Exploring Semantic Consistency and Style Diversity for Domain Generalized Semantic Segmentation

Hongwei Niu[1,2*], **Linhuang Xie**[1*], **Jianghang Lin**[1*], **Shengchuan Zhang**[1†]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China
[2]Institute of Artificial Intelligence, Xiamen University, Fujian, China
{niuhw649, hunterjlin007}@stu.xmu.edu.cn
linhuangxie@gmail.com, zsc_2016@xmu.edu.cn

Abstract

Domain Generalized Semantic Segmentation (DGSS) seeks to utilize source domain data exclusively to enhance the generalization of semantic segmentation across unknown target domains. Prevailing studies predominantly concentrate on feature normalization and domain randomization, these approaches exhibit significant limitations. Feature normalization-based methods tend to confuse semantic features in the process of constraining the feature space distribution, resulting in classification misjudgment. Domain randomization-based methods frequently incorporate domain-irrelevant noise due to the uncontrollability of style transformations, resulting in segmentation ambiguity. To address these challenges, we introduce a novel framework, named SCSD for Semantic Consistency prediction and Style Diversity generalization. It comprises three pivotal components: Firstly, a Semantic Query Booster is designed to enhance the semantic awareness and discrimination capabilities of object queries in the mask decoder, enabling cross-domain semantic consistency prediction. Secondly, we develop a Text-Driven Style Transform module that utilizes domain difference text embeddings to controllably guide the style transformation of image features, thereby increasing inter-domain style diversity. Lastly, to prevent the collapse of similar domain feature spaces, we introduce a Style Synergy Optimization mechanism that fortifies the separation of inter-domain features and the aggregation of intra-domain features by synergistically weighting style contrastive loss and style aggregation loss. Extensive experiments demonstrate that the proposed SCSD significantly outperforms existing state-of-the-art methods. Notably, SCSD trained on GTAV achieved an average of 49.11 mIoU on the four unseen domain datasets, surpassing the state-of-the-art method by +4.08 mIoU.

Introduction

Semantic segmentation, a core task in computer vision, involves assigning a semantic class label to each pixel in an image. Traditionally, models (Strudel et al. 2021; Li et al. 2022; Mi et al. 2022; Hu et al. 2023; Yue et al. 2024) for object detection and image segmentation are trained and evaluated under the assumption that datasets are independent and identically distributed. However, this assumption often fails to hold in real-world scenarios due to variations in lighting, weather conditions, and geographical differences. This challenge has spurred significant research in Domain Adaptive Semantic Segmentation (DASS) (Hoyer, Dai, and Van Gool 2022; Xia et al. 2023; Cheng et al. 2023; Choe et al. 2024), which seeks to minimize the distribution discrepancies between source and target domains by aligning their data distributions. Unlike DASS, which requires access to target domain data during training, an often impractical requirement, Domain Generalized Semantic Segmentation (DGSS) offers a promising alternative. DGSS focuses on learning only from source domain data to enhance the model's ability to generalize to unknown target domains, thus better meeting the challenges posed by real-world applications.

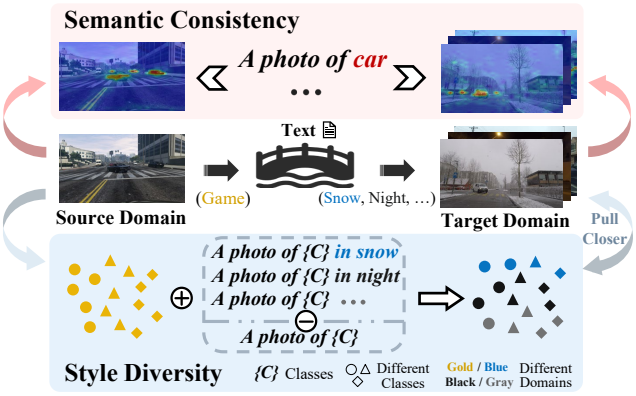


Figure 1: Illustration of our motivations. **Semantic Consistency**: Similarity map between domain-irrelevant general text embeddings and image features in different domains. Cross-domain consistent prediction can be achieved through general text prompts. **Style Diversity**: Simplified version of t-SNE visualization of image features. Domain text difference embeddings are used as style difference features guide the style transformation of image features from source domain to target domain.

*These authors contributed equally.
†Corresponding Author.

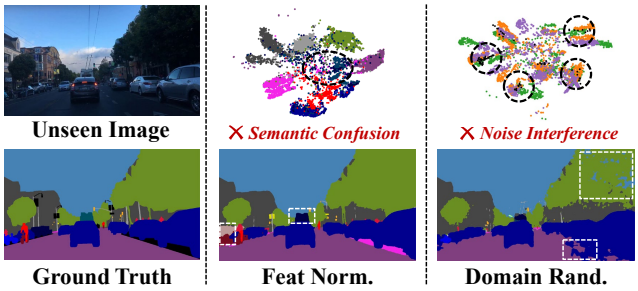


Figure 2: Column 1: Image and ground truth. Columns 2-3: Feature normalization-based and domain randomization-based methods. For both methods, Row 1 shows the t-SNE visualization of feature maps, and Row 2 shows the segmentation result. Semantic Confusion: different colors represent different categories. Noise Interference: different colors represent different domains. **Better view in zoom.**

Existing DGSS methods are generally divided into two categories: feature normalization and domain randomization. Feature normalization-based methods include normalization (Pan et al. 2018; Peng et al. 2022; Huang et al. 2023) and whitening (Pan et al. 2019; Huang et al. 2019; Choi et al. 2021), aim to constrain the distribution of different features to the same space or remove the domain-specific style features, thereby facilitating the learning of domain-invariant semantic content representations. However, as illustrated in the second column of Fig. 2, this often confuse intra-domain semantic features of different categories, resulting in classification misjudgment. Moreover, due to the non-orthogonality of content and style, removing style can also lead to the loss of semantic content. Domain Randomization-based methods (Lee et al. 2022; Chattopadhyay et al. 2023; Fahes et al. 2024) seek to transform the source domain style into multiple other domains to increase style diversity. However, these methods rely on artificially created auxiliary domains (*e.g.*, ImageNet). As depicted in the third column of Fig. 2, this can introduce domain-irrelevant noise, potentially compromising the domain-invariant representations and resulting in segmentation ambiguities.

