This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Xin Zhang[1] Robby T. Tan[1,2]
[1]National University of Singapore [2]ASUS Intelligent Cloud Services
x.zhang@u.nus.edu robby.tan@nus.edu.sg

Abstract

Vision Foundation Models (VFMs) and Vision-Language Models (VLMs) have gained traction in Domain Generalized Semantic Segmentation (DGSS) due to their strong generalization capabilities [1]. However, existing DGSS methods often rely exclusively on either VFMs or VLMs, overlooking their complementary strengths. VFMs (e.g., DINOv2) excel at capturing fine-grained features, while VLMs (e.g., CLIP) provide robust text alignment but struggle with coarse granularity. Despite their complementary strengths, effectively integrating VFMs and VLMs with attention mechanisms is challenging, as the increased patch tokens complicate long-sequence modeling. To address this, we propose MFuser, a novel Mamba-based fusion framework that efficiently combines the strengths of VFMs and VLMs while maintaining linear scalability in sequence length. MFuser consists of two key components: MVFuser, which acts as a co-adapter to jointly fine-tune the two models by capturing both sequential and spatial dynamics; and MTEnhancer, a hybrid attention-Mamba module that refines text embeddings by incorporating image priors. Our approach achieves precise feature locality and strong text alignment without incurring significant computational overhead. Extensive experiments demonstrate that MFuser significantly outperforms state-of-the-art DGSS methods, achieving 68.20 mIoU on synthetic-to-real and 71.87 mIoU on real-to-real benchmarks. The code is available at https://github.com/devinxzhang/MFuser.

1. Introduction

Developing semantic segmentation models that can robustly handle diverse and unseen conditions [6, 54–56] is critical for real-world applications such as autonomous driving, where variations in environment, lighting, and weather [1, 5, 30, 31, 53] can significantly impact performance. Domain Generalized Semantic Segmentation (DGSS) aims for strong performance across unseen domains without relying on target domain data during training. Traditional approaches include normalization and whitening techniques [9, 39], domain randomization methods [22, 61, 63]. Despite these efforts, existing approaches remain suboptimal, as they often rely on conventional backbones pre-trained on limited datasets, which struggle to generalize effectively to the diverse challenges encountered in real-world scenarios.

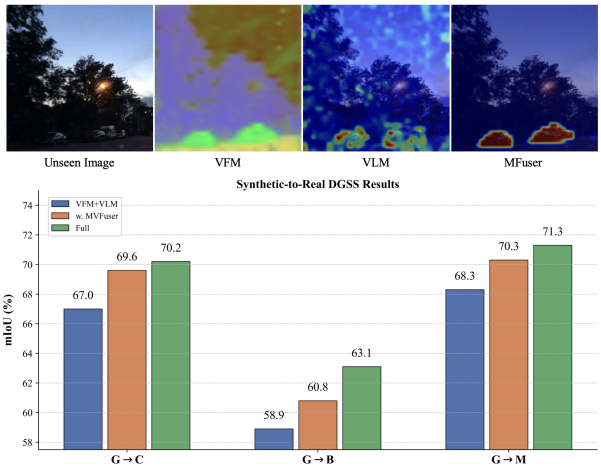


Figure 1. Comparative analysis of the VFM and the VLM features. VFM: Visualization of PCA-computed features from DINOv2 (the first three components of PCA, computed on the image features, serve as color channels), displaying fine-grained details but lacking text alignment. VLM: Image-text similarity map from EVA02-CLIP using the query 'car', demonstrating good alignment with text but insufficient localization of queried objects. MFuser: Our proposed fusion framework integrates VFM and VLM, resulting in unified features that exhibit both precise locality and robust text alignment. Quantitative results on synthetic-to-real DGSS benchmarks further validate our approach, with MFuser consistently achieving the highest mIoU scores across all tasks.

The recent emergence of Vision Foundation Models (VFMs) and Vision Language Models (VLMs) has estab-

[1]In this paper, we refer to foundation models trained solely on visual data as VFMs and those trained on both visual and textual data as VLMs.

lished them as powerful tools for achieving generalization in various domains. Some studies have introduced parameter-efficient fine-tuning (PEFT) methods that effectively adapt these foundation models for DGSS [50, 58]. Additionally, some works leverage diffusion models [43] to generate diverse-style images for training DGSS models [14]. VLMs, in particular, have demonstrated the ability to generalize effectively across varied domains by utilizing text embeddings that provide semantic and domain-invariant representations [40]. This capability has sparked the development of multiple approaches in both image classification [8, 23] and semantic segmentation [14, 36]. However, the specific differences between VFMs and VLMs in the context of DGSS remain underexplored.

VFM features (e.g., DINOv2 [35]) capture strong details at a granular level. In contrast, VLM features (e.g., EVA02-CLIP [15]) struggle to associate text semantics with precise visual regions due to their image-level alignment training. However, this alignment enables VLMs to leverage text embeddings as semantic anchors [37], guiding visual features to remain robust across domain variations. To examine their properties, we perform principal component analysis (PCA) on the DINOv2 features at the final layer. As illustrated in Fig. 1, the PCA-computed features from DINOv2 clearly distinguish between different objects (e.g., cars and trees), even in low-light conditions. Additionally, we apply EVA02-CLIP with the text query 'car'. The activation map also indicates the presence of cars but appears incomplete. This raises an important question: *how can we combine both models to extract features that are both locally precise and text-aligned, enabling effective use of text embeddings for improved generalization?*

An intuitive idea would be to utilize both a VFM and a VLM for training a segmentation model. However, without fine-tuning, foundation models may struggle to adapt to DGSS tasks [50] and VLM text embeddings often fail to align with VFM features, resulting in suboptimal performance. Fully fine-tuning both models, meanwhile, is computationally prohibitive. As such, we propose to introduce additional trainable parameters while keeping the original ones frozen, enabling efficient adaptation. Moreover, combining features from both encoders doubles the patch sequence length, complicating even parameter-efficient fine-tuning methods in handling such long-range sequences. This leads us to our second question: *how can we efficiently adapt and integrate both a VFM and a VLM for DGSS?*

To this end, we propose MFuser, a novel fusion framework based on the State-Space Model (SSM) that efficiently unifies the strengths of VFMs and VLMs. SSMs [16, 66] are well-suited for capturing long-range dependencies with linear computational complexity, making them ideal for jointly adapting VFMs and VLMs with minimal overhead. Following recent advances in text-guided segmentation [36, 57, 65], we build MFuser on the text-queried Mask2Former [7] pipeline, where class text embeddings serve as queries for the segmentation decoder, enabling class-aware feature refinement. Specifically, we introduce **MV**Fuser, a **M**amba-based co-adapter that jointly fine-tunes the two **V**isual models. By taking concatenated patch tokens (features) from both models at each layer, MVFuser models both sequential dynamics and spatial relationships among tokens in parallel. This enables effective interaction between the two feature types, enhancing the granularity of VLM features while also reducing trainable parameters.

