Unleashing the Power of Visual Foundation Models for Generalizable Semantic Segmentation

PeiYuan Tang[1], Xiaodong Zhang[2,3*], Chunze Yang[1], Haoran Yuan[4], Jun Sun[5], Danfeng Shan[1], Zijiang James Yang[4,6*]

[1]School of Computer Science and Technology, Xi'an Jiaotong University
[2]School of Computer Science and Technology, Xidian University
[3]Shaanxi Key Laboratory of Network and System Security, Xidian University
[4]Synkrotron, Inc.
[5]Singapore Management University
[6]University of Science and Technology of China
tangpeiyuan@stu.xjtu.edu.cn, Zhangxiaodong@xidian.edu.cn

Abstract

Deep learning models often suffer from performance degradation in unseen domains, posing a risk for safety-critical applications such as autonomous driving. To tackle this problem, recent studies have leveraged pre-trained Visual Foundation Models (VFMs) to enhance generalization. However, exsiting works mainly focus on designing intricate networks for VFMs, neglecting their inherent strong generalization potential. Moreover, these methods typically perform inference on low-resolution images. The loss of detail hinders accurate predictions in unseen domains, especially for small objects. In this paper, we argue that simply fine-tuning VFMs and leveraging high-resolution images unleash the power of VFMs for generalizable semantic segmentation. Therefore, we design a VFM-based segmentation network (VFMNet) that adapts VFMs to this task with minimal fine-tuning, preserving their generalizable knowledge. Then, to fully utilize high-resolution images, we train a Mask-guided Refinement Network (MGRNet) to refine VFMNet's predictions combining detailed image features. Furthermore, we adopt a two-stage coarse-to-fine inference approach. MGRNet is used to refine the low-confidence regions predicted by VFMNet to obtain fine-grained results. Extensive experiments demonstrate the effectiveness of our method, outperforming state-of-the-art methods by 3.3% on the average mIoU in synthetic-to-real domain generalization.

Code — https://github.com/tpy001/VFMSeg

Introduction

Deep learning has significantly advanced computer vision tasks like semantic segmentation (Chen et al. 2017; Xie et al. 2021; Cheng et al. 2022). These successes ususally rely on the basic assumption that the training and testing data should come from the same distribution. When models are deployed in the real world, they might encounter unseen scenarios outside of their training data. This may lead to significant performance drops, posing a threat to safety-critical applications, such as autonomous driving. Collecting and labeling data for all possible scenarios is the most effective way to solve this problems, but it often requires a significant amount of time and money. Therefore, Domain Generalization (Yang, Gu, and Sun 2023; Fan et al. 2023; Wang et al. 2023b) has been widely investigated to enhance model generalization ability. It aims to train models on labeled data from source domains (e.g., synthetic data) such that it performs well on unseen domains (e.g., real-world data).

Figure 1: Comparison of predictions using VFMs at low and high resolutions. When inferring on low-resolution images, VFMs often struggle with exact boundaries and small objects in the distance. High-resolution inference using a sliding window approach produces fine-grained results.

In the field of Domain Generalization Semantic Segmentation (DGSS), existing works mainly focus on domain-invariant feature learning (Xu et al. 2022; Peng et al. 2022) and data augmentation techniques (Peng et al. 2021; Zhong et al. 2022) to prevent overfitting to the source domains. In recent years, Visual Foundation Models (VFMs) have demonstrated remarkable performance across various computer vision tasks (Radford et al. 2021; Kirillov et al. 2023; Oquab et al. 2023). Through pre-training on large-scale image data, VFMs acquire universal visual representations that can be transfered to other domains. Their strong image understanding and generalization ability make them ideal for developing generalizable models.

Previous works (Wei et al. 2024; Hümmer et al. 2023) have leveraged VFMs for DGSS by employing their encoders as the feature extractor and then training a complex

*Corresponding authors
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

decoder like Mask2Former (Cheng et al. 2022). We observe that while complex decoders enhance model capacity, they potentially lead to overfitting to the source domain, undermining the generalizability of the VFMs. Since VFMs are trained on large-scale data, we hypothesize that their prior knowledge may help recognize long-tail classes in unseen domains. Therefore, We argue that fine-tuning the VFM with a simple decoder can adapt the model to a specific domain while preserving its generalizable prior knowledge.

High-resolution images with rich detials might improve the segmentation result. Howerev, most VFMs struggle with high-resolution images partly due to the length extrapolation problem (Song et al. 2024), i.e. the inconsistency in token length between the training and prediction impairs performance. Therefore, previous methods (Wei et al. 2024) had to perform inference on downsampled images, leading to poor segmentation results, as shown in Fig. 1. Sliding window inference mitigates this problem by dividing the image into fixed-size patches and then making predictions for each patch. But it still faces challenges in generalizing to higher resolutions. As resolution increases, each patch contains less content, leading to a lack of context and degraded performance. For instance, if a bicycle underneath a person does not appear in the patch, it can be challenging to distinguish whether this person is a pedestrian or a cyclist.

To address this issue, we propose leveraging low-resolution semantic predictions to guide inference on high-resolution patches. The low-resolution predictions provide initial labels for each pixel, although they may contain errors such as ignoring small objects. Based on these class priors, the model is trained to retain labels that match the image features and adjust those that do not. Therefore, by extracting useful information from low-resolution semantic predictions, the model can effectively overcome the lack of contextual information.

