Learning Content-Enhanced Mask Transformer for Domain Generalized Urban-Scene Segmentation

Qi Bi, Shaodi You, Theo Gevers

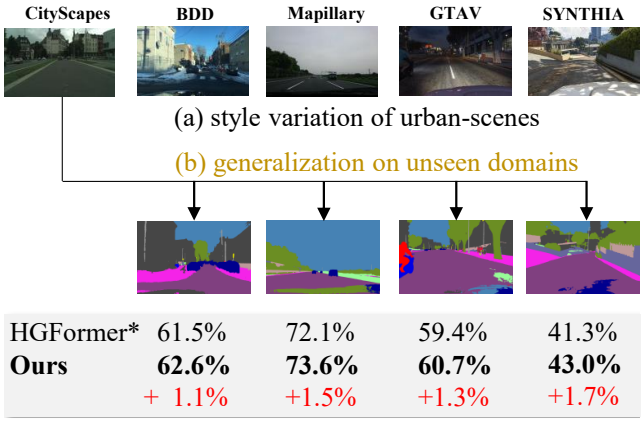
Computer Vision Research Group, University of Amsterdam, Netherlands
{q.bi, s.you, th.gevers}@uva.nl

Abstract

Domain-generalized urban-scene semantic segmentation (USSS) aims to learn generalized semantic predictions across diverse urban-scene styles. Unlike generic domain gap challenges, USSS is unique in that the semantic categories are often similar in different urban scenes, while the styles can vary significantly due to changes in urban landscapes, weather conditions, lighting, and other factors. Existing approaches typically rely on convolutional neural networks (CNNs) to learn the content of urban scenes. In this paper, we propose a Content-enhanced Mask TransFormer (CMFormer) for domain-generalized USSS. The main idea is to enhance the focus of the fundamental component, the mask attention mechanism, in Transformer segmentation models on content information. We have observed through empirical analysis that a mask representation effectively captures pixel segments, albeit with reduced robustness to style variations. Conversely, its lower-resolution counterpart exhibits greater ability to accommodate style variations, while being less proficient in representing pixel segments. To harness the synergistic attributes of these two approaches, we introduce a novel content-enhanced mask attention mechanism. It learns mask queries from both the image feature and its down-sampled counterpart, aiming to simultaneously encapsulate the content and address stylistic variations. These features are fused into a Transformer decoder and integrated into a multi-resolution content-enhanced mask attention learning scheme. Extensive experiments conducted on various domain-generalized urban-scene segmentation datasets demonstrate that the proposed CMFormer significantly outperforms existing CNN-based methods by up to 14.0% mIoU and the contemporary HGFormer by up to 1.7% mIoU. The source code is publicly available at https://github.com/BiQiWHU/CMFormer.

Introduction

Urban-scene semantic segmentation (USSS) is a challenging problem because of the large scene variations due to changing landscape, weather, and lighting conditions (Sakaridis, Dai, and Van Gool 2021; Mirza et al. 2022; Bi, You, and Gevers 2023; Chen et al. 2022). Unreliable USSS can pose a significant risk to road users. Nevertheless, a segmentation model trained on a specific dataset cannot encompass all urban scenes across the globe. As a result, the segmentation model is prone to encountering unfamiliar urban scenes during the inference stage. Hence, domain generalization is essential for robust USSS (Pan et al. 2018; Huang et al. 2019a; Choi et al. 2021), where a segmentation model can effectively extrapolate its performance to urban scenes that it hasn't encountered before (Fig. 1). In contrast to common domain generalization, domain generalized USSS requires special attention because the domain gap is mainly caused by large style variations whereas changes in semantics largely remain consistent (example in Fig. 2).



HGFormer*	61.5%	72.1%	59.4%	41.3%
Ours	**62.6%**	**73.6%**	**60.7%**	**43.0%**
	+ 1.1%	+1.5%	+1.3%	+1.7%

Figure 1: (a) In domain generalized USSS, the domain gap is mainly from the extremely-varied styles. (b) A segmentation model is supposed to show good generalization on unseen target domains.

Existing approaches can be divided into two groups. One group focuses on the style de-coupling. This is usually achieved by a normalization (Pan et al. 2018; Huang et al. 2019a; Peng et al. 2022) or whitening (Pan et al. 2019; Choi et al. 2021; Xu et al. 2022; Peng et al. 2022) transformation. However, the de-coupling methodology falls short as the content is not learnt in a robust way. The other group is based on adverse domain training (Zhao et al. 2022; Lee et al. 2022; Zhong et al. 2022). However, these methods usually do not particularly focus on urban styles and therefore their performance is limited.

Recent work has shown that mask-level segmentation

Figure 2: Domain-generalized USSS demonstrates a distinctive feature of consistent content with diverse styles. An example is given for BDD100K and GTA5.

Transformer (e.g., Mask2Former) (Ding et al. 2023) is a scalable learner for domain generalized semantic segmentation. However, based on our empirical observations, a high-resolution mask-level representation excels at capturing content down to pixel semantics but is more susceptible to style variations. Conversely, its down-sampled counterpart is less proficient in representing content down to pixel semantics but exhibits greater resilience to style variations.

A novel content-enhanced mask attention (CMA) mechanism is proposed. It jointly leverages both mask representation and its down-sampled counterpart, which show complementary properties on content representing and handling style variation. Jointly using both features helps the style to be uniformly distributed while the content to be stabilized in a certain cluster. The proposed CMA takes the original image feature together with its down-sampled counterpart as input. Both features are fused to learn a more robust content from their complementary properties.

The proposed content-enhanced mask attention (CMA) mechanism can be integrated into existing mask-level segmentation Transformer in a learnable fashion. It consists of three key steps, namely, exploiting high-resolution properties, exploiting low-resolution properties, and content-enhanced fusion. Besides, it can also be seamlessly adapted to multi-resolution features. A novel **C**ontent-enhanced **M**ask Trans**former** (CMFormer) is proposed for domain-generalized USSS.

