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

Generalized Foggy-Scene Semantic Segmentation by Frequency Decoupling

Qi Bi Shaodi You Theo Gevers
Computer Vision Research Group, University of Amsterdam
Amsterdam, 1098XH, The Netherlands
{q.bi, s.you, th.gevers}@uva.nl

Abstract

Foggy-scene semantic segmentation (FSSS) is highly challenging due to the diverse effects of fog on scene properties and the limited training data. Existing research has mainly focused on domain adaptation for FSSS, which has practical limitations when dealing with new scenes. In our paper, we introduce domain-generalized FSSS, which can work effectively on unknown distributions without extensive training. To address domain gaps, we propose a frequency decoupling (FreD) approach that separates fog-related effects (amplitude) from scene semantics (phase) in feature representations. Our method is compatible with both CNN and Vision Transformer backbones and outperforms existing approaches in various scenarios.

1. Introduction

Learning robust scene semantic segmentation is critical for many safety-crucial road applications such as autonomous driving [3, 4, 13, 54]. Among a variety of scenes, foggy-scene semantic segmentation (FSSS) deserves special attention [5]. Fog, as a typical bad weather condition, has a variety of atmospheric light and attenuation [6, 18, 33], which poses different levels of confusion on the scene radiance and leads to different levels of appearance ambiguity. Furthermore, the annotation of FSSS is very scarce, compared to clear scenes.

Existing FSSS methods can be summarized into two categories. The first category is the de-foggy based methods, which implement fog removal (dehaze) on top of a segmentation model. These methods need to reconstruct the scene radiance first, which is a highly ill-posed problem. Consequently, the reconstructed scene radiance can still suffer regional degradation (*e.g.* overexposure, color distortion [51, 63]), and the quality is still not sufficient for pixel-wise semantic predictions [42]. The second category consists of curriculum domain adaptation methods that adapt scene radiance from clear conditions (source domain) to conditions affected by atmospheric light and atmosphere scattering (target domain). However, these methods can only generalize to foggy scenes with a similar appearance to the training dataset. As depicted in Fig. 1, the appearance of fog varies significantly from one scene to another.

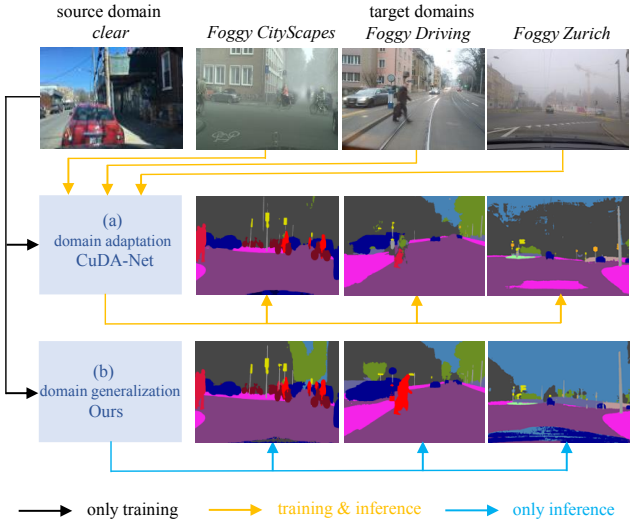

Figure 1. FSSS by: (a) existing curriculum domain adaptation method CuDA-Net [42]; (b) our proposed domain generalized FSSS. Foggy target domains are Foggy-CityScapes [53], Foggy Driving [53], and Foggy Zurich [52].

In this paper, we propose to study a new task: *domain generalized FSSS*, which allows a FSSS model to generalize to arbitrary unseen foggy domains when only trained on a clear source domain (see Fig. 1 for an example). To address this challenge, we propose that the primary goal of domain-generalized FSSS is to separate the scene semantics from the foggy appearance. Our approach focuses on the frequency domain, taking into account the observation that foggy appearance is primarily related to amplitude, while scene semantics are primarily related to phase [22, 30, 40]. Therefore, we propose a novel **Fre**quency **D**ecoupling (FreD) learning scheme.

