Learning Generalized Segmentation for Foggy-Scenes by Bi-directional Wavelet Guidance

Qi Bi, Shaodi You, Theo Gevers

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

Abstract

Learning scene semantics that can be well generalized to foggy conditions is important for safety-crucial applications such as autonomous driving. Existing methods need both annotated clear images and foggy images to train a curriculum domain adaptation model. Unfortunately, these methods can only generalize to the target foggy domain that has seen in the training stage, but the foggy domains vary a lot in both urban-scene styles and fog styles. In this paper, we propose to learn scene segmentation well generalized to foggy-scenes under the domain generalization setting, which does not involve any foggy images in the training stage and can generalize to any arbitrary unseen foggy scenes. We argue that an ideal segmentation model that can be well generalized to foggy-scenes need to simultaneously enhance the content, de-correlate the urban-scene style and de-correlate the fog style. As the content (e.g., scene semantic) rests more in low-frequency features while the style of urban-scene and fog rests more in high-frequency features, we propose a novel bi-directional wavelet guidance (BWG) mechanism to realize the above three objectives in a divide-and-conquer manner. With the aid of Haar wavelet transformation, the low frequency component is concentrated on the content enhancement self-attention, while the high frequency component is shifted to the style and fog self-attention for de-correlation purpose. It is integrated into existing mask-level Transformer segmentation pipelines in a learnable fashion. Large-scale experiments are conducted on four foggy-scene segmentation datasets under a variety of interesting settings. The proposed method significantly outperforms existing directly-supervised, curriculum domain adaptation and domain generalization segmentation methods. Source code is available at https://github.com/BiQiWHU/BWG.

Introduction

Existing foggy-scene semantic segmentation methods usually follow the curriculum domain adaptation paradigm, where both well-annotated clear images and foggy images are involved in the training stage (Truong et al. 2021; Guo et al. 2021; Zhang et al. 2021), so that the scene representation can be progressively adapted to the target foggy domain that has been seen in the training stage (Tsai et al. 2018). Nonetheless, these techniques are solely tailored to adapt to foggy scenes within the target domain, which impose a significant limitation on practical road applications. In real-world scenarios, the need for generalizing to a wide array of unforeseen foggy scenes is important.

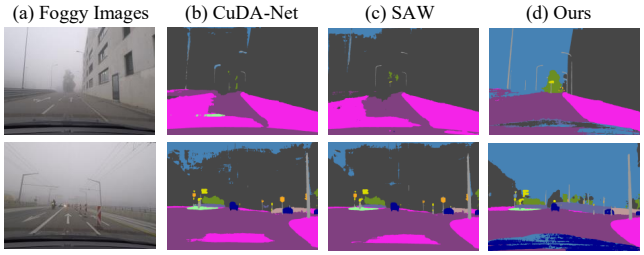

Figure 1: Foggy image (a) segmentation results by: (b) existing curriculum domain adaptation paradigm (e.g. CuDA-Net (Ma et al. 2022)); (c) generic domain generalization paradigm (e.g., SAW (Peng et al. 2022)); (d) our proposed generalization for foggy-scene method BWG.

In this paper, we shift the focus to the domain generalization setting, which does not involve any foggy target domain in the training stage. Ideally, such scene-segmentation model is able to generalize to any unseen foggy target domains. Predicting reliable scene-segmentation under domain generalization setting is plausible, given the effectiveness demonstrated by various domain-generalized segmentation methods in recent years. These methods usually assume that the content is stable while the urban style changes greatly (Choi et al. 2021; Peng et al. 2022; Bi, You, and Gevers 2023a).

However, the challenge becomes more intricate when attempting to generalize to foggy scenes due to the complex nature of image conditions. This complexity can pose difficulties for existing domain-generalized scene segmentation methods (Fig. 1c). Specifically, not only the urban-scene styles but also the foggy styles vary greatly (Ma et al. 2022). Besides, the existence of fog occludes to the scene objects and negatively impairs the content representation (Dai et al. 2020; Wang et al. 2023).

We focus on three key objectives to learn a scene segmentation that can be well generalized to foggy scenes (Ma et al. 2022): 1) decouple the urban-style variation, 2) decouple the foggy style variation, and 3) enhance the content representation caused by fog occlusion. Addressing each of these

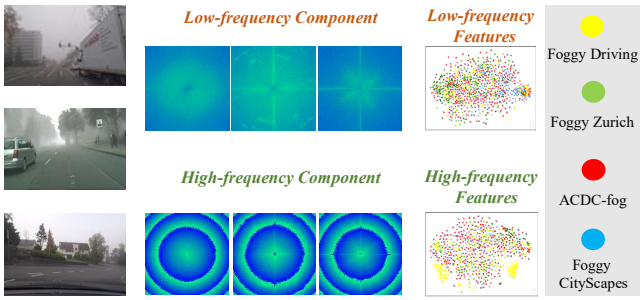

Figure 2: (a) Foggy scenes from different domains. (b) Visualization of low- and high- frequency space, which rests in more content and style information, respectively. (c) t-SNE visualization of low- and high- frequency feature space.

three objectives through a divide-and-conquer approach is straightforward. The pivotal concern lies in devising a viable solution to distinguish urban scene styles and foggy styles from the core urban scene content.

In this paper, we propose to separate these components in a foggy-scene from the frequency domain by the Haar wavelet transformation (Porwik and Lisowska 2004). Handling the style and content separately in the frequency domain has been recognized effective (Yoo et al. 2019; Li et al. 2017). When considering an image representation, the content (such as scene semantics) tends to reside predominantly in the low-frequency components, whereas the style (such as urban landscape, lighting, and weather) is more prominent in the high-frequency components (Bi, You, and Gevers 2023b; Peng et al. 2022; Tjio et al. 2022).

