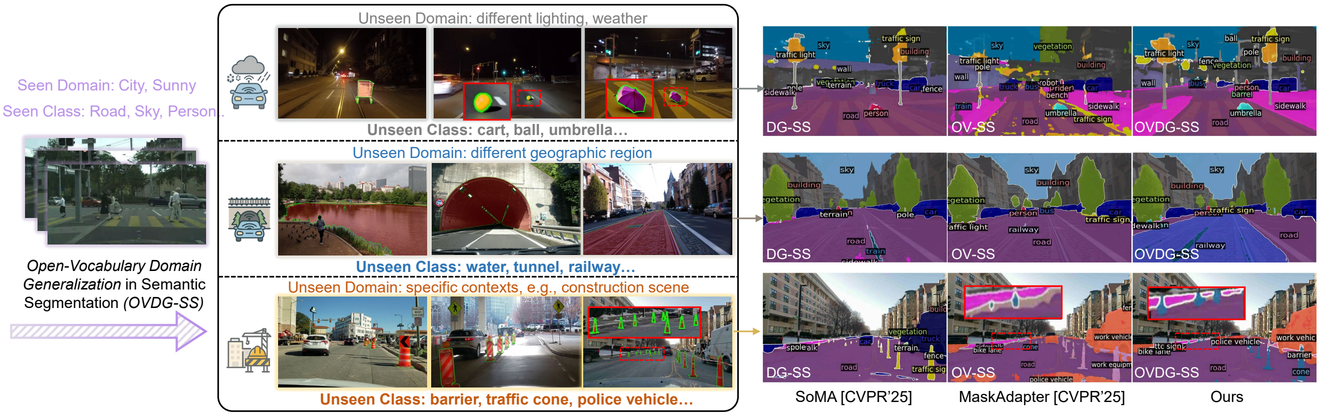
Open-Vocabulary Domain Generalization in Urban-Scene Segmentation

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) (b)

Figure 1. Concept of Open-Vocabulary Domain Generalization in Semantic Segmentation (OVDG-SS). (a) OVDG-SS aims to generalize across both unseen domains (e.g., lighting, weather, geographic region, construction context) and unseen classes (e.g., barrier, cone, railway), moving beyond conventional domain generalization (DG-SS) and open-vocabulary segmentation (OV-SS). (b) Comparison of segmentation results. Traditional DG-SS fails to recognize unseen categories, and OV-SS struggles under domain shifts, while our OVDG-SS effectively handles both unseen domains and classes simultaneously.

Abstract

Domain Generalization in Semantic Segmentation (DG-SS) aims to enable segmentation models to perform robustly in unseen environments. However, conventional DG-SS methods are restricted to a fixed set of known categories, limiting their applicability in open-world scenarios. Recent progress in Vision-Language Models (VLMs) has advanced Open-Vocabulary Semantic Segmentation (OV-SS) by enabling models to recognize a broader range of concepts. Yet, these models remain sensitive to domain shifts and struggle to maintain robustness when deployed in unseen environments, a challenge that is particularly severe in urban-driving scenarios. To bridge this gap, we introduce Open-Vocabulary Domain Generalization in Semantic Segmentation (OVDG-SS), a new setting that jointly addresses unseen domains and unseen categories. We introduce the first benchmark for OVDG-SS in autonomous driving, addressing a previously unexplored problem and covering both synthetic-to-real and real-to-real generalization across diverse unseen domains and unseen categories. In OVDG-SS, we observe that domain shifts often distort text–image correlations in pre-trained VLMs, which hinders the performance of OV-SS models. To tackle this challenge, we propose S^2-Corr, a state-space-driven text–image correlation refinement mechanism that can mitigate domain-induced distortions and produce a more consistent text–image correlation under distribution changes. Extensive experiments on our constructed benchmark demonstrate that the proposed method achieves superior cross-domain performance and efficiency compared to existing OV-SS approaches. The code is available at `https://github.com/DZhaoXd/s2_corr`.

1. Introduction

Domain Generalization in Semantic Segmentation (DG-SS) has long been a key challenge in enabling segmentation models to perform robustly in unseen environments. The recent emergence of large Vision Foundation Models (VFMs) [3, 8] has notably improved the cross-domain robustness of segmentation models [43, 59, 61]. Despite this progress, these models remain limited to seen semantics, as they can only recognize categories present in their training data [63, 66, 67]. This limitation becomes critical for safety

---

✉ Co-corresponding author.

OV-SS Method	Training Data	Seen class	Num.	Shared classes		Performance	
				Dv-19	Dv-58	Dv-19	Dv-58
CAT-Seg [5]	COCO-Stuff [24]	171	118K	17	34	31.6	32.4
CAT-Seg	SYNTHIA [37]	7	9.4K	7	7	43.6	45.3
CAT-Seg	GTA [35]	7	25K	7	7	47.5	48.2
CAT-Seg	Cityscapes [6]	7	3K	7	7	**49.3**	**50.0**
MaskAdapter [21]	COCO-Stuff	171	118K	17	34	30.1	29.6
MaskAdapter	SYNTHIA	7	9.4K	7	7	42.4	43.1
MaskAdapter	GTA	7	25K	7	7	46.5	45.6
MaskAdapter	Cityscapes	7	3K	7	7	**50.7**	**49.3**

Table 1. Effect of training domains on training-based OV-SS performance. Colored rows indicate models trained on driving datasets. Dv-19 and Dv-58 refer to cross-domain driving datasets with 19 and 58 categories.

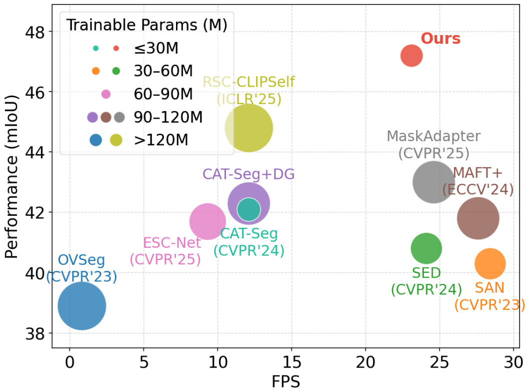


Figure 2. Efficiency and performance comparison on OVDG-SS tasks using EVA02 ViT-B/16 as backbone. FPS is tested on images with a short edge of 480 and a long edge of 960. Our method achieves the best trade-off among generalization ability, speed, and parameter efficiency.

in autonomous driving scenarios, where models trained on urban sunny scenes often fail to detect unseen objects such as barriers or traffic cones appearing at night, in tunnels, or under adverse weather conditions, as illustrated in Fig. 1(a). Such failures restrict the scalability of DG-SS toward truly open-world perception.

