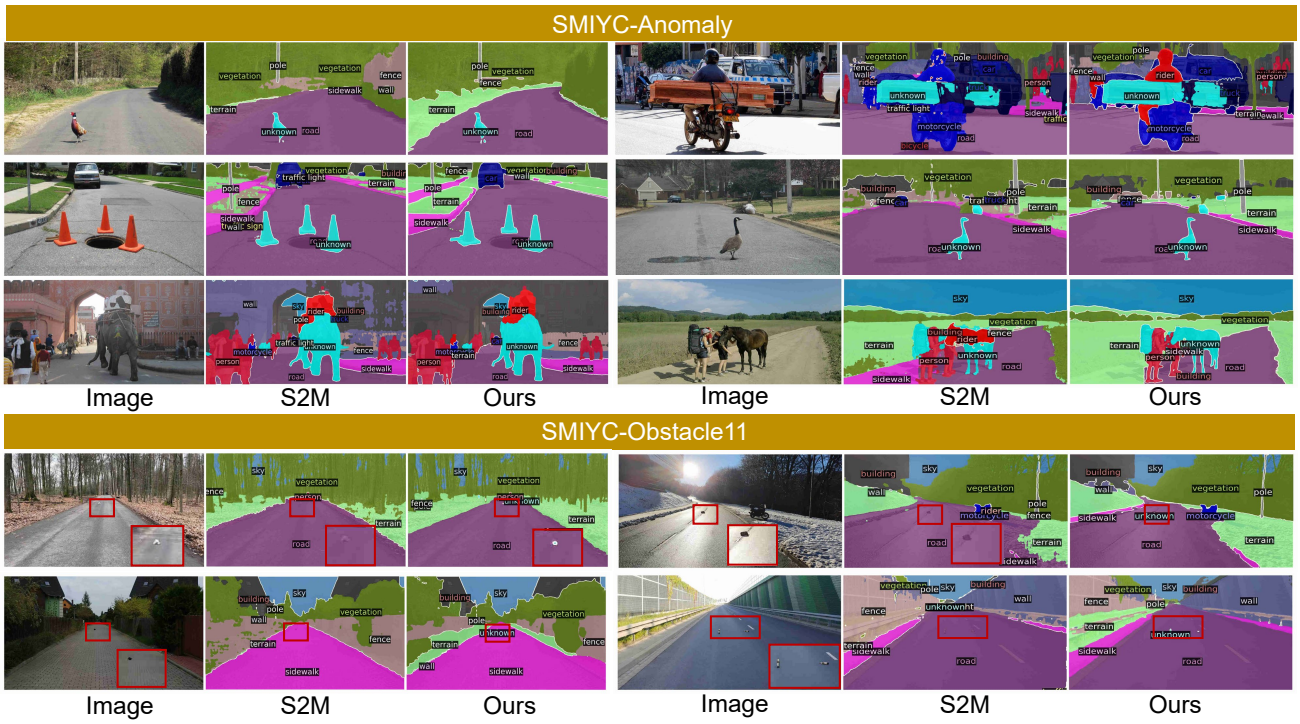

Figure 3. Qualitative comparison with S2M [30] on SMIYC-Anomaly and SMIYC-Obstacle.

segmentation naturally generalizes the goal of OOD segmentation by enabling fine-grained recognition of arbitrary unseen concepts through text prompts, rather than collapsing them into a single "unknown" category. From this perspective, OOD segmentation can be viewed as a special case of open-vocabulary segmentation, and it is thus of interest to evaluate how our generalized formulation performs under standard OOD benchmarks.

To evaluate OOD localization, we leverage the fact that our model can operate on arbitrarily large textual vocabularies without architectural modification. For fair comparison with existing OOD segmentation methods, we construct a test-time vocabulary consisting of the 19 in-domain Cityscapes classes and 150 additional COCO-Stuff concepts. The COCO-Stuff categories are filtered to remove those semantically overlapping with Cityscapes classes, yielding a clean ID–OOD separation. During inference, the model predicts per-pixel labels over this enlarged vocabulary, and any pixel whose predicted label does not belong to the 19 Cityscapes ID classes is treated as OOD. Importantly, no COCO-Stuff pixel annotations or additional OOD-specific supervision are used; only the textual names of the 150 concepts are provided at test time.

To evaluate OOD localization, we leverage the fact that our model is an open-vocabulary segmentation framework that can directly operate on arbitrarily large textual label sets. For fair comparison with existing OOD segmentation methods, we construct a vocabulary consisting of the 19 in-domain Cityscapes classes and 150 additional COCO-Stuff concepts. The COCO-Stuff categories are filtered to remove those that overlap or closely align with Cityscapes classes, ensuring a clean semantic separation. During inference, the

model predicts per-pixel labels over this enlarged vocabulary, and any pixel whose predicted label does not belong to the 19 Cityscapes ID classes is treated as OOD. No additional training or supervision from COCO-Stuff annotations is used; only the textual names of these 150 concepts are provided at test time.

Conventional OOD segmentation works commonly report threshold-based metrics such as FPR95, AUROC, or AUPR. However, these metrics depend heavily on calibrated OOD scores and threshold sweeps, making cross-method comparison unfair when models produce scores with different statistical ranges or when some methods (including ours) do not rely on explicit OOD scoring functions. To avoid such inconsistencies and to ensure a fair and model-agnostic evaluation, we follow the metric protocol of S2M [30], which directly evaluates the predicted OOD masks without requiring any score thresholding. Following [30], we report both IoU and mean F1 for OOD regions on five benchmarks: FS Static [17], FS Lost&Found, SMIYC-Anomaly [1], SMIYC-Obstacle, and RoadAnomaly [13]. In contrast to existing OOD segmentation methods that rely on COCO-Stuff pixel annotations or additional anomaly-supervision signals, our model performs OOD detection *solely* through text-driven vocabulary expansion, without using any OOD segmentation masks during training.

As shown in Table 5, across all datasets, our approach achieves the best overall performance, with notable improvements on SMIYC-Anomaly (+5.4 IoU over S2M) and SMIYC-Obstacle (+8.6 IoU). These results demonstrate that simple vocabulary expansion, enabled by open-vocabulary segmentation, provides a strong and scalable mechanism for OOD detection, outperforming specialized OOD detectors even without additional OOD supervision.

In Fig. 3, we further present qualitative comparisons with S2M on the SMIYC-Anomaly and SMIYC-Obstacle benchmarks. Our method produces cleaner OOD masks with sharper boundaries and better detection of small or thin obstacles. In cluttered or visually challenging scenes, S2M often misclassifies in-domain objects as unknown, whereas our open-vocabulary model preserves more stable predictions. These visual results align with the quantitative gains and demonstrate the improved robustness of our approach.

D.4. More Visualization Comparisons

As shown in Fig. 4, Fig. 5, and Fig. 6, we provide additional visualization comparisons on the CS7 $\rightarrow$ ACDC-41, Mapi-30, and RW-10 benchmarks. These results further verify that our method produces cleaner boundaries, reduces domain-induced noise, and achieves more consistent predictions across diverse unseen domains.

E. Limitations and Outlook

While we introduce the OVDG task setting for the first time and propose a novel enhancement method to mitigate domain-shift challenges in open-vocabulary semantic segmentation, several limitations remain. First, although our correlation refinement strategy effectively suppresses domain-induced noise, more efficient or principled mechanisms may further improve robustness, especially under severe distribution shifts. Second, our approach primarily refines correlations at inference time; strengthening the underlying VLM features to be intrinsically resistant to domain factors is an important direction for future work. Third, the current OVDG benchmark provides only a moderate set of open-vocabulary classes and test scenarios. We plan to substantially expand the benchmark with richer and more diverse categories, particularly those critical for driving safety, to better evaluate open-vocabulary generalization in real-world environments.