Large-scale experiments are conducted with various domain generalized USSS settings, *i.e.*, trained on one dataset from (Richter et al. 2016; Ros et al. 2016; Cordts et al. 2016; Neuhold et al. 2017; Yu et al. 2018) as the source domain, and validated on the rest of the four datasets as the unseen target domains. All the datasets contain the same 19 semantic categories as the content, but vary in terms of scene styles. The experiments show that the proposed CMFormer achieves up to 14.00% mIoU improvement compared to the state-of-the-art CNN based methods (*e.g.*, SAW (Peng et al. 2022), WildNet (Lee et al. 2022)). Furthermore, it demonstrates a mIoU improvement of up to 1.7% compared to the modern HGFormer model (Ding et al. 2023). It also shows state-of-the-art performance on synthetic-to-real and clear-to-adverse generalization.

Our contribution is summarized as follows:

- A content-enhanced mask attention (CMA) mechanism is proposed to leverage the complementary content and style properties from mask-level representation and its down-sampled counterpart.
- On top of CMA, a Content-enhanced Mask Transformer (CMFormer) is proposed for domain generalized urban-scene semantic segmentation.
- Extensive experiments show a large performance improvement over existing SOTA by up to 14.0% mIoU, and HGFormer by up to 1.7% mIoU.

Related Work

Domain Generalization has been studied on no task-specific scenarios in the field of both machine learning and computer vision. Hu *et al*. (Hu and Lee 2022) proposed a framework for image retrieval in an unsupervised setting. Zhou *et al*. (Zhou et al. 2020) proposed a framework to generalize to new homogeneous domains. Qiao *et al*. (Qiao, Zhao, and Peng 2020) and Peng *et al*. (Peng, Qiao, and Zhao 2022) proposed to learn domain generalization from a single source domain. Many other methods have also been proposed (Zhao et al. 2020; Mahajan, Tople, and Sharma 2021; Wang et al. 2020; Chattopadhyay, Balaji, and Hoffman 2020; Segu, Tonioni, and Tombari 2023).

Domain Generalized Semantic Segmentation is more practical than conventional semantic segmentation (Pan et al. 2022; Ji et al. 2021; Li et al. 2021; Ji et al. 2022; Zhou, Yi, and Bi 2021; Ye et al. 2021), which focuses on the generalization of a segmentation model on unseen target domains. Existing methods focus on the generalization of in-the-wild (Piva, de Geus, and Dubbelman 2023), scribble (Tjio et al. 2022) and multi-source images (Kim et al. 2022; Lambert et al. 2020), where substantial alterations can occur in both the content and style.

Domain Generalized USSS focuses on the generalization of driving-scenes (Cordts et al. 2016; Yu et al. 2018; Neuhold et al. 2017; Ros et al. 2016; Richter et al. 2016). These methods use either normalization transformation (*e.g.*, IBN (Pan et al. 2018), IN (Huang et al. 2019a), SAN (Peng et al. 2022)) or whitening transformation (*e.g.*, IW (Pan et al. 2019), ISW (Choi et al. 2021), DIRL (Xu et al. 2022), SAW (Peng et al. 2022)) on the training domain, to enable the model to generalize better on the target domains. Other advanced methods for domain generalization in segmentation typically rely on external images to incorporate more diverse styles (Lee et al. 2022; Zhao et al. 2022; Zhong et al. 2022; Li et al. 2023), and leverage content consistency across multi-scale features (Yue et al. 2019). To the best of our knowledge, all of these methods are based on CNN.

Mask Transformer for Semantic Segmentation uses the queries in the Transformer decoder to learn the masks, *e.g.*, Segmenter (Strudel et al. 2021), MaskFormer (Cheng, Schwing, and Kirillov 2021). More recently, Mask2Former (Cheng et al. 2022) further simplifies the pipeline of MaskFormer and achieves better performance.

Preliminary

Problem Definition Domain generalization can be formulated as a worst-case problem (Li, Namkoong, and Xia 2021; Zhong et al. 2022; Volpi et al. 2018). Given a source domain $\mathcal{S}$, and a set of unseen target domains $\mathcal{T}_1, \mathcal{T}_2, \cdots$, a

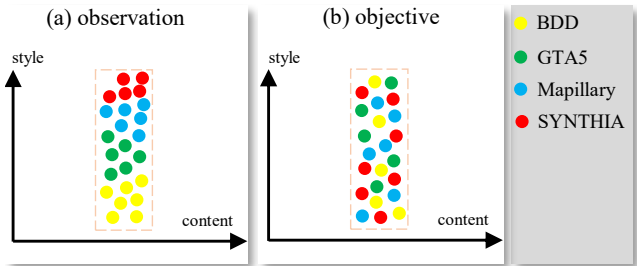


Figure 3: (a) In the domain-generalized USSS setting, within the content-style space, samples from various domains tend to cluster closely along the content dimension while displaying dispersion along the style dimension. (b) An optimal generalized semantic segmentation scenario would involve uniform distribution of styles while maintaining content stability (as indicated by the brown bounding box).

model parameterized by θ with the task-specific loss $\mathcal{L}_{task}$, the generic domain generalization task can be formulated as a worst-case problem, given by

$$\min_{\theta} \quad \sup_{\mathcal{T}:D(\mathcal{S};\mathcal{T}_1,\mathcal{T}_2,\cdots)\leq\rho} \quad \mathbb{E}_T[\mathcal{L}_{task}(\theta;\mathcal{T}_1,\mathcal{T}_2,\cdots)], \qquad (1)$$

where θ denotes the model parameters, $D(\mathcal{S};\mathcal{T}_1,\mathcal{T}_2,\cdots)$ corresponds to the distance between the source $\mathcal{S}$ and target domain $\mathcal{T}$, and ρ denotes the constraint threshold.

Content-style Feature Space Here we analyze the feature space. Figure 3a illustrates that in the context of domain-generalized USSS, samples from distinct domains might exhibit analogous patterns and cluster tightly along the content dimension. Conversely, samples from diverse domains may segregate into separate clusters along the style dimension.