Open-Vocabulary Semantic Segmentation (OV-SS) [19, 51] leverages Vision-Language Models (VLMs) [9, 33] to recognize diverse visual concepts beyond closed-set categories. Most existing OV-SS models [22, 47, 58], trained on COCO-Stuff [24], achieve strong results on generic scenes but suffer sharp performance drops when transferred to domain-specific scenarios such as remote sensing [20] or autonomous driving [10], even when many classes overlap with the target domain (Table 1). To mitigate this degradation, we further train these OV-SS models on several driving-related datasets under a limited-vocabulary setting. As shown in Table 1, models trained on domain-related data (colored rows) exhibit progressively improved performance. When the training domain shifts from synthetic (SYNTHIA, GTA) to real (Cityscapes) and becomes closer to the target distribution, performance consistently improves. *This confirms that OV-SS models are highly sensitive to domain shift and struggle to generalize when the training and target domains are mismatched.*

To this end, we introduce Open-Vocabulary Domain Generalization in Semantic Segmentation (OVDG-SS), a new setting that jointly addresses unseen domains and unseen categories. As illustrated in Fig. 1, OVDG-SS aims to adapt models to new environments while recognizing novel concepts beyond the training distribution. This setting extends OV-SS to real-world applications such as autonomous driving, where robustness to domain shifts and openness to new semantics are both critical for safety.

To facilitate research in this new setting, we construct the first comprehensive OVDG-SS benchmark tailored for urban-driving segmentation. It includes both synthetic-to-real and real-to-real generalization settings and spans three unseen domain types: diverse weather and lighting, geographically distinct regions, and construction-heavy environments. Compared to standard DG-SS benchmarks, our benchmark introduces over 30 additional driving-related categories, greatly enriching the semantic space for OVDG evaluation. Building on this benchmark, we evaluate several representative training-based OV-SS models. Although strong under conventional OV-SS settings [23], these models exhibit degradation when transferred to unseen domains, exposing their limited robustness.

We then analyze why existing OV-SS approaches fail under domain shifts and *find that domain shifts make the initial text–image correlations from VLMs noisy and misaligned*, which in turn severely limits generalization in OVDG-SS. To address this issue, we propose S^2-Corr, a novel **S**tate-**S**pace-driven module that dynamically refines noisy text–image **Corr**elations for robust OVDG-SS. Built upon a selective state-space [7] aggregation baseline, S^2-Corr introduces three key innovations: (i) image- and text-conditioned modulation that injects domain-relevant cues, (ii) a distance-aware decay mechanism that suppresses long-range noise during sequential aggregation, and (iii) a chunk-wise snake scanning strategy that aligns state propagation with the spatial structure. Together, these designs reconstruct cleaner and more coherent text-image correlations under distribution shifts, enabling robust open-vocabulary generalization across unseen domains.

As shown in Fig. 2, our method attains higher mIoU and faster inference speed with fewer trainable parameters compared to existing OV-SS approaches, demonstrating both effectiveness and efficiency. Our main contributions are summarized as follows:

- We analyze and empirically verify the limitations of existing DG-SS and OV-SS paradigms, revealing that neither can effectively generalize to both unseen domains and categories simultaneously.
- We construct the first comprehensive benchmark

for OVDG-SS in driving scenarios, covering both synthetic-to-real and real-to-real OVDG settings with diverse unseen domains and unseen categories.

- We propose S^2-Corr, a novel and efficient state-space-driven correlation refinement module that stabilizes noisy text–image correlations under domain shifts.
- S^2-Corr achieves the strongest overall performance across unseen domains and categories, providing a solid new baseline for future OVDG-SS research.

2. Related Work

Domain Generalized Semantic Segmentation (DGSS) methods mainly fall into two categories. Data augmentation approaches increase domain diversity through image- [55, 60, 64, 68]or feature-level style perturbations [4, 56], or by incorporating additional diverse data [57, 62, 65]. The rise of VFMs has driven the development of Parameter-Efficient Fine-Tuning (PEFT) strategies. Some methods add lightweight adapters [44, 52], while others update only a small set of important parameters [54, 59], enabling efficient adaptation. TQDM [31] also leverages VLMs, but its design remains confined to closed-set DG-SS. Overall, these DG approaches are tailored for VFMs or closed-set settings and do not extend naturally to OVDG-SS.

Open-Vocabulary Semantic Segmentation (OV-SS) mainly evolved into two main paradigms. Training-free methods adapt CLIP to segmentation without pixel-level supervision, typically enhancing its dense predictions through refined spatial–text alignment [29, 34], self-correction of noisy attention maps [2, 16, 41, 69], self-distillation [18, 45], or VFM-based distillation that injects stronger dense priors [14, 42]. Training-based methods enhance OV-SS performance by fine-tuning on generic segmentation datasets such as COCO-Stuff. These approaches generate category-agnostic masks and assign text-based labels [1, 15, 39, 40], improve mask proposal quality [25, 49, 50], or refine cross-modal attention maps more effectively [5, 42]. However, we observe that both paradigms generalize poorly on driving-related OVDG-SS tasks, which motivates us to develop a more robust framework tailored for OVDG.

3. Methodology

Problem Definition. We study the Open-Vocabulary Domain Generalized Semantic Segmentation (OVDG-SS) task. Given a source domain $\mathcal{D}_s = \{(x_i, y_i)\}$ with pixel labels from a base vocabulary of N_d classes, the model is deployed on multiple unseen target domains $\{\mathcal{D}_t^{(k)}\}_{k=1}^{K}$ that contain a much larger open vocabulary with $M \gg N_d$ classes. The objective is to learn a segmentation model that generalizes to unseen domains and recognizes both source classes and novel vocabulary categories.

Overview. We begin by revisiting existing OV-SS approaches and analyzing why they fail under domain shifts (Sec. 3.1). Next, we establish a strong baseline by applying a state-space model to aggregate text–image correlations (Sec. 3.2). Finally, we introduce S^2-Corr module and explain how it improves upon this baseline (Sec. 3.3). The framework of our method is shown in Fig. 3.

3.1. Revisiting and Analyzing OV-SS

In this work, we use CAT-Seg [5] as our baseline, owing to its simple correlation-based design and strong robustness in OV-SS. Specifically, given an image–text pair, CAT-Seg extracts visual features $\mathbf{F}_v \in \mathbb{R}^{HW \times d}$ and textual class embeddings $\mathbf{F}_t \in \mathbb{R}^{N_C \times d}$ from a vision–language backbone (e.g., CLIP), where H and W are spatial dimensions, d is the feature dimension, and N_C is the number of classes. The initial correlation map is

$$\mathbf{C} = \mathrm{Norm}(\mathbf{F}_v \mathbf{F}_t^\top) \in \mathbb{R}^{HW \times N_C}. \tag{1}$$

To reduce noise and misaligned activations [41], CAT-Seg performs a two-stage refinement.