References

[1] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. *arXiv preprint arXiv:2104.14812*, 2021. 6

[2] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 1, 4, 10, 11, 12

[3] 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

[4] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16918–16927, June 2021. 4, 5

[5] Hao Fang et al. Eva-clip: Improving vision-language models with masked modeling. *arXiv preprint arXiv:2303.13495*, 2023. 2

[6] Anurag Ghosh, Shen Zheng, Robert Tamburo, Khiem Vuong, Juan Alvarez-Padilla, Hailiang Zhu, Michael Cardei, Nicholas Dunn, Christoph Mertz, and Srinivasa G Narasimhan. Roadwork: A dataset and benchmark for learning to recognize, observe, analyze and drive through work zones. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6132–6142, 2025. 1, 2

[7] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dginstyle: Domain-generalizable semantic segmentation with

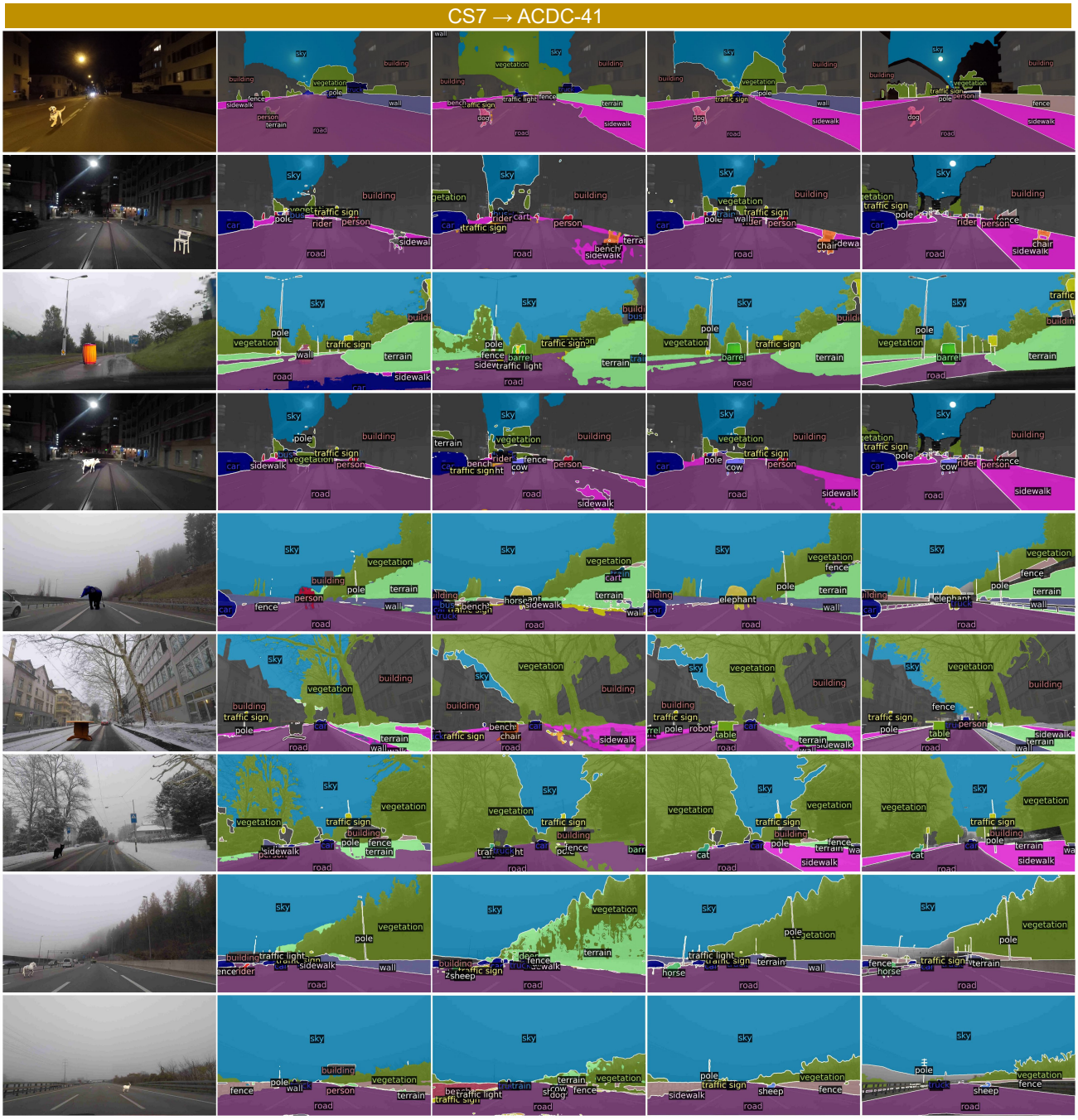

Figure 4. More qualitative comparisons on unseen domains under the OVDG-SS setting on ACDC-41 dataset.

image diffusion models and stylized semantic control. In *European Conference on Computer Vision*, 2024. 10

[8] Siyu Jiao, Hongguang Zhu, Jiannan Huang, Yao Zhao, Yunchao Wei, and Shi Humphrey. Collaborative vision-text representation optimizing for open-vocabulary segmentation. In *European Conference on Computer Vision*, 2024. 4, 10, 11, 12

[9] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pages 143–160. Springer, 2024. 10

[10] Minhyeok Lee, Suhwan Cho, Jungho Lee, Sunghun Yang, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Effective sam combination for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Com-*

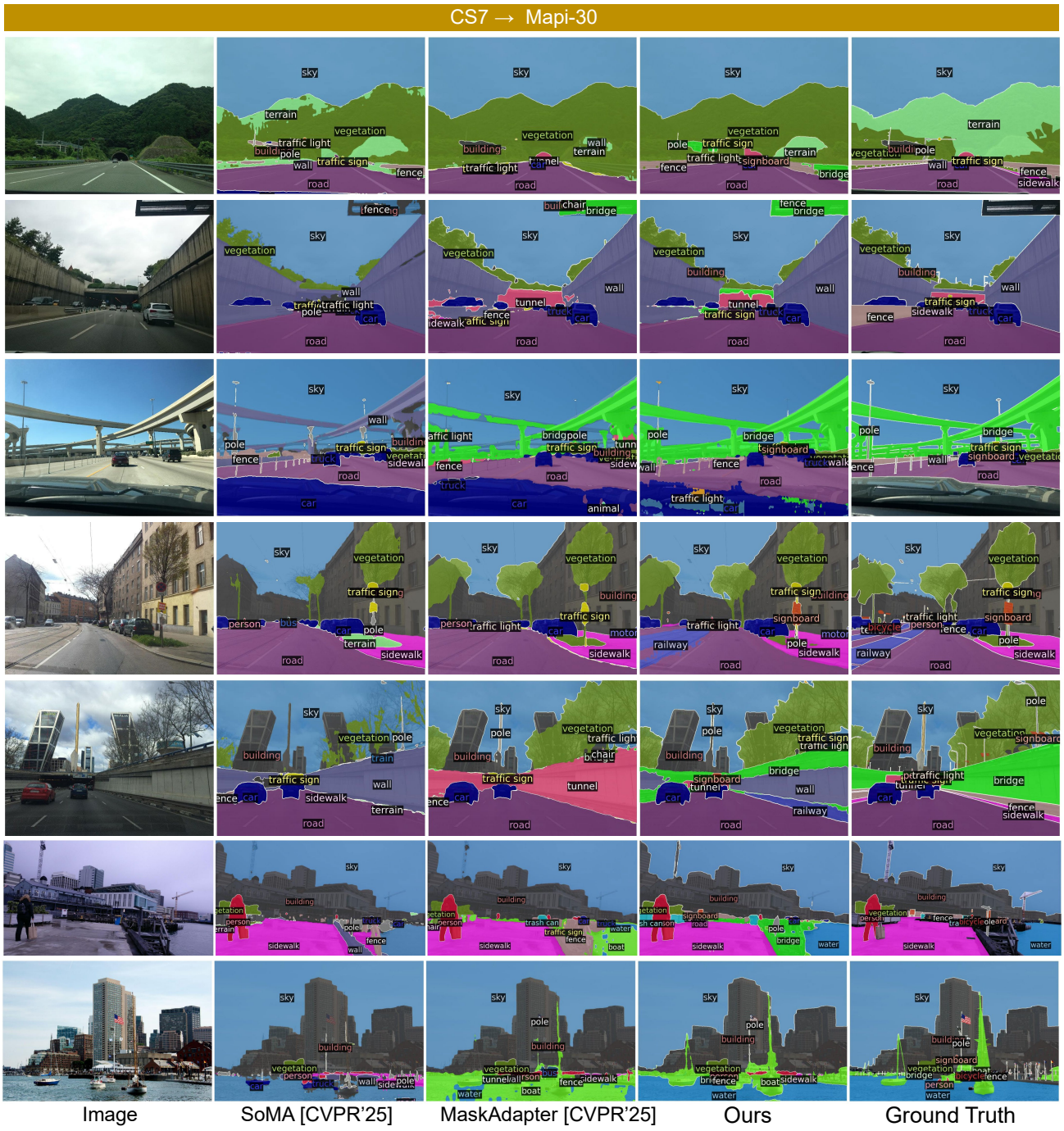

Figure 5. More qualitative comparisons on unseen domains under the OVDG-SS setting on Mapi-30 datasets.