An optimal and adaptable segmentation representation should achieve content stability while simultaneously exhibiting resilience in the face of significant style variations. Illustrated in Figure 3b, our objective is to cultivate a content-style space wherein: 1) samples from diverse domains can occupy analogous positions along the content dimension, and 2) samples can be uniformly dispersed across the style dimension. Both learning objectives allow us to therefore minimize the domain gap.

Overall Idea Recent work has shown that mask-level segmentation Transformer (e.g., Mask2Former) (Ding et al. 2023) is a scalable learner for domain generalized semantic segmentation. However, we empirically observe that, a mask-level representation is better at representing content, but more sensitive to style variations (similar to Fig. 3a); its low-resolution counterpart, on the contrary, is less capable to represent content, but more robust to the style variations (similar to the style dimensions in Fig. 3b).

Overall, the mask representation and its down-sampled counterpart shows complementary properties when handling samples from different domains. Thus, it is natural to jointly leverage both mask representation and its down-sampled counterparts, so as to at the same time stabilize the content and be insensitive to the style variation.

Difference between Existing Pipelines Existing methods usually focus on decoupling the styles from urban scenes, so that along the style dimension the samples from different domains are more uniformly distributed.

In contrast, the proposed method intends to leverage the content representation ability of mask-level features and the style handling ability of its down-sampled counterpart, so as to realize the aforementioned learning objective.

Methodology

Recap on Mask Attention

Recent studies show that the mask-level pipelines (Strudel et al. 2021; Cheng, Schwing, and Kirillov 2021; Cheng et al. 2022) have stronger representation ability than conventional pixel-wise pipelines for semantic segmentation, which can be attributed to the mask attention mechanism.

It learns the query features as the segmentation masks by introducing a mask attention matrix based on the self-attention mechanism. Let $\mathbf{F}_l$ and $\mathbf{X}_l$ denote the image features from the image decoder and the features of the l^{th} layer in a Transformer decoder, respectively. When $l = 0$, $\mathbf{X}_0$ refers to the input query features of the Transformer decoder.

The key $\mathbf{K}_l$ and value $\mathbf{V}_l$ on $\mathbf{F}_{l-1}$ are computed by linear transformations f_K and f_V, respectively. Similarly, the query $\mathbf{Q}_l$ on $\mathbf{X}_{l-1}$ is computed by linear transformation f_Q. Then, the query feature $\mathbf{X}_l$ is computed by

$$\mathbf{X}_l = \text{softmax}(\mathcal{M}_{l-1} + \mathbf{Q}_l\mathbf{K}_l^{\mathsf{T}})\mathbf{V}_l + \mathbf{X}_{l-1}, \qquad (2)$$

where $\mathcal{M}_{l-1} \in \{0,1\}^{N\times H_l W_l}$ is a binary mask attention matrix from the resized mask prediction of the previous $(l-1)^{th}$ layer, with a threshold of 0.5. $\mathcal{M}_0$ is binarized and resized from $\mathbf{X}_0$. It filters the foreground regions of an image, given by

$$\mathcal{M}_{l-1}(x,y) = \begin{cases} 0 & \text{if } \mathcal{M}_{l-1}(x,y)\text{=1} \\ \text{-}\infty & \text{else} \end{cases}. \qquad (3)$$

Exploiting High-Resolution Properties

Highlighted within the green block in Figure 4, our empirical observations reveal that the high-resolution mask representation exhibits the following characteristics: 1) greater proficiency in content representation, and 2) reduced susceptibility to domain variation. Achieving uniform mixing of samples from four domains presents a challenge.

To leverage the properties from high-resolution mask representations, we use the self-attention mechanism to exploit the amplified content representation from $\mathbf{X}_l$. Let $\mathbf{Q}_{\mathbf{X}_l}$, $\mathbf{V}_{\mathbf{X}_l}$ and $\mathbf{K}_{\mathbf{X}_l}$ denote its query, value and key, and d_k denotes their dimension. Then, the self-attention is computed as

$$\text{Attention}(\mathbf{Q}_{\mathbf{X}_l},\mathbf{K}_{\mathbf{X}_l},\mathbf{V}_{\mathbf{X}_l}) = \text{Softmax}(\frac{\mathbf{Q}_{\mathbf{X}_l}\mathbf{K}_{\mathbf{X}_l}}{\sqrt{d_k}})\mathbf{V}_{\mathbf{X}_l}, \quad (4)$$

where Softmax denotes the softmax normalization function, and the final output is denoted as $\bar{\mathbf{X}}_l$.