Spatial Aggregation. To lift the correlation map into a d_f-dimensional embedding space, CAT-Seg applies a learnable projection $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_{N_C}] \in \mathbb{R}^{d_f \times N_C}$. The lifted embeddings are computed by a class-wise broadcasted multiplication: $\mathbf{E}_{i,j,:} = C_{ij}\, \mathbf{p}_j, \mathbf{E} \in \mathbb{R}^{HW \times N_C \times d_f}$, which can be compactly written as $\mathbf{E} = \mathbf{C} \odot \mathbf{P}^\top$. Spatial refinement is then applied independently for each class using a shared cross-attention:

$$\mathbf{E}^{\text{spa}} = \mathrm{CrossAttn}_\theta(\mathbf{E},\ \mathrm{Neigh}(\mathbf{E}))\,, \tag{2}$$

where θ are class-shared parameters and $\mathrm{Neigh}(\mathbf{E})$ gathers local spatial neighborhoods (e.g., shifted windowed regions as in [27]). This produces the spatially aggregated embeddings $\mathbf{E}^{\text{spa}} \in \mathbb{R}^{HW \times N_C \times d_f}$.

Class-wise Aggregation. To model inter-class relationships at each spatial location, CAT-Seg applies another cross-attention across the class dimension:

$$\mathbf{E}^{\text{cls}} = \mathrm{CrossAttn}_\phi^{\text{cls}}(\mathbf{E}^{\text{spa}}, \mathrm{Classes}(\mathbf{E}^{\text{spa}}))\,, \tag{3}$$

where ϕ are shared parameters and $\mathrm{Classes}(\mathbf{E}^{\text{spa}})$ gathers class embeddings at the same spatial position. The resulting $\mathbf{E}^{\text{cls}} \in \mathbb{R}^{HW \times N_C \times d_f}$ serves as the final refined correlation map, which is decoded into pixel-wise predictions.

Why do OV-SS method perform poorly in OVDG-SS? Although CAT-Seg performs well in standard OV-SS settings, its performance drops sharply under OVDG-SS. We attribute this to two main factors. (1) *Heavy correlation map noise induced by domain shift.* As shown in Fig. 4, large domain shifts corrupt the initial text–image correlation map $\mathbf{C}$,

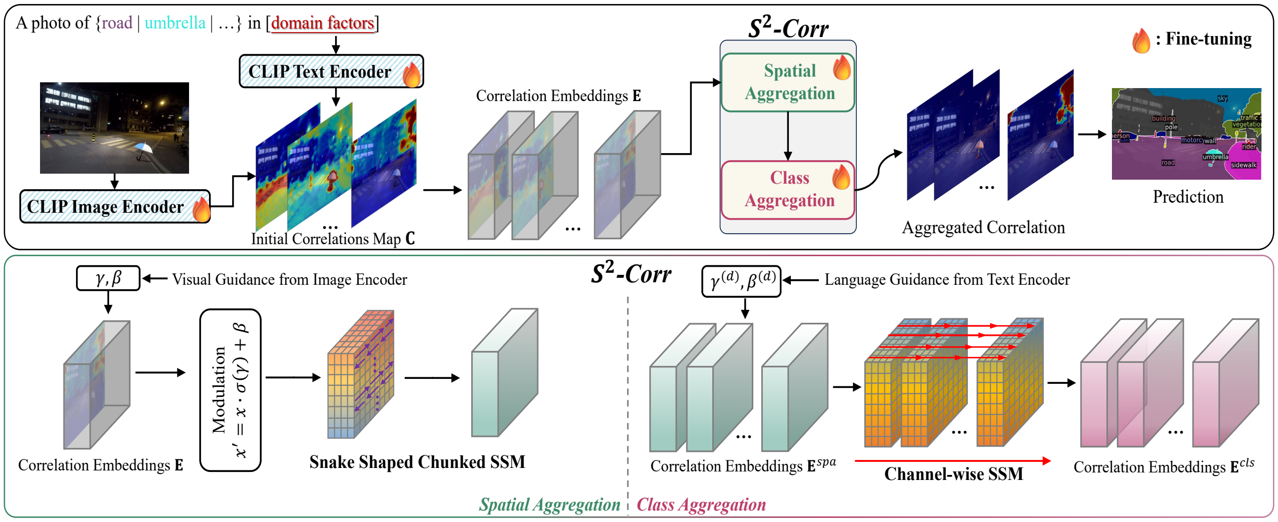


Figure 3. Overview of the proposed S^2-Corr. The upper part shows the CLIP-based encoding and correlation aggregation pipeline. The lower part illustrates our S^2-Corr, which refines text–image correlations using a specially designed chunked State-Space Models (SSM) aggregation scheme.

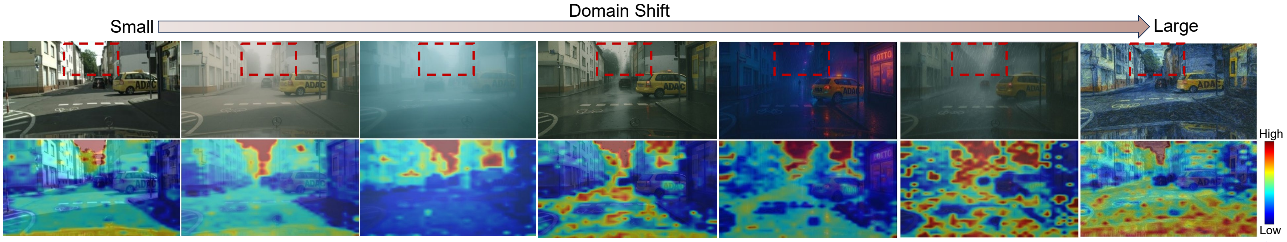


Figure 4. Effect of domain shift on text–image correlations of the class "*sky*" from EVA02 model [8]. Color map ranges from **blue (low correlation)** to **red (high correlation)**. As the domain shift increases from left to right, the initial correlation maps become progressively noisier, with incorrect activations spreading across irrelevant regions.

resulting in biased and spatially inconsistent activations that hinder downstream correlation refinement. (2) *Noise propagation via cross-attention.* As in Eq. 2, CAT-Seg refines class-wise correlation embeddings through attention-based aggregation. Under domain shift, the correlation map $\mathbf{C}$ contains many corrupted activations, which enter the cross-attention as noisy keys and values and distort the attention weights. These errors then propagate to neighboring positions within each class, and subsequent class aggregation further amplifies such domain-specific noise.

3.2. Refining Correlation with State-Space Models

To construct a more robust baseline for correlation aggregation, we replace the attention-based aggregation in CAT-Seg with a selective state-space model (SSM) [7, 11], which processes correlations in a sequential manner.