In this work, we focus on the task of domain generalization semantic segmentation (DGSS) and introduce a novel framework to fine-tune VFMs and enable high-resolution inference. To achieve this, we design a simple yet effective VFM-based segmentation network (VFMNet) to learn the global layout and content of input images, alongside a Mask-Guided Refinement Network (MGRNet) that focuses on details. The coarsed mask predicted by VFMNet is used as a guidance for MGRNet to refine high-resolution image features. During inference, we employ a two-stage coarse-to-fine approach to effectively combine high-resolution and low-resolution predictions. The main contributions of this work are summarized as follows:

- We design a simple yet effective network to utilize VFMs for sematntic segmentation while preserving their rich generalizable knowledge.
- We introduce a multi-scale training framework along with a coarse-to-fine sliding window inference method to enable high-resolution inference.
- Extensive experiments are conducted on various benchmarks and backbones to validate the effectiveness of our method and demonstrate the superior performance of our approach over the state-of-the-art methods.

Related Works

Domain Generalized Semantic Segmentation

Domain Generalization aims to train a generalizable model on labeled source domains such that it performs well across multiple domains. In Domain Generalization Semantic Segmentation (DGSS), existing methods can be divided into two categories: (1) domain-invariant feature learning forces the model to learn domain-agnostic features. Some approaches (Pan et al. 2018; Peng et al. 2022; Choi et al. 2021) leverage Instance Normalization (IN) or Instance Whitening (IW) to standardize global features. Other approaches (Huang et al. 2023; Yang, Gu, and Sun 2023) project images into a feature space to reduce style variations. This method effectively removes domain-specific statistics, but it is only implemented on simple backbones like ResNet, leaving its effectiveness on other transformer-based backbones unclear. (2) Data augmentation has proven to be a simple and effective technique in DGSS. (Peng et al. 2021; Zhong et al. 2022) randomizes the style or texture of images, increasing the diversity of training data. Other approaches (Jia et al. 2024; Benigmim et al. 2024) leverage generative models like Stable Diffusion (Rombach et al. 2022) to synthesize new images. However, this method significantly increases the training time, and the model's performance is unstable as it depends on the quantity and quality of the generated data.

Visual Foundation Models

Visual Foundation Models (VFMs) are base models trained on large-scale image data in a self-supervised or semi-supervised manner (Bommasani et al. 2021). By being pretrained on millions of images, they acquire general knowledge, allowing them to easily adapt to various downstream visual tasks (Zhang et al. 2023; Wang et al. 2023a; Hümmer et al. 2023). CLIP (Radford et al. 2021) is a vision-language model that learns high-quality visual representations through contrastive learning with large-scale image-text pairs. SAM (Kirillov et al. 2023) is an interactive image segmentation model trained on a large-scale segmentation dataset, demonstrating zero-shot segmentation ability even for unseen objects. DINOv2 (Oquab et al. 2023) is pretrained on carefully curated datasets with self-supervised learning methods, achieving general visual representations without task-specific annotations.

Due to the superior performance of VFMs, many recent works have utilized them for generalized segmentation. Rein (Wei et al. 2024) proposed an effective fine-tuning method that maintains the generalization ability of VFMs. CLOUDS (Benigmim et al. 2024) designs a framework that combines multiple VFMs to leverage their strengths. (Pak et al. 2024) design a textual query-driven transformer that leverages domain-invariant semantic knowledge from text to enhance generalization. However, most previous methods ignore the challenges faced by VFM when processing high-resolution images. In this paper, we address this issue by developing a novel framework.

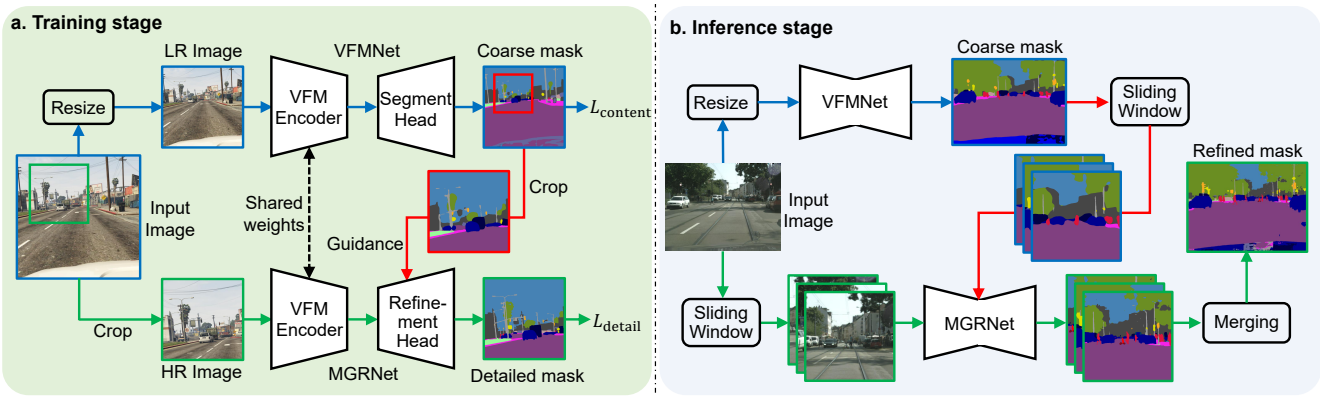

Figure 2: The overall framework of our method consisting of VFM-based Segmentation Network (VFMNet) and Masked-guided Refinement Network (MGRNet). (a) VFMNet and MGRNet are trained on resized and cropped images, respectively. Both networks share the same encoder from the VFM but use different decoders. To introduce contextual information, the coarse mask from VFMNet is used as a class prior for MGRNet. (b) During inference, a two-stage coarse-to-fine inference method combines high-resolution features with low-resolution predictions for fine-grained results.