To further ensure cross-modality consistency between the fused visual features and VLM **T**ext embeddings, we introduce **MT**Enhancer. MTEnhancer employs a hybrid attention-**M**amba architecture, leveraging the strengths of both model families. Visual features are used as conditional inputs within MTEnhancer, enabling effective sequence modeling that produces text embeddings closely related to visual content, resulting in image-conditioned text embeddings. Extensive experiments across diverse DGSS settings demonstrate that the proposed MFuser consistently outperforms existing state-of-the-art methods, achieving superior results in both synthetic-to-real and real-to-real scenarios. Contributions can be summarized into three aspects:

- We propose a novel fusion framework, MFuser, to collaborate arbitrary pairs of VFMs and VLMs for DGSS, integrating the strengths of both without introducing significant computational overhead.
- We present MVFuser, a Mamba-based co-adapter that enables joint fine-tuning of VFMs and VLMs, bridging the gap between these models and enhancing their complementary feature interactions. Additionally, we introduce MTEnhancer, a hybrid attention-Mamba module that refines text embeddings with visual priors, ensuring superior cross-modal consistency and robust alignment.
- Extensive experiments show the proposed MFuser consistently outperforms state-of-the-art methods, achieving 68.20 mIoU on synthetic-to-real and 71.87 mIoU on real-to-real benchmarks.

2. Related Works

Domain Generalized Semantic Segmentation Domain Generalized Semantic Segmentation (DGSS) aims to develop models capable of generalizing to unseen domains without relying on target domain data during training. Common approaches include meta-learning, which exposes models to diverse tasks to learn features that are robust to domain shifts [25]; data augmentation techniques, such as style transfer and synthetic data creation, to introduce extensive visual diversity [4]; instance normalization and whitening [21, 38, 39, 52], which encourages the model to foucs on domain-invariant features rather than domain-specific styles. Some works also explore to design new architec-

tures based on transformers [12, 20]. Recently, increasing attention has been paid to leveraging foundation models to enhance generalization [37, 50, 58]. Efforts have been taken to harness generative foundation models to creat new images [2], parameter-efficiently fine-tune VFMs [50], leverage textual semantics to guide invariance learning [37], etc. However, the complementary potential of combining VFMs and VLMs remains largely underexplored.

Foundation Models Foundation models represent a transformative approach in deep learning, focusing on pre-training networks on a vast collection of unlabeled images. This pre-training equips the model with strong general representation capabilities, allowing it to be fine-tuned effectively for various downstream tasks. Initially popularized in Natural Language Processing (NLP), this paradigm has also drawn increasing attention in computer vision. In this paper, we refer to the vision-only pre-trained models as Vision Foundation Models (VFMs) including DINO [3] and DINOv2 [35], iBOT [64], MAE [19], SAM [27], etc. Vision-language pre-trained models are referred to as Vision Language Models (VLMs), which include CLIP [40], EVA02-CLIP [15, 49], SIGLIP [60], etc. There are also generative foundation models such as Stable Diffusion [43, 51]. We focus on effectively combining VFMs and VLMs for DGSS.

State Space Models for Visual Applications State-space models (SSMs) [17, 47] have emerged as promising alternatives for capturing long-range dependencies, offering linear scalability with sequence length. Building on the foundational S4 model [17], which introduced deep state-space modeling, SSMs have found applications across a range of fields, including Natural Language Processing (NLP) [33], computer vision [66], medical applications [45]. Mamba [16] extended S4 by introducing a hardware-aware design and a selective scan mechanism, leading to the development of a selective SSM called the S6 model. More recently, VMamba [66] emerged as a fully Mamba-based architecture for vision tasks, while other studies [18] explored hybrid models combining Mamba and transformers. Unlike previous SSM-based efforts that primarily focus on creating entire backbone architectures, we take a different approach by designing Mamba-based adapters to efficiently fine-tune pre-trained VFMs and VLMs. This method enhances the adaptability and performance of VFMs and VLMs across various domains, leveraging Mamba's efficiency to optimize existing models rather than training from scratch.

3. Preliminary

Domain Generalized Segmantic Segmentation Given the source images $\mathcal{X}^S = \{x_i^S\}_{i=1}^{N_S}$ with corresponding ground truth masks $\mathcal{Y}^S = \{y_i^S\}_{i=1}^{N_S}$ where N_S denotes the number of source images, and a segmentation model M, composed of a visual encoder E followed by a segmentation decoder D, namely $M = D \circ E$, domain generalized semantic segmentation (DGSS) aims to train the network to generalize to unknown target domains. With the advancements in foundation models, recent DGSS methods increasingly leverage their strong generalization capabilities to design effective visual encoders [50, 58].

Semantic Segmentation with Text Queries Recent segmentation frameworks like Mask2Former [7], utilize a query-based mechanism where learnable object queries serve as dynamic pointers to direct the model's focus on relevant regions. Building on this, recent studies have increasingly leveraged the image-text alignment capabilities of Vision Language Models (VLMs) to design text-based queries [11, 29, 32, 36, 57, 65]. The text embeddings produced by VLMs have been found to be inherently domain-invariant, capturing semantic information that remains consistent across various contexts and visual styles. This domain invariance stems from the VLM training process, which associates textual descriptions with diverse visual inputs, effectively disentangling semantic content from domain-specific features. The domain invariance of text embeddings forms a basis for promoting the domain generalization of visual features. In this paper, we follow a similar pipeline which utilizes the text embeddings of each class as the queries in a Mask2Former decoder. Formally, the visual encoder E_V^{VLM} of a VLM serves as the encoder of the segmentation model, the aligned text encoder E_T^{VLM} generates class embeddings $q_t = [t^1, t^2, ..., t^C]$ for each class label name $\{\mathrm{class_k}\}_{\mathrm{k=1}}^{\mathrm{C}}$. q_t will be used to design queries or conditional queries of the decoder [36, 57, 65].