The emergence of Visual-Language Models (VLMs), such as CLIP (Radford et al. 2021), represents a significant development in multimodal learning. These models map images and text into a unified representation space, achieving modality alignment by comparing the similarities and differences between various image-text pairs. Such capabilities offer new perspectives for the Domain Generalized Semantic Segmentation (DGSS) field. By harnessing the inherent alignment properties of VLMs, the text modality can effectively act as a conduit to bridge the domain gap between source and target domains. Specifically, we explore two properties of the text modality in VLMs for enhancing DGSS: **1) Semantic Consistency**: Despite the significant variability in data distribution across domains, semantic categories remain consistent. For example, "a car in game" and "a car in snow" are classified under the same semantic category. As illustrated in the pink section of Fig. 1, the alignment capabilities of VLMs between image and text modalities enable a general text prompt (*e.g.*, "a photo of car") to consistently correspond with the target object (car) across various domains (*e.g.*, game, snow, night, etc.). Thus, the text modality can achieve semantically consistent predictions across domains while preserving semantic content. **2) Style Diversity**: By aligning images and texts within a unified representation space, VLMs also learn the differences between various image-text pairs. As shown in the blue section of Fig. 1, a simplified version of t-SNE visualization (see Experiments for more details) of the original image features and the image features weighted by text difference embeddings demonstrates that differences in text prompt embeddings across domains can be leveraged to diversify the style of image features.

Based on these observations, we introduced a novel framework named **SCSD**, which is designed to explore **S**emantic **C**onsistency and **S**tyle **D**iversity for DGSS. It comprises three carefully designed innovative components that fully exploit the potential of the text modality. Firstly, we propose Semantic Query Booster (SQB), which leverages semantic consistency between image and text modalities to establish cross-modal semantic associations and aggregate relevant semantic features. By enhancing the semantic discernment of object queries within the mask decoder, SQB facilitates robust predictions of semantic consistency across different domains. Secondly, we introduce a Text-Driven Style Transformation (TDST) module that mines the style diversity of the text modality. By utilizing the difference between text embedding vectors from specific domain prompts and general domain prompts as domain difference embeddings and mapping them to be style difference features, this module can controllably guide the transformation of the low-frequency amplitude spectrum of image features, thereby achieving cross-domain style transformation and enhancing inter-domain style diversity. Lastly, to prevent the collapse of similar domain feature spaces, we introduce a Style Synergy Optimization mechanism that fortifies the separation of inter-domain features and the aggregation of intra-domain features by synergistically weighting style contrastive loss and style aggregation loss.

We conduct comprehensive experiments in both single-source and multi-source settings to demonstrate that SCSD exhibits superior generalization compared to existing DGSS methods. Notably, SCSD outperforms the state-of-the-art methods by up to +4.87 and +2.92 mIoU on unseen Mapillary and ACDC domains, respectively.

Related Work

Domain Generalized Semantic Segmentation

To address the challenges of domain shift and the absence of target domain data, several domain generalization (DG) methods (Shu et al. 2021; Kang et al. 2022; Chen et al. 2023; Dayal et al. 2024) have been extensively studied. They aim to train the model using data from a single or multiple related but different source domains so that it can be generalized to any out-of-distribution target domain.

Domain Generalization Semantic Segmentation (DGSS) extends the domain generalization to a more challenging fine-grained segmentation task. Existing methods primarily focus on normalization (Pan et al. 2018; Peng et al. 2022; Huang et al. 2023; Ahn et al. 2024), whitening (Pan et al. 2019; Huang et al. 2019; Choi et al. 2021), and domain randomization (Yue et al. 2019; Peng et al. 2021; Zhong et al. 2022; Wu et al. 2022; Lee et al. 2022; Kim, Kim, and Kim 2023; Chattopadhyay et al. 2023; Hoyer, Dai, and Van Gool 2023; Fahes et al. 2024; Niemeijer et al. 2024; Jia et al. 2025). For example, SPC-Net (Huang et al. 2023) introduce style projection and semantic clustering to achieve better representation. TLDR (Kim, Kim, and Kim 2023) learns texture and shape features to mitigate the domain gap problem. BlindNet (Ahn et al. 2024) decouples content and style through covariance alignment and semantic consistency contrastive learning. DIDEX (Niemeijer et al. 2024) employs a diffusion model to generate pseudo target domains with diverse text prompts. CMFormer (Bi, You, and Gevers 2024) introduces a content-enhanced mask attention mechanism and multi-resolution feature fusion strategy to improve the model's adaptability to style variations.

Vision-Language Models

Visual-Language Models (VLMs) such as CLIP (Radford et al. 2021) leverage web-scale image-text pairs to align visual and textual modalities through contrastive learning, thereby demonstrating strong zero-shot generalization capabilities (Qu, Wang, and Qi 2023; Gong et al. 2024; Zhang et al. 2024; Lin et al. 2024; Qu et al. 2024) and robustness to natural distribution shifts (Ming et al. 2022; Tu, Deng, and Gedeon 2024) in various downstream tasks. For example, FC-CLIP (Yu et al. 2024) leverages the powerful open-vocabulary classification capabilities of the frozen CNN-based CLIP, combined with a mask generator, to significantly enhance the performance of open-vocabulary panoptic segmentation. CLIP-RC (Zhang et al. 2024) explores the use of regional cues to translate image-level knowledge into pixel-level understanding, enabling zero-shot semantic segmentation. Our work further explores the potential of text modality for DGSS.

Method

Overview

The overall framework of our proposed SCSD is depicted in Fig. 3, consisting of three key components: Semantic Query Booster (SQB), Text-Driven Style Transform (TDST), and Style Synergy Optimization (SSO). Initially, for a given input image I, multi-scale image features $F_i \in \mathbb{R}^{D\times\frac{H}{2^i}\times\frac{W}{2^i}}$ for $i \in \{2,3,4,5\}$ are extracted using the CLIP image encoder. Prompt text embeddings $E_t \in \mathbb{R}^{C\times D}$, corresponding to C categories from the training dataset, are derived from the CLIP text encoder. Then, SQB leverages the semantic consistency between image and text modalities to boost object queries Q into semantic queries $\hat{Q}$. TDST mines text style diversity and achieves latter three scale image features style transformation through style difference features obtained by domain difference embeddings mapping. Finally, in the mask decoder, the semantic queries undergo multiple interactions with the multi-scale image features and continuous refinement, resulting in class and mask predictions. The overall loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{seg} + \mathcal{L}_{s}, \tag{1}$$

where $\mathcal{L}_{cls}$ is the cross-entropy loss and $\mathcal{L}_{seg}$ includes the binary cross-entropy loss and dice loss. $\mathcal{L}_s$ includes style contrastive loss $\mathcal{L}_{sc}$ and style aggregation loss $\mathcal{L}_{sa}$ in SSO.