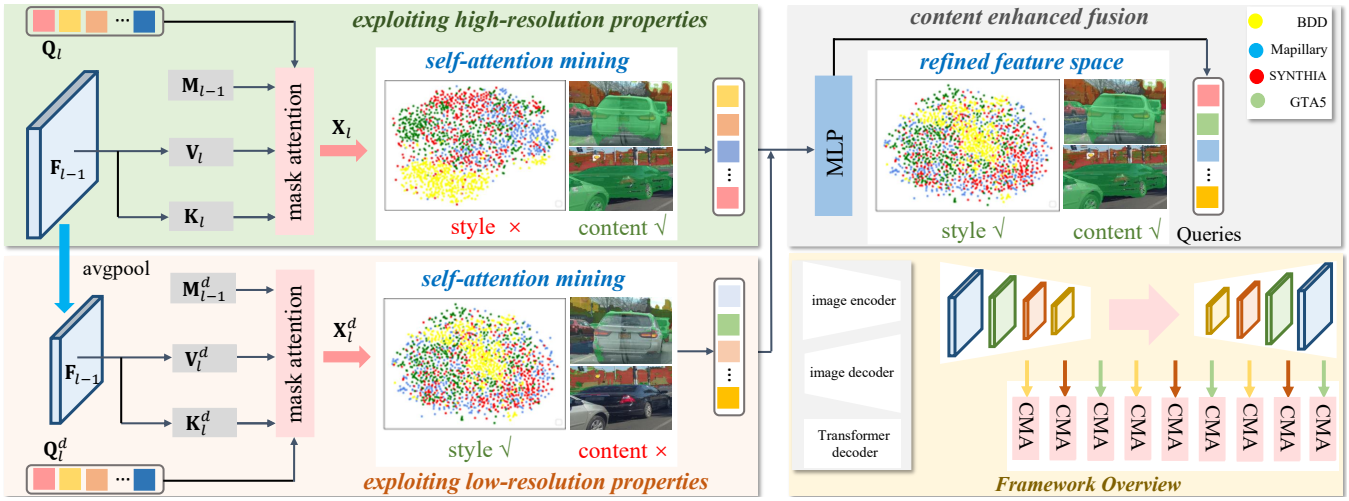


Figure 4: (a) The proposed Content-enhanced Mask Attention (CMA) consists of three key steps, namely, exploiting high-resolution properties (in green), exploiting low-resolution properties (in brown), and content enhanced fusion (in gray). (b) Framework overview (in yellow) of the proposed **C**ontent-enhanced **M**ask Trans**Former** (CMFormer) for domain generalized semantic segmentation. The image decoder is directly inherited from the Mask2Former (Cheng et al. 2022).

Exploiting Low-Resolution Properties

As shown in the brown block of Fig. 4, the low-resolution mask-level representation has the following properties: 1) less qualified to represent the content; 2) more capable to handle the style variation. In the feature space, samples from different domains are more uniformly distributed. We propose to build a low-resolution mask representation derived from its high-resolution counterpart. This approach capitalizes on the attributes of the low-resolution representation to effectively address domain variations.

The low-resolution counterpart $\mathbf{F}_l^d$ is computed by average pooling avgpool from the original image feature $\mathbf{F}_l$ by

$$\mathbf{F}_l^d = \text{avgpool}(\mathbf{F}_l), \tag{5}$$

where the width and height of $\mathbf{F}_l$ is both twice the width and height of $\mathbf{F}_l^d$.

Similarly, the key and value from $\mathbf{F}_l^d$ are computed by linear transformations, and can be denoted as $\mathbf{K}_l^d$ and $\mathbf{V}_l^d$, respectively. The query from $\mathbf{X}_{l-1}^d$ is also computed by linear transformation, and can be denoted as $\mathbf{K}_l^d$. The mask attention on the low-resolution feature $\mathbf{X}_l^d$ is computed as

$$\mathbf{X}_l^d = \text{softmax}(\mathcal{M}_{l-1}^d + \mathbf{Q}_l \mathbf{K}_l^{d\mathsf{T}})\mathbf{V}_l^d + \mathbf{X}_{l-1}^d. \tag{6}$$

To exploit the properties from the low-resolution mask representation $\mathbf{X}_l^d$, we use the self-attention mechanism. Let $\mathbf{Q}_{\mathbf{X}_l^d}$, $\mathbf{V}_{\mathbf{X}_l^d}$ and $\mathbf{K}_{\mathbf{X}_l^d}$ denote its query, value and keys. Then, the self-attention is computed by

$$\text{Attention}(\mathbf{Q}_{\mathbf{X}_l^d}, \mathbf{K}_{\mathbf{X}_l^d}, \mathbf{V}_{\mathbf{X}_l^d}) = \text{Softmax}(\frac{\mathbf{Q}_{\mathbf{X}_l^d}\mathbf{K}_{\mathbf{X}_l^d}}{\sqrt{d_k}})\mathbf{V}_{\mathbf{X}_l^d}. \tag{7}$$

The final output is denoted as $\tilde{\mathbf{X}}_l^d$. It inherits the characteristics of the low-resolution mask representation, which is adept at accommodating style variations while being less resilient in capturing pixel-level intricacies.

Content-enhanced Fusion

Our idea is to leverage the complementary properties of mask-level representation and its down-sample counterpart, so as to enhance both the pixel-wise representing and style variation handing (shown in the gray box of Fig. 4). The joint use of both representations aids the segmentation masks in concentrating on scene content while reducing sensitivity to style variations.

To this end, we fuse both representations $\tilde{\mathbf{X}}_l$ and $\tilde{\mathbf{X}}_l^d$ in a simple and straight-forward way. The fused feature $\mathbf{X}_l^{final}$ serves as the final output of the l^{th} Transformer decoder, and it is computed as

$$\mathbf{X}_l^{final} = h_l([\tilde{\mathbf{X}}_l, \tilde{\mathbf{X}}_l^d]), \tag{8}$$

where $[\cdot,\cdot]$ represents the concatenation operation, and $h_l(\cdot)$ refers to a linear layer.

Network Architecture and Implementation Details

The overall framework is shown in the yellow box of Fig. 4. The Swin Transformer (Liu et al. 2021) is used as the backbone. The pre-trained backbone from ImageNet (Deng et al. 2009) is utilized for initialization.

The image decoder from (Cheng et al. 2022) uses the off-the-shelf multi-scale deformable attention Transformer (MSDeformAttn) (Zhu et al. 2021) with the default setting in (Zhu et al. 2021; Cheng et al. 2022). By considering the image features from the Swin-Based encoder as input, every 6 MSDeformAttn layers are used to progressively up-sample the image features in $\times 32$, $\times 16$, $\times 8$, and $\times 4$, respectively. The $1/4$ resolution feature map is fused with the features from the Transformer decoder for dense prediction.

The Transformer decoder is also directly inherited from Mask2Former (Cheng et al. 2022), which has 9 self-attention layers in the Transformer decoder to handle the $\times 32$, $\times 16$ and $\times 8$ image features, respectively.