For spatial aggregation, given the correlation embeddings $\mathbf{E} \in \mathbb{R}^{HW \times N_C \times d_f}$, we flatten the spatial grid into a 1D sequence under a fixed scan order π (e.g., row-major):

$$\mathbf{x}_t = \mathbf{E}_{\pi(t),:} \in \mathbb{R}^{d_f}, \qquad t = 1, \dots, T,\ T = HW,$$

where $\mathbf{x}_t$ represents the correlation embedding at the t-th spatial position. Instead of explicit token-to-token interaction through cross-attention, we refine the sequence using selective recurrent state updates governed by a continuous-time state-space model,

$$\mathbf{h}_t = \mathbf{A}_t\, \mathbf{h}_{t-1} + \mathbf{B}_t\, \mathbf{x}_t, \qquad \mathbf{y}_t = \mathbf{W}_t\, \mathbf{h}_t + \mathbf{U}_t\, \mathbf{x}_t, \quad (4)$$

where $\mathbf{A}_t, \mathbf{B}_t, \mathbf{W}_t, \mathbf{U}_t$ are input-dependent parameters generated by lightweight linear projections:

$$\mathbf{A}_t = \sigma(\mathbf{W}_a \mathbf{x}_t + \mathbf{b}_a), \quad \mathbf{B}_t = \sigma(\mathbf{W}_b \mathbf{x}_t + \mathbf{b}_b), \quad (5)$$

and $\sigma(\cdot)$ denotes the sigmoid activation, $\mathbf{A}_t$ acts as a decay gate controlling how much past information is preserved, and $\mathbf{B}_t$ modulates the injection of new input information. Reshaping $\{\mathbf{y}_t\}_{t=1}^{T}$ back to the spatial layout yields the refined correlation embedding $\tilde{\mathbf{C}} \in \mathbb{R}^{HW \times N_c \times d_f}$.

For class aggregation, we adopt the same selective SSM formulation but process the category embedding $\tilde{\mathbf{C}}_{i,:} \in \mathbb{R}^{N_C \times d_f}$ at each spatial position sequentially. To maintain permutation consistency, classes are arranged in a fixed index order during training and inference.

Why is SSM-based Aggregation Better? We summarize its advantages in two aspects: (1) Unlike attention, which explicitly mixes every token with every other token, the selective SSM aggregates information through a controlled recurrent state update. The decay gate $\mathbf{A}_t$ dynamically regulates how much of the previous state $\mathbf{h}_{t-1}$ should be preserved or forgotten. When correlations at step $t-1$ contain noise, a small decay value $\mathbf{A}_t \approx 0$ effectively forgets the unreliable state, preventing noisy patterns from propagating forward. Meanwhile, the input gate $\mathbf{B}_t$ determines how strongly the current token $\mathbf{x}_t$ should influence the state, allowing clean and informative correlations to be injected when needed. (2) The class-wise SSM refines inter-class dependencies in a more efficient manner and avoids the quadratic complexity $\mathcal{O}(N_C^2)$ inherent in attention. This linear-time formulation scales favorably when the open-vocabulary set is large, offering both higher efficiency (Table 5) and better generalization to unseen categories.

3.3. S^2-Corr

Building upon the above enhanced baseline, we further propose a complete correlation refinement framework named S^2-Corr. It is designed to suppress domain-shift-induced noise from three perspectives: (1) modulation before aggregation, (2) a decay mechanism in the state-space model, and (3) a snake-shaped scanning strategy. Specifically, the first component injects informative cues into the correlation embeddings to guide the aggregation, while the latter two aim to suppress long-range noisy dependencies that arise during sequential propagation.

Modulation Before Aggregation. Before performing spatial aggregation, we inject image-specific cues into the correlation embeddings $\mathbf{E}_{\pi(t),:}$ through a lightweight modulation step. Given the corresponding image features $\mathbf{F}_{\pi(t)}$, the correlation representation is adjusted as

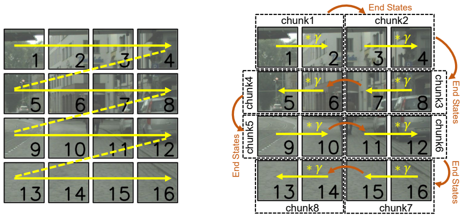
$$\hat{\mathbf{E}}_{\pi(t),:} = \mathbf{E}_{\pi(t),:} \odot (1 + \gamma_{\pi(t)}) + \beta_{\pi(t)}, \tag{6}$$

where $(\gamma_{\pi(t)}, \beta_{\pi(t)})$ are modulation factors derived from $\mathbf{F}_{\pi(t)}$ via a linear projection. This operation injects image-dependent context and improves spatial consistency before the sequence aggregation.

Similarly, before class-wise aggregation, the category embeddings $\tilde{\mathbf{C}}_{i,:} \in \mathbb{R}^{N_C \times d_f}$ are refined using multi-domain textual prompts. We construct several domain-related templates such as "a photo of a cat at night" or "a photo of a cat in the rain", which are encoded by a pre-trained text encoder to obtain domain-aware text features $\mathbf{t}^{(d)}$. Each text feature produces a pair of modulation vectors $(\gamma^{(d)}, \beta^{(d)})$ to adjust the category embeddings:

$$\hat{\mathbf{C}}_{i,:} = \tilde{\mathbf{C}}_{i,:} \odot (1 + \gamma^{(d)}) + \beta^{(d)}. \tag{7}$$

This text-driven modulation injects domain-specific semantics into the class representation, allowing the subsequent correlation fusion to better adapt to domain conditions.



(a) VMamba

(b) Ours

Figure 5. Comparison of scanning strategies in state-space correlation aggregation between VMamba [26] and ours. Our method introduces a learnable geometric decay ($*\gamma$) to suppress long-range noise within each chunk and uses a snake-shaped scanning strategy that preserves spatial continuity by passing end states between adjacent chunks.

Learnable Geometric Decay Prior. While the dynamic gate $\mathbf{A}_t$ in Eq. (5) enables data-dependent adaptation, it may still carry the risk of propagating long-range noise (See Fig. 4 last column) under domain shifts. To mitigate this, we introduce a learnable geometric decay prior,

$$\mathbf{A}_t^{\text{eff}} = \sigma(\mathbf{w})\, \sigma(\mathbf{W}_a \mathbf{x}_t + \mathbf{b}_a) + \big(1 - \sigma(\mathbf{w})\big)\, \boldsymbol{\gamma}, \tag{8}$$

where $\boldsymbol{\gamma} \in (0,1)^K$ is a head-specific spatial decay prior assigning each channel a base attenuation rate, and $\sigma(\mathbf{w})$ is a learnable coefficient balancing the data-driven gate and the geometric prior. The final update becomes

$$\mathbf{h}_t = \mathbf{A}_t^{\text{eff}} \odot \mathbf{h}_{t-1} + \mathbf{B}_t \odot \mathbf{x}_t, \tag{9}$$

where $\odot$ denotes element-wise multiplication. This design preserves a geometric attenuation pattern $\|\partial \mathbf{h}_t / \partial \mathbf{h}_{t-d}\| \propto (\mathbf{A}_t^{\text{eff}})^d$, while allowing its decay rate to be learned from data, thus acting as a robust learnable geometric decay mechanism.