4. Proposed Method

In this section, we introduce the Mamba-based foundation models fuser (MFuser), a framework designed to integrate an arbitrary VFM with a CLIP-like VLM using a Mask2Former decoder for DGSS. Fig. 2 illustrates the overall architecture of MFuser. MFuser enhances feature locality while leveraging domain-invariant semantic knowledge provided by text embeddings to effectively constrain visual representations. The core components of this framework include MVFuser and MTEnhancer. MVFuser jointly fine-tunes the visual encoders of both models in a parameter-efficient manner, fusing their features to maximize synergy. MTEnhancer enriches the text queries by incorporating visual features, enhancing semantic alignment and feature robustness.

4.1. MVFuser

Due to the large number of parameters in the VFM and VLM visual encoders, fully fine-tuning all parameters is impractical. Instead, we propose the introduction of additional

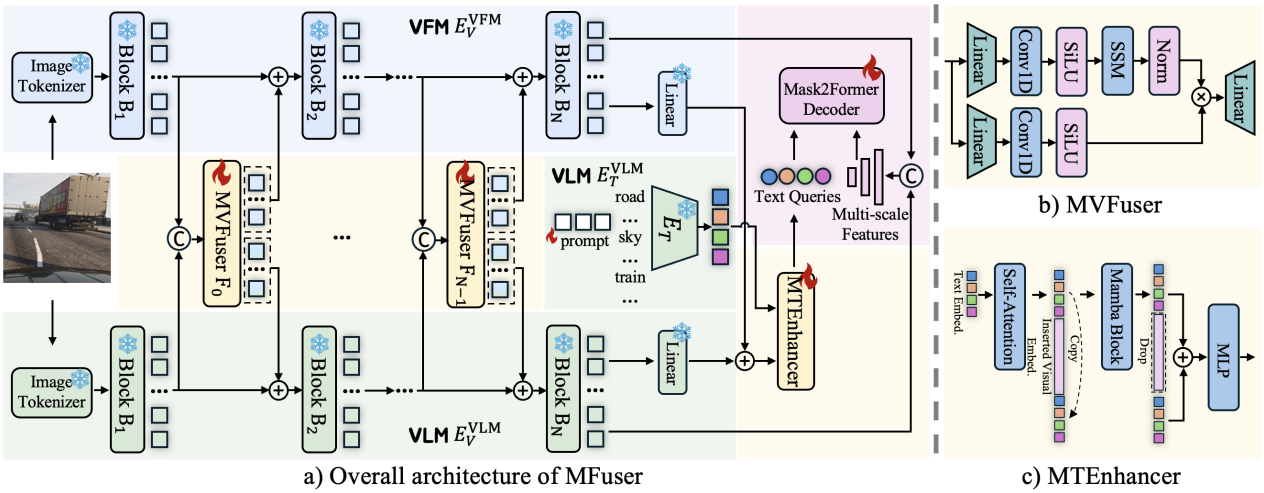


Figure 2. Overall architecture of MFuser. MFuser takes inputs through both VFM and VLM visual encoders. Features from each encoder layer are concatenated and refined in MVFuser, which captures sequential and spatial dependencies in parallel. The refined features are then added back to the original features and passed to the next layer. MTEnhancer strengthens text embeddings of each class by integrating visual features through a hybrid attention-Mamba mechanism. The enhanced text embeddings serve as object queries for the Mask2Former decoder, alongside multi-scale visual features. During training, only MVFusers, MTEnhancers, and the segmentation decoder are trainable while the VFM and VLM remain frozen, preserving their generalization ability and enabling efficient training. Note that skip connections between each block of MTEnhancer are omitted for clarity.

modules, MVFuser, to refine visual features while keeping the original encoder parameters frozen.

This design offers several advantages. First, the distinct characteristics of the two visual encoders could be compromised by full fine-tuning, whereas adapter-style fine-tuning preserves their original strengths while mitigating their weaknesses. Second, refining features from both encoders through a shared MVFuser encourages effective interaction between the two feature types.

Specifically, the visual encoders of VFMs and VLMs are composed of an image tokenizer layer and N consecutively connected transformer blocks $\{B_i\}_{i=1}^{N}$. The image tokenizer layer first converts a 2D image into flatten patch tokens $x_p \in \mathbb{R}^{T\times D}$, where T represents the length of the patch sequence and D denotes the feature dimension.

Normally, x_p is input into the transformer blocks to calculate features. The process is as follows:

$$x_1 = B_1(x_p), x_i = B_i(x_{i-1}), \quad (1)$$

where x_i is the token features output by Block B_i. The features for VFM and VLM can be denoted as x_i^{VFM}, and x_i^{VLM}, respectively.

As stated, x_i^{VFM} exhibits finer granularity, from which x_i^{VLM} can benefit through the interaction. We propose inserting the MVFuser at each block to bridge the two visual encoders, encouraging layer-wise interaction of the two models. MVFuser receives both x_i^{VFM} and x_i^{VLM} as input, the learned feature offsets are then added back to x_i^{VFM} and x_i^{VLM}, respectively, enabling multi-level feature refinement where one MVFuser refines the features from both encoders:

$$[\Delta x_i^{\text{VFM}}; \Delta x_i^{\text{VLM}}] = \text{MVFuser}([x_i^{\text{VFM}}; x_i^{\text{VLM}}]), \quad (2)$$

$$x_i^{\text{VFM}\prime} = x_i^{\text{VFM}} + \Delta x_i^{\text{VFM}}, x_i^{\text{VLM}\prime} = x_i^{\text{VLM}} + \Delta x_i^{\text{VLM}}, \quad (3)$$

where Δx_i^{VFM} and Δx_i^{VLM} are the learned feature offsets for VFM and VLM, respectively. $x_i^{\text{VFM}\prime}$ and $x_i^{\text{VLM}\prime}$ symbolize the refined features.