Method

Problem Definition

In this work, we focus on single-source domain generalization for semantic segmentation. We define the source domain S as a set of image-label pairs $S = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^{H\times W\times 3}$ denotes an RGB image, and $y_i \in \{0,1\}^{H\times W\times N}$ is the one-hot encoded class label for each pixel, where N is the number of classes.

Our goal is to train a generalizable semantic segmentation model on a single source domain S such that it performs well on various unseen domians $T = \{T_1, T_2, ...T_m\}$. We first present the overall framework and our main ideas, and discuss details such as model architectures in later sections.

Framework Overview

We aim to adapt VFMs to this tasks while enabling high-resolution inference. To achieve this, we train two networks: one to capture global context and the other to focus on local details. Their predictions are then combined during inference, as shown in Fig. 2. Specifically, we design the VFM-based segmentation network (VFMNet) and the Masked-guided Refinement Network (MGRNet), denoted as f_θ and g_ϕ, respectively.

VFMNet is trained on low-resolution images, focusing on learning the global layout and content of the image. Conversely, MGRNet is trained on high-resolution image crops to capture fine-grained details. To supplement the missing context, we input the coarse segmentation mask of VFMNet to MGRNet as the class prior. This allows MGRNet to effectively refine the image features guided by the mask.

We share the weights of the VFM's encoder between the two networks. This not only reduces the number of trainable parameters but also facilitates learning of both global and local features.

Training. We consider semantic segmentation as a pixel-wise classification task and employ the cross-entropy loss to train two networks, as shown in the following equations:

$$\mathcal{L}_{seg}(y,\hat{y}) = -\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{N} y_{i,j,c}\log \hat{y}_{i,j,c} \tag{1}$$

We denote the loss of VFMNet as content loss $\mathcal{L}_c$ and the loss of MGRNet as detail loss $\mathcal{L}_d$. The total loss is a weighted sum of individual network losses:

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_d \tag{2}$$

where λ is a hyperparameter that balances the contribution of the detail loss relative to the content loss.

Inference. Since VFMNet excels at capturing global context, while MGRNet focuses on local details, we propose a two-stage inference strategy to leverage both strengths, as illustrated in Fig. 2.

Given an input image $x_i \in \mathbb{R}^{H\times W\times 3}$, the first stage involves generating the coarse prediction $\hat{y}$ using VFMNet f_θ on the resized low-resolution image, defined as:

$$\hat{y} = f_\theta(\text{resize}(x_i)) \tag{3}$$

In the second stage, we first divide both the input image x_i and the coarse prediction $\hat{y}$ into overlapped patches $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_k\}$, $\hat{y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_k\}$, using a sliding window approach. Then MGRNet refines the coarse prediction $\hat{y_k}$ for each patch $\mathcal{P}_k$. However, not all patches need to be refined. We only process the patches with low confidence, as these may contain fine-grained details not captured by VFMNet. The confidence of each patch is calculated based on the softmax probabilities of the segmentation logits.

$$p_k^{(i,j,c)} = \frac{\exp(\hat{y}_k^{(i,j,c)})}{\sum_{n=1}^{N}\exp(\hat{y}_k^{(i,j,n)})} \tag{4}$$

$$C_k = \frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w}\left[\max_c p_k^{(i,j,c)} > \theta\right] \tag{5}$$