Chunk-wise Snake Scanning. To make the sequential update consistent with the 2D spatial layout, we divide the flattened sequence $\{\mathbf{x}_t\}_{t=1}^T$ into non-overlapping chunks of equal length W, each corresponding to a local region in the image grid. Instead of scanning chunks strictly in a row-major order, which causes discontinuities at row boundaries, we adopt a snake-shaped traversal that alternates the scanning direction between consecutive rows. Within each chunk, the state is updated sequentially following Eq. (9), and the final hidden state of the current chunk is passed to the next one, $\mathbf{h}_{k+1}^{\text{init}} \leftarrow \mathbf{h}_k^{\text{end}}$. This snake scanning strategy maintains spatial adjacency between rows and ensures smooth feature propagation across chunk boundaries. The illustration of our scanning strategy is shown in Fig. 5.

4. Experiments

4.1. Dataset and Evaluation

Training Set. We use two datasets that have seven common driving categories (road, sidewalk, building, vegetation, sky, person, car) as source domains. *CS-7* is constructed

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) [48]	EVA02+DINOV2 ViT-L	Training-Free	25.3	28.5	27.7	27.2	31.9	34.0	26.9	21.6	28.6
ClearCLIP(ECCV2024) [16]	EVA02 ViT-L	Training-Free	26.7	28.3	27.0	27.3	33.5	34.0	26.6	21.6	28.9
ProxyCLIP (ECCV'24) [17]	EVA02+SAM ViT-L	Training-Free	31.0	38.2	40.0	36.4	38.5	52.2	33.5	27.2	37.9
CAT-Seg (CVPR'24) [5]	EVA02 ViT-L/14	COCO	30.3	31.9	36.1	32.8	33.2	42.8	28.6	27.6	33.1
DeCLIP (CVPR'25) [42]	EVA02+DINOV2 ViT-L	COCO	32.1	31.8	38.1	34.0	34.2	47.0	32.0	28.3	35.4
OVSeg (CVPR'23) [23]	EVA02 [8] ViT-B/16	CS-7	36.6	40.9	45.8	41.1	41.2	48.3	31.7	32.6	38.5
SAN (CVPR'23) [49]		CS-7	37.6	39.4	44.0	40.3	44.3	46.8	35.4	30.7	39.3
CAT-Seg (CVPR'24) [5]		CS-7	38.9	44.0	47.6	43.5	47.6	51.6	38.2	36.5	43.5
MAFT+(ECCV'24) [13]		CS-7	40.4	43.5	50.2	44.7	46.4	48.8	38.1	35.4	42.2
ESC-Net (CVPR'25) [18]		CS-7	40.4	43.6	45.9	43.3	47.3	49.2	40.1	34.5	42.7
MaskAdapter(CVPR'25) [21]		CS-7	38.9	47.7	49.9	45.5	48.3	51.2	39.8	36.1	43.8
CLIPSelf (ICLR'24) [45]		CS-7+COCO	43.6	45.6	47.9	45.7	49.4	54.3	39.6	36.5	45.0
RSC-CLIPSelf (ICLR'25) [32]		CS-7+COCO	41.5	46.3	50.2	46.0	50.0	53.0	40.1	35.1	44.5
CAT-Seg+AdvStyle [68]		CS-7	37.2	43.2	44.6	41.7	48.0	47.9	36.0	36.5	42.1
CAT-Seg+DGInStyle [12]		CS-7	42.5	49.8	49.4	44.6	46.1	53.7	36.8	36.0	43.2
S^2-Corr (Ours)		CS-7	**44.6**	**50.1**	**56.2**	**50.3**	**52.3**	**58.6**	**42.0**	**38.6**	**47.9**
CAT-Seg (CVPR'24) [5]	EVA02 ViT-L/14	CS-7	48.5	48.5	50.8	49.3	59.2	60.0	41.1	39.5	50.0
MaskAdapter(CVPR'25) [21]		CS-7	48.1	49.4	54.6	50.7	59.7	59.7	41.8	36.1	49.3
CLIPSelf (ICLR'24) [45]		CS-7+COCO	51.1	53.0	55.7	53.3	61.3	60.3	44.7	39.8	51.5
CAT-Seg+DGInStyle [12]		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**

Table 2. Comparison of different OV-SS methods across different backbones under the *Real-to-Real* OVDG-SS setting. Dv-19 groups ACDC-19, BDD-19, and Mapi-19, while Dv-58 groups ACDC-41, BDD-41, Mapi-30, and RW-10.

from Cityscapes [6] by selecting these seven classes from the original 19 categories, yielding 2,975 real urban images. *GTA-7* is the synthetic data from GTAV [35], containing 24,999 rendered images with the same label space.

Testing Set. We conduct evaluations on seven target domains. *ACDC-19* [38] contains 1,600 real images under extreme weather and follows the Cityscapes 19-class vocabulary. *BDD-19* [53] provides 1,000 images with diverse illumination and weather, also using the 19-class vocabulary. *Mapi-19* is from the Mapillary validation set [30], mapped to the same 19-class label space for evaluation. *Mapi-30* is built from the Mapillary training and validation sets [30] by merging 65 categories into 30 coarse road-type classes, including bridges, tunnels, railways, and waterways (3,943 images). *RW-10* comes from the ROADWork dataset [10], merged into 10 construction-related classes including workers, cones, work vehicles, and equipment (2,098 images).

To introduce more open-vocabulary objects, we further construct *ACDC-41* and *BDD-41* using Stable Diffusion 2.1 [36] inpainting on ACDC and BDD images. The edited objects mimic additional entities commonly appearing on roads (e.g., daily objects, animals, obstacles). We manually filter distorted results and refine the segmentation labels, producing 1,000 high-quality synthetic images for each dataset with 41 categories. (Details in Appendix). We refer {ACDC-19, BDD-19 and Mapi-19} as Dv-19 set which includes 19 classes while refer {ACDC-41, BDD-41, Mapi-30 and RW-10} as Dv-58 set which totally includes 58 classes.

Task Construction. We study OVDG-SS under two evaluation paradigms: *Synthetic-to-Real.* We train on synthetic dataset *GTA-7* with large gaps to real imagery, and evaluate on all real-world datasets (Dv-19 set and Dv-58 set). *Real-to-Real.* We train on *CS-7* and evaluate on the same target datasets to assess generalization from clean urban scenes to diverse real conditions.