Technically, we propose a bi-directional wavelet-guided guidance (BWG) for this task (Fig. 1d). First, we represent the content, fog style and urban scene style by three independent self-attention modules. For each module, the Haar wavelet transformation allows us to decompose the high-frequency component from low-frequency component. Then, we concentrate all the low-frequency component to the content enhancement module, and shift all the high-frequency components to the fog-style and urban-style module. Afterwards, both high-frequency representations are implemented with instance normalization to decouple the impact of urban-style and fog-style variation.

Extensive experiments are conducted to generalize to foggy scenes. Using CityScapes (Cordts et al. 2016) as source domain, the proposed method is compared with existing domain generalized and domain adaptation methods on four foggy benchmarks, namely, ACDC-fog (Sakaridis, Dai, and Van Gool 2021), Foggy-Zurich (Sakaridis et al. 2018), Foggy-Driving (Sakaridis, Dai, and Van Gool 2018) and Foggy-CityScapes (Sakaridis, Dai, and Van Gool 2018). Besides, some state-of-the-art directly-supervised and foundation model based segmentation methods (Kirillov et al. 2023; Wang et al. 2023) are also compared for reference. Rigorous ablation studies and generalization to other adverse weather conditions are also validated.

Our contribution can be summarized as follows.

- We propose to learn segmentation generalizable to foggy scenes under the domain generalization setting. It is more practical, general and applicable to real-world scenarios than prior curriculum domain adaptation works.
- We propose a bi-directional wavelet guided self-attention (BWG) mechanism. It handles the content enhancement, urban-style de-correlation and fog de-correlation in a divide-and-conquer manner.
- The proposed BWG is integrated into the mask-level Transformer segmentation models in a learnable fashion.
- The proposed BWG outperforms existing state-of-the-art domain generalized segmentation methods by upto 11.8% mIoU on Foggy Zurich and curriculum domain adaptation methods by upto 16.7% mIoU on ACDC-fog.

Related Work

Foggy-scene Semantic Segmentation has been extensively studied. Existing works tackle this problem under the paradigm of curriculum domain adaptation, which uses the clear images and foggy images as source domain and target domain, respectively. Some typical works include AdSegNet (Tsai et al. 2018), ADVENT (Vu et al. 2019), DISE (Chang et al. 2019), CCM (Li et al. 2020), SAC (Araslanov and Roth 2021), ProDA (Zhang et al. 2021), DMLC (Guo et al. 2021), DACS (Truong et al. 2021), CMAda3+ (Dai et al. 2020) and CuDA-Net (Ma et al. 2022). On the other hand, although some recent unsupervised domain adaptation techniques (e.g. DAFormer (Hoyer, Dai, and Van Gool 2022), Refign-DAFormer (Brüggemann et al. 2023)) have been proposed, they are not specially designed for foggy scenes.

Fog Removal enhances visibility. Earlier de-fog works model the degraded image as a combination between the background image and weather effect layer (Li, Cheong, and Tan 2019; Li, Tan, and Cheong 2020). More recent works follow the all-in-one paradigm (Valanarasu, Yasarla, and Patel 2022; Yang et al. 2023). However, semantic segmentation on de-fogged images still shows an significant inferior performance compared with the curriculum domain adaptation methods (Dai et al. 2020; Ma et al. 2022).

Domain Generalized Semantic Segmentation (Pan et al. 2019; Choi et al. 2021; Peng et al. 2022; Huang et al. 2023; Tjio et al. 2022; Lee et al. 2022; Ding et al. 2023; Bi, You, and Gevers 2023b) is more challenging than conventional semantic segmentation (Pan et al. 2022; Ji et al. 2021; Li et al. 2021; Ji et al. 2022), which focuses on the generalization ability of a segmentation model on unseen target domains. These methods usually assume the content is stable and the domain gap is caused by the style variation of urban landscape. However, this assumption does not fully describe the complexity of foggy-scene formulation. Despite the style variation caused by urban landscape, the foggy style also varies a lot. More importantly, the fog poses severe occlusion, which harms the completeness of content information.

Preliminary

Generalized Segmentation for Foggy Scenes Given clean scenes as source domain $\mathcal{S}$, and foggy scenes as an un-

seen target domain $\mathcal{T}$. Given a semantic segmentation model parameterized by θ and the segmentation loss $\mathcal{L}_{seg}$, generalized segmentation for foggy scenes can be formulated as

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

where $D(\mathcal{S},\mathcal{T})$ denotes the distance between the clean-scene source domain $\mathcal{S}$ and foggy-scene target domain $\mathcal{T}$, and ρ denotes the constraint threshold.

Theoretical Analysis from Frequency Domain Handling the style and content separately in the frequency domain has been recognized effective (Yoo et al. 2019; Li et al. 2017). We analyze the low-frequency and high- frequency component (Fig. 2b) from different foggy target domains (Fig. 2a). After transforming the low- and high- frequency component into the spatial space, the t-SNE visualization shows that the low-frequency features are more robust to handle style variations than high-frequency features. Cross-domain samples are more uniformly distributed when using low-frequency features (Fig. 2c).

Haar Wavelet Transformation Haar wavelet pooling (Porwik and Lisowska 2004) enables the separation from the low-frequency component to high-component. It has four kernels, namely, LL^{T}, LH^{T}, HL^{T}, HH^{T}, given by

$$L^{\mathrm{T}} = \frac{1}{\sqrt{2}}[1 \quad 1], H^{\mathrm{T}} = \frac{1}{\sqrt{2}}[-1 \quad 1]. \quad (2)$$

The low-frequency component LL preserves more content information (e.g. scene semantics) for foggy scenes. Instead, the high-frequency components LH, HL and HH contain more style information (e.g. urban landscape, foggy density) for foggy scenes.