In our approach, we utilize a semantic encoder and a fog encoder. For both encoders, we split their feature repre-

sentations into phase and amplitude components using Fast Fourier Transformation (FFT). Then, we transfer the amplitude component from the semantic encoder to the fog encoder branch, and vice versa, transferring the phase component from the fog encoder to the semantic encoder branch. This allows the semantic encoder to emphasize the phase component [22, 30, 40], while the fog encoder can focus on the amplitude component [55, 67, 68]. Lastly, we apply instance normalization to the amplitude component before performing the Inverse Fast Fourier Transformation (IFFT). This ensures that the scene representation remains stable even in the presence of varying fog density.

We conduct extensive experiments to validate the effectiveness of the proposed method. As common practice, CityScapes [14] is used as the clear source domain, ACDC-fog [54], Foggy-Zurich [52], Foggy-Driving [53] and Foggy-CityScapes [53] are used as the target domains. The proposed method is able to outperform all existing paradigms including: 1. existing domain generalized which are not foggy scenes focused [13, 16, 23, 24, 28, 38, 46–48, 71], 2. Foggy scene domain adaptation [1, 8, 15, 17, 21, 29, 31, 42, 57, 58, 61, 70], 3. directly-supervised [12, 64] and 4. de-fog based methods [2, 8, 18, 49, 50, 60, 69]. Furthermore, extensive ablation analysis and visualization further validate the effectiveness of the proposed method.

Our contribution can be summarized as follows.

- We introduce a novel task, domain-generalized FSSS, which is more practical than domain-adaptive FSSS.
- We propose an innovative **Fre**quency **D**ecoupling (FreD) learning scheme for this task. It improves semantics by focusing on the phase component and separates the fog impact by normalizing the amplitude component.
- The proposed method is versatile to both CNN and ViT backbones, and outperforms existing domain generalized segmentation methods upto 4.1% mIoU on Foggy-CityScapes and curriculum domain adaptation methods upto 14.4% mIoU on ACDC-fog.

2. Related Work

Foggy-Scene Semantic Segmentation (FSSS) holds unique challenge than other semantic segmentation tasks [25–27, 32, 45]. Direct segmentation on de-fogged images still suffers from regional degradation [42] as many other types of degraded image do [9, 20]. The dominant trend is to approach this problem as curriculum domain adaptation (*e.g.*, AdSegNet [58], ADVENT [61], ProDA [70], DMLC [17], SAC [1], Refign-DAFormer [7], DACS [57], DISE [8], CCM [31], CuDA-Net [42], CMAda3+ [15], DAFormer [21]). The concept is to utilize both clear and foggy scenes as input, allowing the semantic representation from clear scenes to adapt to the appearance in foggy scenes.

In contrast to domain generalization, this approach has a limitation. For each new scene, the model needs to undergo training again. In practice, foggy scenes are highly dynamic, resulting in significant domain gaps between them. Consequently, an adapted foggy scene may not perform well in a new foggy scene.

Fog Removal is an important and well studied task, which aims to restore the appearance of fog as if it is fog-free. Previous methods, whether model-driven or data-driven, typically assume that the degraded image consists of both the background layer and the weather effect layer [34, 35]. Some typical de-fog works include DCP [18], MSCNN [49], Non-local [2], DCPDN [69], GFN [50] and DISE [8]. In the past few years, weather removal is usually implemented in an all-in-one paradigm [60, 65, 72], which aims to remove multiple types of weather by a single model.

However, as reported in recent works [15, 42], applying scene segmentation models directly to de-fogged images results in a notable decrease in performance compared to domain adaptation-based scene segmentation methods. The localized degradation of de-fogged images presents difficulties for existing scene segmentation models in accurately determining pixel-wise semantics [51, 63].

Domain Generalized Semantic Segmentation mainly focuses on driving scenes, where the domain gap comes from varied urban landscapes. Earlier works usually leverage instance normalization [23, 46, 48] or instance whitening transformation [13, 47, 48] to decouple the impact of cross-domain style variation. More recently, style hallucination becomes another feasible solution to improve the cross-domain generalization [16, 24, 28, 56]. These approaches augment the variety of styles using images from real-world scenarios, allowing the segmentation model to encounter a broader range of styles during training [3–5].