MVFuser acts two roles: 1) refines x_i^{VFM} and x_i^{VLM} to generate more task-specific features; 2) interacts between two kinds of features to complement each's weaknesses. A natural idea to capture inter-token relationship is to employ self-attention mechanism. However, the sequence length is doubled with the features from the two encoders. Applying the attention mechanism in transformers for adaptation is inefficient due to the quadratic increase in computational complexity with token count. While introducing learnable tokens and applying cross-attention between learnable tokens and patch token features can reduce this computational cost, it struggles to capture inter-token dependencies effectively. To address these challenges, we design a fusion module based on state-space models for efficient long-range sequence modeling.

Core of the MVFuser The architecture of MVFuser is shown in Fig. 2. Token features from both encoders are concatenated to form the input to MVFuser. Following a bottleneck design, MVFuser first projects the concatenated token

features to a lower-dimensional space, models inter-token dependencies, and then projects them back to the original feature dimension.

We modify the original Mamba block to encourage the two branches to capture the sequential dynamics and spatial relationships respectively in parallel.

$$x_i^{(\text{seq})} = \text{SSM}(\text{conv}(\text{proj}([x_i^{\text{VFM}}; x_i^{\text{VLM}}]))), \quad (4)$$

$$x_i^{(\text{spa})} = \text{conv}(\text{proj}([x_i^{\text{VFM}}; x_i^{\text{VLM}}])). \quad (5)$$

Note that we omit the activation and normalization layers for clarity. Finally, a gating mechanism is applied between the outputs of the two branch to improve generalization, followed by a projection layer to recover the feature dimension.

$$[\Delta x_i^{\text{VFM}}; \Delta x_i^{\text{VLM}}] = \text{proj}(x_i^{(\text{seq})} \otimes x_i^{(\text{spa})}), \quad (6)$$

where $\otimes$ denotes the element-wise multiplication.

4.2. MTEnhancer

Text embeddings have been utilized as queries in semantic segmentation by framing the task as a matching problem between representative class queries and image patch features, or by serving as the initial object queries for the Mask2Former decoder. This approach leverages the domain-invariant semantic information embedded in text to enhance the model's ability to accurately identify and segment relevant regions within an image [57, 65]. Unlike previous methods, which typically assume that visual features and text embeddings are already aligned in a pre-trained VLM, our approach enhances the original text embeddings from a VLM by incorporating the fused visual priors through the proposed MTEnhancer. MTEnhancer is designed to enriches text embeddings by modeling their relationships with fused image tokens.

As illustrated in Fig. 2, MTEnhancer is a hybrid architecture combining an attention block, a conditional Mamba block, and an MLP, leveraging the strengths of diverse model architectures. The attention block encodes inter-class relationships, while the conditional Mamba block integrates image tokens into the text embeddings. While the Mamba block excels at processing long token sequences, its use in cross-attention mechanisms remains largely unexplored. To efficiently leverage the unidirectional scan order inherent to Mamba, we propose concatenating two copies of text embeddings at both sides of the image token, together they serve as the input of the Mamba block. Each block within MTEnhancer is implemented with residual connections.

$$q_t = q_t + \text{Attention}(q_t), \quad (7)$$

$$[\Delta q_t; \Delta x_v; \Delta q_t^{\text{copy}}] = \text{Mamba}([q_t; x_v; q_t^{\text{copy}}]), \quad (8)$$

$$q_t = q_t + \Delta q_t + \Delta q_t^{\text{copy}}, \quad (9)$$

$$q_t = q_t + \text{MLP}(q_t), \quad (10)$$

where x_v represents the fused visual features output by the encoders' final heads. q_t is denoted without distinguishing between updates throughout the process. We adopt the approach of using enhanced text embeddings q_t as object queries for a Mask2Former decoder [36, 57].

Training Objective We train the framework with the prediction-level segmentation loss together with the feature-level alignment loss. For the segmentation loss, we follow the standard Mask2Former [7]:

$$\mathcal{L}_{\text{seg}} = \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}, \quad (11)$$

where $\mathcal{L}_{\text{bce}}$, $\mathcal{L}_{\text{dice}}$, $\mathcal{L}_{\text{cls}}$ represent the binary cross-entropy loss and the dice loss for the predicted masks, and the cross-entropy loss for each queried proposal, respectively.

Additionally, we enforce a pixel-level vision-language alignment using a pixel-text alignment loss to ensure that textual semantics are precisely mapped to corresponding image regions [41]. The experiments involve three VLMs: CLIP, EVA02-CLIP, and SIGLIP. We apply SoftMax loss for CLIP and EVA02-CLIP, and Sigmoid loss for SIGLIP, consistent with the loss functions used during each VLM's original training. Therefore, the overall training loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{align}}. \quad (12)$$

5. Experiments

5.1. Settings

Datasets We evaluate the performance of MFuser on both synthetic-to-real, clear-to-adverse-weather, and real-to-real scenarios are involved. As synthetic datasets, GTAV [42] contains 12,403, 6,382, and 6181 images for training, validation, and testing, respectively, at a resolution of 1914×1052. As real-world datasets, Cityscapes [10] comprises 2,975 images for training and 500 images for validation, with a resolution of 2048×1024. BDD100K [59] includes 7,000 and 1,000 images for training and validation, each at 1280×1024 resolution. Mapillary [34] consists of 18,000 training and 2,000 validation images, with varying resolutions across the dataset. We also include the clear-to-adverse-weather generalization in the supplement.

Network Architecture To make a comprehensive evaluation of the proposed MFuser, we employ the VFM of DINOv2 [35], and VLMs including CLIP [40], EVA02-CLIP [49], SIGLIP [60]. For the segmentation decoder, we follow tqdm [37] which modifies a standard Mask2Former decoder by replacing the randomly initialized object queries with the enhanced class embeddings. Thus, the text object queries are set to 19 to match the number of classes.