Following the default setting of MaskFormer (Cheng, Schwing, and Kirillov 2021) and Mask2Former (Cheng et al. 2022), the final loss function $\mathcal{L}$ is a linear combination of the binary cross-entropy loss $\mathcal{L}_{ce}$, dice loss $\mathcal{L}_{dice}$, and the classification loss $\mathcal{L}_{cls}$, given by

$$\mathcal{L} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{dice}\mathcal{L}_{dice} + \lambda_{cls}\mathcal{L}_{cls}, \quad (9)$$

with hyper-parameters $\lambda_{ce} = \lambda_{dice} = 5.0, \lambda_{cls} = 2.0$ as the default of Mask2Former without any tuning. The Adam optimizer is used with an initial learning rate of 1×10^{-4}. The weight decay is set 0.05. The training terminates after 50 epochs.

Experiment

Dataset & Evaluation Protocols

Building upon prior research in domain-generalized USSS, our experiments utilize five different semantic segmentation datasets. Specifically, CityScapes (Cordts et al. 2016) provides 2,975 and 500 well-annotated samples for training and validation, respectively. These driving-scenes are captured in Germany cities with a resolution of 2048×1024. BDD-100K (Yu et al. 2018) also provides diverse urban driving scenes with a resolution of 1280×720. 7,000 and 1,000 fine-annotated samples are provided for training and validation of semantic segmentation, respectively. Mapillary (Neuhold et al. 2017) is also a real-scene large-scale semantic segmentation dataset with 25,000 samples. SYNTHIA (Ros et al. 2016) is large-scale synthetic dataset, and provides 9,400 images with a resolution of 1280×760. GTA5 (Richter et al. 2016) is a synthetic semantic segmentation dataset rendered by the GTAV game engine. It provides 24,966 simulated urban-street samples with a resolution of 1914×1052. We use C, B, M, S and G to denote these five datasets.

Following prior domain generalized USSS works (Pan et al. 2018, 2019; Choi et al. 2021; Peng et al. 2022), the segmentation model is trained on one dataset as the source domain, and is validated on the rest of the four datasets as the target domains. Three settings include: 1) G to C, B, M, S; 2) S to C, B, M, G; and 3) C to B, M, G, S. mIoU (in percentage %) is used as the validation metric. All of our experiments are performed three times and averaged for fair comparison. All the reported performance is directly cited from prior works under the ResNet-50 backbone (Pan et al. 2018, 2019; Choi et al. 2021; Peng et al. 2022).

Existing domain generalized USSS methods are included for comparison, namely, IBN (Pan et al. 2018), IW (Pan et al. 2019), Iternorm (Huang et al. 2019b), DRPC (Yue et al. 2019), ISW (Choi et al. 2021), GTR (Peng et al. 2021), DIRL (Xu et al. 2022), SHADE (Zhao et al. 2022), SAW (Peng et al. 2022), WildNet (Lee et al. 2022), AdvStyle (Zhong et al. 2022), SPC (Huang et al. 2023), and HGFormer (Ding et al. 2023).

Comparison with State-of-the-art

GTA5 Source Domain Table 1 reports the performance on target domains of C, B, M and S, respectively. The proposed CMFormer shows a performance improvement of 10.66%, 9.45%, 14.00% and 12.46% compared to existing state-of-the-art CNN based methods on each target domain, respectively. These outcomes demonstrate the feature generalization ability of the proposed CMFormer. Notice that the source domain GTA5 is a synthetic dataset, while the target domains are real images. It further validates the performance of the proposed method.

Method	Trained on GTA5 (G)			
	→ C	→ B	→ M	→ S
IBN	33.85	32.30	37.75	27.90
IW	29.91	27.48	29.71	27.61
Iternorm	31.81	32.70	33.88	27.07
DRPC	37.42	32.14	34.12	28.06
ISW	36.58	35.20	40.33	28.30
GTR	37.53	33.75	34.52	28.17
DIRL	41.04	39.15	41.60	-
SHADE	44.65	39.28	43.34	-
SAW	39.75	37.34	41.86	30.79
WildNet	44.62	38.42	46.09	31.34
AdvStyle	39.62	35.54	37.00	-
SPC	44.10	40.46	45.51	-
CMFormer (Ours)	**55.31**	**49.91**	**60.09**	**43.80**

Table 1: G → {C, B, M, S} setting. Performance comparison between the proposed CMFormer and existing domain generalized USSS methods. '-': The metric is either not reported or the official source code is not available. Evaluation metric mIoU is given in (%).

SYNTHIA Source Domain Table 2 reports the performance. The proposed CMFormer shows a 5.67%, 8.73% and 11.49% mIoU performance gain against the best CNN based methods, respectively. However, on the BBD-100K (B) dataset, the semantic-aware whitening (SAW) method (Peng et al. 2022) outperforms the proposed CMFormer by 1.80% mIoU. Nevertheless, the proposed CMFormer still outperforms the rest methods. The performance gain of the proposed CMFormer when trained on SYNTHIA dataset is not as significant as it is trained on CityScapes or GTA5 dataset. The explanation may be that the SYNTHIA dataset has much fewer samples than GTA5 dataset, *i.e.*, 9400 v.s. 24966, and a transformer may be under-trained.

CityScapes Source Domain Table 3 reports the performance. As HGFormer only reports one decimal results (Ding et al. 2023), we also report one decimal results when compared with it. The proposed CMFormer (with Swin-Base backbone) shows a performance gain of 6.32%, 10.43%, 9.50% and 12.11% mIoU on the B, M, G and S dataset against the state-of-the-art CNN based method. As BDD100K dataset contains many nigh-time urban-street images, it is particularly challenging for existing domain generalized USSS methods. Still, a performance gain of 6.32% is observed by the proposed CMFormer.

On the other hand, when comparing ours with the contemporary HGFormer with the Swin-Large backbone, it shows an mIoU improvement of 1.1%, 1.5%, 1.3% and 1.7% on the B, M, G and S target domain, respectively.