To the best of our knowledge, domain-generalized FSSS has not been previously explored. In addition to coping with diverse urban environments, domain-generalized FSSS faces challenges related to varying lighting and attenuation conditions.

3. Theoretical Analysis

3.1. Problem Definition

Fog and semantic segmentation A foggy scene is usually formulated as [18, 37, 43]:

$$\mathbf{I}(\mathbf{x}) = \mathbf{J}(\mathbf{x}) \cdot e^{-\beta d(\mathbf{x})} + \mathbf{A}(1 - e^{-\beta d(\mathbf{x})}), \qquad (1)$$

where $\mathbf{I}$ is the observed image, $\mathbf{J}$ is the scene radiance (namely the fog-free scene), $\mathbf{A}$ is the global atmospheric light, β is the scattering coefficient of the atmosphere (namely the fog density), and d is the scene depth.

For the task of foggy-scene semantic segmentation (FSSS), given a segmentation model $f(\cdot)$ which learns the

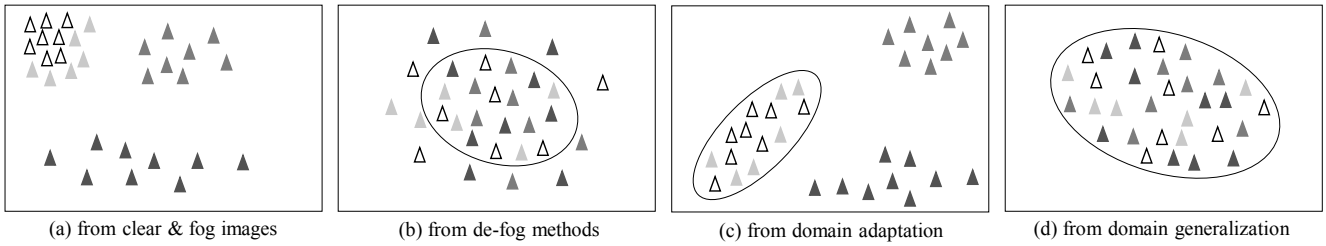

Figure 2. Learning semantic representation of foggy scenes: (a) If only learn on clear images, the more the fog is, the more scattered the feature is; (b) de-fogged images cannot be perfectly reverse the scattered representation, because de-fog is an ill-posed problem by definition; (c) domain adaptation may learn a clustered representation of each foggy domain, however, the learned representation does not generalize to an unseen foggy domain. (d) Domain generalized FSSS (our approach), which aims to learn a generalized presentation that decouples the fog.

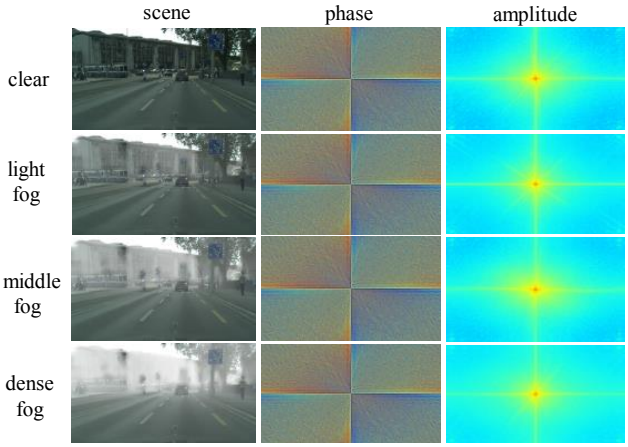

Figure 3. When considering the same driving scene captured under varying fog densities, we observe that, after undergoing Fast Fourier Transformation (FFT), the phase component—containing abundant semantics—remains similar, while the amplitude component, associated with fog densities, exhibits notable differences. Here the light, medium and dense fog density is defined in Foggy CityScapes [53].

mapping between the observed intensity $\mathbf{I}$ and the semantic category y, the relation between semantic prediction and fog can be described as

$$y(\mathbf{x}) = f(\mathbf{I}(\mathbf{x})) = f(\mathbf{J}(\mathbf{x}) \cdot e^{-\beta d(\mathbf{x})} + \mathbf{A}(1 - e^{-\beta d(\mathbf{x})})). \quad (2)$$

Note that when $\beta = 0$, namely fog-free, Eq. 2 falls back to the canonical urban scene semantic segmentation (USSS) given by

$$y(\mathbf{x}) = f(\mathbf{J}(\mathbf{x})), \quad (3)$$

which is a well-studied topic with numerous methods [10, 12, 39, 64] and large scale datasets [14, 44, 66].