Implementation Details We keep the parameters of the VFM and VLM frozen and only train the MVFuser, MTEnhancer and the segmentation decoder. We use the same

training configuration on all VLM alternatives and both generalization setups. We also apply prompt tuning for the text encoder, similar to [37]. All experiments are conducted with the input size of 512×512, a batch size of 2 and learning rate of 1e-4. Following [37, 50], AdamW optimizer is employed with a linear warm-up over $t_{\text{warm}} = 1.5k$ iterations, followed by a linear decay. Standard augmentations for segmentation tasks are applied, including random scaling, random cropping, random flipping, and color jittering. All experiments are conducted on one 24GB RTX A5000.

5.2. Comparison with State-of-The-Art Methods

We compare our MFuser with existing DGSS methods on two setups: synthetic-to-real (G→{C, B, M}) and real-to-real (C→{B, M}). Three VLMs are involved together with DINOv2, namely CLIP, EVA02-CLIP, and SIGLIP, all of *Large* types. We mainly compare with recent foundation model-based approaches, including CLOUDS [2], VLTseg [24], Rein [50], SET [58], and tqdm [37]. Several conventional methods are also involved. We provide results on Synthia [44] and ACDC [46] in the supplement.

Synthetic-to-Real Generalization Tab. 1 compares the performance of the proposed MFuser with existing state-of-the-art DGSS methods under the synthetic-to-real setup. For each combination of the VFM and VLMs, we consistently outperform the existing methods on all benchmarks by a large margin. In particular, our MFuser with the EVA02-CLIP model improves the G→B benchmark by 1.49 mIoU. On average, we achieve 2.15 mIoU better than the state-of-the-art. Our proposed MFuser remains excellent performance using different VFM and VLM combinations, showing the versatility of our framework. To better understand how the proposed MFuser improves the feature generalization, Fig. 3 shows the qualitative comparison with the most recent methods, Rein [50] and tqdm [37]. Our method identifies fine-grained differences more effectively.

Real-to-Real Generalization As shown in Tab. 2, we compare the performance of MFuser with existing state-of-the-art DGSS methods under the real-to-real setting. MFuser largely surpasses the existing methods with all three VLMs. Specifically, we improve the C→B benchmark by 0.74 mIoU, and the C→M benchmark by 1.7 mIoU. An overall improvement of 1.43 mIoU is achieved.

5.3. In-Depth Analysis

Efficiency Analysis MVFuser is more efficient than *self*-attention-based adapters, which have quadratic complexity in modeling inter-patch relationships. To evaluate this, we replace MVFuser with 3 self-attention-based adapters while keeping all other components intact: self-attn(concat.): $\text{attn}(\text{q}, \text{k}, \text{v}{=}\text{concat}(\text{F}_{\text{VFM}}, \text{F}_{\text{VLM}}))$; self-attn(separate): $\{\text{attn}(\text{q}{=}\text{F}_{\text{VFM}}, \text{k}, \text{v}{=}\text{F}_{\text{VLM}}), \text{attn}(\text{q}{=}\text{F}_{\text{VLM}}, \text{k}, \text{v}{=}\text{F}_{\text{VFM}})\}$.

Table 1. Performance comparison (mIoU in %) under the synthetic-to-real setting (G→{C, B, M}). DINOv2 [35] is used as the VFM for all MFuser variants, showing only the applied VLMs. Our method is marked in gray. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Backbone	synthetic-to-real			
		G→C	G→B	G→M	Avg.
SAN-SAW [39]	RN101	45.33	41.18	40.77	42.43
WildNet [28]	RN101	45.79	41.73	47.08	44.87
SHADE [61]	RN101	46.66	43.66	45.50	45.27
TLDR [26]	RN101	47.58	44.88	48.80	47.09
FAMix [13]	RN101	49.47	46.40	51.97	49.28
SHADE [62]	MiT-B5	53.27	48.19	54.99	52.15
IBAFormer [48]	MiT-B5	56.34	49.76	58.26	54.79
VLTSeg [24]	CLIP-B	47.50	45.70	54.30	49.17
CLOUDS [2]	ConvNeXt-L	60.20	57.40	67.00	61.50
VLTSeg [24]	EVA02-L	65.60	58.40	66.50	63.50
Rein [50]	EVA02-L	65.30	60.50	64.90	63.60
Rein [50]	DINOv2-L	66.40	60.40	66.10	64.30
SET [58]	DINOv2-L	68.06	61.64	67.68	65.79
tqdm [37]	EVA02-L	68.88	59.18	70.10	66.05
MFuser	CLIP-L	**71.24**	61.08	71.14	67.82
MFuser	SIGLIP-L	71.10	61.19	**71.71**	68.00
MFuser	EVA02-L	70.19	**63.13**	71.28	**68.20**

Table 2. Performance comparison (mIoU in %) under the real-to-real setting (C→{B, M}). DINOv2 [35] is used as the VFM for all MFuser variants, showing only the applied VLMs. Our method is marked in gray. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Backbone	real-to-real		
		B	M	Avg.
SAN-SAW [39]	RN101	54.73	61.27	58.00
WildNet [28]	RN101	47.01	50.94	48.98
SHADE [61]	RN101	50.95	60.67	55.81
HGFormer [12]	Swin-L	61.50	72.10	66.80
VLTSeg [24]	EVA02-L	64.40	76.40	70.40
Rein [50]	EVA02-L	64.10	69.50	66.80
Rein [50]	DINOv2-L	65.00	72.30	68.65
SET [58]	DINOv2-L	65.07	75.67	70.37
tqdm [37]	EVA02-L	64.72	76.15	70.44
MFuser	SIGLIP-L	65.44	77.97	71.71
MFuser	CLIP-L	65.58	**78.10**	71.84
MFuser	EVA02-L	**65.81**	77.93	**71.87**

Table 3. Efficiency analysis. The experiments are conducted with DINOv2 and EVA02-CLIP models under the G→{C, B, M} settings. The best results are highlighted in **bold**.