puter Vision and Pattern Recognition (CVPR), pages 26081–26090, June 2025. 10, 11, 12

[11] Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, and Xinggang Wang. Mask-adapter: The devil is in the masks for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14998–15008, June 2025. 4, 10, 11, 12

[12] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 10, 11

[13] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis.

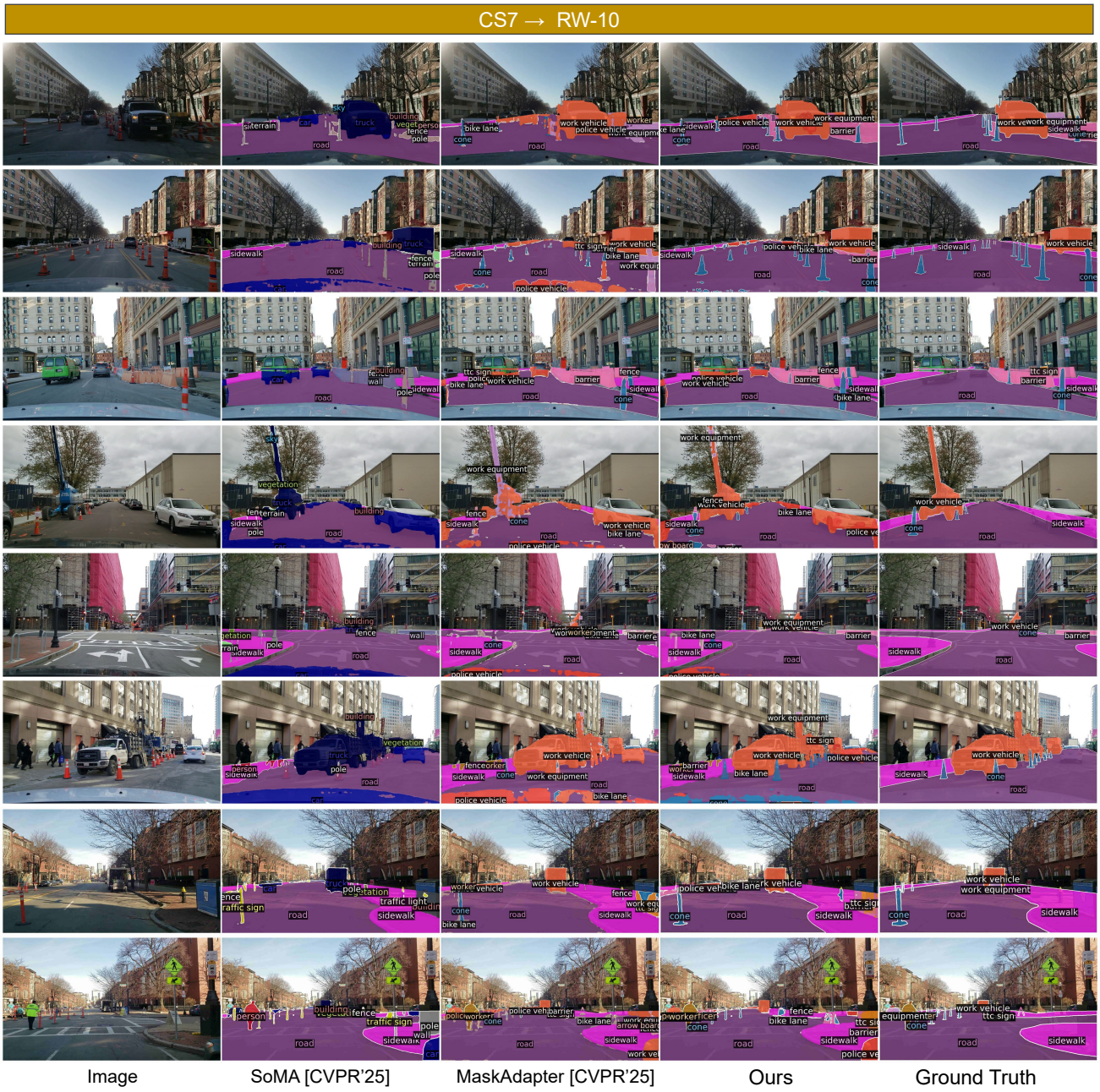

Figure 6. More qualitative comparisons on unseen domains under the OVDG-SS setting on RW-10 datasets.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2152–2161, 2019. 6

[14] Yuyuan Liu, Choubo Ding, Yu Tian, Guansong Pang, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1151–1161, October 2023. 5

[15] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 1

[16] 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, 2024. 4

[17] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 IEEE/RSJ International Conference on Intel-*

Method	Backbone	Training Data	Dv-19				Dv-58				
			ACDC-19	BDD-19	Mapi-19	Ave.	ACDC-41	BDD-41	Mapi-30	RW-10	Ave.
CLIP-DINOiser(ECCV2024) [26]	CLIP+DinoV2 ViT-L/14	Training-Free	25.3	28.5	27.7	27.2	31.9	34.0	26.9	21.6	28.6
ClearCLIP(ECCV2024) [9]	CLIP ViT-L/14	Training-Free	26.7	28.3	27.0	27.3	33.5	34.0	26.6	21.6	28.9
ProxyCLIP (ECCV'24)	CLIP+SAM ViT-L/14	Training-Free	31.0	38.2	40.0	36.4	38.5	52.2	33.5	27.2	37.9
OVSeg (CVPR'23) [12]	EVA02 ViT-L/14	CS-7	41.9	42.6	43.9	42.8	51.2	49.3	33.7	35.7	42.5
SAN (CVPR'23) [27]		CS-7	46.6	47.0	48.3	47.3	57.5	56.9	38.0	34.8	46.8
CAT-Seg (CVPR'24) [2]		CS-7	48.5	48.5	50.8	49.3	59.2	60.0	41.1	39.5	50.0
MAFT+(ECCV'24) [8]		CS-7	47.4	46.4	54.9	49.6	58.4	57.9	42.2	39.0	49.4
ESC-Net (CVPR'25) [10]		CS-7	46.3	47.9	53.0	49.1	57.1	59.2	43.8	36.5	49.2
MaskAdapter(CVPR'25) [11]		CS-7	48.1	49.4	54.6	50.7	59.7	59.7	41.8	36.1	49.3
CLIPSelf (ICLR'24) [25]		CS-7+CoCo	51.1	53.0	55.7	53.3	61.3	60.3	44.7	39.8	51.5
RSC-CLIPSelf (ICLR'25) [18]		CS-7+CoCo	50.7	50.0	55.7	52.1	61.3	59.6	44.6	36.4	50.5
CAT-Seg+AdvStyle [31]		CS-7	45.1	48.0	49.0	47.4	56.0	56.4	40.8	38.6	48.0
CAT-Seg+DGInStyle [7]		CS-7	50.4	51.7	50.4	50.8	60.2	60.8	43.5	38.7	50.8
S^2-Corr (Ours)		CS-7	**54.3**	**53.1**	**60.0**	**55.8**	**62.0**	**61.7**	**47.4**	**41.9**	**53.2**
OVSeg (CVPR'23) [12]	EVA02 ViT-L/14	GTA-7	37.0	43.7	44.4	41.9	50.5	54.9	38.1	35.6	44.8
SAN (CVPR'23) [27]		GTA-7	39.1	46.6	47.7	44.1	50.8	55.9	39.1	32.9	44.7
CAT-Seg (CVPR'24) [2]		GTA-7	42.4	47.9	50.8	47.5	53.6	59.7	41.1	38.2	48.2
MAFT+(ECCV'24) [8]		GTA-7	39.4	45.6	49.3	45.2	50.7	56.5	37.6	36.7	46.4
ESC-Net (CVPR'25) [10]		GTA-7	41.1	45.8	51.0	46.3	51.5	57.3	38.7	35.1	45.7
MaskAdapter(CVPR'25) [11]		GTA-7	40.4	47.2	50.4	46.5	52.1	55.9	37.1	37.1	45.6
CLIPSelf (ICLR'24) [25]		GTA-7+CoCo	43.1	50.0	50.1	47.7	51.2	59.7	40.6	39.7	48.0
RSC-CLIPSelf (ICLR'25) [18]		GTA-7+CoCo	42.6	49.7	51.9	48.0	52.0	59.1	39.8	38.4	47.8
CAT-Seg+AdvStyle [31]		GTA-7	41.0	46.3	50.6	46.0	51.8	59.2	40.7	35.5	46.8
CAT-Seg+DGInStyle [7]		GTA-7	43.9	48.9	51.0	47.9	53.7	59.0	37.9	38.0	47.2
S^2-Corr (Ours)		GTA-7	**44.2**	**50.3**	**53.3**	**49.9**	**55.4**	**59.8**	**41.5**	**40.8**	**49.4**