Difference from Existing Pipelines Fig. 3a summarizes the pipeline of existing foggy-scene segmentation methods under the domain adaptation setting. Both clear source domain and fog target domain are involved in training.

Fig. 3b outlines the workflow of generic domain generalized segmentation. During training, only the clear source domain is utilized. While these methods demonstrate the ability to generalize to diverse, unseen target domains, their emphasis lies primarily in decoupling urban styles. They are not explicitly tailored to represent foggy scenes.

Fig. 3c summarizes the proposed pipeline, which intends to learn generalized scene-segmentation for foggy-scenes. It only involves clear source domain in the training stage. The proposed framework implements urban style decoupling, content enhancement and foggy style decoupling.

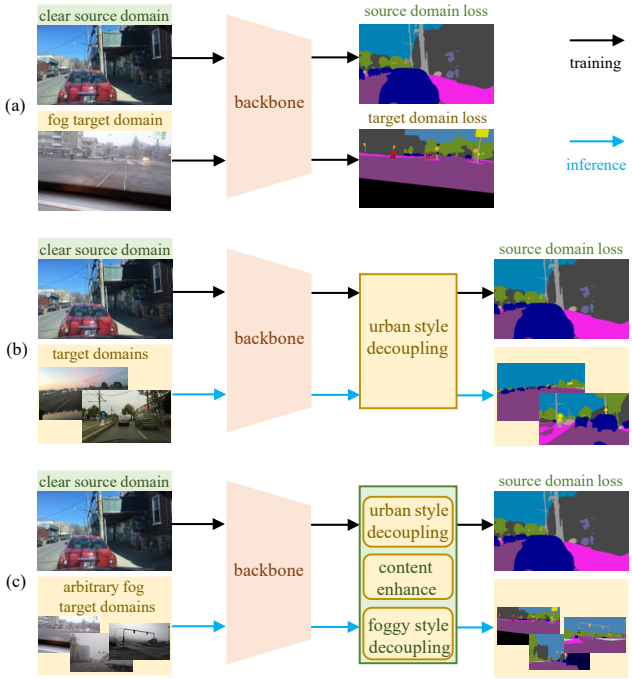

Figure 3: Difference of foggy-scene segmentation pipelines between: (a) existing curriculum domain adaptation setting; (b) generic domain generalization setting; and (c) the proposed generalized segmentation to foggy-scene setting.

Methodology

Triplet Self-attention Representing

Three key objectives to learn segmentation that can be well generalized to foggy-scenes are content enhancement, fog-style decoupling and urban-style decoupling. It is intuitive to realize these objectives in a divide-and-conquer manner. So, we use three self-attention modules to represent the content, fog style and urban style, respectively. The self-attention mechanism is adapted in our framework not only because its strong representation ability and its long-range dependency mining, but also because it can seamlessly integrated to Transformer segmentation backbones.

Before our triplet self-attention representing, we use the mask attention to encode the foggy-scene features from backbone. Compared with conventional pixel-level segmentation methods, mask attention based segmentation (Cheng et al. 2022; Cheng, Schwing, and Kirillov 2021) has stronger scene representation. Given the image feature $\mathbf{F}_l \in \mathbb{R}^{(W_l \cdot H_l) \times C_F}$ to input into the l^{th} layer of a Transformer decoder, its key, value, and query counterpart $\mathbf{K}_l \in \mathbb{R}^{(W_l \cdot H_l) \times C}$, $\mathbf{V}_l \in \mathbb{R}^{(W_l \cdot H_l) \times C}$ and $\mathbf{Q}_l \in \mathbb{R}^{N \times C}$ can be computed by linear transformations f_K, f_V and f_Q, respectively. Then, the mask attention computes the features $\mathbf{X}_l \in \mathbb{R}^{N \times C}$, given by

$$\mathbf{X}_l = \mathrm{softmax}(\mathcal{M}_{l-1} + \mathbf{Q}_l \mathbf{K}_l^{\mathsf{T}})\mathbf{V}_l + \mathbf{X}_{l-1}, \quad (3)$$

where $\mathcal{M}_{l-1} \in \{0,1\}^{N \times H_l W_l}$ is a binary mask attention matrix from the $(l-1)^{th}$ layer, with a threshold of 0.5. $\mathcal{M}_0$ is binarized and resized from $\mathbf{X}_0$. The mask can highlight the foreground regions and suppress the background of an image, which has been reported effective to enhance the feature representation for scene segmentation (Cheng et al. 2022; Cheng, Schwing, and Kirillov 2021).

Then, the learnt mask query $\mathbf{X}_l$ is fed into three parallel self-attention components S, C and F to decouple the urban-style, enhance the content and decouple the fog-style. The output is denoted as $\mathbf{X}_l^s$, $\mathbf{X}_l^c$ and $\mathbf{X}_l^f$, respectively.