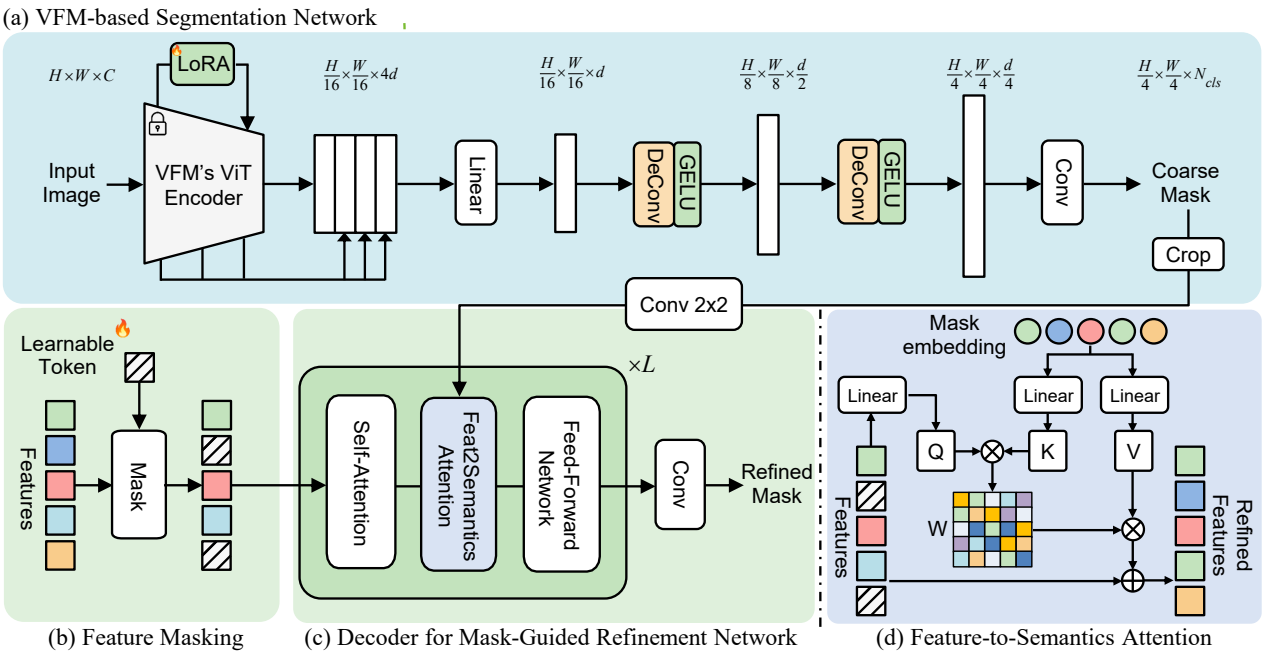

Figure 3: The architecture of the proposed VFMNet and MGRNet. (a) VFMNet fine-tunes the ViT encoder of VFM using LoRA and employs a deconvolution-based decoder. (b) To prevent MGRNet from over-relying on high-resolution features, some feature tokens are randomly masked with a learnable token. (c) The decoder of MGRNet uses the feature-to-semantics attention to combine high-resolution features with low-resolution semantic predictions. (d) Details of the feature-to-semantics attention, where image features serve as queries to retrieve semantic information from the coarse mask. Note that MGRNet shares the encoder with VFMNet and fuses multi-scale features using a linear layer, which we have omitted here for brevity.

where $p_k^{(i,j,c)}$ defines the softmax probabilities, $[\cdot]$ denotes the Iverson bracket, θ is a predefined threshold. C_k indicates the certainty, reflecting the proportion of high-confidence pixels within patch $\mathcal{P}_k$.

Then, the MGRNet g_ϕ refines the predictions $\hat{y}_k$ from the VFMNet based on a predefined confidence threshold C_τ:

$$\tilde{y}_k = \begin{cases} \hat{y}_k & \text{if } C_k > C_\tau \\ g_\phi(\mathcal{P}_k, \hat{y}_k) & \text{otherwise} \end{cases} \tag{6}$$

VFM-Based Segmentation Network (VFMNet)

To leverage the VFM for DGSS, we need to adapt the model for the segmentation task with minimal fine-tuning in order to preserve its pre-trained generalizable knowledge. Therefore, we design a simple and effective segmentation network based on encoder-decoder structure.

Encoder. We utilize the encoder of VFM and fine-tune it using the Parameter-Efficient Fine-Tuning (PEFT) method. This approach efficiently adapts the VFMs while preserving their generalizable knowledge. Here we employ Low-Rank Adaptation (LoRA) (Hu et al. 2022) for its effectiveness and negligible impact on inference speed.

Specifically, LoRA freezes the weight matrix $W \in \mathbb{R}^{d \times k}$ in the original model and introduces extra trainable low-rank matrices $A \in \mathbb{R}^{r \times k}$ and $B \in \mathbb{R}^{d \times r}$ to update the weight. The updated weight W' is given by:

$$W' = W + \frac{\alpha}{r} BA \tag{7}$$

where α is a scaling factor that controls the contribution of the original weight and the LoRA weight.

Decoder. Most VFMs produce single-scale and low-resolution features, making them less effective for semantic segmentation. To address this problem, we propose a simple and effective decoder, as shown in Fig. 3 (a). We begin by extracting multi-scale features F_i from different depths of the VFM backbone (at depths of 1/4, 1/2, 3/4, and full depth). Each feature map, of size $h \times w \times c$, is then concatenated along the channel dimension and passed through a linear layer to produce the fused features. Next, two transposed convolutions, each with a kernel size of 2 and a stride of 2, are applied to upsample the feature maps. Finally, a 1x1 convolution layer generates the segmentation mask $\hat{y}$.

Mask-Guided Refinement Network (MGRNet)

We train the MGRNet on high-resolution image crops to capture fine-grained details. However, the lack of context hinders its ability to recognize large objects. To address this, we feed the low-resolution coarse mask into MGRNet as class priors. The MGRNet then generates refined output based on high-resolution image features and the low-resolution coarse mask.