Table 6. Comparison of different OV-SS methods using ViT-L/14 backbone under OVDG-SS setting.

ligent Robots and Systems (IROS), pages 1099–1106. IEEE, 2016. 6

[18] Congpei Qiu, Yanhao Wu, Wei Ke, Xiuxiu Bai, and Tong Zhang. Refining clip's spatial awareness: A visual-centric perspective, 2025. 10, 11, 12

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*, 2021. 2

[20] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 102–118, 2016. 2

[21] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021. 1, 2

[22] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *Computer Vision – ECCV 2022*, pages 246–263, Cham, 2022. Springer Nature Switzerland. 4, 5

[23] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pages 315–332. Springer, 2024. 2

[24] 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. 4

[25] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023. 2, 10, 11, 12

[26] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation, 2024. 10

[27] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15546–15561, 2023. 10, 11, 12

[28] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1, 2

[29] Seokju Yun, Seunghye Chae, Dongheon Lee, and Youngmin Ro. Soma: Singular value decomposed minor components adaptation for domain generalizable representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25602–25612, 2025. 4

[30] Wenjie Zhao, Jia Li, Xin Dong, Yu Xiang, and Yunhui Guo. Segment every out-of-distribution object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3910–3920, June 2024. 5, 6

[31] Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial style augmentation for domain general-