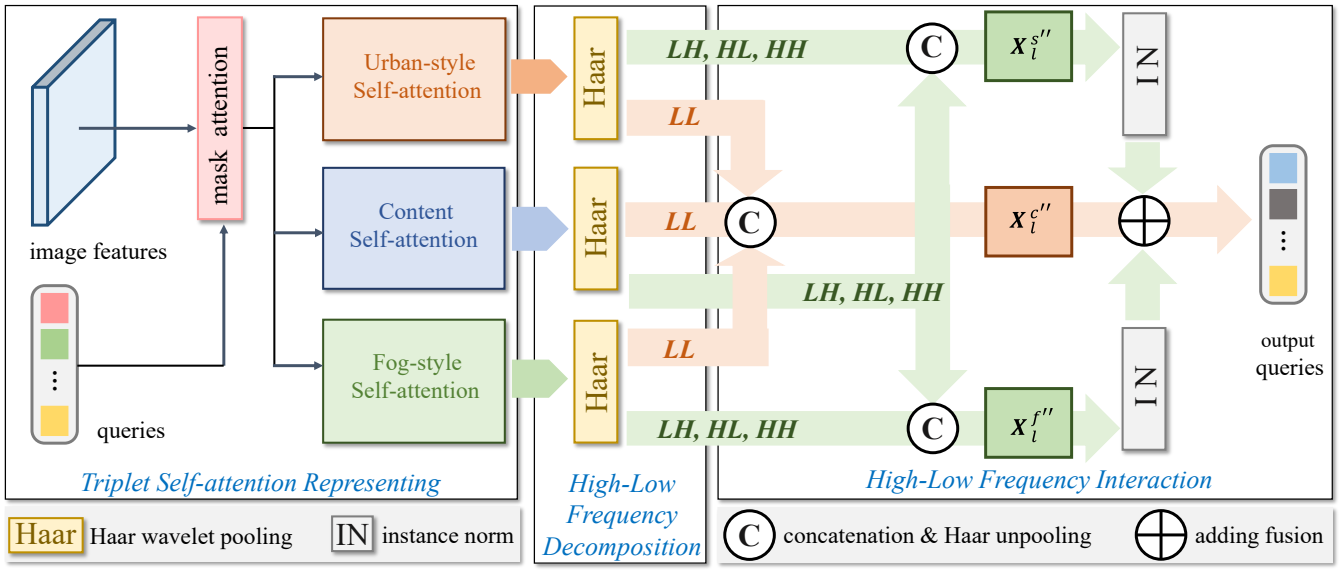

Figure 4: Technique framework overview. The proposed bi-directional wavelet guided self-attention (BWG) implements content enhancement, style de-correlation and fog density de-correlation. Four low- and high- frequency components from the Haar wavelet transformation are denoted as LL, LH, HL and HH, respectively.

High-Low Frequency Decomposition

A nature divide-and-conquer way to learn segmentation that is well generalized to foggy scenes is to let the content enhancement representation $\mathbf{X}_l^c$ focus the low-frequency component, and to let the foggy and urban-scene representation $\mathbf{X}_l^s$ and $\mathbf{X}_l^f$ focus on the high-frequency component.

To realize this objective, it is necessary to at first separate the low frequency component from the high frequency component for $\mathbf{X}_l^c$, $\mathbf{X}_l^s$ and $\mathbf{X}_l^f$, respectively. The Haar wavelet transformation allows us to decompose $\mathbf{X}_l^s$ into one low-pass component LL and three high-pass component LH, HL, HH. Take $\mathbf{X}_l^s$ as an example, the decomposition is:

$$\mathbf{X}_l^{s,LL} = \mathbf{X}_l^s \otimes LL^{\mathrm{T}}, \tag{4}$$

$$\mathbf{X}_l^{s,LH} = \mathbf{X}_l^s \otimes LH^{\mathrm{T}}, \tag{5}$$

$$\mathbf{X}_l^{s,HL} = \mathbf{X}_l^s \otimes HL^{\mathrm{T}}, \tag{6}$$

$$\mathbf{X}_l^{s,HH} = \mathbf{X}_l^s \otimes HH^{\mathrm{T}}, \tag{7}$$

where $\otimes$ denotes the filter operation.

For $\mathbf{X}_l^c$ and $\mathbf{X}_l^f$, similarly we can get $\mathbf{X}_l^{c,LL}$, $\mathbf{X}_l^{c,LH}$, $\mathbf{X}_l^{c,HL}$, $\mathbf{X}_l^{c,HH}$ and $\mathbf{X}_l^{f,LL}$, $\mathbf{X}_l^{f,LH}$, $\mathbf{X}_l^{f,HL}$, $\mathbf{X}_l^{f,HH}$. For simplicity and clarity, in this paper, we directly use the notations of spatial domain to state the operations in frequency domain, which avoids to involve complicated notations and equations in spatial-frequency transformation.

High-Low Frequency Interaction

After the decomposition of high and low frequency information, the rest step is to: 1) allow the content enhancement branch to only focus on the low-frequency component, so that the scene semantics are more well-represented; 2) allow the fog branch and urban scene style branch to focus on the high-frequency component, so that the foggy style information and urban-scene style information is more depicted.

For the content enhancement branch, all the low frequency components from the other two branches, namely, $\mathbf{X}_l^{s,LL}$ and $\mathbf{X}_l^{f,LL}$, are merged together with its original low frequency component, given by

$$\mathbf{X}_l^{c'} = [\mathbf{X}_l^{c,LL}, \mathbf{X}_l^{c,LL}, \mathbf{X}_l^{s,LL}, \mathbf{X}_l^{f,LL}], \tag{8}$$

where $[\cdot,\cdot]$ denotes the concatenation operation followed by Haar wavelet unpooling.