We first obtain the coarse mask $\hat{y}$ by cropping the low-resolution prediction from the VFMNet. To match the size of the image features, we process the mask through two 2x2 convolutional layers with a stride of 2 to obtain the mask

Backbone	Method	Trained on GTAV				Trained on Citys		
		Citys	BDD	Map	Avg.	BDD	Map	Avg.
ResNet101	SAN-SAW (Peng et al. 2022)	45.33	41.18	40.77	42.43	54.73	61.27	58.00
	WildNet (Lee et al. 2022)	45.79	41.73	47.08	44.87	50.94	58.79	54.87
	SHADE (Zhao et al. 2022)	46.66	43.66	45.50	45.27	50.95	60.67	55.81
	TLDR (Kim, Kim, and Kim 2023)	47.58	44.88	48.80	47.09	–	–	–
	FAMix (Fahes et al. 2024)	49.47	46.40	51.97	49.28	–	–	–
Swin-L	HGFormer (Ding et al. 2023)	–	–	–	–	61.50	72.10	66.80
	CMFormer (Bi, You, and Gevers 2024)	55.31	49.91	60.09	55.10	62.60	73.60	68.10
CLIP-L	VLTSeg (Hümmer et al. 2023)	55.60	52.70	59.60	55.97	–	–	–
	Rein (Wei et al. 2024)	57.10	54.70	60.50	57.43	–	–	–
	Ours	**62.31**	**56.09**	**66.47**	**61.62**	60.62	73.27	66.95
SAM-H	Rein (Wei et al. 2024)	59.60	52.00	62.10	57.90	–	–	–
	Ours	**64.05**	**55.59**	**67.71**	**62.45**	59.79	70.83	65.31
EVA02-L	VLTSeg (Hümmer et al. 2023)	65.60	58.40	66.50	63.50	64.40	76.40	70.40
	Rein (Wei et al. 2024)	65.30	60.50	64.90	63.57	64.10	69.50	66.80
	Ours	**69.53**	**61.14**	**69.97**	**66.88**	**64.70**	**76.43**	**70.56**
DINOv2-L	Rein† (Wei et al. 2024)	69.20	60.65	70.16	66.67	65.00	76.09	70.54
	Ours	**73.87**	**62.91**	**73.52**	**70.10**	**66.16**	**77.08**	**71.62**

Table 1: Performance comparison with existing domain generalized methods. The results are shown in mIoU, with the best ones highlighted. '-' indicates no results reported in the paper or no official code available to reproduce the results. † indicates that we reproduced the results using the official pre-trained checkpoints on the same testing size.

embedding Y. High-resolution features F are extracted using a LoRA-based VFM encoder and fused with a linear layer. We then apply feature-to-semantics attention to integrate both features, as shown in Fig. 3 (d). This process can be described as follows:

$$Q = \text{Linear}(F), K = \text{Linear}(Y), V = \text{Linear}(Y) \quad (8)$$

$$W = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right) \quad (9)$$

Here, three different linear layers are used to project F and Y into the same dimension. The matrix W reflects the similarity between the image features and the coarse mask. We then fuse the high-resolution features F and the low-resolution mask Y as follows:

$$F' = F + W \cdot V \quad (10)$$

The feature-to-semantics attention allows the model to extract valuable information from the low-resolution coarse mask, thereby refining the high-resolution features.

Feature Masking

During the training process, we observed that the MGRNet might exhibit an overreliance on fine-grained high-resolution features, potentially neglecting the coarse mask. To mitigate this bias and enhance performance, we implemented a feature masking strategy. Specifically, we randomly mask a portion of feature tokens and replacing them with a learnable token.

Let $F \in \mathbb{R}^{H \times W \times D}$ denote the high-resolution features and $T \in \mathbb{R}^{1 \times D}$ denote the learnable masking token. We generate a random binary mask $M \in \{0,1\}^{H \times W}$ as follows:

$$M = [v > p] \quad \text{where } v \sim U(0,1) \quad (11)$$

Here, $[\cdot]$ denotes the Iverson bracket, and $U(0,1)$ is the uniform distribution between 0 and 1, p is the masking ratio. The masked feature map $\hat{F}$ is then computed by:

$$\hat{F} = F \odot M + T \odot (1 - M) \quad (12)$$

Experiments

Experimental Setups

Datasets. Following previous studies (Wei et al. 2024), we evaluate our method on both synthetic and real-world datasets. The synthetic dataset is GTAV (Richter et al. 2016), which contains 24,966 street-view images rendered by a computer game engine with the resolution of 1914x1052. For real-world datasets, we use Cityscapes (Cordts et al. 2016), a large-scale semantic segmentation dataset for autonomous driving, with 2,975 training images and 500 validation images, all with a resolution of 2048×1024. BDD100K (Yu et al. 2020) is another realworld dataset that contains diverse urban driving scene images with the resolution of 1280×720. The last real-world dataset we use is Mapillary (Neuhold et al. 2017), which consists of high-resolution images with a minimum resolution of 1920×1080 collected from around the world. BDD100K and Mapillary provide 1000 and 2000 validation images, respectively. For brevity, we refer to Cityscapes, BDD100K, and Mapillary as Citys, BDD, and Map, respectively

Visual Foundation Models. We conduct experiments using CLIP (Radford et al. 2021), EVA02 (Fang et al. 2024), SAM (Kirillov et al. 2023), and DINOv2 (Oquab et al. 2023) as backbones to evaluate the effectiveness of our method. Following previous approaches (Wei et al. 2024), we employ

Config	Citys	BDD	Map	Avg.
Baseline	59.80	54.83	61.57	58.73
+ VFMNet	72.16	**62.98**	71.88	69.00
+ MGRNet	72.94	62.56	72.56	69.35
+ Feat.Mask	**73.87**	62.91	**73.52**	**70.10**

Table 2: Ablation study of different components. Components are sequentially incorporated to show their impact. The baseline model uses a frozen DINOv2 backbone with a linear decoder and is trained on GTAV.

the ViT-L architecture for all models except SAM, which uses the ViT-H architecture.