Methods	Rd	Sw	Bldg	Veg	Sky	Per	Car	Wal	Fen	Pol	TL	TS	Ter	Rid	Trk	Bus	Trn	Mot	Bic	mIoU
CS-7→ACDC-19 ViT-B/16																				
OVSeg (CVPR'23) [12]	83.3	37.5	85.5	69.1	80.5	43.8	68.2	29.3	21.0	8.9	8.6	7.8	13.0	2.3	42.6	35.7	9.4	16.5	32.9	36.6
SAN (CVPR'23) [27]	86.4	34.5	82.6	73.6	96.5	49.8	71.2	31.0	18.2	7.5	7.0	5.7	14.3	0.1	46.1	32.7	9.0	17.2	30.9	37.6
CAT-Seg (CVPR'24) [2]	82.2	39.8	84.8	73.5	92.0	49.3	78.9	34.9	20.4	8.8	8.5	7.2	15.4	2.3	44.6	34.6	11.1	18.4	32.0	38.9
CLIPSelf (ICLR'24) [25]	84.9	38.9	73.1	72.0	89.1	45.4	74.7	27.9	31.6	24.9	24.5	32.8	18.3	14.5	40.3	35.5	12.1	31.8	36.5	42.6
RSC-CLIPSelf (ICLR'25) [18]	74.4	39.1	83.0	74.4	73.9	44.9	69.7	26.7	27.5	23.4	27.5	28.6	16.5	12.5	38.2	30.2	11.6	29.9	36.2	40.4
MAFT+(ECCV'24) [8]	79.8	40.4	81.2	72.3	75.1	49.6	72.3	30.1	28.8	24.8	27.9	29.3	17.3	13.1	35.7	32.3	11.2	31.9	35.8	41.5
ESC-Net (CVPR'25) [10]	74.4	39.1	83.0	74.4	73.9	44.9	69.7	26.7	27.5	23.4	27.5	28.6	16.5	12.5	38.2	30.2	11.6	29.9	36.2	40.4
MaskAdapter(CVPR'25) [11]	73.8	40.1	73.9	67.2	68.2	45.4	67.7	27.5	25.5	24.1	25.2	28.7	14.5	11.1	40.8	28.2	12.2	30.3	35.0	38.9
S^2-Corr	84.2	41.3	85.4	78.1	90.7	54.6	80.2	43.4	25.6	24.0	18.6	32.4	36.9	7.0	48.8	30.7	2.7	23.3	39.2	44.6
CS-7→BDD-19 ViT-B/16																				
OVSeg (CVPR'23) [12]	76.1	53.9	78.3	71.9	86.5	45.8	87.6	24.3	40.9	7.2	2.5	2.3	20.2	1.3	31.2	79.8	0.0	35.6	32.2	40.9
SAN (CVPR'23) [27]	82.6	54.6	87.9	64.0	95.9	60.5	65.3	17.4	33.6	5.9	0.1	0.0	18.0	0.0	31.1	73.5	0.0	36.9	22.1	39.4
CAT-Seg (CVPR'24) [2]	88.9	60.2	83.8	81.7	92.9	61.0	86.7	26.2	38.5	7.4	2.3	2.5	20.2	1.8	40.5	73.7	0.0	36.8	30.2	44.0
CLIPSelf (ICLR'24) [25]	79.9	63.6	83.9	80.2	98.4	66.1	96.4	24.4	40.7	6.6	2.0	2.4	21.3	1.9	45.5	85.4	0.0	35.9	31.6	45.6
RSC-CLIPSelf (ICLR'25) [18]	80.2	69.5	77.7	86.1	83.9	55.8	81.3	29.6	40.3	7.1	2.3	2.8	21.1	2.0	43.2	71.6	0.0	37.2	34.3	43.5
MAFT+(ECCV'24) [8]	91.5	66.8	87.7	89.0	96.2	61.5	89.4	27.8	39.4	7.4	2.1	2.6	20.1	2.0	46.0	79.4	0.0	37.2	33.8	46.3
ESC-Net (CVPR'25) [10]	85.7	63.0	88.5	80.4	89.8	57.8	87.1	27.2	39.8	6.9	1.8	2.6	19.6	1.8	41.5	68.8	0.0	33.3	33.3	43.6
MaskAdapter(CVPR'25) [11]	97.3	69.2	88.8	88.6	102.3	68.0	99.1	26.7	40.1	7.5	2.2	2.7	19.0	2.2	43.7	76.0	0.0	38.8	34.9	47.7
S^2-Corr	93.0	62.4	85.0	84.1	93.4	60.7	87.1	31.9	41.8	28.5	10.2	23.5	32.2	4.6	43.1	75.0	0.0	53.8	41.9	50.1
CS-7→Mapi-19 ViT-B/16																				
OVSeg (CVPR'23) [12]	78.3	56.4	88.1	90.3	81.3	56.8	85.3	38.8	33.7	13.8	5.4	15.0	16.2	2.0	48.1	50.7	24.1	51.8	34.3	45.8
SAN (CVPR'23) [27]	83.1	45.3	75.8	89.6	96.6	66.6	76.9	29.5	33.1	12.1	5.1	12.1	16.6	0.5	52.0	51.9	12.0	42.8	33.7	44.0
CAT-Seg (CVPR'24) [2]	87.3	51.5	84.7	83.8	96.5	65.0	85.2	38.0	34.4	15.2	6.3	14.0	17.0	2.3	49.4	59.2	23.8	52.1	39.5	47.6
CLIPSelf (ICLR'24) [25]	77.0	59.7	93.2	109.1	105.1	70.1	81.4	49.7	40.5	13.0	7.2	15.2	20.3	2.5	54.7	51.5	27.7	52.6	43.0	51.2
RSC-CLIPSelf (ICLR'25) [18]	82.3	59.9	89.5	99.6	114.3	76.5	73.3	45.4	36.7	13.0	7.4	14.8	19.1	2.2	59.8	47.7	23.9	51.2	36.4	50.2
MAFT+(ECCV'24) [8]	81.5	61.0	85.6	101.0	108.0	74.4	75.9	45.9	36.5	13.8	7.6	15.4	20.6	2.4	55.6	51.6	25.7	51.4	39.0	50.2
ESC-Net (CVPR'25) [10]	72.3	56.9	76.2	92.6	92.1	72.7	69.5	46.3	34.7	13.1	7.0	14.4	20.7	2.3	53.2	46.0	24.6	44.0	33.2	45.9
MaskAdapter(CVPR'25) [11]	83.2	63.7	86.2	95.6	103.7	68.8	76.7	42.3	36.0	14.9	6.8	13.7	20.5	2.4	59.4	53.5	24.7	54.7	40.6	49.9
S^2-Corr	89.8	53.4	86.9	87.3	97.2	68.8	87.9	50.6	44.5	36.6	18.2	45.2	46.7	7.4	65.5	64.9	21.0	51.9	44.9	56.2
CS-7→ACDC-19 ViT-L/14																				
SAN (CVPR'23) [27]	84.2	60.0	80.6	83.5	88.1	60.1	89.0	36.7	27.5	10.0	8.9	20.0	13.3	4.7	58.2	37.6	42.3	42.1	39.4	46.6
CAT-Seg (CVPR'24) [2]	89.8	64.8	88.0	83.4	95.4	61.5	85.1	40.2	27.9	9.9	8.5	20.0	13.0	4.9	63.0	39.1	41.3	46.6	38.9	48.5
CLIPSelf (ICLR'24) [25]	94.2	70.4	91.8	88.8	101.5	66.0	88.8	43.0	27.7	10.7	8.8	21.9	13.1	5.0	64.9	42.7	42.5	46.6	41.7	51.1
RSC-CLIPSelf (ICLR'25) [18]	88.6	66.8	96.9	90.4	103.3	63.3	88.2	41.4	25.9	11.0	8.8	23.2	13.3	4.7	60.8	45.0	42.6	46.0	42.9	50.7
MAFT+(ECCV'24) [8]	83.5	66.5	83.3	80.6	90.8	59.8	91.4	35.5	26.3	10.3	8.4	21.8	13.8	4.5	56.7	37.7	42.3	44.6	42.5	47.4
ESC-Net (CVPR'25) [10]	86.1	61.0	80.4	79.6	86.7	61.7	86.4	36.8	27.8	9.6	8.9	20.4	13.0	4.6	56.1	38.4	40.0	43.2	38.2	46.3
MaskAdapter(CVPR'25) [11]	81.0	65.1	87.9	90.3	96.5	60.0	87.5	35.0	25.9	10.3	9.0	19.9	13.0	4.8	60.0	40.0	43.5	42.0	43.1	48.1
S^2-Corr	91.9	67.2	89.2	84.0	95.4	62.7	87.0	51.9	39.5	37.0	39.6	47.4	20.0	6.4	68.5	35.1	20.2	41.9	46.3	54.3
CS-7→BDD-19 ViT-L/14																				
SAN (CVPR'23) [27]	86.5	67.1	82.4	82.1	93.7	66.7	91.2	28.0	31.2	6.9	5.8	8.2	10.9	0.6	55.9	73.8	0.9	60.0	40.4	47.0
CAT-Seg (CVPR'24) [2]	92.0	66.7	86.1	84.3	94.6	68.7	89.7	27.7	34.0	7.0	5.5	8.6	11.0	0.6	58.9	81.5	0.9	62.9	41.4	48.5
CLIPSelf (ICLR'24) [25]	92.6	69.7	88.4	87.6	105.7	74.4	92.4	28.2	38.3	7.5	5.4	8.9	12.3	0.6	55.7	81.1	1.1	60.0	38.7	49.9
RSC-CLIPSelf (ICLR'25) [18]	88.4	66.8	78.2	77.9	93.4	66.0	92.4	28.2	29.8	7.0	5.8	8.2	11.1	0.6	55.9	72.7	0.9	57.9	40.1	46.4
MAFT+(ECCV'24) [8]	86.2	69.8	87.1	85.0	91.8	71.0	90.2	29.8	33.9	17.1	15.5	28.7	11.4	0.6	59.3	83.4	1.0	52.6	36.4	50.0
ESC-Net (CVPR'25) [10]	90.8	70.2	83.0	83.4	90.6	72.8	89.2	30.1	33.9	7.1	5.9	8.2	11.0	0.6	59.7	69.9	1.0	63.3	40.3	47.9
MaskAdapter(CVPR'25) [11]	91.7	64.9	78.7	90.5	94.8	61.0	91.9	26.4	28.7	26.3	15.6	28.4	11.0	0.6	57.0	71.5	0.9	60.9	37.5	49.4
S^2-Corr	94.9	67.2	86.9	84.7	94.6	69.1	90.4	30.8	40.0	32.5	29.3	38.0	16.6	0.4	51.6	77.0	0.2	58.8	46.6	53.1
CS-7→Mapi-19 ViT-L/14																				
OVSeg (CVPR'23) [12]	81.4	54.5	71.7	74.9	83.1	61.3	84.8	35.7	29.2	9.7	8.8	21.0	7.8	5.1	54.0	65.8	23.8	44.1	24.7	44.3
SAN (CVPR'23) [27]	90.0	62.1	78.8	79.4	89.3	66.8	82.7	42.5	32.7	10.8	10.2	22.9	8.5	5.9	59.3	68.8	26.9	53.6	26.6	48.3
CAT-Seg (CVPR'24) [2]	91.3	62.9	86.6	86.0	97.0	71.6	86.5	44.8	33.5	11.4	10.7	24.7	9.1	6.0	61.1	73.7	27.2	54.7	27.3	50.8
CLIPSelf (ICLR'24) [25]	90.3	68.4	82.8	88.2	95.1	78.5	86.7	47.0	46.7	13.7	11.8	58.2	9.7	25.8	61.5	75.4	29.0	47.7	41.9	55.7
RSC-CLIPSelf (ICLR'25) [18]	90.4	72.5	85.9	89.1	96.9	77.0	85.9	44.6	44.2	12.9	11.0	54.5	9.3	26.7	55.9	72.0	28.3	47.0	39.4	54.9
MAFT+(ECCV'24) [8]	91.0	70.5	86.3	89.1	95.3	79.6	88.9	46.6	45.5	12.6	11.3	55.4	9.4	26.1	59.0	70.2	28.2	51.6	41.0	55.7
ESC-Net (CVPR'25) [10]	84.0	65.9	86.6	88.6	91.0	74.5	84.3	45.4	42.0	11.7	10.6	55.0	8.8	24.8	55.7	65.2	26.4	47.2	39.5	53.0
MaskAdapter(CVPR'25) [11]	88.9	68.3	86.5	89.7	96.9	77.7	89.9	44.4	44.0	11.9	10.9	54.6	8.9	24.9	55.8	69.4	26.6	46.7	41.6	54.6
S^2-Corr	92.9	63.8	88.8	87.5	97.6	74.0	90.0	56.5	46.5	39.1	40.5	59.0	20.4	19.3	71.6	71.8	17.7	53.9	50.1	60.0

Table 7. Per-class results for OVDG-SS. Green entries denote seen classes, while blue entries correspond to unseen classes.

ized urban-scene segmentation. *Advances in neural information processing systems*, 35:338–350, 2022. 10