4.2. Implementation Details

We implement our framework in Detectron2 [46]. The model is optimized using AdamW [28] with a learning rate of 2×10^{-4} for the aggregation module and 2×10^{-6} for the EVA-CLIP encoders. The correlation embedding dimension is set to 128, and the aggregation module uses 2 spatial blocks and 2 upsampling stages following CAT-Seg [5]. The number of chunks is set to 16 and γ is set as 0.8. For vision encoders, the input resolution is 512 for ViT-B/16 and 448 for ViT-L/14, resulting in a 32×32 token grid for both. In the visual encoder, only selected attention projection layers are updated, and in the text encoder, only projection weights within residual blocks are trainable. Our model updates only 26M and 76.8M parameters using ViT-B/16 and ViT-L/14, respectively. With a batch size of 4, we train the model by 20k iterations a single NVIDIA RTX 3090 GPU. **Notably, our method requires only about 2 hours training for EVA-CLIP ViT-B and 4 hours training for EVA-CLIP ViT-L** while achieving superior generalization on OVDG-SS.

4.3. Comparison with State of The Art

Summary. In Tables 2 and 3, we compare our method with state-of-the-art training-free methods, open-vocabulary methods, and "open-vocabulary + domain generalization [68]" methods. Clearly, our method consistently achieves the best performance across all target datasets

Method	Backbone	Training Data	Dv-19					Dv-58				
			CS-19	*ACDC-19*	*BDD-19*	*Mapi-19*	Ave.	*ACDC-41*	*BDD-41*	*Mapi-30*	*RW-10*	Ave.
OVSeg (CVPR'23) [23]	EVA02 ViT-B/16	GTA-7	41.2	34.9	42.5	45.1	40.9	45.1	52.6	35.5	32.9	41.5
SAN (CVPR'23) [49]		GTA-7	43.3	36.7	42.0	48.1	42.5	47.7	49.0	37.8	30.3	41.2
CAT-Seg (CVPR'24) [5]		GTA-7	43.7	37.6	45.4	48.8	43.9	51.2	56.2	39.3	35.7	45.6
MAFT+(ECCV'24) [13]		GTA-7	45.7	38.5	43.7	50.2	44.5	49.1	48.3	39.9	31.8	42.3
ESC-Net (CVPR'25) [18]		GTA-7	42.2	35.9	44.6	48.1	42.7	49.4	52.4	37.9	33.3	43.3
MaskAdapter(CVPR'25) [21]		GTA-7	46.0	36.4	45.7	50.1	44.6	50.1	49.2	38.1	32.7	42.5
CLIPSelf (ICLR'24) [45]		GTA-7+COCO	48.1	37.6	47.9	51.1	46.2	50.6	52.1	40.4	34.6	44.4
RSC-CLIPSelf (ICLR'25) [32]		GTA-7+COCO	46.2	37.9	46.0	51.6	45.4	51.1	50.7	40.3	32.0	43.5
CAT-Seg+AdvStyle [68]		GTA-7	44.9	38.2	43.8	51.1	44.5	49.2	54.2	39.1	34.3	44.2
CAT-Seg+DGInStyle [12]		GTA-7	44.4	39.1	48.1	51.4	45.8	50.7	55.9	39.6	34.0	45.1
S^2-Corr (Ours)		GTA-7	**49.3**	**40.4**	**48.6**	**54.5**	**48.2**	**51.3**	**56.4**	**42.4**	**36.7**	**46.7**
CAT-Seg (CVPR'24) [5]	EVA02 ViT-L/14	GTA-7	48.7	42.4	47.9	50.8	47.5	53.6	59.7	41.1	38.2	48.2
MaskAdapter(CVPR'25) [21]		GTA-7	47.9	40.4	47.2	50.4	46.5	52.1	55.9	37.1	37.1	45.6
CLIPSelf (ICLR'24) [45]		GTA-7+COCO	47.8	43.1	50.0	50.1	47.8	51.2	59.7	40.6	39.7	48.0
CAT-Seg+DGInStyle [12]		GTA-7	46.0	43.9	48.9	51.0	47.5	53.7	59.0	37.9	38.0	47.2
S^2-Corr (Ours)		GTA-7	**52.2**	**44.2**	**50.3**	**53.3**	**49.9**	**55.4**	**59.8**	**41.5**	**40.8**	**49.4**

Table 3. Comparison of different OV-SS methods across various backbones under the *synthetic-to-real* OVDG-SS setting.

and backbones. Overall, OVDG exhibits strong generalization in both the Real-to-Real (CS-7 → Dv-19 / Dv-58) and Synthetic-to-Real (GTA-7 → Dv-19 / Dv-58) settings, achieving clear improvements over prior OV-SS approaches under both small- and large-vocabulary transfers.

Results of Real-to-Real OVDG-SS. With ViT-B/16, S^2-Corr achieves 50.3% on Dv-19, surpassing the previous best by 4.3 points and outperforming CAT-Seg and MaskAdapter by 6.8 and 4.8 points. On Dv-58, it reaches 47.9%, improving the strongest prior result by 2.9 points with consistent gains across ACDC-41, BDD-41, and Mapi-30. With ViT-L/14, S^2-Corr attains 55.8% on Dv-19, improving the best existing result by 2.5 points and exceeding CAT-Seg and MaskAdapter by 6.5 and 5.1 points. On Dv-58, it reaches 53.2%, outperforming the strongest baseline by 1.7 points while maintaining similarly stable improvements across all target domains.

Results of Synthetic-to-Real OVDG-SS. The GTA-to-real gap makes open-vocabulary transfer particularly challenging. With ViT-B/16, on Dv-19, S^2-Corr achieves 48.2%, outperforming the previous best (46.2%) by 2.0 points and exceeding CAT-Seg and MaskAdapter by 4.3 and 3.6 points. On Dv-58, it reaches 46.7%, improving the strongest prior method by 1.1 points with consistent gains across ACDC-41, BDD-41, and Mapi-30. With ViT-L/14, S^2-Corr attains 49.9% on Dv-19, surpassing the previous best by 1.7 points and maintaining improvements across all real target domains. On Dv-58, it reaches 49.4%, exceeding the strongest baseline by 1.2 points while sustaining clear gains on ACDC-41, BDD-41, Mapi-30, and RW-10.

4.4. Ablation Study and Analysis

Note	Methods	ViT-B/16		ViT-L/14		Ave.
		Dv-19	Dv-58	Dv-19	Dv-58	
Base	Cross-attention	43.5	43.5	49.3	50.0	46.6
#1	Seletive SSM	45.6	44.1	50.7	50.5	47.7
#2	+Modulation	47.6	45.3	52.1	50.9	49.0
#3	+Geometric decay	48.3	46.4	53.2	51.8	49.9
#4	+Chunk	49.6	47.3	55.3	52.7	51.2
#5	+Snake scanning	**50.3**	**47.9**	**55.8**	**53.2**	**51.8**

Table 4. Ablation study of the proposed method under different design components on CS7 → Dv-19 and Dv-58 across backbones. Method #1 replaces cross-attention by selective SSM. Methods of #2–#5 are progressively added on top of #1.