Domain generalized FSSS in this paper, as illustrated in Fig. 1, is defined as to only train on Eq. 3 setting. It directly does inference on various non-trivial Eq. 2 settings.

On top of the challenges of fog-free domain generalized USSS [13, 23, 46–48, 48], this new task has its unique challenges because each unseen domain (foggy scenes) has its atmospheric light $\mathbf{A}$, fog density β, and distribution of depth $d(\cdot)$. Furthermore, labelled FSSS data is very scarce.

3.2. Analysis of Existing & Proposed Paradigm

Here, we further analyze how fog impacts the semantic representation.

Direct Inference As illustrated in Fig. 2a, when trained on a clear source domain $\mathcal{D}^{(\mathcal{S})} = \{\mathbf{x}_i^{(S)}\}_{i=1}^{N_{(\mathcal{S})}}$, and doing inference on a variety of foggy target domains $\mathcal{D}^{(\mathcal{T}_1)} = \{\mathbf{x}_i^{(\mathcal{T}_1)}\}_{i=1}^{N_{(\mathcal{T}_1)}}, \cdots, \mathcal{D}^{(\mathcal{T}_j)} = \{\mathbf{x}_i^{(\mathcal{T}_j)}\}_{i=1}^{N_{(\mathcal{T}_j)}}, \cdots$. For a clear source domain $\mathcal{D}^{(\mathcal{S})}$, we have $\beta = 0$. The segmentation model is learnt from the image appearance only impacted by the scene radiance $\mathbf{J}$ in Eq. 3. For a foggy target domain $\mathcal{D}^{(\mathcal{T}_j)}$, the image appearance is also impacted by atmospheric light $\mathbf{A}^{(\mathcal{T}_j)}$ and atmosphere scattering $\beta^{(\mathcal{T}_j)}$. In this paradigm, the semantics is only learnt from the appearance determined by the scene radiance $\mathbf{J}$. However, the presence of fog in the target domains brings in the impact of atmospheric light $\mathbf{A}$ and attenuation $e^{-\beta d(\mathbf{x})}$, which poses a natural ambiguity of the scene radiance $\mathbf{J}$ reflected on the image appearance. This phenomenon renders the semantic space and poses challenge to cluster effectively. In general, as fog density increases, the learned semantic representation becomes more scattered.

De-fog For de-fog methods which implement scene segmentation on the de-fogged images, the input of the segmentation model $f(\cdot)$ in the inference stage is the reconstructed scene radiance $\hat{\mathbf{J}}$. Finding an optimal reconstructed scene radiance $\hat{\mathbf{J}}$ is an inverse operation of Eq. 1, which is highly ill-posed (many more unknowns than constraints). Consequently, the regional degradation of $\hat{\mathbf{J}}$ can confuse the segmentation model $f(\cdot)$ pre-trained on clear scenes. Difficulties remain to cluster the semantics (illustrated in Fig. 2b).

Domain Adaption For domain adaptation scene segmentation methods, the key idea is to take images from the clear

source domain $\mathcal{D}^{(\mathcal{S})}$ and a foggy target domain $\mathcal{D}^{(\mathcal{T}_1)}$ as input. Such methods can learn a stable mapping between $\mathbf{J}^{(\mathcal{S})}$ and $\mathbf{J}^{(\mathcal{T}_1)}$, atmospheric light $\mathbf{A}^{(\mathcal{T}_1)}$ and atmosphere scattering $\beta^{(\mathcal{T}_1)}$. However, when inference on other foggy domain $\mathbf{J}^{(\mathcal{T}_j)}$ that has not seen before, the different atmospheric light $\mathbf{A}^{(\mathcal{T}_j)}$ and atmosphere scattering $\beta^{(\mathcal{T}_j)}$ pose different levels of ambiguity and confusion on the scene radiance $\mathbf{J}^{(\mathcal{T}_j)}$, which has not encountered during training. Consequently, it can still be difficult for the domain adaptation methods to correctly cluster the semantics from unseen fog domains (in Fig. 2c).