	Rd	Sw	Bldg	Veg	Sky	Per	Car	Wal	Fen	Pol	TL	TS	Ter	Rid	Trk	Bus	Trn	Mot	Bic	Bag	Bal	Bar	Bnc	Brd	Btl	Crt	Cat	Chr	Cow	Der	Dog	Drn	Elp	Hat	Hrs	Rbt	Shp	Tbl	Toy	Umb	Zbr	mIoU
CS-7→ACDC-41 ViT-B/16																																										
SAN (CVPR'23) [27]	82.5	35.4	83.6	65.2	76.2	37.9	73.9	25.0	12.1	7.5	6.6	4.0	13.9	0.0	41.2	32.1	6.7	17.9	17.1	62.1	79.3	56.8	4.0	41.4	40.1	41.1	62.2	44.7	68.3	52.9	60.1	44.5	80.4	80.4	68.9	60.9	62.9	6.6	28.3	67.5	66.3	44.4
CAT-Seg (CVPR'24) [2]	83.5	42.4	81.4	65.3	85.4	39.6	71.9	29.4	14.3	8.9	7.9	5.6	15.0	1.4	46.5	33.6	10.1	20.7	17.6	69.0	76.2	60.4	7.1	49.4	44.1	51.3	75.3	53.3	66.9	51.7	68.3	49.3	76.6	83.3	67.1	62.2	71.2	8.1	30.3	74.6	76.2	47.6
CLIPSelf (ICLR'24) [25]	88.2	42.0	80.6	73.9	92.0	41.3	72.0	29.5	14.7	9.2	9.0	6.3	15.1	1.5	50.2	33.7	10.5	20.0	17.1	72.5	78.9	68.8	7.6	45.4	46.7	53.3	72.1	55.4	63.8	51.2	73.0	46.5	84.0	86.9	63.3	69.9	80.6	7.8	30.0	73.7	86.0	49.4
RSC-CLIPSelf (ICLR'25) [18]	88.2	44.0	88.5	74.5	92.6	40.1	69.3	28.8	16.1	8.7	8.8	6.1	15.1	1.6	50.1	37.0	10.5	21.7	18.2	75.1	74.8	76.3	8.1	48.1	45.1	57.4	72.8	58.2	63.3	54.7	74.9	49.7	79.1	82.1	60.3	73.1	80.6	7.7	32.2	70.9	84.2	50.0
MAFT+(ECCV'24) [8]	80.0	40.9	74.0	66.6	81.6	36.3	68.2	29.8	13.5	8.7	8.4	5.7	13.1	1.5	46.8	29.8	10.4	20.2	15.8	68.3	78.8	68.4	6.9	45.4	46.7	48.1	64.9	48.6	61.7	45.8	72.1	47.7	76.9	87.9	59.7	66.1	80.7	6.8	27.1	67.8	75.9	46.4
ESC-Net (CVPR'25) [10]	86.0	41.3	70.6	74.5	84.4	37.8	72.2	26.2	13.3	8.4	8.9	6.4	15.4	1.5	48.4	33.1	10.0	18.5	17.3	64.8	78.2	68.4	6.8	45.7	40.9	53.6	70.0	56.0	64.5	47.2	74.1	42.0	86.3	79.7	57.2	67.4	78.9	7.8	27.9	67.2	78.9	47.3
MaskAdapter(CVPR'25) [11]	76.3	48.6	82.6	74.8	83.6	45.3	71.7	37.0	24.5	8.7	1.9	1.5	18.2	0.3	39.7	63.7	0.0	41.2	24.2	72.8	50.8	65.8	21.3	66.7	40.7	58.4	49.8	42.8	74.7	66.2	58.2	45.4	74.3	83.9	64.7	60.4	82.8	13.2	20.9	58.8	64.7	48.3
S^2-Corr	84.9	42.8	82.1	69.5	81.7	38.2	72.7	34.9	17.6	17.7	16.8	25.8	32.8	1.2	50.8	28.8	4.0	23.0	25.1	57.8	82.7	67.2	81.8	3.1	70.9	80.5	54.7	79.0	66.1	60.0	66.4	41.7	70.2	77.6	72.7	70.0	62.2	27.0	65.0	62.9	75.4	52.3
CS-7→BDD-41 ViT-B/16																																										
SAN (CVPR'23) [27]	86.7	49.9	80.3	71.4	86.2	42.8	70.7	32.9	24.7	6.7	0.0	0.0	15.8	0.0	32.6	72.9	0.0	36.4	23.0	69.6	47.0	65.6	19.6	58.4	35.3	57.2	53.2	37.0	72.1	63.4	56.9	45.0	72.6	78.0	68.6	58.2	77.3	12.3	21.7	55.0	62.8	46.8
CAT-Seg (CVPR'24) [2]	87.5	50.8	81.3	77.4	92.2	45.5	78.4	36.0	26.3	9.4	2.1	1.5	20.2	0.3	38.9	72.9	0.0	44.1	26.3	78.2	55.1	75.2	21.5	69.2	42.3	66.7	56.4	44.2	73.1	64.5	64.9	50.3	76.4	86.6	72.0	67.8	90.0	14.9	23.9	60.5	70.2	51.6
CLIPSelf (ICLR'24) [25]	92.3	49.8	77.5	77.1	94.1	55.2	81.0	34.2	31.5	9.0	2.4	1.3	20.7	0.3	39.6	72.8	0.0	43.5	30.1	91.3	51.8	82.8	25.6	76.3	39.4	68.4	62.0	47.2	76.3	77.4	69.4	52.3	76.9	98.2	71.4	76.6	94.9	16.4	22.6	63.3	75.4	54.3
RSC-CLIPSelf (ICLR'25) [18]	90.9	53.0	80.7	74.1	88.9	45.1	85.6	35.3	28.6	9.8	2.2	1.5	19.2	0.3	42.4	76.8	0.0	48.1	29.0	77.3	60.0	74.2	22.1	73.1	44.3	69.8	61.4	45.1	73.0	71.1	62.2	51.4	72.3	86.2	76.8	70.5	93.0	14.3	22.9	63.6	74.9	53.0
MAFT+(ECCV'24) [8]	81.6	47.6	80.7	79.0	81.7	44.5	73.7	35.8	22.9	8.4	2.1	1.5	19.7	0.3	35.7	67.8	0.0	39.5	27.1	80.3	55.2	65.4	20.6	63.8	37.7	58.7	49.4	45.4	71.2	60.9	60.8	44.9	75.7	86.4	63.2	60.3	90.7	15.3	23.9	55.4	65.8	48.8
ESC-Net (CVPR'25) [10]	88.0	47.8	83.7	70.5	94.5	46.8	73.2	33.8	24.5	8.4	1.9	1.5	18.1	0.3	37.1	70.9	0.0	45.1	25.3	70.2	49.3	66.0	19.9	62.1	38.2	58.0	53.7	40.1	74.8	64.8	58.5	47.1	67.4	88.8	67.8	69.7	81.0	14.6	24.3	57.1	72.0	49.2
MaskAdapter(CVPR'25) [11]	83.7	50.9	79.5	70.6	94.2	49.3	76.1	32.6	28.2	9.2	2.2	1.3	21.5	0.3	34.9	66.4	0.0	42.0	26.5	80.0	52.7	75.2	23.4	70.6	41.3	64.9	52.9	45.7	77.4	67.4	61.8	46.7	77.8	85.9	70.0	69.6	91.9	14.9	22.9	60.1	76.8	51.2
S^2-Corr	90.6	54.3	82.5	80.6	92.8	40.5	78.9	41.8	29.6	28.5	10.7	17.2	32.4	1.8	46.2	77.5	0.0	65.1	30.0	60.6	86.2	81.6	80.2	23.2	77.5	72.6	69.4	62.4	78.3	67.1	74.7	44.0	77.0	80.7	79.4	75.6	88.4	25.1	62.9	58.6	75.6	58.6