For the style de-correlation branch, after shifting its low frequency component $\mathbf{X}_l^{s,LL}$ to the content branch, the high frequency components from the content branch ($\mathbf{X}_l^{c,LH}$, $\mathbf{X}_l^{c,HL}$, $\mathbf{X}_l^{c,HH}$) are fused into it, given by

$$\mathbf{X}_l^{s'} = [\mathbf{X}_l^{s,LH}, \mathbf{X}_l^{s,HL}, \mathbf{X}_l^{s,HH}, \mathbb{E}[\mathbf{X}_l^{c,LH}, \mathbf{X}_l^{c,HL}, \mathbf{X}_l^{c,HH}]]. \tag{9}$$

The implementation on the foggy branch is similar. After shifting its low frequency component $\mathbf{X}_l^{f,LL}$ to the content branch, the high frequency components from the content branch ($\mathbf{X}_l^{c,LH}$, $\mathbf{X}_l^{c,HL}$, $\mathbf{X}_l^{c,HH}$) are fused into it, given by

$$\mathbf{X}_l^{f'} = [\mathbf{X}_l^{f,LH}, \mathbf{X}_l^{f,HL}, \mathbf{X}_l^{f,HH}, \mathbb{E}[\mathbf{X}_l^{c,LH}, \mathbf{X}_l^{c,HL}, \mathbf{X}_l^{c,HH}]]. \tag{10}$$

Finally, the high frequency representation $\mathbf{X}_l^{s'}$ and $\mathbf{X}_l^{f'}$ are implemented with the instance normalization, which has been reported effective to decouple the impact of styles. In this way, the segmentation representation can be more robust to the variance of fog and urban-scene landscape. Take $\mathbf{X}_l^{s'} \in \mathbb{R}^{N\times C}$ as an example, the instance normalization is implemented through channel-wise, given by

$$\mathbf{X}_{l,N,c}^{s''} = \frac{\mathbf{X}_{l,N,c}^{s'} - \mu}{\sigma + \epsilon} \cdot \gamma + \beta, \tag{11}$$

$$\mu = \frac{1}{C}\sum_{c=1}^{C}\mathbf{X}_{l,N,c}^{s'}, \sigma = \sqrt{\frac{1}{C}\sum_{i=1}^{C}(\mathbf{X}_{l,N,c}^{s'} - \mu)^2}, \quad (12)$$

where $c = 1, 2, \cdots, C$.

For the l^{th} transformer layer, the normalized $\mathbf{X}_l^{s''}$, $\mathbf{X}_l^{f''}$ and $\mathbf{X}_l^{c'}$ are fused together by adding for the rest processing.

Framework Overview & Implementation Details

Fig. 4 gives an overview of the proposed framework. The overall framework follows the mask-level segmentation Transformer paradigm (Cheng et al. 2022; Cheng, Schwing, and Kirillov 2021). The image encoder uses a backbone of Swin-base Transformer (Liu et al. 2021), with a pre-trained weight on ImageNet. The image decoder is directly inherited from the Mask2Former (Cheng et al. 2022), where the image features are up-sampled from $\times 32$ resolution to $\times 16$, $\times 8$ and $\times 4$ resolution, respectively.

For the Transformer decoder, it takes the $\times 32$, $\times 16$ and $\times 8$ resolution image features from the decoder as input in a progressive way, which is of the same way as the original Mask2Former (Cheng et al. 2022). The Transformer decoder has nine of the proposed bi-directional wavelet guided self-attention (BWG) components. Finally, the learnt Transformer queries from the Transformer decoder are fused with the $\times 4$ resolution image features for predictions.

All the loss and hyper-parameter settings keep the same as the original Mask2Former (Cheng et al. 2022) without any additional fine-tuning. By default, 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.

Experiments and Analysis

Datasets

CityScapes (Cordts et al. 2016) is a commonly-used semantic segmentation datasets for driving-scenes. It has 2965 training samples and 500 validation samples, with 19 common scene categories in driving-scenes.

Clear CityScapes (Sakaridis, Dai, and Van Gool 2018) is a subset of CityScapes. It consists of 498 training samples from the clear condition.

Foggy-CityScapes (Sakaridis, Dai, and Van Gool 2018) contains 550 synthetic foggy images in total, including 498 training images and 52 testing images. Each of the 498 training samples has three different types of synthetic fog layers, with light-, medium- and dense- density.

Foggy Zurich (Sakaridis et al. 2018) contains 3,808 real-world foggy road scenes from the Zurich city. For light and medium foggy conditions, it has 1,552 images and 1,498 images, respectively. In addition, it has 40 images with labels that are compatible with Cityscapes.

Foggy Driving (Sakaridis, Dai, and Van Gool 2018) has 101 real-world foggy road-scenes images. Among them, 33 images are finely annotated and the rest 68 images are coarsely annotated. Following (Sakaridis, Dai, and Van Gool 2018), they are only used for testing.

Adverse Conditions Dataset with Correspondences (ACDC) (Sakaridis, Dai, and Van Gool 2021) has 4006 driving-scene segmentation samples under adverse conditions. 1000 of them are foggy images. Data split for training, validation and testing is 4:1:5.

Comparison with Domain Generalization Methods

The proposed method is compared with state-of-the-art domain generalized segmentation methods, including IBNet (Pan et al. 2018), Iternorm (Huang et al. 2019), SW (Pan et al. 2019), ISW (Choi et al. 2021), SHADE (Zhao et al. 2022), SAW (Peng et al. 2022), WildNet (Lee et al. 2022), SPC (Huang et al. 2023) and HGFormer (Ding et al. 2023). DIRL (Xu et al. 2022) and AdvStyle (Zhong et al. 2022) are not involved for comparison due to neither unavailable source code nor official performance report. In addition, three directly-supervised segmentation methods (RefineNet (Lin et al. 2017), SegFormer (Xie et al. 2021), Mask2Former (Cheng et al. 2022)) and two recent foundational model based segmentation methods (SAM-fine-tune (Kirillov et al. 2023), SAM-SSA-fine-tune (Wang et al. 2023)) are reported.

Following the evaluation protocols of above domain generalized segmentation methods, CityScapes, under the clear imaging condition, is used as the source domain. Four foggy datasets, namely, Foggy-CityScapes, Foggy Zurich, Foggy Driving and ACDC-fog, are used as unseen target domains for only inference stage.