Implementation Details. Our implementation is based on the MMSegmentation framework. We use the AdamW optimizer with learning rates of 1e-5 for the backbone and 1e-4 for all decode heads. Training is conducted for 40,000 iterations with a batch size of 2 and crop size of 512x512. We employ basic data augmentation techniques including random cropping, random horizontal flipping, photo-metric transformation and rare class sampling (Hoyer, Dai, and Van Gool 2022). During training, we set $\lambda = 1.0$, $r = \alpha = 32$, and $p = 0.2$. During inference, we use a sliding window approach with a window size of 512x512 and a stride of 320. The θ and C_τ are set to 0.968 and 0.8 respectively.

Comparison with State-of-the-Art Methods

Compared Methods. We compare our method with several DGSS methods: SAN-SAW (Peng et al. 2022), WildNet (Lee et al. 2022), SHADE (Zhao et al. 2022), TLDR (Kim, Kim, and Kim 2023), FAMix (Fahes et al. 2024), HGFormer (Ding et al. 2023), CMFormer (Bi, You, and Gevers 2024), VLTSeg (Hümmer et al. 2023), and Rein (Wei et al. 2024).

Main Results. We compare our method with existing methods in two generalization settings: GTA $\rightarrow$ Citys + BDD + Map, and Citys $\rightarrow$ BDD + Map, as shown in Table 1. Our method with DINOv2 backbone outperforms existing approaches in both synthetic-to-real (trained on GTAV) and real-to-real (trained on Citys) generalization, achieving average mIoU improvements of 3.4% and 1.1% over the state-of-the-art, respectively. This demonstrates the effectiveness of our approach. We find that our method shows a more significant improvement in synthetic-to-real generalization compared to real-to-real generalization. We hypothesize that this is due to the larger domain gap between synthetic and real data, which requires more prior knowledge for effective generalization. It also shows that our approach can fully leverage the capabilities of VFM to bridge this gap effectively.

Comparison with Various VFM Backbones. As shown in Table 1, our method can effectively integrate with different VFMs and consistently outperforms Rein, demonstrating its effectiveness. Notably, using DINOv2 and Eva 02 as the backbone yields the better results, highlighting its strong generalization capabilities. Howerev, the results of CLIP and SAM are relatively poor. This may be due to the fact that

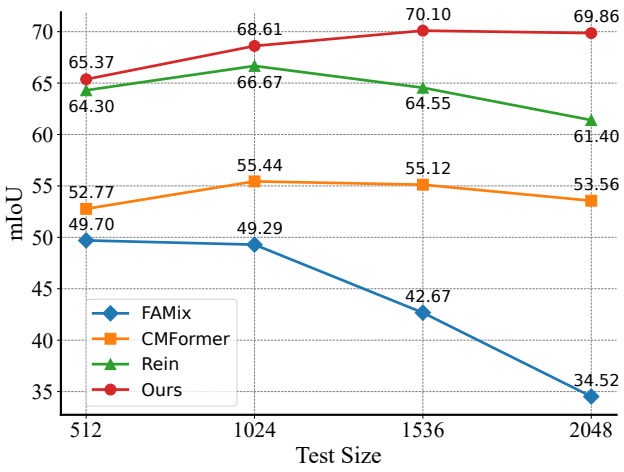

Figure 4: The average mIoU comparison under GTAV $\rightarrow$ Citys + BDD + Map generalization at different test resolutions . The short edge of the input image is scaled to 512, 1024, 1536, 2048, respectively.

Decoder	Citys	BDD	Map	Avg.
Linear	71.17	61.75	70.89	67.94
SegFormer	70.72	62.24	71.16	68.04
Mask2Former	71.22	60.12	71.59	67.64
Ours	**72.16**	**62.98**	**71.88**	**69.00**

Table 3: Ablation study of different decoders for VFMNet. Models are trained on GTAV with the DINOv2 backbone.

CLIP tends to extract features with richer semantics, which leads to the neglect of details. On the other hand, SAM itself is a backbone designed specifically for segmentation, which may focus too much on details, potentially hindering the generalization ability of the model.

Inference Under Different Resolutions. We conducted inference at various resolutions by scaling the short edge of images to 512, 1024, 1536, and 2048, comparing the performance of different methods. As shown in Fig 4, our method consistently achieved the best performance across all resolutions. For most methods, increasing resolution initially improves performance, but later leads to degradation. This partly due to a mismatch between training and inference resolutions, preventing effective generalization to higher resolutions. In contrast, the performance of our method improves with resolution increases, with the highest resolution achieving a 4.7% improvement compared to the lowest resolution. This demonstrates the effectiveness of our model in adapting to increasing image resolutions.