CS-7→ACDC-41 ViT-L/14																																										
SAN (CVPR'23) [27]	76.7	50.5	79.5	75.7	79.1	50.4	71.8	35.5	31.3	38.7	29.1	47.5	15.6	6.4	59.9	36.6	36.6	43.3	30.7	65.7	77.6	75.4	65.4	19.3	68.2	78.8	56.7	74.1	67.2	46.2	72.5	64.4	75.0	79.9	70.6	71.7	58.4	62.6	64.1	75.6	74.9	57.5
CAT-Seg (CVPR'24) [2]	89.3	62.0	85.6	79.6	93.0	51.1	78.1	39.8	24.7	8.9	6.8	18.3	12.2	2.4	71.5	37.0	35.7	42.4	34.9	63.1	88.7	77.6	81.9	15.2	64.4	86.4	75.8	80.8	68.8	60.8	72.8	70.3	84.1	84.1	65.8	79.8	54.9	56.9	60.0	86.4	76.3	59.2
CLIPSelf (ICLR'24) [25]	82.1	56.3	84.0	75.3	89.9	51.5	76.0	39.4	31.0	40.0	32.9	47.8	16.3	6.4	66.6	39.1	40.9	44.0	34.7	68.4	81.1	85.0	73.0	21.6	69.7	85.1	61.1	75.9	75.4	47.8	73.4	70.2	81.1	80.8	77.4	78.8	61.1	62.7	68.6	80.8	79.6	61.3
RSC-CLIPSelf (ICLR'25) [18]	87.0	57.3	85.0	76.9	90.3	53.5	74.2	39.6	33.3	36.9	33.8	43.8	16.1	6.1	65.3	39.8	40.7	47.3	34.9	69.8	83.0	86.8	68.9	21.2	70.2	77.6	64.8	81.8	71.3	46.3	69.0	66.9	85.2	78.7	80.6	79.1	64.0	66.1	67.4	76.9	75.5	61.3
MAFT+(ECCV'24) [8]	84.4	59.6	84.8	76.1	91.8	49.6	77.5	38.8	24.3	8.5	6.7	18.7	12.2	2.5	67.8	37.4	34.9	41.4	33.0	60.1	88.2	79.8	77.8	14.8	63.7	84.5	77.3	80.2	68.3	61.1	73.0	72.2	86.3	83.3	64.1	80.8	53.4	54.5	57.7	86.2	77.8	58.4
ESC-Net (CVPR'25) [10]	88.3	60.7	81.9	74.1	86.8	49.8	76.8	36.9	22.9	8.3	6.6	17.0	11.8	2.3	68.8	36.3	35.4	39.8	32.5	61.8	85.4	76.2	75.9	14.2	60.3	84.3	73.4	80.3	63.9	57.8	70.8	68.0	80.1	84.9	64.8	78.4	52.6	53.5	57.1	85.5	76.0	57.1
MaskAdapter(CVPR'25) [11]	85.1	63.7	81.1	76.9	91.4	55.5	83.8	42.9	27.2	9.8	7.5	19.3	11.7	2.5	76.3	39.3	38.3	41.8	33.5	60.8	98.3	85.0	89.7	15.3	64.7	82.8	70.9	85.9	67.9	65.4	69.0	67.4	87.9	84.3	63.4	79.3	55.3	55.1	62.3	79.8	67.6	59.7
S^2-Corr	90.2	63.5	86.1	79.3	91.8	50.4	79.4	49.1	33.6	33.0	33.5	40.0	18.4	4.4	72.4	32.4	22.2	37.3	41.6	69.7	89.0	85.7	90.8	15.8	76.0	89.7	73.4	81.7	64.8	59.1	74.2	68.2	85.7	86.1	67.7	79.6	46.3	52.6	66.4	86.6	75.8	62.0
CS-7→BDD-41 ViT-L/14																																										
SAN (CVPR'23) [27]	85.9	44.8	70.6	67.3	82.6	44.9	70.2	42.5	31.9	34.2	25.9	37.8	14.0	9.2	45.6	76.2	0.7	51.1	31.8	57.0	80.5	67.7	72.6	47.6	64.8	71.6	72.5	61.8	73.9	50.4	59.2	55.0	70.0	74.3	66.3	74.5	69.5	63.9	69.4	73.6	68.9	56.9
CAT-Seg (CVPR'24) [2]	90.7	58.6	83.3	81.0	93.8	51.4	81.3	40.0	24.9	7.4	5.0	7.1	11.8	0.3	57.1	82.1	1.5	56.3	30.0	57.4	89.4	78.3	82.4	43.0	76.4	79.8	80.1	66.4	79.0	65.4	73.3	63.8	78.2	91.4	78.8	76.5	84.0	29.3	67.1	91.1	66.0	60.0
CLIPSelf (ICLR'24) [25]	89.8	49.2	77.5	71.2	86.2	50.4	72.2	43.0	34.3	37.5	29.4	39.2	14.3	9.8	49.8	82.5	3.8	55.2	33.7	61.1	81.8	71.9	73.2	47.9	72.9	73.1	73.8	65.8	76.9	54.2	66.8	56.7	70.7	81.8	69.2	80.9	78.2	66.6	71.8	77.4	70.1	60.3
RSC-CLIPSelf (ICLR'25) [18]	86.6	57.4	82.3	77.9	88.5	48.3	80.7	39.4	24.0	7.1	4.8	6.6	10.9	0.3	52.8	76.0	1.5	55.9	29.3	55.6	86.2	72.4	76.6	39.6	70.5	78.0	77.9	65.1	78.8	63.3	69.8	63.7	76.4	91.8	74.9	75.8	82.0	28.4	62.0	90.7	63.8	57.9
MAFT+(ECCV'24) [8]	89.5	57.3	81.8	77.5	95.7	49.6	80.8	39.7	24.0	7.3	4.7	6.9	11.3	0.3	55.5	84.5	1.5	53.9	30.7	56.5	85.8	79.2	78.8	43.6	72.5	76.8	81.5	64.4	80.0	62.0	75.4	64.9	78.7	86.7	75.9	74.5	85.1	29.8	64.6	93.5	66.4	59.2
ESC-Net (CVPR'25) [10]	86.3	57.6	79.5	80.9	94.9	50.9	83.0	40.4	23.6	7.5	5.1	7.1	11.8	0.3	58.3	78.1	1.4	53.3	28.3	57.0	91.4	76.3	80.2	41.9	78.2	80.1	82.2	67.5	80.5	61.7	74.4	63.1	79.3	86.9	80.2	78.5	84.6	29.5	69.0	92.3	64.2	59.7
MaskAdapter(CVPR'25) [11]	87.8	52.7	88.9	89.8	92.8	50.7	85.0	42.0	24.7	7.1	4.9	7.0	11.9	0.3	55.5	82.2	1.7	54.7	31.9	61.4	87.2	79.2	79.2	45.4	73.4	85.0	70.3	64.5	75.1	69.2	72.7	67.7	75.2	81.6	76.1	76.1	79.5	32.1	64.5	88.0	70.4	59.6
S^2-Corr	92.1	59.3	84.0	81.6	93.9	49.3	82.1	43.2	30.8	32.7	25.3	29.9	14.8	0.4	47.3	78.9	0.3	46.6	33.4	56.4	89.3	83.3	82.1	43.8	76.4	81.9	77.6	67.1	77.2	64.2	75.7	68.8	77.1	90.5	78.8	80.8	76.3	28.9	70.6	90.4	67.3	61.7