Semantic Query Booster

To achieve a consistent understanding of cross-domain semantic knowledge, we propose the SQB, as illustrated in the pink section of Fig. 3. It leverages domain-irrelevant general text embeddings to establish semantic associations with image features across any domains. This enables the object queries Q in the mask decoder to capture the concept of cross-domain semantic consistency and integrate the rich discriminative features learned from various domains, thereby enhancing these semantic awareness and discrimination capabilities of object queries Q. Specifically, the last layer of image features $F_5 \in \mathbb{R}^{D\times H'\times W'}$ is processed through attention pooling to obtain dense visual features $F_v \in \mathbb{R}^{H'\times W'\times D}$, where H' and W' represent $\frac{H}{32}$ and $\frac{W}{32}$, respectively. This allows for the aggregation of global context information while maintaining the feature scale, thereby providing fine-grained feature representations that more effectively support dense visual tasks. The semantic similarity map $S \in \mathbb{R}^{H'\times W'\times C}$ between the dense visual features F_v and the text embeddings $E_t \in \mathbb{R}^{C\times D}$ is computed as: $S = F_v \cdot E_t^T$. It encodes the correlation between each pixel and all categories, which enhance the cross-domain semantic awareness capabilities of the object queries Q, allowing for the correction of semantic bias within the model across different domains and achieving cross-domain semantic consistency prediction. Subsequently, the most relevant category index $G \in \mathbb{R}^{H'\times W'}$ for each pixel is determined by performing a maximum operation along the category dimension of the semantic similarity map S. The corresponding category embedding is retrieved from the text embeddings E_t using the category index G, and a semantic aggregation map $S_a \in \mathbb{R}^{H'\times W'\times D}$ tailored for pixel-level classification is constructed:

$$G = \arg\max_{C} S, \tag{2}$$

$$S_a = E_t[G], \tag{3}$$

where $E_t[G]$ denotes the selection of the indices in G along the first dimension of E_t. The semantic aggregation map S_a aggregates text semantic embeddings at the pixel level to model a fine-grained semantic space, thereby facilitating the object queries Q to capture rich semantic information and improving semantic discrimination capabilities. Both the semantic similarity map S and the semantic aggregation map S_a are mapped to the same channel dimensions as the object queries Q through a MLP respectively. Finally, a set of learnable object queries Q interact sequentially with the semantic similarity map S and the semantic aggregation map S_a

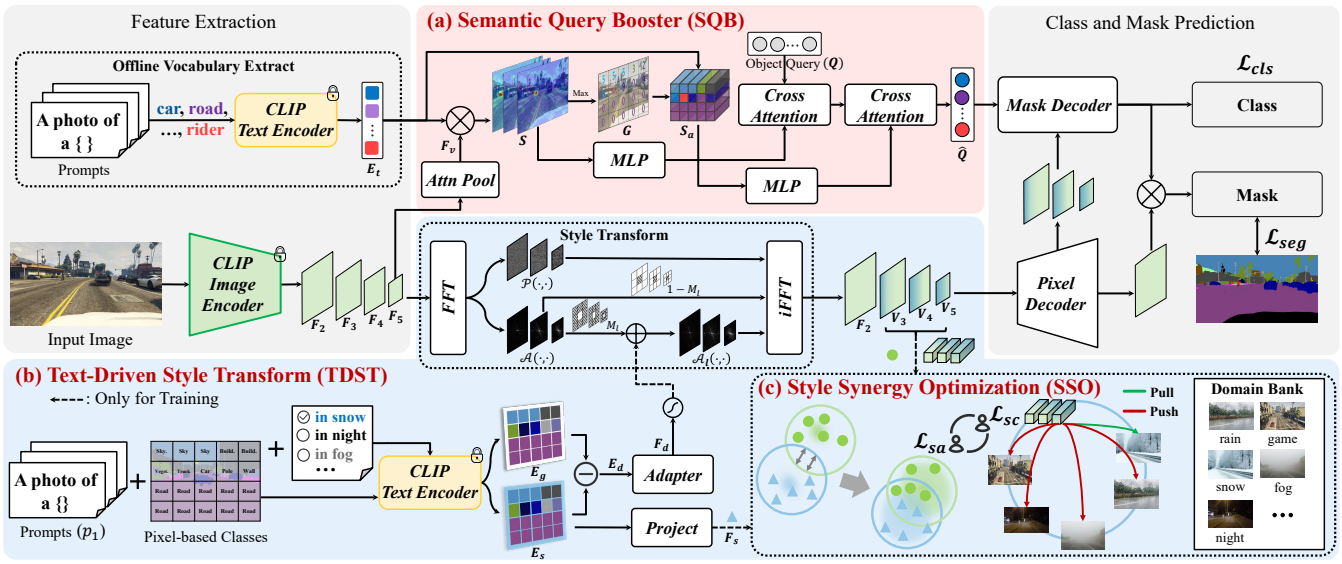


Figure 3: Overview of our proposed SCSD. The main contribution are: (a) Semantic Query Booster enhances object queries for cross-domain semantic consistency prediction. (b) Text-Driven Style Transform leverages the diversity of the text modality to facilitate the style transformation of image features. (c) Style Synergy Optimization fortifies the separation of inter-domain features and the aggregation of intra-domain features through the collaboration of two loss functions (*i.e.*, $\mathcal{L}_{sc}$ and $\mathcal{L}_{sa}$).

through cross-attention mechanisms to generate the semantic queries $\hat{Q} \in \mathbb{R}^{N \times D}$ that possess both semantic awareness and discrimination capabilities for cross-domain consistency prediction.