	Params. (M)	FLOPs (G)	C	B	M	Avg.
self-attn (concat.)	4.20	98.64	70.24	62.31	71.11	67.89
self-attn (separate)	8.40	71.08	69.68	61.91	70.85	67.48
bi-deform-attn	3.35	34.65	69.46	61.17	70.11	66.91
MVFuser	1.67	17.21	**70.19**	**63.13**	**71.26**	**68.19**

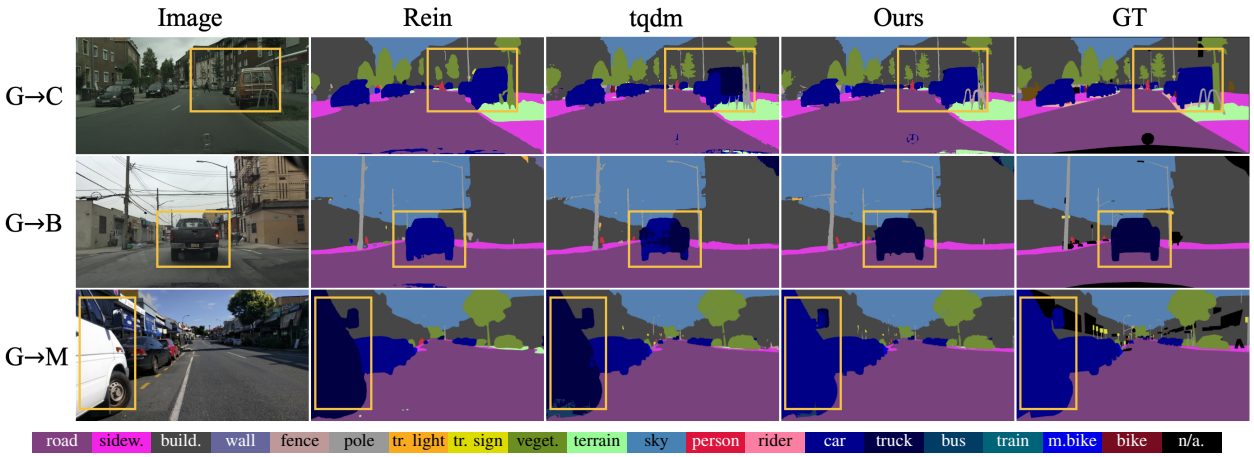


Figure 3. Qualitative results on unseen target domains under the G→{C, B, M} setting. MFuser is compared with Rein [50] and tqdm [37].

bi-deform-attn applies self-attn(concat.) using bidirectional deformable self-attention from Deformable DETR [67]. Tab. 3 summarizes efficiency and results, with parameters and *FLOPs* per adapter (using DeepSpeed package, batch size=2). MVFuser achieves the best while significantly reducing parameters and *FLOPs*.

Foundation Model Ensemble It is natural to consider ensembling multiple foundation models to enhance performance. To rigorously assess the effectiveness of the proposed MFuser, we address the following questions: *1) Is simply combining multi-encoder features sufficient to achieve the desired results? 2) Can any parameter-efficient fine-tuning method alone achieve comparable results?*

To answer the first question, we replaced the MVFuser with a simple concatenation of features from the VFM and VLM visual encoders. We also evaluated using only the VFM or VLM visual features independently. As shown in Tab. 4, merely concatenating the features from both encoders does not yield satisfactory results and even performs worse than using only VFM or VLM features alone. This occurs because the frozen VFM features are not aligned with the text queries when both are input into the decoder. Additionally, the alignment between VLM visual features and text queries is compromised when the VLM features are mixed with the VFM features.

Furthermore, fully fine-tuning both encoders is challenging. For example, fully fine-tuning the EVA02-CLIP visual encoder alone requires 4×80GB A100 GPUs for 20 hours, as reported in [37], which imposes a significant computational burden—let alone the cost of fine-tuning two encoders simultaneously. Alternatively, our MFuser keeps the original VFM and VLM parameters fixed and introduces an additional fusion block, MVFuser, which acts as a bridge between the two foundation models. By optimizing only the MVFuser, we not only adapt the features of both encoders to be more effective but also facilitate interactions between them. Consequently, our method provides a more efficient and effective approach for promoting DGSS with foundation models, achieving the best performance with only 15 hours of training on a single 24GB GPU. Fig. 4 shows that our proposed MVFuser significantly improves the localization and robustness of the features.

Table 4. Ablation studies on the vision feature fusion under the G→{C, B, M} setting. DINOv2 and EVA02-CLIP are applied as the VFM and the VLM, respectively. w.o finetune: directly concatenate features of the two encoders; Conv: utilize convolution layers for fusion; Cross-Attention: implement cross-attention in [50] for fusion. The best results are highlighted in **bold**.

Fusion Choice	C	B	M	Avg.
VFM-only	67.68	60.82	66.89	65.13
VLM-only	68.26	60.02	70.18	66.15
w.o Fintune	66.96	58.88	68.25	64.70
Convolution	69.28	61.45	69.78	66.83
Cross-Attention	69.67	60.52	70.43	66.87
Sep. MVFuser	69.57	62.88	70.59	67.68
MVFuser	**70.19**	**63.13**	**71.28**	**68.20**

To answer the second question, we implement two alternative adapters to fine-tune the two encoders, based on convolution and attention mechanisms, respectively. For the convolution-based adapter, we first reshape the 1D patch sequence into a 2D feature map and then employ an architecture similar to the spatial branch of the MVFuser, replacing 1D convolutions with 2D convolutions. The attention-based adapter reimplements Rein [50] to jointly fine-tune both encoders using a single set of learnable tokens through cross-attention. We do not include a self-attention-based adapter due to its quadratic computational cost with respect to the number of tokens, which makes it impractical. As shown in Table 4, our Mamba-based MVFuser

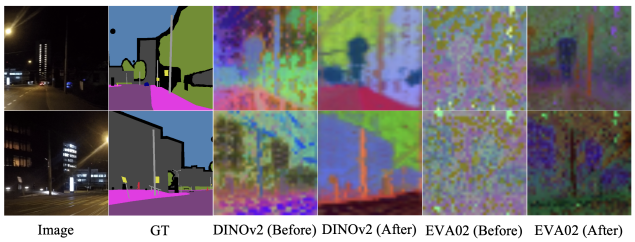


Figure 4. PCA visualization of features from DINOv2 and EVA02-CLIP, illustrating how MVFuser-based adaptation refines their distributions before and after tuning.

significantly outperforms both the convolution-based and attention-based adapters. This is understandable, as the convolution-based adapter captures only local information, while cross-attention struggles to model token dependencies. Conversely, the Mamba-based MVFuser efficiently captures sequential dynamics with linear complexity.