Table 8. Per-class results for OVDG-SS in CS-7→ACDC-41 and BDD-41.

	Rd	Sw	Bldg	Veg	Sky	Per	Car	Wal	Fen	Pol	TL	TS	Ter	Rid	Trk	Bus	Trn	Mot	Bic	Chr	Brg	Tun	Wtr	Snw	Snd	Anm	Sbd	Rwy	Bot	Tc	mIoU
CS-7→Map-30 ViT-B/16																															
SAN (CVPR'23) [27]	69.3	44.7	67.6	87.1	101.5	61.2	84.6	21.2	18.2	10.8	4.7	6.0	13.2	0.4	36.4	52.5	15.5	51.0	43.6	29.4	27.1	20.2	31.0	42.0	18.0	38.6	1.0	16.0	10.5	39.8	35.4
CAT-Seg (CVPR'24) [2]	78.0	43.0	81.6	82.5	95.8	62.7	83.9	30.7	20.4	13.3	6.9	8.1	17.8	2.5	45.0	54.9	18.8	48.5	42.3	31.2	32.0	44.2	32.3	43.9	19.3	37.1	3.1	15.7	11.1	40.4	38.2
CLIPSelf (ICLR'24) [25]	81.1	45.8	82.4	81.4	108.2	70.0	88.6	30.3	23.1	14.0	6.7	7.9	17.9	2.7	44.8	52.5	20.7	51.4	46.3	29.7	32.2	32.9	32.8	46.6	18.8	41.3	3.1	16.5	11.7	45.1	39.6
RSC-CLIPSelf (ICLR'25) [18]	80.1	45.6	88.4	86.1	98.0	65.4	82.8	29.9	25.6	13.7	6.2	7.7	16.8	3.1	49.0	48.5	18.9	55.1	49.5	29.5	34.5	44.2	37.0	51.8	17.6	43.3	2.9	15.4	13.3	42.8	40.1
MAFT+(ECCV'24) [8]	82.6	43.1	79.3	76.9	110.2	71.9	87.9	30.1	22.6	12.5	5.9	7.8	16.1	2.6	45.4	49.7	18.1	46.7	44.4	27.7	29.6	33.3	32.3	41.9	16.8	38.5	3.0	15.8	12.0	39.5	38.1
ESC-Net (CVPR'25) [10]	83.9	47.3	73.1	80.4	109.9	69.9	91.2	28.7	24.2	13.8	6.5	8.1	17.9	2.6	49.7	60.2	20.9	46.8	49.5	30.2	31.5	35.6	36.9	47.6	19.3	40.5	3.4	16.3	12.1	43.9	40.1
MaskAdapter(CVPR'25) [11]	79.5	44.0	77.5	80.0	106.2	72.5	96.9	28.5	22.1	14.0	6.8	7.4	18.6	2.7	47.0	57.6	21.0	48.9	48.2	31.1	31.1	34.4	33.8	48.5	18.3	41.3	3.3	16.3	11.3	43.9	39.8
S^2-Corr	80.5	42.4	84.3	86.0	96.3	66.1	86.6	36.3	24.9	22.6	21.8	6.2	36.1	6.1	60.8	57.2	4.3	47.9	44.5	27.5	38.1	19.7	63.1	26.9	46.3	16.1	15.2	10.3	36.0	49.5	42.0
CS-7→Map-30 ViT-L/14																															
SAN (CVPR'23) [27]	85.8	49.3	70.5	81.7	94.9	73.1	79.4	20.7	19.1	36.9	28.7	44.3	8.5	10.4	28.6	34.3	29.6	31.3	25.4	19.9	29.7	75.1	38.9	6.6	10.6	16.2	23.2	23.2	25.3	18.9	38.0
CAT-Seg (CVPR'24) [2]	81.7	52.2	83.7	84.6	96.1	68.8	85.3	34.9	24.7	9.3	9.8	19.2	7.9	2.9	57.8	64.8	28.5	54.3	28.4	36.4	39.9	59.4	45.6	15.6	27.9	4.8	10.2	15.7	39.1	44.6	41.1
CLIPSelf (ICLR'24) [25]	85.0	47.2	80.4	76.8	87.4	68.3	87.4	34.4	35.3	32.2	29.8	41.3	11.8	17.0	51.0	58.0	38.7	54.8	33.3	39.4	44.3	72.6	54.4	11.7	11.6	13.7	28.5	17.8	39.6	37.8	44.7
RSC-CLIPSelf (ICLR'25) [18]	81.9	50.6	77.7	81.6	86.2	68.0	85.9	37.1	37.0	32.5	29.3	41.2	12.6	18.2	53.0	56.1	36.1	52.6	34.3	38.1	46.7	76.1	54.9	11.1	11.5	13.3	17.2	28.9	31.4	37.9	44.6
MAFT+(ECCV'24) [8]	87.9	55.7	81.4	87.0	104.6	76.2	87.3	36.7	25.7	10.2	10.4	20.0	8.8	3.0	54.6	69.6	27.2	51.0	31.2	35.2	39.2	59.7	44.1	14.9	28.1	4.6	9.7	16.5	42.1	43.1	42.2
ESC-Net (CVPR'25) [10]	94.8	55.6	84.9	87.5	98.2	78.3	102.7	34.8	26.8	11.1	9.7	22.5	9.3	3.4	59.6	71.1	32.8	52.2	32.9	39.7	35.9	58.4	46.0	15.6	27.9	4.7	10.5	17.6	46.8	43.8	43.8
MaskAdapter(CVPR'25) [11]	81.5	56.4	76.7	76.8	100.1	80.3	96.4	34.2	23.9	10.4	9.7	20.4	8.5	3.0	54.5	65.9	29.6	51.2	34.3	34.8	36.2	58.1	46.0	15.1	29.3	4.1	10.3	17.1	44.5	43.5	41.8
S^2-Corr	82.8	53.0	86.8	86.2	96.9	72.1	89.1	39.4	34.5	24.0	9.9	12.0	14.1	8.4	68.8	64.2	23.9	56.2	44.2	45.4	51.3	66.1	50.1	19.5	42.4	25.7	28.0	21.0	46.6	58.2	47.4

Table 9. Per-class results for OVDG-SS in CS-7→Mapi-30.

	Rd	Sw	Bar	PV	WV	PO	Wkr	Con	ABd	TTC	mIoU
CS-7→RW-10 ViT-B/16											
SAN (CVPR'23) [27]	80.3	37.9	58.7	3.8	41.6	28.7	32.4	6.8	3.7	12.7	30.7
CAT-Seg (CVPR'24) [2]	89.9	62.8	57.3	20.9	48.1	22.9	24.7	10.0	4.5	24.2	36.5
CLIPSelf (ICLR'24) [25]	88.1	62.9	58.8	20.0	49.5	23.5	24.8	9.5	4.5	23.8	36.5
RSC-CLIPSelf (ICLR'25) [18]	85.6	61.8	54.9	20.9	46.7	23.5	23.4	9.8	4.3	23.4	35.4
MAFT+(ECCV'24) [8]	89.4	55.9	51.4	20.9	44.8	20.8	23.8	9.7	4.3	23.8	34.5
ESC-Net (CVPR'25) [10]	91.7	59.9	54.9	21.0	48.2	21.8	24.7	9.8	4.3	24.8	36.1
MaskAdapter(CVPR'25) [11]	87.4	58.0	54.0	20.6	47.4	22.7	22.7	9.8	4.2	23.7	35.1
S^2-Corr	86.8	63.1	67.8	10.2	45.6	19.8	39.1	24.3	2.4	27.4	38.6
CS-7→RW-10 ViT-L/14											
SAN (CVPR'23) [27]	81.8	60.7	54.4	18.7	49.1	22.2	23.8	9.4	4.5	23.7	34.8
CAT-Seg (CVPR'24) [2]	87.4	68.4	71.7	4.1	48.1	22.6	32.3	33.8	5.3	21.8	39.5
CLIPSelf (ICLR'24) [25]	96.9	76.6	57.8	19.7	47.6	28.5	27.3	8.8	5.3	29.3	39.8
RSC-CLIPSelf (ICLR'25) [18]	92.2	76.2	58.5	19.7	46.7	28.0	27.3	8.5	5.1	28.2	39.0
MAFT+(ECCV'24) [8]	88.8	63.6	55.9	19.5	50.9	23.3	25.3	9.7	4.6	22.9	36.5
ESC-Net (CVPR'25) [10]	85.8	67.2	54.5	18.9	43.7	25.8	25.7	8.0	4.8	26.1	36.1
MaskAdapter(CVPR'25) [11]	85.4	64.1	59.9	18.3	48.4	24.2	25.5	8.9	4.8	24.8	36.4
S^2-Corr	83.2	68.7	72.9	2.4	47.7	21.5	39.0	40.8	16.3	26.4	41.9

Table 10. Per-class results for OVDG-SS in CS-7→RW10.