Text-Driven Style Transform

The visual-language models are able to map language and visual information to each other in a unified representation space due to its inherent multimodal alignment capabilities. Building on it, we leverage the difference features between domain text embeddings to construct style features tailored to each domain. These style features are used to guide the transformation of image feature styles, ensuring that only the style is altered without affecting the content.

Initially, we introduce a triplet domain prompts set $P = \{p_1, p_2, p_3\}$, which consists of the following:

- **General domain prompts** p_1, such as "a photo of {Class}","This is a photo of a {Class}", etc.
- **Conditional domain prompts** p_2, such as "ϕ", "in snow", "in night", "in fog", etc, where ϕ represents a null character used to maintain the style of the source domain.
- **Specific domain prompts** p_3, such as "a photo of {Class} {Domain}", "This is a photo of a {Class} {Domain}", etc. The {Domain} is derived from the conditional domain prompts.

To transfer the semantic knowledge of CLIP to pixel-level tasks, domain prompts based on each pixel category are constructed. Specifically, for each image in the a batch, the specific domain prompts p_3 for the current image are generated by combining the category of each pixel from the ground truth mask with a randomly selected conditional domain prompt p_2. The general domain prompts p_1 are constructed directly from the category of each pixel in the ground truth mask. These two sets of prompts are then separately fed into the text encoder, and the resulting embeddings are averaged along the prompt dimension to obtain the corresponding specific domain embeddings $E_s \in \mathbb{R}^{H \times W \times C}$ and general domain embeddings $E_g \in \mathbb{R}^{H \times W \times C}$. The domain difference embeddings $E_d \in \mathbb{R}^{H \times W \times C}$ are obtained through element-wise subtraction as follows:

$$E_d = E_s - E_g = \mathcal{E}(p_3) - \mathcal{E}(p_1), \tag{4}$$

where $\mathcal{E}(\cdot)$ denotes the CLIP text encoder. To transform the pixel-level text embeddings into the image feature space, we introduce a domain style adapter that maps the domain difference embeddings E_d to the style difference features $F_d \in \mathbb{R}^{H \times W \times C}$. It can be implemented using any unbiased adapter (Chen et al. 2024), or alternatively, with the simplest 1×1 convolution structure. However, directly integrating the style difference features F_d with image features may alter the semantic content of the features, leading to prediction bias. To address this issue, a simple and effective alternative, the Fourier Transform, is worth considering.

Previous research (Yang and Soatto 2020; Xu et al. 2021) has demonstrated the effectiveness of Fourier Transform-based data augmentation strategies in enhancing model generalization capabilities. Additionally, since the low-frequency components of the amplitude spectrum encode the overall style and global features of an image, while the phase spectrum encodes the semantic content and spatial distribution. Therefore, unlike conventional data augmentation, our objective is to use the style difference features as controllable style components within the low-frequency amplitude spectrum of the multi-scale features of images and guide the transformation of feature styles in a controlled manner. For

brevity, we concentrate exclusively on single layer image features $f \in \mathbb{R}^{D\times H\times W}$ to illustrate our method. First, Fast Fourier Transform (FFT) is applied to decompose the image features f into the amplitude spectrum $\mathcal{A}(\cdot,\cdot)$ and the phase spectrum $\mathcal{P}(\cdot,\cdot)$, as follows:

$$\begin{aligned}\mathcal{A}(u,v) &= \sqrt{\mathrm{Re}(\mathcal{F}(u,v))^2 + \mathrm{Im}(\mathcal{F}(u,v))^2},\\ \mathcal{P}(u,v) &= \tan^{-1}\left(\frac{\mathrm{Im}(\mathcal{F}(u,v))}{\mathrm{Re}(\mathcal{F}(u,v))}\right),\end{aligned} \tag{5}$$

where H and W are the height and width of the image features, x and y are the pixel coordinates in the spatial domain, and u and v are correspond to the frequency coordinates in the frequency domain. $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ denote the real and imaginary parts of the Fourier spectrum $\mathcal{F}(\cdot,\cdot)$. To enhance the interpretability of the spectrum, the amplitude spectrum is centralized to concentrate the low-frequency components at the center of the spectrum. Additionally, a low-frequency mask $M_l \in \{0,1\}$ is constructed to effectively extract the low-frequency components, as defined by:

$$M_l(x,y) = \mathbb{1}_{(x,y)\in[\frac{H}{2}(1-\alpha):\frac{H}{2}(1+\alpha),\frac{W}{2}(1-\alpha):\frac{W}{2}(1+\alpha)]}, \tag{6}$$

where α is the ratio of low-frequency components. Next, to prevent the introduction of uncontrollable noise resulting from excessive style differences, the style difference features F_d are numerically constrained by the hyperbolic tangent function. The activated style difference features are then weighted and summed with the low-frequency components of the amplitude spectrum, and combined with the amplitude spectrum of the high-frequency components to obtain the composite amplitude spectrum $\mathcal{A}_c(\cdot,\cdot)$, as follows:

$$\begin{aligned}\mathcal{A}_l(u,v) &= M_l(x,y)\cdot\mathcal{A}(u,v)\cdot(1+\beta\cdot\tanh(F_d)), \\ \mathcal{A}_c(u,v) &= (1-M_l(x,y))\cdot\mathcal{A}(u,v) + M_l(x,y)\cdot\mathcal{A}_l(u,v),\end{aligned} \tag{7}$$

where β is the intensity of style control. For high-level image features, which contain more abstract characteristics, a stronger style control is applied to alter the overall style. Conversely, for low-level image features, which encompass finer details, a weaker style control is employed to better preserve the original content. Finally, the composite amplitude spectrum $\mathcal{A}_c(\cdot,\cdot)$ and the original phase spectrum $\mathcal{P}(\cdot,\cdot)$ are merged and transformed using the inverse Fast Fourier Transform (iFFT) to obtain the style visual features $V \in \mathbb{R}^{D\times H\times W}$ after style transformation:

$$V(x,y) = \mathcal{F}^{-1}([\mathcal{A}_c(u,v), \mathcal{P}(u,v)]). \tag{8}$$

Style Synergy Optimization

Domain condition prompts are the key factors in guiding style transformation. However, a domain collapse issue arises: similar domain condition prompts (such as "in rain" and "in hail") are combined and encoded into similar style difference features. This leads to different style visual features being projected into the same style space distribution, thereby weakening the model's ability to learn and represent a diverse range of domain styles. To address this problem, we introduced SSO mechanism that synergistically weights the style contrastive loss and style aggregation loss based on their changing trends, further strengthening inter-domain feature separation and intra-domain feature aggregation. Specifically, a set of learnable domain bank $D \in \mathbb{R}^{K\times D}$ initialized by domain conditional prompts embeddings are used to store style vectors for each domain, where K is the number of conditional domain prompts. The style contrastive loss is designed to pull the style visual features closer to the style vectors of the same domain while pushing them farther from the style vectors of other domains. Consequently, style contrastive loss is computed between the globally average-pooled style visual features $\overline{V} \in \mathbb{R}^{1\times D}$ after global average pooling and the domain bank D as follows:

$$\mathcal{L}_{sc} = -\frac{1}{L}\sum_{l=1}^{L}\log\left(\frac{e^{(\overline{V}_l\cdot D^+)/\tau}}{e^{(\overline{V}_l\cdot D^+)/\tau} + \sum_{j=1}^{N} e^{(\overline{V}_l\cdot D_j^-)/\tau}}\right), \tag{9}$$

where L is number of feature layers and N is the number of negative domain samples. Through this bidirectional optimization, the domain bank can continuously enhance the distinction between style vectors of different domains. In turn, this process also constrains the style visual features produced by style transformations guided by similar domains, ensuring effective inter-domain feature separation and preventing domain collapse.

To further enhance style alignment, specific domain embeddings are first projected into the image feature space to obtain specific domain style features F_s. Both the specific domain style features F_s and the style visual features V are then normalized to better measure their correlation. Finally, the relationship is constrained using an L2 loss function:

$$\mathcal{L}_{\mathrm{sa}} = \frac{1}{LHW}\sum_{l=1}^{L}\sum_{h=1}^{H}\sum_{w=1}^{W} w_l\cdot\|V_{(h,w)}^{(l)} - F_{s,(h,w)}\|_2^2, \tag{10}$$

where w_l is weight coefficients that change with the number of layers. To encourage mutual cooperation between the two losses, our synergy weighting strategy is designed to adaptively adjust the weight of one loss based on changes in the other, as shown below:

$$w = \begin{cases} w_{init}\times(1-\lambda\times\Delta\mathcal{L}), & \text{if } \Delta\mathcal{L} > 0 \\ w_{init}\times(1+\lambda\times|\Delta\mathcal{L}|), & \text{otherwise}\end{cases} \tag{11}$$

where w_{init} is the initial weight coefficient, λ is the hyperparameter used for scaling the change. The idea behind synergy optimization is that when one loss increases, it should take precedence in the optimization process, reducing the contribution of the other loss. As the loss decreases, the weight of the other loss should be increased to promote collaboration between the two losses. According to Eq. 11, the weight coefficients for each component are determined, resulting in the overall loss as follows:

$$\mathcal{L}_s = w_{sa}\cdot\mathcal{L}_{sc} + w_{sc}\cdot\mathcal{L}_{sa}. \tag{12}$$

Experiments

Settings & Implementation Details

Datasets. We conduct the experiments on two synthetic datasets and four real-world datasets to evaluate the generalization capability of our method. For synthetic datasets,

Method	Trained on GTAV (G)				
	→C	→B	→M	→S	Avg.
IBN (Pan et al. 2018)	33.85	32.30	37.75	27.90	32.95
Iternorm (Huang et al. 2019)	31.81	32.70	33.88	27.07	31.37
ISW (Choi et al. 2021)	36.58	35.20	40.33	28.30	35.10
SHADE (Zhao et al. 2022)	44.65	39.28	43.34	28.41	38.92
SAN-SAW (Peng et al. 2022)	39.75	37.34	41.86	30.79	37.44
WildNet (Lee et al. 2022)	44.62	38.42	46.09	31.34	40.12
HRDA* (Hoyer et al. 2023)	39.63	38.69	42.21	-	-
TLDR (Kim et al. 2023)	46.51	42.58	46.18	36.30	42.89
DPCL (Yang et al. 2023)	44.87	40.21	46.74	-	-
BlindNet (Ahn et al. 2024)	45.72	41.32	47.08	31.39	41.38
FAMix (Fahes et al. 2024)	48.15	**45.61**	52.11	34.23	45.03
DGInStyle* (Jia et al. 2025)	46.89	42.81	50.19	-	-
SCSD (Ours)	**51.72**	44.67	**56.98**	**43.08**	**49.11**

Table 1: **Single-source setting trained on GTAV.** Mean IoU(%) comparison of different methods. "-" indicates the metric is not reported or the official source code is not available. * means the method using ResNet-101 as the backbone. The **best** and second best results are emphasized.

Method	Trained on Two Datasets (G+S)			
	→ C	→ B	→ M	Avg.
IBN (Pan et al. 2018)	35.55	32.18	38.09	35.27
ISW (Choi et al. 2021)	37.69	34.09	38.49	36.76
SHADE (Zhao et al. 2022)	47.43	40.30	47.60	45.11
Pin the memory (Kim et al. 2022)	44.51	38.07	42.70	41.76
AdvStyle (Zhong et al. 2022)	39.29	39.26	41.14	39.90
TLDR (Kim et al. 2023)	48.83	42.58	47.80	46.40
SPC-Net (Huang et al. 2023)	46.36	43.18	48.23	45.92
FAMix (Fahes et al. 2024)	49.41	**45.51**	51.61	48.84
SCSD (Ours)	**52.43**	45.25	**56.58**	**51.42**

Table 2: **Multi-source setting trained on GTAV + SYNTHIA.** Mean IoU(%) comparison of different methods with ResNet-50 backbone.