From Synthetic Domain to Real Domain We also test

Method	Trained on SYNTHIA (S)			
	→ C	→ B	→ M	→ G
IBN	32.04	30.57	32.16	26.90
IW	28.16	27.12	26.31	26.51
DRPC	35.65	31.53	32.74	28.75
ISW	35.83	31.62	30.84	27.68
GTR	36.84	32.02	32.89	28.02
SAW	38.92	**35.24**	34.52	29.16
AdvStyle	37.59	27.45	31.76	-
CMFormer (Ours)	**44.59**	33.44	**43.25**	**40.65**

Table 2: S → {C, B, M, G} setting. Performance comparison between the proposed CMFormer and existing domain generalized USSS methods. '-': The metric is either not reported or the official source code is not available. Evaluation metric mIoU is given in (%).

Method	Backbone	Trained on Cityscapes (C)			
		→ B	→ M	→ G	→ S
IBN	Res50	48.56	57.04	45.06	26.14
IW	Res50	48.49	55.82	44.87	26.10
Iternorm	Res50	49.23	56.26	45.73	25.98
DRPC	Res50	49.86	56.34	45.62	26.58
ISW	Res50	50.73	58.64	45.00	26.20
GTR	Res50	50.75	57.16	45.79	26.47
DIRL	Res50	51.80	-	46.52	26.50
SHADE	Res50	50.95	60.67	48.61	27.62
SAW	Res50	52.95	59.81	47.28	28.32
WildNet	Res50	50.94	58.79	47.01	27.95
HGFormer	Swin-T	53.4	66.9	51.3	33.6
Ours	Swin-B	**59.27**	**71.10**	**58.11**	**40.43**
HGFormer	Swin-L	61.5	72.1	59.4	41.3
Ours	Swin-L	**62.6**	**73.6**	**60.7**	**43.0**

Table 3: C → {B, M, G, S} setting. Performance comparison between the proposed CMFormer and existing domain generalized USSS methods. '-': the metric is either not reported or the official source code is not available. Evaluation metric mIoU is given in (%). †: HGFormer only reports one decimal results (Ding et al. 2023).

the generalization ability of the CMFormer when trained on the synthetic domains (G+S) and validated on the three real-world domains B, C and M, respectively. The results are shown in Table 4. The proposed CMFormer significantly outperforms the instance normalization based (IBN (Pan et al. 2018)), whitening transformation based (ISW (Choi et al. 2021)) and adversarial domain training based (SHADE (Zhao et al. 2022), AdvStyle (Zhong et al. 2022)) methods by >10% mIoU.

From Clear to Adverse Conditions we further validate the proposed CMFormer's performance on the adverse conditions dataset with correspondance (ACDC) (Sakaridis, Dai, and Van Gool 2021). We set the fog, night, rain and snow as four different unseen domains, and directly use the model pre-trained on CityScapes for inference. The results are shown in Table 5. It significantly outperforms existing domain generalized segmentation methods (Pan et al. 2018; Huang et al. 2019a; Pan et al. 2019; Choi et al. 2021; Li et al. 2023) by up to 10.3%, 0.5%, 11.6%, 11.1% on the fog, night, rain and snow domains, respectively.

Backbone	Trained on Two Synthetic Domains (G+S)			
	→ Citys	→ BDD	→ MAP	mean
Res50	35.46	25.09	31.94	30.83
IBN	35.55	32.18	38.09	35.27
ISW	37.69	34.09	38.49	36.75
SHADE	47.43	40.30	47.60	45.11
AdvStyle	39.29	39.26	41.14	39.90
SPC	46.36	43.18	48.23	45.92
Ours	**59.70**	**53.36**	**61.61**	**58.22**

Table 4: Generalization of the proposed CMFormer when trained on two synthetic datasets and generalized on real domains. Evaluation metric mIoU is presented in (%).

Method	Trained on Cityscapes (C)			
	→ Fog	→ Night	→ Rain	→ Snow
IBN	63.8	21.2	50.4	49.6
IW	62.4	21.8	52.4	47.6
ISW	64.3	24.3	56.0	49.8
ISSA	67.5	33.2	55.9	53.2
Ours	**77.8**	**33.7**	**67.6**	**64.3**

Table 5: Generalization of the proposed CMFormer to the adverse condition domains (rain, fog, night and snow) on ACDC dataset (Sakaridis, Dai, and Van Gool 2021).

Content Enhancement			Trained on CityScapes (C)			
×32	×16	×8	→ B	→ M	→ G	→ S
			55.43	66.12	55.05	38.19
✓			56.17	67.55	55.42	38.83
✓	✓		58.10	69.72	55.54	39.41
✓	✓	✓	**59.27**	**71.10**	**58.11**	**40.43**

Table 6: Ablation studies on each component of the proposed CMFormer. ×32, ×16 and ×8 denote the image features of ×32, ×16 and ×8 resolution. ✓ refers to the content enhancement is implemented. Evaluation metric mIoU.

Ablation Studies

On Content-enhancement of Each Resolution Table 6 reports the performance of the proposed CMFormer when ×32, ×16 and ×8 image features are or are not implemented with content enhancement. The content enhancement on a certain resolution feature allows the exploiting of its low-resolution properties. When no image features are implemented with content enhancement, CMFormer degrades into a Mask2Former (Cheng et al. 2022) which only includes the high-resolution properties. When only implementing content enhancement on the ×32 image feature, the down-sampled ×128 image feature may propagate little content information to the segmentation mask, and only a performance gain of 0.74%, 1.43%, 0.37% and 0.64% on B, M, G and S target domain is observed. When further implementing content enhancement on the ×16 image feature,