Domain Generalized FSSS (proposed) We consider the domain generalization paradigm for FSSS, which aims to generalize to arbitrary unseen foggy domains $\mathcal{D}^{(\mathcal{T}_1)}$, $\cdots$, $\mathcal{D}^{(\mathcal{T}_j)}$, $\cdots$, when only learns from a clear source domain $\mathcal{D}^{(\mathcal{S})}$. Its key idea is to decouple the impact of atmospheric light and atmosphere scattering from different fog domains, which we denote as $\mathbf{A}^{(\mathcal{T}_1)}$, $\beta^{(\mathcal{T}_1)}$, $\cdots$, $\mathbf{A}^{(\mathcal{T}_j)}$, $\beta^{(\mathcal{T}_j)}$, $\cdots$. In this way, the domain gap is alleviated, and the semantic space can be robust to the variation of fog densities (illustrated in Fig. 2d).

3.3. Frequency Domain Analysis

We use Fast Fourier Transformation (FTT) for frequency analysis. As shown in Fig. 3, the fog density, which is determined by atmospheric light $\mathbf{A}$ and atmosphere scattering β, reflects more from the amplitude component [55, 67, 68]. In contrast, the phase component, which reflects the scene objects [22, 30, 40], is relatively stable from different domains that have levels of fog densities.

To summarize, the domain gap of FSSS is more reflected in the amplitude component. The composition between phase and amplitude provides a feasible solution to address this task in a divide-and-conquer way.

Overall Idea Fig. 4 gives an overview of the proposed frequency decoupling method. Given the representation outputted from the semantic encoder E_S and the fog encoder E_F, we decompose them into phase and amplitude component. Then, the phase components from both encoders, denoted as P_S and P_F, are concentrated to the semantic encoder branch. At the same time, the amplitude components from both encoders, denoted as A_S and A_F, are shifted to the fog encoder branch for decoupling.

4. Methodology

Fig. 5 shows the full pipeline of the proposed **Fre**quency **D**ecoupling (FreD) method. It consists of three key steps, namely, dual self-attention encoding, phase-amplitude interaction and deep amplitude decoupling.

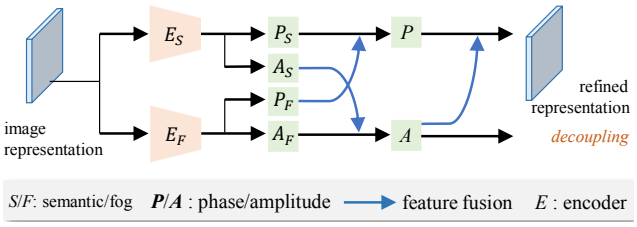

Figure 4. Overall idea to learn domain generalized FSSS by frequency decoupling. E_S/E_F: semantic /fog encoder; P/A: phase / amplitude component; Blue arrow refers to feature fusion.

4.1. Dual Self-Attention Encoding

Before studying the semantic and fog representation in the frequency space, a pre-requisite is to build both representations from the image encoder. The recently developed mask attention method, as demonstrated in [11, 12], has exhibited superior capabilities in capturing contextual information within image representations compared to traditional convolutional neural networks. In contrast to existing FSSS methods that rely on domain adaptation and use CNNs to represent semantics and fog, our approach introduces dual mask-attention encoding to construct semantic and fog representations.