GTAV (G) (Richter et al. 2016) is a game synthetic dataset that contains 24,966 images with a resolution of 1914×1052. SYNTHIA (S) (Ros et al. 2016) is a large-scale synthetic dataset that contains 9,400 images with a resolution of 1280×760. For real-world datasets, Cityscapes (C) (Cordts et al. 2016), BDD-100K (B) (Yu et al. 2020) and Mapillary (M) (Neuhold et al. 2017) contain 2,975, 7,000 and 18,000 training images and 500, 1,000 and 2,000 validation images, respectively. ACDC (Sakaridis, Dai, and Van Gool 2021) is a challenging dataset of driving scenes under adverse weather conditions, including: night (AN), snow (AS), rain (AR) and fog (AF) with 106, 100, 100 and 100 validation images respectively.

Evaluation Protocols. Following prior works (Lee et al. 2022; Huang et al. 2023; Fahes et al. 2024), the model is trained on the source domain dataset and evaluated on other datasets as the target domain. Three settings include: 1) G→{C, B, M, S}; 2) C→{B, M, G, S}; and 3) G+S→{C, B, M}. We use the mIoU (%) metric for evaluation. In addition, we report the average mIoU on the target domain datasets.

Implementation Details. For fair comparison, we use the same ResNet-50 as the prior methods (Lee et al. 2022; Yang, Gu, and Sun 2023; Fahes et al. 2024) as the backbone of the segmentation model, which is initialized with CLIP pre-trained weights and perform architecture surgery following (Li et al. 2023). For the pixel decoder and mask decoder, we follow the default settings in (Ding et al. 2023; Bi, You, and Gevers 2024). The ratio of low-frequency components α is set to 0.15 in Eq. 6. The intensity of style control β is set to [1.0, 2.0, 4.0] in Eq. 7. The weight coefficients w_l is set to [0.2, 0.5, 1.0] in Eq. 10. The initial weight coefficient w_{init} is set to 1, and the hyperparameter λ in Eq. 11 is set to 0.3. The model is trained for 200K iterations with a batch size of 2. All experiments are conducted on a single machine equipped with two NVIDIA 3090 GPUs, each with 24 GB of memory. For more implementation details, reference to the *Appendix*.

Method	Trained on GTAV (G)				
	→ AN	→ AS	→ AR	→ AF	Avg.
ISW (Choi et al. 2021)	6.32	29.97	33.02	32.56	25.47
SHADE (Zhao et al. 2022)	8.18	30.38	35.44	36.87	27.72
WildNet (Lee et al. 2022)	8.27	30.29	36.32	35.39	27.57
SiamDoGe (Wu et al. 2022)	10.60	30.71	35.84	36.45	28.40
TLDR (Kim et al. 2023)	13.13	36.02	38.89	40.58	32.16
FAMix (Fahes et al. 2024)	14.96	37.09	38.66	40.25	32.74
SCSD (Ours)	**15.06**	**41.37**	**42.77**	**43.43**	**35.66**

Table 3: **Adverse Condition setting trained on GTAV.** Mean IoU(%) comparison of different methods with ResNet-50 backbone.

Components	Trained on GTAV (G)				
	→ C	→ B	→ M	→ S	Avg.
baseline	49.16	42.19	53.73	40.54	46.41
+ SQB	49.52 (↑ 0.36)	43.28 (↑ 1.09)	55.17 (↑ 1.44)	41.75 (↑ 1.21)	47.43 (↑ 1.02)
++ TDST	50.70 (↑ 1.18)	44.04 (↑ 0.76)	55.71 (↑ 0.54)	42.83 (↑ 1.08)	48.32 (↑ 0.89)
+++ SOO	**51.72** (↑ 1.02)	**44.67** (↑ 0.63)	**56.98** (↑ 1.27)	**43.08** (↑ 0.25)	**49.11** (↑ 0.79)

Table 4: Ablation studies on each component of SCSD.

Comparison With the State-of-the-Art Methods

Single-source setting. Tab. 1 shows the generalization performance of the models under G→{C, B, M, S} setting. Our method achieves an average of 49.11 mIoU on four target domain datasets, yielding an improvement of +4.08 mIoU over the state-of-the-art DGSS methods.

Multi-source setting. Tab. 2 shows the generalization performance of the models under G+S→{C, B, M} setting. Even when trained exclusively on synthetic datasetss, our method demonstrates a significantly superior generalization capabilities, consistently outperforming other methods across all real-world datasets, with an average improvement of +2.58 mIoU compared to the state-of-the-art methods.

Adverse condition settings. In Tab. 3, we further validate the generalization performance of SCSD on the challenging adverse condition dataset (Sakaridis, Dai, and Van Gool 2021). Specifically, our method improves mIoU by +0.10, +4.28, +4.11, and +3.18 on night (AN), snow (AS), rain (AR) and fog (AF), respectively. This further demonstrates the strong generalization capability of SCSD.

SSO			Trained on GTAV (G)				
$\mathcal{L}_{sc}$	$\mathcal{L}_{sa}$	w	$\rightarrow$ C	$\rightarrow$ B	$\rightarrow$ M	$\rightarrow$ S	Avg.
-	-	-	50.70	44.04	55.71	42.83	48.32
✓	-	-	50.47	44.25	55.94	42.87	48.38
-	✓	-	49.86	44.38	54.56	41.72	47.63
✓	✓	-	50.83	**45.21**	56.50	42.49	48.76
✓	✓	✓	**51.72**	44.67	**56.98**	**43.08**	**49.11**

Table 5: Ablation studies on SSO.

Ablation Studies

Key Components. We study the individual contribution of each proposed component to the overall performance. Tab. 4 shows the average mIoU improvement on SCSD as we progressively integrate the three key components. Specifically, the model with SQB (row 2) shows better results compared to the baseline model (row 1), with an average gain of +1.02 mIoU. This indicates that semantic consistency is crucial for cross-domain robust prediction. The TDST (row3) introduced only during the training phase improves the performance by +0.89 mIoU on average, which demonstrates the importance of style diversity for the model to generalize to the wide range of domains. Finally, SSO (row4) is added to further improve the performance by +0.79 mIoU on average.
Style Synergy Optimization. In Tab. 5, we find that when only the style aggregation loss is applied (row 3), it merely focuses on aggregating style visual features and specific domain style features. As a result, it fails to distinguish between the styles of similar domains, which exacerbates the negative impact of domain collapse and leads to performance degradation. When the style contrastive loss and synergy weighting strategy are gradually introduced (row 4 and 5), the two losses work together to strengthen the separation of inter-domain features and the aggregation of intra-domain features, and the performance is further improved.