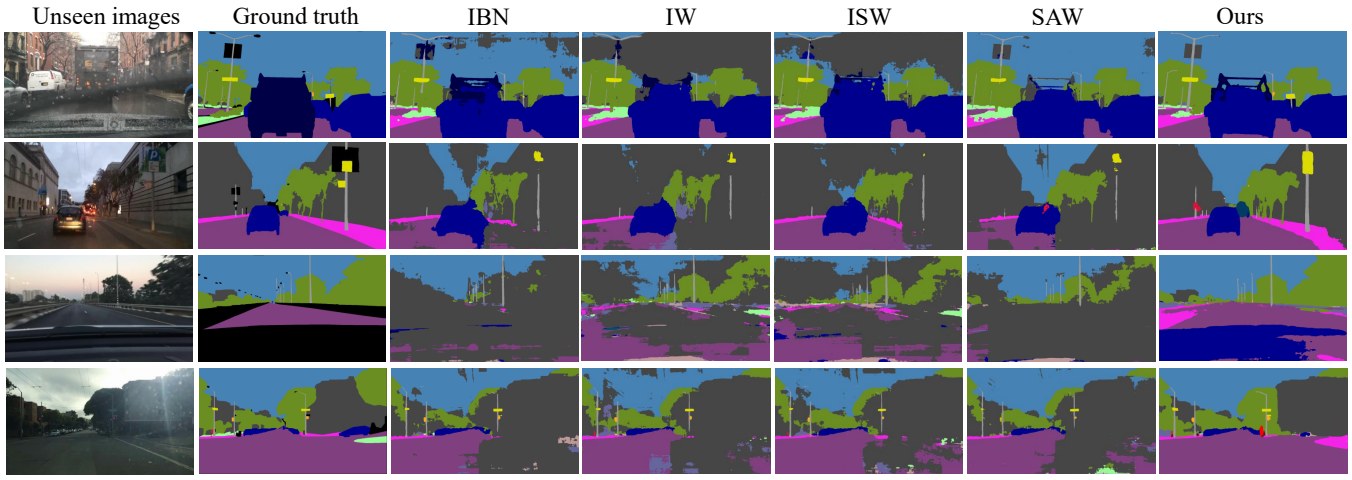


Figure 5: Unseen domain segmentation prediction of existing CNN based domain generalized semantic segmentation methods (IBN (Pan et al. 2018), IW (Pan et al. 2019), ISW (Choi et al. 2021), SAW (Peng et al. 2022)) and the proposed CMFormer under the C $\rightarrow$ B, M, G, S setting.

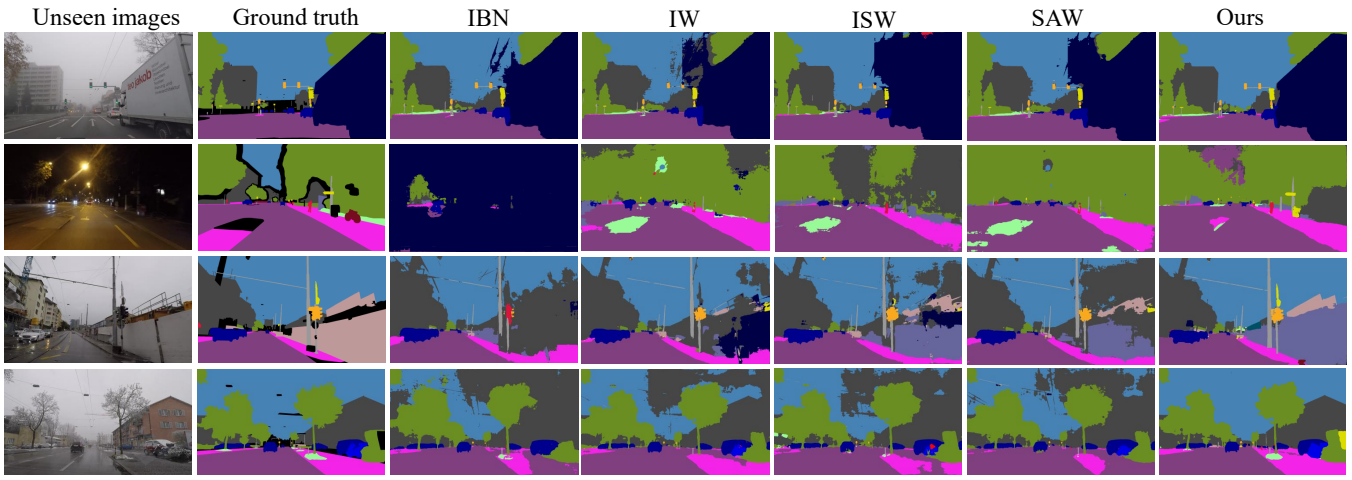


Figure 6: Unseen domain segmentation prediction of existing CNN based domain generalized semantic segmentation methods (IBN (Pan et al. 2018), IW (Pan et al. 2019), ISW (Choi et al. 2021), SAW (Peng et al. 2022)) and the proposed CMFormer under the C $\rightarrow$ adverse domain setting.

the enhanced content information begins to play a role, and an additional performance gain of 1.93%, 2.17%, 0.12% and 0.58% is observed. Then, the content enhancement on the $\times 8$ image feature also demonstrates a significant impact on the generalization ability.

Quantitative Segmentation Results

Some segmentation results on the C $\rightarrow$ B, M, G, S setting and C $\rightarrow$ adverse domain setting are visualized in Fig. 5 and 6. Compared with the CNN based methods, the proposed CMFormer shows a better segmentation prediction, especially in terms of the completeness of objects.

Conclusion

In this paper, we explored the feasibility of adapting the mask Transformer for domain-generalized urban-scene semantic segmentation (USSS). To address the challenges of style variation and robust content representation, we proposed a content-enhanced mask attention (CMA) mechanism. This mechanism is designed to capture more resilient content features while being less sensitive to style variations. Furthermore, we integrate it into a novel framework called the **C**ontent-enhanced **M**ask Trans**Former** (CMFormer). Extensive experiments on multiple settings demonstrated the superior performance of CMFormer compared to existing domain-generalized USSS methods.

Boarder Social Impact. The proposed method has the potential to enhance the accuracy and reliability of semantic segmentation models, thereby contributing to safer and more efficient autonomous systems. Overall, the proposed content-enhanced mask attention mechanism offers promising advancements in domain-generalized USSS.