In our implementation of MVFuser, VFM features are concatenated before VLM visual features, aiming to enhance VLM features through Mamba's sequential modeling. To evaluate this, we implemented separate MVFuser for DINOv2 and EVA02-CLIP, disentangling their connection. It can be observed from Tab. 4 that this leads to performance drops, demonstrating the effectiveness of feature interaction. We provide more insights into MVFuser's effectiveness in the supplement.

Foundation Model Choices It remains uncertain whether the performance gain arises from the complementary effects between the VFM and the VLM, or if any two foundation models could achieve similar results. Our method is based on the premise that, while both VFMs and VLMs demonstrate strong robustness, they possess distinct properties due to their different training principles. Consequently, MFuser leverages these differences to complementarily enhance the model's generalization capabilities.

To verify this, we conduct experiments using two VLMs, where the additional VLM serves as the VFM by utilizing only its visual encoder. Two combinations are tested: "SIGLIP + EVA02-CLIP" and "CLIP + EVA02-CLIP" with EVA02-CLIP functioning as the VLM while SIGLIP or CLIP acts as the VFM. Evaluation is conducted under the G→{C, B, M} setting, and results are presented in Tab. 5. Both combinations show slight performance improvements over the "VLM-only" in Tab. 4, yet they fall significantly short of any "VFM + VLM" pairing. This suggests that the complementary effects between VFMs and VLMs are much more significant than those observed among VLMs alone. Additional evaluations on other VFMs beyond DINOv2 are provided in the supplement.

Table 5. Ablation studies on the used foundation models. VFM + VLM: only the visual encoder is used when a VLM serve as the VFM. The experiments are conducted under the G→{C, B, M} setting. The best results are highlighted in **bold**.

VFM + VLM	C	B	M	Avg.
SIGLIP + EVA02	68.48	60.98	69.26	66.24
CLIP + EVA02	68.78	61.17	70.21	66.72
DINOv2 + CLIP	**71.24**	61.08	71.14	67.82
DINOv2 + SIGLIP	71.10	61.19	**71.71**	68.00
DINOv2 + EVA02	70.19	**63.13**	71.28	**68.20**

Table 6. Ablation studies on the text embeddings enhancement. Experiments use DINOv2 and EVA02-CLIP under the G→{C, B, M} settings. The best results are highlighted in **bold**.

Enhancement Choice	C	B	M	Avg.
w.o. Enhance	69.57	60.83	70.32	66.91
w.o. Hybrid	69.62	61.90	70.67	67.40
Cross-Attention	69.88	61.26	70.78	67.31
MTEnhancer	**70.19**	**63.13**	**71.28**	**68.20**

Text Queries Enhancement Solely using class names to obtain text embeddings for each class may not adequately adapt to diverse image types. Encoding image-specific information with text embeddings has been a common practice. In this section, we evaluate the effectiveness of the proposed MTEnhancer under the "G→{C, B, M}" setting using DINOv2 and EVA02-CLIP. As demonstrated in Tab. 6, the advantages provided by MTEnhancer are evident. Notably, the hybrid architecture that incorporates self-attention with the conditional Mamba proves to be effective. Furthermore, MTEnhancer outperforms the approach of utilizing cross-attention to encode visual priors.

6. Conclusions

In this work, we proposed MFuser, a novel fusion framework designed to integrate VFMs and VLMs for DGSS. By leveraging the complementary strengths of VFMs and VLMs, MFuser addresses the challenges of increased patch tokens through efficient, scalable fusion with linear complexity. The framework incorporates two key components: MVFuser, which jointly fine-tunes VFMs and VLMs to enhance feature interaction, and MTEnhancer, which refines text embeddings using image priors for better alignment and robustness. Extensive experimental results demonstrate that MFuser achieves precise feature localization and robust text alignment while outperforming state-of-the-art DGSS methods across various benchmarks. The study underscores the potential of combining VFMs and VLMs to achieve superior generalization capabilities in semantic segmentation tasks, and highlights MFuser's effectiveness in advancing DGSS by improving generalization to unseen domains without adding significant computational overhead.

References

[1] Yihao Ai, Yifei Qi, Bo Wang, Yu Cheng, Xinchao Wang, and Robby T Tan. Domain-adaptive 2d human pose estimation via dual teachers in extremely low-light conditions. In European Conference on Computer Vision, pages 221–239. Springer, 2024. 1

[2] Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lathuilière. Collaborating foundation models for domain generalized semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3108–3119, 2024. 3, 6

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9650–9660, 2021. 3

[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. 2

[5] Tingting Chen, Beibei Lin, Yeying Jin, Wending Yan, Wei Ye, Yuan Yuan, and Robby T Tan. Dual-rain: Video rain removal using assertive and gentle teachers. In European Conference on Computer Vision, pages 127–143. Springer, 2024. 1

[6] Zitan Chen, Zhuang Qi, Xiao Cao, Xiangxian Li, Xiangxu Meng, and Lei Meng. Class-level structural relation modeling and smoothing for visual representation learning. In Proceedings of the 31st ACM International Conference on Multimedia, pages 2964–2972, 2023. 1

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1290–1299, 2022. 2, 3, 5

[8] Junhyeong Cho, Gilhyun Nam, Sungyeon Kim, Hunmin Yang, and Suha Kwak. Promptstyler: Prompt-driven style generation for source-free domain generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15702–15712, 2023. 2

[9] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11580–11590, 2021. 1

[10] 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. 5

[11] Anurag Das, Xinting Hu, Li Jiang, and Bernt Schiele. Mta-clip: Language-guided semantic segmentation with mask-text alignment. In European Conference on Computer Vision, pages 39–56. Springer, 2024. 3

[12] Jian Ding, Nan Xue, Gui-Song Xia, Bernt Schiele, and Dengxin Dai. Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15413–15423, 2023. 3, 6

[13] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. A simple recipe for language-guided domain generalized segmentation. arXiv preprint arXiv:2311.17922, 2023. 6