Qualitative Analysis

Benefits from Semantic Query Booster. In Fig. 4, we visualize the mask predictions of the baseline and SCSD outputs at different layers of the mask decoder. In the second column of Fig. 4, we observe that when dealing with challenging domains, such as snow, whether it's the initial mask output by the first layer of the mask decoder or the refined mask from the sixth layer, the baseline consistently confuses the "vegetation" and "building" categories (refer to the white box in the lower-left corner). Additionally, the "truck" consistently lacks a clear mask prediction (refer to the upper-right white box). In the third column of Fig. 4, it is incredible that after being equipped with SQB, SCSD generates the accurate initial masks even in the first layer, allowing for further refinement in subsequent layers. We consider this improvement is due to the object queries learning semantic associations between the visual and textual modalities, enabling cross-domain consistency in predictions, even in unseen extreme domains.
Distribution analysis. We conducted t-SNE visualizations of the image features used for classification to analyze the

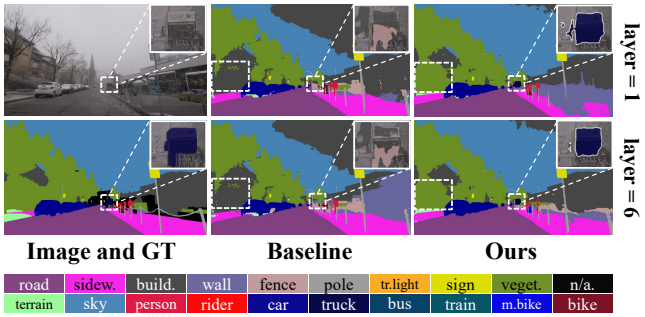
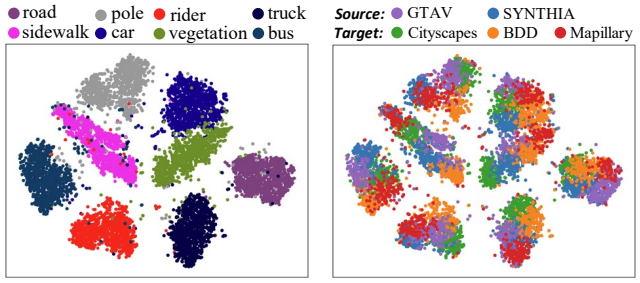


Figure 4: Column 1: Image and ground truth (GT). Columns 2-3: Prediction masks output by baseline without SQB and our SCSD at the first and sixth layers of the mask decoder.



(a) Class distribution (b) Domain distribution

Figure 5: t-SNE visualization of image features. Colors denote different categories in (a) and different domains in (b).

effectiveness of our SCSD. Fig. 5a demonstrates that our method effectively enables the model to distinguish between different categories, significantly reducing semantic confusion caused by the convergence of features from different categories within the same domain. Furthermore, as shown in Figure 5b, the model successfully clusters samples from different domains according to their respective categories. This success can be attributed to our proposed TDST module, which allows the model to learn domain-specific style knowledge through domain text difference features, even without the use of additional domain data. See the *Appendix* for more qualitative analysis.

Conclusion

In this paper, we propose a novel Semantic Consistency and Style Diversity (SCSD) framework to addresses the limitations of existing methods in Domain Generalized Semantic Segmentation (DGSS). To effectively leverage semantic consistency between image and text modalities, and the style diversity of the text modality, we introduce three innovative components: the Semantic Query Booster (SQB), the Text-Driven Style Transform (TDST) module, and the Style Synergy Optimization (SSO) mechanism. Extensive experiments on multiple benchmarks datasets demonstrated the significant performance improvements achieved by SCSD, setting new state-of-the-art results for DGSS.

Acknowledgments

This work was supported by National Science and Technology Major Project (No.2022ZD0118201), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No.U21B2037, No.U22B2051, No.U23A20383, No.U21A20472, No.62176222, No.62176223, No.62176226, No.62072386, No.62072387, No.62072389, No.62002305 and No.62272401), and the Natural Science Foundation of Fujian Province of China (No.2021J06003, No.2022J06001).

References

Ahn, W.-J.; Yang, G.-Y.; Choi, H.-D.; and Lim, M.-T. 2024. Style Blind Domain Generalized Semantic Segmentation via Covariance Alignment and Semantic Consistence Contrastive Learning. In *CVPR*.

Bi, Q.; You, S.; and Gevers, T. 2024. Learning content-enhanced mask transformer for domain generalized urban-scene segmentation. In *AAAI*.

Chattopadhyay, P.; Sarangmath, K.; Vijaykumar, V.; and Hoffman, J. 2023. Pasta: Proportional amplitude spectrum training augmentation for syn-to-real domain generalization. In *ICCV*.

Chen, H.; Tao, R.; Zhang, H.; Wang, Y.; Li, X.; Ye, W.; Wang, J.; Hu, G.; and Savvides, M. 2024. Conv-adapter: Exploring parameter efficient transfer learning for convnets. In *CVPR*.

Chen, L.; Zhang, Y.; Song, Y.; van den Hengel, A.; and Liu, L. 2023. Domain Generalization via Rationale Invariance. In *ICCV*.

Cheng, Y.; Wei, F.; Bao, J.; Chen, D.; and Zhang, W. 2023. Adpl: Adaptive dual path learning for domain adaptation of semantic segmentation. *TPAMI*.

Choe, S.-A.; Shin, A.-H.; Park, K.-H.; Choi, J.; and Park, G.-M. 2024. Open-Set Domain Adaptation for Semantic Segmentation. In *CVPR*.

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 *CVPR*.

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 *CVPR*.

Dayal, A.; KB, V.; Cenkeramaddi, L. R.; Mohan, C.; Kumar, A.; and N Balasubramanian, V. 2024. MADG: margin-based adversarial learning for domain generalization. In *NeurIPS*.

Ding, J.; Xue, N.; Xia, G.-S.; Schiele, B.; and Dai, D. 2023. HGFormer: Hierarchical Grouping Transformer for Domain Generalized Semantic Segmentation. In *CVPR*.

Fahes, M.; Vu, T.-H.; Bursuc, A.; Pérez, P.; and de Charette, R. 2024. A Simple Recipe for Language-guided Domain Generalized Segmentation. In *CVPR*.