Table 1 reports the performance. The proposed method outperforms the second-best by 6.8%, 11.8%, 7.2% and 9.6% on ACDC-fog, Foggy Zurich, Foggy Driving and Foggy-CityScapes, respectively. Besides, compared with the original Mask2Former, the proposed method leads to a performance gain of 3.4%, 1.9%, 3.1% and 3.6% on ACDC-fog, Foggy Zurich, Foggy Driving and Foggy-CityScapes.

Comparison with Domain Adaptation Methods

The proposed method is also compared with existing foggy-scene segmentation methods which are under the curriculum domain adaptation paradigm, namely, AdSegNet (Tsai et al. 2018), ADVENT (Vu et al. 2019), DISE (Chang et al. 2019), CCM (Li et al. 2020), SAC (Araslanov and Roth 2021), ProDA (Zhang et al. 2021), DMLC (Guo et al. 2021), DACS (Truong et al. 2021), CMAda3+ (Dai et al. 2020), CuDA-Net (Ma et al. 2022), FIFO (Lee, Son, and Kwak 2022) and DAFormer (Hoyer, Dai, and Van Gool 2022).

Following the evaluation protocols of the above curriculum domain adaptation methods, Clear-CityScapes, which has 498 samples under the clear condition, is used as the source domain. For our BWG, ACDC-fog, Foggy-Zurich and Foggy-driving are used as unseen target domains in inference stage only. These domain adaptation methods take more advantage as they need Clear Zurich (CZ) and Foggy Zurich (FZ) as additional training data.

Table 2 reports the performance. On ACDC-fog, it significantly outperforms existing cumulative domain adaptation methods by at least 12.2% mIoU. Also, it outperforms all these methods on Foggy-Zurich, e.g., 1.0% mIoU gain against CuDA-Net, 4.8% mIoU gain against DAFormer. On Foggy-driving, it outperforms all the methods except CMAda3+ (Dai et al. 2020), CuDA-Net (Ma et al. 2022) and CumFormer (Wang et al. 2023).

Method	Backbone	Source Domain: Cityscapes (2965 images)			
		→ ACDC-Fog	→ Foggy Zurich	→ Foggy-driving	→ Foggy-CityScapes
RefineNet (Lin et al. 2017)	Res-101	46.4	34.6	35.8	-
SAM-fine-tune (Kirillov et al. 2023)	ViT-L	41.7	37.2	38.5	-
SAM-SSA (Wang et al. 2023)	ViT-L	46.5	35.8	50.9	-
SegFormer (Xie et al. 2021)	MiT-B2	59.2	43.9	46.6	75.5
Mask2Former (Cheng et al. 2022)	Swin-B	73.3	49.4	51.1	73.8
IBNet (Pan et al. 2018)	Res-50	63.8	33.4	45.5	66.5
Iternorm (Huang et al. 2019)	Res-50	63.3	35.2	44.6	66.9
SW (Pan et al. 2019)	Res-50	62.4	34.1	45.8	66.4
ISW (Choi et al. 2021)	Res-50	64.3	36.1	46.2	66.6
SHADE (Zhao et al. 2022)	Res-50	61.4	39.5	42.0	65.8
SAW (Peng et al. 2022)	Res-50	64.0	37.3	47.0	67.8
WildNet (Lee et al. 2022)	Res-50	64.7	39.2	42.6	64.4
SPC (Huang et al. 2023)	Res-50	68.0	39.3	43.5	64.7
HGFormer (Ding et al. 2023)	Swin-L	69.9	-	-	-
ISSA (Li et al. 2023)	MiT-B2	67.5	-	-	-
Ours	Swin-B	**76.7**(+6.8)	**51.3**(+11.8)	**54.2**(+7.2)	**77.4**(+9.6)

Table 1: Comparison with existing domain generalized segmentation methods and directly-supervised methods. Evaluation metric mIoU is in %. '-': either no official code or no performance report.

Method	Backbone	Addition Data			Source Domain: Clear-Cityscapes (498 images)		
		CZ	FZ	Oth.	→ ACDC-Fog	→ Foggy Zurich	→ Foggy-driving
AdSegNet (Tsai et al. 2018)	Res-101	✓	✓	✗	31.8	26.1	37.6
ADVENT (Vu et al. 2019)	Res-101	✓	✓	✗	32.9	24.5	36.1
DISE (Chang et al. 2019)	Res-101	✓	✓	✗	42.4	40.7	45.2
CCM (Li et al. 2020)	Res-101	✓	✓	✗	-	35.8	42.6
SAC (Araslanov and Roth 2021)	Res-101	✓	✓	✗	-	37.0	43.4
ProDA (Zhang et al. 2021)	Res-101	✓	✓	✗	38.4	37.8	41.2
DMLC (Guo et al. 2021)	Res-101	✓	✓	✗	-	33.5	32.6
DACS (Truong et al. 2021)	Res-101	✓	✓	✗	-	28.7	35.0
CMAda3+ (Dai et al. 2020)	RefineNet	✓	✓	✓	-	46.8	49.8
FIFO (Lee, Son, and Kwak 2022)	RefineNet	✓	✓	✓	54.1	48.4	50.7
CuDA-Net (Ma et al. 2022)	Res-101	✓	✓	✗	55.6	48.2	52.7
DAFormer (Hoyer, Dai, and Van Gool 2022)	MiT-B5	✓	✓	✗	48.9	44.4	-
CumFormer (Wang et al. 2023)	MiT-B5	✓	✓	✗	60.7	-	56.2
Ours	Swin-B	✗	✗	✗	**72.9**	**49.2**	46.9
		✓	✗	✗	**73.7**	**50.7**	49.3
		✓	✓	✗	**74.3**	N/A	52.3
		✓	✓	✓	**77.4**(+16.7)	N/A	**57.6**(+1.4)

Table 2: Comparison with foggy-scene cumulative domain adaptation methods. Evaluation metric mIoU is in %. '-': either no official code or no performance report. N/A: the result is not meaningful under our domain generalization setting.