[14] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. A simple recipe for language-guided domain generalized segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 23428–23437, 2024. 2

[15] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. Image and Vision Computing, page 105171, 2024. 2, 3

[16] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752, 2023. 2, 3

[17] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396, 2021. 3

[18] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. arXiv preprint arXiv:2407.08083, 2024. 3

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16000–16009, 2022. 3

[20] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In European Conference on Computer Vision, pages 372–391. Springer, 2022. 3

[21] Lei Huang, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Iterative normalization: Beyond standardization towards efficient whitening. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4874–4883, 2019. 2

[22] Wei Huang, Chang Chen, Yong Li, Jiacheng Li, Cheng Li, Fenglong Song, Youliang Yan, and Zhiwei Xiong. Style projected clustering for domain generalized semantic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3061–3071, 2023. 1

[23] Zeyi Huang, Andy Zhou, Zijian Ling, Mu Cai, Haohan Wang, and Yong Jae Lee. A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11685–11695, 2023. 2

[24] Christoph Hümmer, Manuel Schwonberg, Liangwei Zhong, Hu Cao, Alois Knoll, and Hanno Gottschalk. Vltseg: Simple transfer of clip-based vision-language representations for domain generalized semantic segmentation. arXiv preprint arXiv:2312.02021, 2023. 6

[25] Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, and Kwanghoon Sohn. Pin the memory: Learning to generalize semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4350–4360, 2022. 2

[26] Sunghwan Kim, Dae-hwan Kim, and Hoseong Kim. Texture learning domain randomization for domain generalized segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 677–687, 2023. 6

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4026, 2023. 3

[28] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9936–9946, 2022. 6

[29] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767, 2023. 3

[30] Beibei Lin, Yeying Jin, Wending Yan, Wei Ye, Yuan Yuan, and Robby T Tan. Nighthaze: Nighttime image dehazing via self-prior learning. arXiv preprint arXiv:2403.07408, 2024. 1

[31] Beibei Lin, Yeying Jin, Wending Yan, Wei Ye, Yuan Yuan, Shunli Zhang, and Robby T Tan. Nightrain: Nighttime video deraining via adaptive-rain-removal and adaptive-correction. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 3378–3385, 2024. 1

[32] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499, 2023. 3

[33] Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. Long range language modeling via gated state spaces. arXiv preprint arXiv:2206.13947, 2022. 3

[34] 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. 5

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023. 2, 3, 5, 6

[36] Byeonghyun Pak, Byeongju Woo, Sunghwan Kim, Dae-hwan Kim, and Hoseong Kim. Textual query-driven mask transformer for domain generalized segmentation. arXiv preprint arXiv:2407.09033, 2024. 2, 3, 5

[37] Byeonghyun Pak, Byeongju Woo, Sunghwan Kim, Dae-hwan Kim, and Hoseong Kim. Textual query-driven mask transformer for domain generalized segmentation. In European Conference on Computer Vision, pages 37–54. Springer, 2025. 2, 3, 5, 6, 7

[38] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In Proceedings of the european conference on computer vision (ECCV), pages 464–479, 2018. 2

[39] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2594–2605, 2022. 1, 2, 6

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021. 2, 3, 5

[41] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 18082–18091, 2022. 5

[42] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pages 102–118. Springer, 2016. 5

[43] 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. 2, 3

[44] 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. 6

[45] Jiacheng Ruan and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491, 2024. 3

[46] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The Adverse Conditions Dataset with Correspondences for semantic driving scene understanding. In ICCV, 2021. 6

[47] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. arXiv preprint arXiv:2208.04933, 2022. 3

[48] Qiyu Sun, Huilin Chen, Meng Zheng, Ziyan Wu, Michael Felsberg, and Yang Tang. Ibaformer: Intra-batch attention transformer for domain generalized semantic segmentation. arXiv preprint arXiv:2309.06282, 2023. 6

[49] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389, 2023. 3, 5

[50] 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. 2, 3, 6, 7

[51] Yihang Wu, Xiao Cao, Kaixin Li, Zitan Chen, Haonan Wang, Lei Meng, and Zhiyong Huang. Towards better text-to-image generation alignment via attention modulation. arXiv preprint arXiv:2404.13899, 2024. 3

[52] Qi Xu, Liang Yao, Zhengkai Jiang, Guannan Jiang, Wenqing Chu, Wenhui Han, Wei Zhang, Chengjie Wang, and Ying Tai. Dirl: Domain-invariant representation learning for generalizable semantic segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 2884–2892, 2022. 2

[53] Weilong Yan, Robby T. Tan, Bing Zeng, and Shuaicheng Liu. Deep homography mixture for single image rolling shutter correction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 9868–9877, 2023. 1

[54] Xin Yang, Michael Bi Mi, Yuan Yuan, Xin Wang, and Robby T Tan. Object detection in foggy scenes by embedding depth and reconstruction into domain adaptation. In Proceedings of the Asian Conference on Computer Vision, pages 1093–1108, 2022. 1

[55] Xin Yang, Wending Yan, Yuan Yuan, Michael Bi Mi, and Robby T Tan. Semantic segmentation in multiple adverse weather conditions with domain knowledge retention. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 6558–6566, 2024.

[56] Xin Yang, Yan Wending, Michael Bi Mi, Yuan Yuan, and Robby Tan. End-to-end video semantic segmentation in adverse weather using fusion blocks and temporal-spatial teacher-student learning. Advances in Neural Information Processing Systems, 37:141000–141020, 2025. 1

[57] Jong Chul Ye, Yujin Oh, et al. Otseg: Multi-prompt sinkhorn attention for zero-shot semantic segmentation. In The 18th European Conference on Computer Vision, ECCV 2024. European Computer Vision Association (ECVA), 2024. 2, 3, 5

[58] 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. 2, 3, 6

[59] 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. 5

[60] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11975–11986, 2023. 3, 5

[61] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In European conference on computer vision, pages 535–552. Springer, 2022. 1, 6

[62] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning: A unified framework for visual domain generalization. IJCV, 2023. 6

[63] 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. 1

[64] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832, 2021. 3

[65] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11175–11185, 2023. 2, 3, 5

[66] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417, 2024. 2, 3

[67] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159, 2020. 7