Ablation Studies

Analysis of the Key Components. The proposed method integrates three key components: VFMNet, MGRNet, and a feature masking strategy. In our ablation study, our baseline is a frozen DINOv2 encoder with a linear head. We

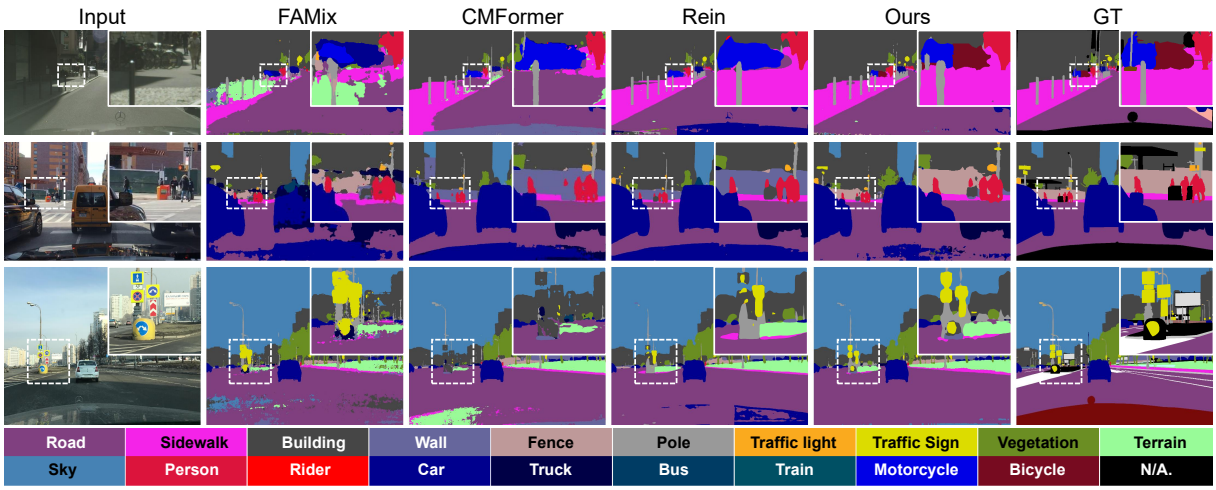

Figure 5: Qualitative Comparison under GTAV → Citys + BDD + Map generalization setting.

then incorporated different components sequentially to show their impact. The results are shown in Table 2. First, the frozen DINOv2 only achieves 58.7% mIoU, highlighting the limitations of using pre-trained models without adaptation. Then, it is observed that our VFMNet allows effective adaptation for semantic segmentation tasks, improving performance by 10.3% compared to the baseline. Furthermore, MGRNet boosts the performance of VFMs by 0.35%, demonstrating the effectiveness of combining high-resolution features with low-resolution predictions. Additionally, the introduction of the feature masking strategy led to a further performance increase of 0.75%, highlighting its critical role.

Comparison of Different Decoders in VFMNet. We evaluated several decoder designs for VFMNet, with the results shown in Table. 3. Notably, even a simple linear head achieved 67.9% mIoU after fine-tuning the backbone with LoRA. Interestingly, more complex decoder heads do not necessarily improve performance, e.g., Mask2Former yielded similar results to the linear head. Our decoder uses deconvolution to upsample image features, creating high-resolution features while preserving the pre-trained knowledge of VFMs. These experimental results validate the effectiveness of our decoder design.

Sensitivity to Hyper-parameters

Impact of Mask Ratio and Detail Loss. As shown in Table. 4, we investigated the impact of mask ratio and detail loss weights on performance. The optimal result was achieved with a detail loss of 1.0, indicating that detail loss and content loss are equally important. For the mask ratio, a value of 0.2 yields the best performance. Higher mask ratios may reduce information from high-resolution images, causing the model to rely too much on low-resolution predictions, while lower mask ratios can excessively focus on high-resolution features, neglecting class priors from low-resolution predictions. This balance highlights the need to integrate information from both high and low resolutions for optimal performance.

(a) Choice of mask ratio p

p	0	0.1	**0.2**	0.3	0.4
Avg. mIoU	69.35	69.56	**70.10**	69.87	69.33

(b) Choice of detail loss weight λ

λ	0.2	0.5	**1.0**	1.5	2.0
Avg. mIoU	69.27	69.61	**70.10**	69.63	69.48

Table 4: Ablation study on mask ratio and detail loss weight.

Quantitative Results

As shown in Fig 5, we present a visual comparison of the segmentation results on the GTAV → {Citys + BDD + Map} generalization setting. It can be observed that our approach exhibits a strong ability to identify small objects in the distance. This clearly demonstrates that our method can fully leverage high resolution, as well as the importance of high-resolution images for accurate semantic segmentation.

Conclusion

In this work, we explored the benefits of leveraging VFMs for generalizable semantic segmentation. We first proposed a simple yet effective network, VFMNet, to adapt VFMs to this task while preserving their generalizable knowledge. Then, we designed MGRNet to capture details guided by the prediction of VFMNet. Additionally, we proposed a two-stage inference method to enhance inference on high-resolution images. Extensive experiments demonstrate the superiority of our method over the state-of-the-art on various benchmarks. In the future, we will apply this method to other non-foundation models and further explore more efficient ways to perform inference on high-resolution images.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under grants No. 62232008 and No. 62032010, National Natural Science Foundation of China (Youth Program) under Grant No. 62402367, Fundamental Research Funds for the Central Universities under Grant No.20101247556.