Gong, Y.; Zhong, Z.; Qu, Y.; Luo, Z.; Ji, R.; and Jiang, M. 2024. Cross-modality perturbation synergy attack for person re-identification. *arXiv preprint arXiv:2401.10090*.

Hoyer, L.; Dai, D.; and Van Gool, L. 2022. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*.

Hoyer, L.; Dai, D.; and Van Gool, L. 2023. Domain adaptive and generalizable network architectures and training strategies for semantic image segmentation. *TPAMI*.

Hu, J.; Huang, L.; Ren, T.; Zhang, S.; Ji, R.; and Cao, L. 2023. You only segment once: Towards real-time panoptic segmentation. In *CVPR*.

Huang, L.; Zhou, Y.; Zhu, F.; Liu, L.; and Shao, L. 2019. Iterative normalization: Beyond standardization towards efficient whitening. In *CVPR*.

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 *CVPR*.

Jia, Y.; Hoyer, L.; Huang, S.; Wang, T.; Van Gool, L.; Schindler, K.; and Obukhov, A. 2025. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *ECCV*.

Kang, J.; Lee, S.; Kim, N.; and Kwak, S. 2022. Style neophile: Constantly seeking novel styles for domain generalization. In *CVPR*.

Kim, J.; Lee, J.; Park, J.; Min, D.; and Sohn, K. 2022. Pin the memory: Learning to generalize semantic segmentation. In *CVPR*.

Kim, S.; Kim, D.-h.; and Kim, H. 2023. Texture learning domain randomization for domain generalized segmentation. In *ICCV*.

Lee, S.; Seong, H.; Lee, S.; and Kim, E. 2022. Wildnet: Learning domain generalized semantic segmentation from the wild. In *CVPR*.

Li, L.; Zhou, T.; Wang, W.; Li, J.; and Yang, Y. 2022. Deep hierarchical semantic segmentation. In *CVPR*.

Li, Y.; Wang, H.; Duan, Y.; and Li, X. 2023. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*.

Lin, J.; Shen, Y.; Wang, B.; Lin, S.; Li, K.; and Cao, L. 2024. Weakly supervised open-vocabulary object detection. In *AAAI*.

Mi, P.; Lin, J.; Zhou, Y.; Shen, Y.; Luo, G.; Sun, X.; Cao, L.; Fu, R.; Xu, Q.; and Ji, R. 2022. Active teacher for semi-supervised object detection. In *CVPR*.

Ming, Y.; Cai, Z.; Gu, J.; Sun, Y.; Li, W.; and Li, Y. 2022. Delving into out-of-distribution detection with vision-language representations. In *NeurIPS*.

Neuhold, G.; Ollmann, T.; Rota Bulo, S.; and Kontschieder, P. 2017. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*.

Niemeijer, J.; Schwonberg, M.; Termöhlen, J.-A.; Schmidt, N. M.; and Fingscheidt, T. 2024. Generalization by Adaptation: Diffusion-Based Domain Extension for Domain-Generalized Semantic Segmentation. In *WACV*.

Pan, X.; Luo, P.; Shi, J.; and Tang, X. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*.

Pan, X.; Zhan, X.; Shi, J.; Tang, X.; and Luo, P. 2019. Switchable whitening for deep representation learning. In *ICCV*.

Peng, D.; Lei, Y.; Hayat, M.; Guo, Y.; and Li, W. 2022. Semantic-aware domain generalized segmentation. In *CVPR*.

Peng, D.; Lei, Y.; Liu, L.; Zhang, P.; and Liu, J. 2021. Global and local texture randomization for synthetic-to-real semantic segmentation. *TIP*.

Qu, Y.; Dai, S.; Li, X.; Lin, J.; Cao, L.; Zhang, S.; and Ji, R. 2024. Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In *ACM MM*.

Qu, Y.; Wang, Y.; and Qi, Y. 2023. Sg-nerf: Semantic-guided point-based neural radiance fields. In *ICME*.

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 *ICML*.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *ECCV*.

Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2021. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*.

Shu, Y.; Cao, Z.; Wang, C.; Wang, J.; and Long, M. 2021. Open Domain Generalization with Domain-Augmented Meta-Learning. In *CVPR*.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *ICCV*.

Tu, W.; Deng, W.; and Gedeon, T. 2024. A closer look at the robustness of contrastive language-image pre-training (clip). In *NeurIPS*.

Wu, Z.; Wu, X.; Zhang, X.; Ju, L.; and Wang, S. 2022. Siamdoge: Domain generalizable semantic segmentation using siamese network. In *ECCV*.

Xia, R.; Zhao, C.; Zheng, M.; Wu, Z.; Sun, Q.; and Tang, Y. 2023. Cmda: Cross-modality domain adaptation for nighttime semantic segmentation. In *ICCV*.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A fourier-based framework for domain generalization. In *CVPR*.

Yang, L.; Gu, X.; and Sun, J. 2023. Generalized semantic segmentation by self-supervised source domain projection and multi-level contrastive learning. In *AAAI*.

Yang, Y.; and Soatto, S. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*.

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 *CVPR*.

Yu, Q.; He, J.; Deng, X.; Shen, X.; and Chen, L.-C. 2024. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *NeurIPS*.

Yue, P.; Lin, J.; Zhang, S.; Hu, J.; Lu, Y.; Niu, H.; Ding, H.; Zhang, Y.; Jiang, G.; Cao, L.; et al. 2024. Adaptive Selection based Referring Image Segmentation. In *ACM MM*.

Yue, X.; Zhang, Y.; Zhao, S.; Sangiovanni-Vincentelli, A.; Keutzer, K.; and Gong, B. 2019. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *ICCV*.

Zhang, Y.; Guo, M.-H.; Wang, M.; and Hu, S.-M. 2024. Exploring Regional Clues in CLIP for Zero-Shot Semantic Segmentation. In *CVPR*.

Zhao, Y.; Zhong, Z.; Zhao, N.; Sebe, N.; and Lee, G. H. 2022. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *ECCV*.

Zhong, Z.; Zhao, Y.; Lee, G. H.; and Sebe, N. 2022. Adversarial style augmentation for domain generalized urban-scene segmentation. In *NeurIPS*.