References

Bi, Q.; You, S.; and Gevers, T. 2023. Interactive Learning of Intrinsic and Extrinsic Properties for All-Day Semantic Segmentation. *IEEE Transactions on Image Processing*, 32: 3821–3835.

Chattopadhyay, P.; Balaji, Y.; and Hoffman, J. 2020. Learning to balance specificity and invariance for in and out of domain generalization. In *European Conference on Computer Vision*, 301–318. Springer.

Chen, W.-T.; Huang, Z.-K.; Tsai, C.-C.; Yang, H.-H.; Ding, J.-J.; and Kuo, S.-Y. 2022. Learning Multiple Adverse Weather Removal via Two-Stage Knowledge Learning and Multi-Contrastive Regularization: Toward a Unified Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17653–17662.

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 (CVPR)*, 1290–1299.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34: 17864–17875.

Choi, S.; Jung, S.; Yun, H.; Kim, J.; 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 (CVPR)*, 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 conference on computer vision and pattern recognition*, 3213–3223.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

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.

Hu, C.; and Lee, G. H. 2022. Feature Representation Learning for Unsupervised Cross-Domain Image Retrieval. In *European Conference on Computer Vision*, 529–544. Springer.

Huang, L.; Zhou, Y.; Zhu, F.; Liu, L.; and Shao, L. 2019a. Iterative Normalization: Beyond Standardization towards Efficient Whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4874–4883.

Huang, L.; Zhou, Y.; Zhu, F.; Liu, L.; and Shao, L. 2019b. Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 4874–4883.

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.

Ji, W.; Li, J.; Bi, Q.; Liu, J.; Cheng, L.; et al. 2022. Promoting Saliency From Depth: Deep Unsupervised RGB-D Saliency Detection. In *International Conference on Learning Representations*.

Ji, W.; Yu, S.; Wu, J.; Ma, K.; Bian, C.; Bi, Q.; Li, J.; Liu, H.; Cheng, L.; and Zheng, Y. 2021. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12341–12351.

Kim, J.; Lee, J.; Park, J.; Min, D.; and Sohn, K. 2022. Pin the memory: Learning to generalize semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4350–4360.

Lambert, J.; Liu, Z.; Sener, O.; Hays, J.; and Koltun, V. 2020. MSeg: A composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2879–2888.

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.

Li, J.; Ji, W.; Bi, Q.; Yan, C.; Zhang, M.; Piao, Y.; Lu, H.; et al. 2021. Joint semantic mining for weakly supervised RGB-D salient object detection. *Advances in Neural Information Processing Systems*, 34: 11945–11959.

Li, M.; Namkoong, H.; and Xia, S. 2021. Evaluating model performance under worst-case subpopulations. *Advances in Neural Information Processing Systems*, 34: 17325–17334.

Li, Y.; Zhang, D.; Keuper, M.; and Khoreva, A. 2023. Intra-Source Style Augmentation for Improved Domain Generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 509–519.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 10012–10022.

Mahajan, D.; Tople, S.; and Sharma, A. 2021. Domain generalization using causal matching. In *International Conference on Machine Learning*, 7313–7324. PMLR.

Mirza, M. J.; Masana, M.; Possegger, H.; and Bischof, H. 2022. An Efficient Domain-Incremental Learning Approach to Drive in All Weather Conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3001–3011.

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.

Pan, J.; Bi, Q.; Yang, Y.; Zhu, P.; and Bian, C. 2022. Label-efficient hybrid-supervised learning for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2026–2034.

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 (ECCV)*, 464–479.

Pan, X.; Zhan, X.; Shi, J.; Tang, X.; and Luo, P. 2019. Switchable Whitening for Deep Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1863–1871.

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.

Peng, X.; Qiao, F.; and Zhao, L. 2022. Out-of-Domain Generalization From a Single Source: An Uncertainty Quantification Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Piva, F. J.; de Geus, D.; and Dubbelman, G. 2023. Empirical Generalization Study: Unsupervised Domain Adaptation vs. Domain Generalization Methods for Semantic Segmentation in the Wild. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 499–508.

Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12556–12565.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *European conference on computer vision*, 102–118. Springer.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3234–3243.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2021. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10765–10775.

Segu, M.; Tonioni, A.; and Tombari, F. 2023. Batch normalization embeddings for deep domain generalization. *Pattern Recognition*, 135: 109115.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7262–7272.

Tjio, G.; Liu, P.; Zhou, J. T.; and Goh, R. S. M. 2022. Adversarial semantic hallucination for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 318–327.

Volpi, R.; Namkoong, H.; Sener, O.; Duchi, J. C.; Murino, V.; and Savarese, S. 2018. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31.

Wang, S.; Yu, L.; Li, C.; Fu, C.-W.; and Heng, P.-A. 2020. Learning from extrinsic and intrinsic supervisions for domain generalization. In *European Conference on Computer Vision*, 159–176. Springer.

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.

Ye, Q.; Shen, X.; Gao, Y.; Wang, Z.; Bi, Q.; Li, P.; and Yang, G. 2021. Temporal cue guided video highlight detection with low-rank audio-visual fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7950–7959.

Yu, F.; Xian, W.; Chen, Y.; Liu, F.; Liao, M.; Madhavan, V.; and Darrell, T. 2018. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5): 6.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2100–2110.

Zhao, S.; Gong, M.; Liu, T.; Fu, H.; and Tao, D. 2020. Domain generalization via entropy regularization. *Advances in Neural Information Processing Systems*, 33: 16096–16107.

Zhao, Y.; Zhong, Z.; Zhao, N.; Sebe, N.; and Lee, G. H. 2022. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, 535–552. Springer.

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*.

Zhou, B.; Yi, J.; and Bi, Q. 2021. Differential convolution feature guided deep multi-scale multiple instance learning for aerial scene classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4595–4599.

Zhou, K.; Yang, Y.; Hospedales, T.; and Xiang, T. 2020. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, 561–578. Springer.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*.