Ablation Study on Components. Table 4 reports ablations on CS-7 → Dv-19 and Dv-58 using both ViT-B/16 and ViT-L/14. Three observations emerge. *First*, replacing the Swin-style cross-attention [5] with our selective SSM baseline already yields clear gains (#1 vs. Base), showing the advantage of sequential correlation aggregation over windowed attention. *Second*, each enhancement further strengthens the baseline: modulation improves correlation modeling (#2), and geometric decay brings additional gains by attenuating unreliable long-range interactions (#3). *Third*, the noise-suppression components—geometric decay and the chunk mechanism—deliver the largest improvements, particularly under the larger Dv-58 vocabulary. The final snake-scanning variant (#5) achieves the best results across all settings, confirming the importance of controlling long-range propagation while preserving spatial continuity. Overall, these ablations verify that each component contributes to a stronger correlation-refinement pipeline, together forming a robust solution for OVDG-SS.

Correlation Aggregation Analysis. Fig. 6 shows correlation refinement under domain shifts. CAT-Seg often yields diffuse or misaligned responses, especially on unseen classes such as tunnel and railway. S^2-Corr produces clearer and more localized correlations that better follow object structures, even in low-visibility scenes. For seen and unseen classes alike, S^2-Corr suppresses noise, aligns with semantic boundaries, and ultimately improves OVDG-SS.

Qualitative Results. Fig. 7 shows that DG-SS methods fail

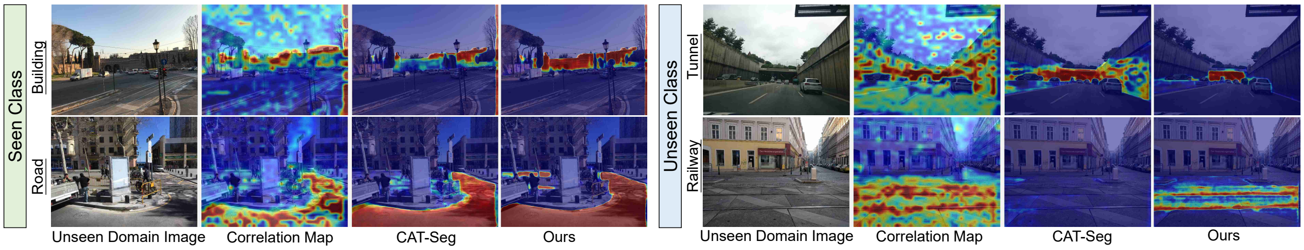


Figure 6. Comparison of text–image correlation aggregation on seen and unseen classes from unseen domains. Our method yields clearer and more localized text–image correlations than CAT-Seg [5] and Corelation Map (initially obtained by Eq. 1), improving both seen and unseen predictions under domain shifts.

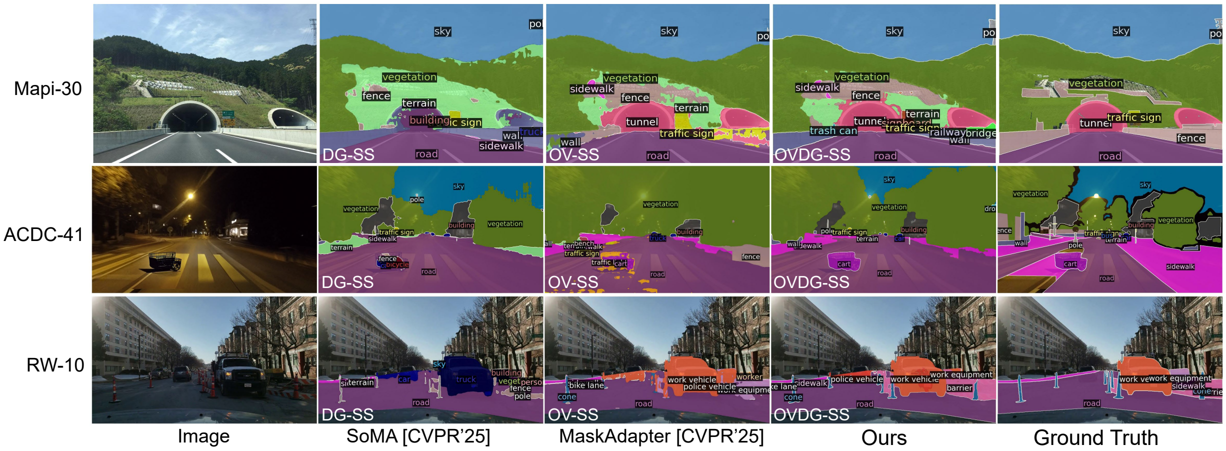


Figure 7. Qualitative comparisons on unseen domains under the OVDG-SS setting, with DG-SS method [54] and OV-SS method [21].

Model	FPS vocab19	FPS vocab58	FPS vocab150	GPU Mem. (GB)	Training Time (min)
CAT-Seg [5]	15.4	10.6	5.7	13.8	180
ESC-Net [18]	15.0	9.9	5.1	15.7	220
Seletive SSM	12.3	9.4	6.4	16.8	240
S^2-Corr (Ours)	**26.1**	**22.2**	**18.3**	**9.2**	**140**

Table 5. Inference and training efficiency of competing methods using a ViT-B/16 backbone. FPS is evaluated under different test vocabulary sizes.

on unseen terrains, often misclassifying critical structures such as tunnels. OV-SS methods offer broader vocabulary but remain unstable under large domain shifts. In contrast, OVDG-SS delivers more accurate and consistent segmentation across diverse unseen domains, aligning well with true scene semantics.

Efficiency Analysis. As shown in Table 5, our method achieves clear efficiency advantages over existing correlation-based OV-SS models such as CAT-Seg and ESC-Net. When the vocabulary size increases, the throughput of CAT-Seg and ESC-Net drops sharply to 5.7 and 5.1 FPS, whereas our model still maintains 18.3 FPS. Our GPU memory usage is also significantly lower at 9.2 GB, compared with 13.8 GB and 15.7 GB, and the training time decreases from 180–220 minutes to 140 minutes. These results demonstrate the superior computational efficiency and scalability of our design under large-vocabulary settings. Although native SSMs scan all tokens sequentially and incur considerable overhead that can make them slower than CAT-Seg under large vocabularies, our chunk-wise aggregation preserves high parallelism and avoids unnecessary long-range interactions. This leads to substantially faster inference and much lower memory consumption. Additional efficiency comparisons are provided in Fig. 2.

More detailed Analysis on hyper-parameters, vocabulary sizes, training and benchmark are provided in the appendix.