Components			Trained on CityScapes	
C	S	F	→ ACDC-Fog	→ Foggy-driving
✓			73.4	51.1
✓	✓		75.2	52.9
✓	✓	✓	76.7	54.2

Table 3: Ablation studies on the content, style and foggy encoder C, S, F in BWG. Evaluation metric mIoU is in %.

S2C	F2C	C2S	C2F	ACDC-Fog	Foggy-driving
				73.4	51.1
✓				74.4	52.5
✓	✓			75.3	53.6
✓	✓	✓		76.1	53.9
✓	✓	✓	✓	76.7	54.2

Table 4: Ablation studies on each frequency interaction. $S2C$ and $F2C$: shift LL from S and F to C. $C2S$ and $C2F$: shift HL, LH and HH from C to S and F. Metric mIoU.

Ablation Studies

On Each Branch The proposed BWG has three encoders to handle the content enhancement, style de-correlation and fog de-correlation, which we denote as C, S and F, respectively. Table 3 reports the impact of each encoder on the generalization performance. When there are both C and S encoders, to keep only one single variable, the wavelet transformations are kept in this experiment setting. The style de-correlation encoder and the fog de-correlation encoder contribute to 1.8% and 1.5% mIoU gain on ACDC-fog, 1.8% and 1.3% mIoU gain on Foggy-driving.

On Low-high Frequency Interaction The proposed BWG has four operations to interact the low and high fre-

Method	Category	Backbone	Trained on Cityscapes (C) & Inferred on Foggy-CityScapes (FC)			
			→ FC-light	→ FC-medium	→ FC-dense	mean
DeepLabv3+ (Chen et al. 2018)	DS	ResNet-101	67.1	65.2	61.6	63.4
SegFormer (Xie et al. 2021)		MiT-B2	70.5	66.1	62.3	65.3
Mask2Former (Cheng et al. 2022)		Swin-B	76.2	74.5	70.8	73.7
IBNet (Pan et al. 2018)	DG	ResNet-50	72.4	67.9	59.5	66.6
Iternorm (Huang et al. 2019)		ResNet-50	72.0	68.3	60.7	66.9
SW (Pan et al. 2019)		ResNet-50	73.3	69.4	61.7	66.5
ISW (Choi et al. 2021)		ResNet-50	72.1	67.9	60.1	66.7
Ours		Swin-B	**79.6**(+6.3)	**78.1**(+8.7)	**74.6**(+12.9)	**77.4** (+10.5)

Table 5: Sensitivity analysis of the proposed method and the backbone on the foggy density. CityScapes as the source domain. Foggy CityScapes -light, -medium and -dense are used as unseen target domains, respectively. Evaluation metric mIoU is in %. The mean mIoU (denoted as mean) on Foggy-CityScapes is not a simple average of three kinds of densities.

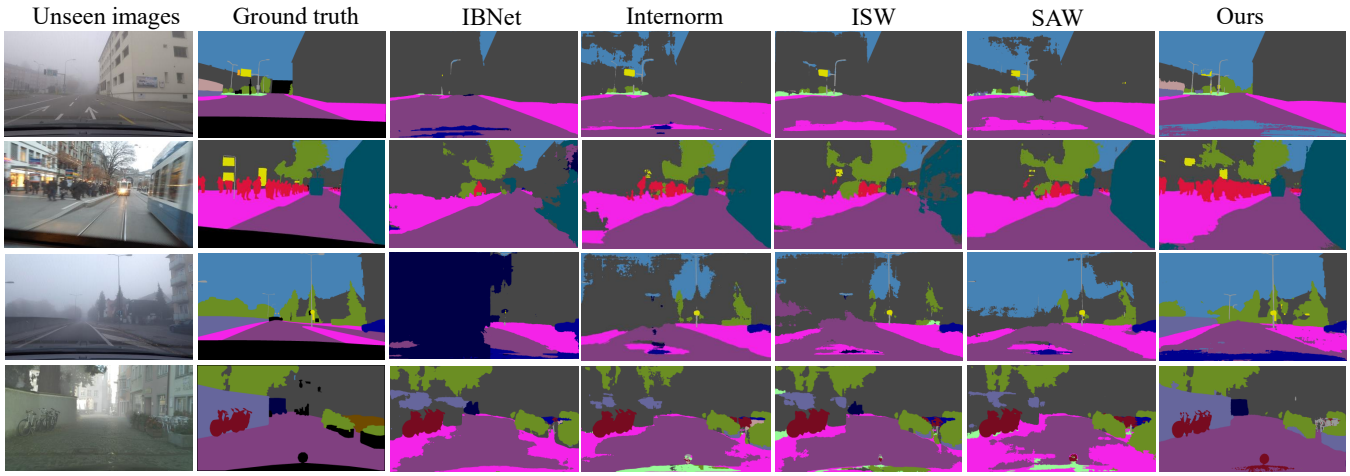

Figure 5: Visualized segmentation predictions from the proposed method (denoted as Ours) and the state-of-the-art methods IBNet (Pan et al. 2018), Internorm (Huang et al. 2019), ISW (Choi et al. 2021) and SAW (Peng et al. 2022).

quency information between the content, style and fog encoder, which we denote as $S2C$, $F2C$, $C2S$ and $C2F$. Table 4 reports the impact of each operation. All these operations positively contribute to generalized representation for foggy-scenes, with 1.0%, 0.9%, 0.8% and 0.6% mIoU gain on ACDC-fog, and 1.4%, 1.1%, 0.3% and 0.3% mIoU gain on Foggy-driving. Generally, concentrating the low-frequency information to the content branch ($S2C$, $F2C$) has a more significant impact than the operation on high-frequency information ($C2S$ and $C2F$).