References

Benigmim, Y.; Roy, S.; Essid, S.; Kalogeiton, V.; and Lathuilière, S. 2024. Collaborating Foundation models for Domain Generalized Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3108–3119.

Bi, Q.; You, S.; and Gevers, T. 2024. Learning content-enhanced mask transformer for domain generalized urban-scene segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 819–827.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 834–848.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1290–1299.

Choi, S.; Jung, S.; Yun, H.; Kim, J. T.; Kim, S.; and Choo, J. 2021. 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*, 11580–11590.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3213–3223.

Ding, J.; Xue, N.; Xia, G.-S.; Schiele, B.; and Dai, D. 2023. Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15413–15423.

Fahes, M.; Vu, T.-H.; Bursuc, A.; Pérez, P.; and de Charette, R. 2024. A Simple Recipe for Language-guided Domain Generalized Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23428–23437.

Fan, Q.; Segu, M.; Tai, Y.-W.; Yu, F.; Tang, C.-K.; Schiele, B.; and Dai, D. 2023. Towards robust object detection invariant to real-world domain shifts. In *International Conference on Learning Representations*.

Fang, Y.; Sun, Q.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2024. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 105171.

Hoyer, L.; Dai, D.; and Van Gool, L. 2022. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9924–9935.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, W.; Chen, C.; Li, Y.; Li, J.; Li, C.; Song, F.; Yan, Y.; and Xiong, Z. 2023. Style projected clustering for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3061–3071.

Hümmer, C.; Schwonberg, M.; Zhong, L.; Cao, H.; Knoll, A.; and Gottschalk, H. 2023. VLTSeg: Simple transfer of CLIP-based vision-language representations for domain generalized semantic segmentation. *arXiv preprint arXiv:2312.02021*.

Jia, Y.; Hoyer, L.; Huang, S.; Wang, T.; Gool, L. V.; Schindler, K.; and Obukhov, A. 2024. DGInStyle: Domain-Generalizable Semantic Segmentation with Image Diffusion Models and Stylized Semantic Control. In *Proceedings of the European Conference on Computer Vision*, 91–109.

Kim, S.; Kim, D.-h.; and Kim, H. 2023. Texture learning domain randomization for domain generalized segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 677–687.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Lee, S.; Seong, H.; Lee, S.; and Kim, E. 2022. Wildnet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9936–9946.

Neuhold, G.; Ollmann, T.; Rota Bulo, S.; and Kontschieder, P. 2017. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, 4990–4999.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Pak, B.; Woo, B.; Kim, S.; hwan Kim, D.; and Kim, H. 2024. Textual Query-Driven Mask Transformer for Domain Generalized Segmentation. In *Proceedings of the European Conference on Computer Vision*, 37–54.

Pan, X.; Luo, P.; Shi, J.; and Tang, X. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision*, 484–500.

Peng, D.; Lei, Y.; Hayat, M.; Guo, Y.; and Li, W. 2022. Semantic-aware domain generalized segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2594–2605.

Peng, D.; Lei, Y.; Liu, L.; Zhang, P.; and Liu, J. 2021. Global and local texture randomization for synthetic-to-real semantic segmentation. *IEEE Transactions on Image Processing*, 30: 6594–6608.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *Proceedings of the European Conference on Computer Vision*, 102–118.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Song, Y.; Zhou, Q.; Li, X.; Fan, D.-P.; Lu, X.; and Ma, L. 2024. BA-SAM: Scalable Bias-Mode Attention Mask for Segment Anything Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3162–3173.

Wang, D.; Zhang, J.; Du, B.; Xu, M.; Liu, L.; Tao, D.; and Zhang, L. 2023a. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. In *Advances in Neural Information Processing Systems*, 8815–8827.

Wang, P.; Zhang, Z.; Lei, Z.; and Zhang, L. 2023b. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3769–3778.

Wei, Z.; Chen, L.; Jin, Y.; Ma, X.; Liu, T.; Ling, P.; Wang, B.; Chen, H.; and Zheng, J. 2024. 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*, 28619–28630.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, 12077–12090.

Xu, Q.; Yao, L.; Jiang, Z.; Jiang, G.; Chu, W.; Han, W.; Zhang, W.; Wang, C.; and Tai, Y. 2022. Dirl: Domain-invariant representation learning for generalizable semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2884–2892.

Yang, L.; Gu, X.; and Sun, J. 2023. Generalized semantic segmentation by self-supervised source domain projection and multi-level contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10789–10797.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2636–2645.

Zhang, J.; Herrmann, C.; Hur, J.; Polania Cabrera, L.; Jampani, V.; Sun, D.; and Yang, M.-H. 2023. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. In *Advances in Neural Information Processing Systems*, 45533–45547.

Zhao, Y.; Zhong, Z.; Zhao, N.; Sebe, N.; and Lee, G. H. 2022. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 837–853.

Zhong, Z.; Zhao, Y.; Lee, G. H.; and Sebe, N. 2022. Adversarial style augmentation for domain generalized urban-scene segmentation. In *Advances in Neural Information Processing Systems*, 338–350.