5. Conclusion

In this work, we introduce Open-Vocabulary Domain Generalization Semantic Segmentation (OVDG-SS) as a unified and realistic setting under both unseen domains and unseen categories. Through a new autonomous-driving benchmark, we uncover a core weakness of VLM-based OV-SS models, *i.e.,* their text–image correlations degrade sharply under domain shifts. To address this challenge, we propose S^2-Corr, a state-space correlation refinement mechanism that significantly improves cross-domain robustness while preserving open-vocabulary flexibility. We expect OVDG-SS and S^2-Corr to provide a practical step toward building more adaptable and reliable open-world segmentation systems for dynamic real-world environments.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants No. 62572166 and No. 62402157, and by the Fundamental Research Funds for the Central Universities under Grant No. JZ2025HGTB0219. It was also supported by the EU Horizon projects ELIAS (No. 101120237) and ELLIOT (No. 101214398), and by the FIS project GUIDANCE (No. FIS2023-03251).

References

[1] Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3689–3698, 2024. 3

[2] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2024. 3

[3] Mathilde Caron et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[4] Prithvijit Chattopadhyay, Kartik Sarangmath, Vivek Vijaykumar, and Judy Hoffman. Pasta: Proportional amplitude spectrum training augmentation for syn-to-real domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19288–19300, 2023. 3

[5] 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. 2, 3, 6, 7, 8

[6] 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, 6

[7] Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024. 2, 4

[8] Hao Fang et al. Eva-clip: Improving vision-language models with masked modeling. *arXiv preprint arXiv:2303.13495*, 2023. 1, 4, 6

[9] 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*, 149:105171, 2024. 2

[10] 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. 2, 6

[11] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 4

[12] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *European Conference on Computer Vision*, 2024. 6, 7

[13] 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. 6, 7

[14] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24905–24916, 2025. 3

[15] Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 143–164. Springer, 2024. 3

[16] 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. 3, 6

[17] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 70–88. Springer, 2024. 6

[18] 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 Computer Vision and Pattern Recognition (CVPR)*, pages 26081–26090, June 2025. 3, 6, 7, 8

[19] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 2

[20] Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10545–10556, 2025. 2

[21] 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. 2, 6, 7, 8

[22] Yunheng Li, Yuxuan Li, Quansheng Zeng, Wenhai Wang, Qibin Hou, and Ming-Ming Cheng. Unbiased region-language alignment for open-vocabulary dense prediction. *arXiv preprint arXiv:2412.06244*, 2024. 2

[23] 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. 2, 6, 7

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2

[25] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3500, 2024. 3

[26] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model, 2024. 5

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2017. 6

[29] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023. 3

[30] 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. 6

[31] Byeonghyun Pak, Byeongju Woo, Sunghwan Kim, Daehwan Kim, and Hoseong Kim. Textual query-driven mask transformer for domain generalized segmentation. In *European Conference on Computer Vision*, pages 37–54. Springer, 2024. 3

[32] Congpei Qiu, Yanhao Wu, Wei Ke, Xiuxiu Bai, and Tong Zhang. Refining clip's spatial awareness: A visual-centric perspective, 2025. 6, 7

[33] 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

[34] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5571–5584, 2023. 3

[35] 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, 6

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 6

[37] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 2

[38] 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. 6

[39] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 139–156. Springer, 2024. 3

[40] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13171–13182, 2024. 3

[41] 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. 3

[42] Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. Declip: Decoupled learning for open-vocabulary dense perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14824–14834, 2025. 3, 6

[43] 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 (CVPR)*, pages 28619–28630, June 2024. 1

[44] 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. 3

[45] 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. 3, 6, 7

[46] Yuxin Wu et al. Detectron2. 2019. https://github.com/facebookresearch/detectron2. 6

[47] 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. *ECCV*, 2024. 2

[48] 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. 6

[49] 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. 3, 6, 7

[50] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2945–2954, 2023. 3

[51] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model, 2022. 2

[52] 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. 3

[53] 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. 6

[54] 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. 3, 8

[55] Qi Zang, Shuang Wang, Dong Zhao, Wanqing Li, Zining Wang, Dou Quan, Fei Liang, and Licheng Jiao. Boosting generalization of semantic segmentation with unseen style seeking-based meta-learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9):16113–16127, 2025. 3

[56] Qi Zang, Shuang Wang, Dong Zhao, Zhun Zhong, Biao Hou, and Licheng Jiao. Joint style and layout synthesizing: Toward generalizable remote sensing semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2024. 3

[57] Qi Zang, Jiayi Yang, Shuang Wang, Dong Zhao, Wenjun Yi, and Zhun Zhong. Changediff: A multi-temporal change detection data generator with flexible text prompts via diffusion model. *AAAI*, 39:9763–9771, Apr. 2025. 3

[58] Dengke Zhang, Fagui Liu, and Quan Tang. Corrclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2411.10086*, 2024. 2

[59] Dong Zhao, Jinlong Li, Shuang Wang, Mengyao Wu, Qi Zang, Nicu Sebe, and Zhun Zhong. Fishertune: Fisher-guided robust tuning of vision foundation models for domain generalized segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15043–15054, June 2025. 1, 3

[60] Dong Zhao, Shuang Wang, Qi Zang, Licheng Jiao, Nicu Sebe, and Zhun Zhong. Stable neighbor denoising for source-free domain adaptive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23416–23427, June 2024. 3

[61] Dong Zhao, Shuang Wang, Qi Zang, Dou Quan, Xiutiao Ye, and Licheng Jiao. Towards better stability and adaptability: Improve online self-training for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11733–11743, June 2023. 1

[62] Dong Zhao, Shuang Wang, Qi Zang, Dou Quan, Xiutiao Ye, Rui Yang, and Licheng Jiao. Learning pseudo-relations for cross-domain semantic segmentation. In *ICCV*, pages 19191–19203, October 2023. 3

[63] Dong Zhao, Qi Zang, Nan Pu, Shuang Wang, Nicu Sebe, and Zhun Zhong. Secov2: Semantic connectivity-driven pseudo-labeling for robust cross-domain semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):10378–10395, 2025. 1

[64] Dong Zhao, Qi Zang, Shuang Wang, Nicu Sebe, and Zhun Zhong. Pseudo-sd: Pseudo controlled stable diffusion for semi-supervised and cross-domain semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22393–22403, October 2025. 3

[65] 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. 3

[66] Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Prototypical hash encoding for on-the-fly fine-grained category discovery. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 101428–101455. Curran Associates, Inc., 2024. 1

[67] Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 41–58, Cham, 2025. Springer Nature Switzerland. 1

[68] Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial style augmentation for domain generalized urban-scene segmentation. *Advances in neural information processing systems*, 35:338–350, 2022. 3, 6, 7

[69] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*, pages 696–712. Springer, 2022. 3