On Foggy Density We further test the sensitivity of the proposed method on the foggy density. Foggy-CityScapes dataset, as the unseen target domain, provides foggy maps under the light-, medium- and dense- densities, which we denote as FC-light, FC-medium and FC-dense. CityScapes is used as the source domain. Table 5 reports the results. On all foggy densities, the proposed method outperforms the existing methods significantly. Also, it outperforms the original Mask2Former by 3.4%, 3.6% and 3.8% mIoU on the light, medium and dense fog densities, respectively.

Visualization

Fig. 5 provides some visualized segmentation predictions on ACDC-fog, Foggy-Zurich, Foggy-Driving and Foggy-CityScapes dataset, when using CityScapes as the source domain. The proposed method shows a more reasonable and more reliable inference compared with existing methods.

Conclusion

Robust foggy-scene segmentation is crucial for autonomous driving, but existing curriculum domain adaptation methods can only adapt to the foggy domain seen in the training stage. In this paper, we tackle this challenge under the domain generalization setting. We aim to learn a segmentation Transformer that can be well generalized to arbitrary unseen foggy scenes. Technically, we propose a bi-directional wavelet guidance (BWG) mechanism, which simultaneously handles the content enhancement, style de-correlation and fog de-correlation for foggy scenes in a divide-and-conquer manner. Extensive experiments show that the proposed method significantly outperforms existing directly-supervised, domain adaptation and domain generalization segmentation methods under a variety of settings.

Limitation Discussion. The fog de-correlation encoder is data-driven rather than physics-driven. However, its effectiveness has been demonstrated by the superior performance than the scenario when only using the rest two encoders for content enhancement and style de-correlation.

References

Araslanov, N.; and Roth, S. 2021. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15384–15394.

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

Bi, Q.; You, S.; and Gevers, T. 2023b. Learning Content-enhanced Mask Transformer for Domain Generalized Urban-Scene Segmentation. *arXiv preprint arXiv:2307.00371*.

Brüggemann, D.; Sakaridis, C.; Truong, P.; and Van Gool, L. 2023. Refign: Align and refine for adaptation of semantic segmentation to adverse conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3174–3184.

Chang, W.-L.; Wang, H.-P.; Peng, W.-H.; and Chiu, W.-C. 2019. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1900–1909.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, 801–818.

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

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*, 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.

Dai, D.; Sakaridis, C.; Hecker, S.; and Van Gool, L. 2020. Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *International Journal of Computer Vision*, 128: 1182–1204.

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.

Guo, X.; Yang, C.; Li, B.; and Yuan, Y. 2021. Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3927–3936.

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

Huang, L.; Zhou, Y.; Zhu, F.; Liu, L.; and Shao, L. 2019. 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.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

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.

Lee, S.; Son, T.; and Kwak, S. 2022. Fifo: Learning fog-invariant features for foggy scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18911–18921.

Li, G.; Kang, G.; Liu, W.; Wei, Y.; and Yang, Y. 2020. Content-consistent matching for domain adaptive semantic segmentation. In *European Conference on Computer Vision*, 440–456.

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, R.; Cheong, L.-F.; and Tan, R. T. 2019. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1633–1642.

Li, R.; Tan, R. T.; and Cheong, L.-F. 2020. All in one bad weather removal using architectural search. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, 3175–3185.

Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; and Yang, M.-H. 2017. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30.

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.

Lin, G.; Milan, A.; Shen, C.; and Reid, I. 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1925–1934.

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*, 10012–10022.

Ma, X.; Wang, Z.; Zhan, Y.; Zheng, Y.; Wang, Z.; Dai, D.; and Lin, C.-W. 2022. Both style and fog matter: Cumulative domain adaptation for semantic foggy scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18922–18931.

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 *European Conference on Computer Vision*, 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*, 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.

Porwik, P.; and Lisowska, A. 2004. The Haar-wavelet transform in digital image processing: its status and achievements. *Machine graphics and vision*, 13(1/2): 79–98.

Sakaridis, C.; Dai, D.; Hecker, S.; and Van Gool, L. 2018. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *European Conference on Computer Vision*, 687–704.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126: 973–992.

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*, 10765–10775.

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.

Truong, T.-D.; Duong, C. N.; Le, N.; Phung, S. L.; Rainwater, C.; and Luu, K. 2021. Bimal: Bijective maximum likelihood approach to domain adaptation in semantic scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8548–8557.

Tsai, Y.-H.; Hung, W.-C.; Schulter, S.; Sohn, K.; Yang, M.-H.; and Chandraker, M. 2018. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7472–7481.

Valanarasu, J. M. J.; Yasarla, R.; and Patel, V. M. 2022. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2353–2363.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2517–2526.

Wang, Z.; Zhang, Y.; Ma, X.; Yu, Y.; Zhang, Z.; Jiang, Z.; and Cheng, B. 2023. Semantic Segmentation of Foggy Scenes Based on Progressive Domain Gap Decoupling. *TechRxiv*.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

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

Yang, Z.; Huang, J.; Chang, J.; Zhou, M.; Yu, H.; Zhang, J.; and Zhao, F. 2023. Visual Recognition-Driven Image Restoration for Multiple Degradation with Intrinsic Semantics Recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14059–14070.

Yoo, J.; Uh, Y.; Chun, S.; Kang, B.; and Ha, J.-W. 2019. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9036–9045.

Zhang, P.; Zhang, B.; Zhang, T.; Chen, D.; Wang, Y.; and Wen, F. 2021. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12414–12424.

Zhao, Y.; Zhong, Z.; Zhao, N.; Sebe, N.; and Lee, G. H. 2022. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *European Conference on Computer Vision*, 535–552.

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*.