Given an image encoder (*e.g.*, ResNet-50, Swin-Transformer), we have an image representation from a certain block, denoted as $\mathbf{F} \in \mathbb{R}^{(W \cdot H) \times C}$. Assume it is fed into the l^{th} layer of the mask attention decoder, its key $\mathbf{K}_l \in \mathbb{R}^{(W \cdot H) \times C}$, value $\mathbf{V}_l \in \mathbb{R}^{(W \cdot H) \times C}$ and query $\mathbf{Q}_l \in \mathbb{R}^{N \times C}$ is computed by three linear transformations f_{K_l}, f_{V_l} and f_{Q_l}, respectively. Then, the mask queries $\mathbf{X}_l \in \mathbb{R}^{N \times C}$ are computed as

$$\mathbf{X}_l = \mathrm{softmax}(\mathcal{M}_{l-1} + \mathbf{Q}_l \mathbf{K}_l^{\mathsf{T}}) \mathbf{V}_l + \mathbf{X}_{l-1}, \quad (4)$$

where softmax is the softmax function, and $\mathcal{M}_{l-1} \in \{0,1\}^{N \times (W \cdot H)}$ is a binary matrix from the mask queries $\mathbf{X}_{l-1}$ of the $(l-1)^{th}$ layer, with a threshold of 0.5. Further, $\mathcal{M}_0$ is binarized and resized from $\mathbf{X}_0$.

The mask queries $\mathbf{X}_l$ are later on fed into the self-attention mechanism, which aims to exploit the long-range dependencies throughout the scene. It is computed as

$$\mathrm{Attention}(\mathbf{Q}_{\mathbf{X}_l}, \mathbf{K}_{\mathbf{X}_l}, \mathbf{V}_{\mathbf{X}_l}) = \mathrm{Softmax}(\frac{\mathbf{Q}_{\mathbf{X}_l} \mathbf{K}_{\mathbf{X}_l}}{\sqrt{d_k}}) \mathbf{V}_{\mathbf{X}_l}, \quad (5)$$

where $\mathbf{Q}_{\mathbf{X}_l}$, $\mathbf{K}_{\mathbf{X}_l}$ and $\mathbf{V}_{\mathbf{X}_l}$ denote the query, key and value of $\mathbf{X}_l$, each of which can be computed by a linear transformation. Further, Softmax denotes the softmax function.

On top of Eq. 5, we design two self-attention based encoders to transfer the mask queries $\mathbf{X}_l$ into the semantic and fog representation, respectively. Given the semantic encoder E_S, the mask queries $\mathbf{X}_l^S \in \mathbb{R}^{N \times C}$ for semantic representation can be computed according to Eq. 5. Similarly, given the fog encoder E_F, the mask queries $\mathbf{X}_l^F \in \mathbb{R}^{N \times C}$ for fog representation are also computed according to Eq. 5.

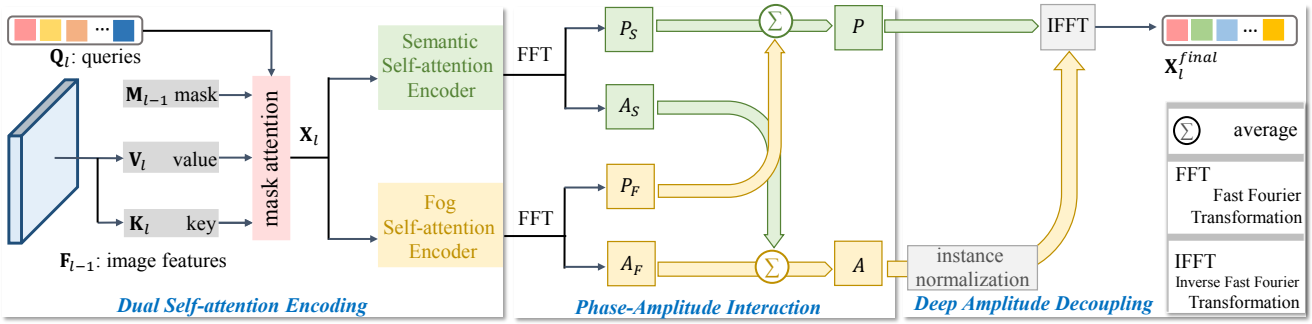

Figure 5. Framework overview of the proposed **Fre**quency **D**ecoupling (FreD) learning for domain generalized foggy-scene semantic segmentation (FSSS). Given an image feature $\mathbf{F}_l$ from the image encoder, the proposed FreD consists of three key steps, namely, dual self-attention encoding (in Sec. 4.1), phase-amplitude interaction (in Sec. 4.2) and deep amplitude decoupling (in Sec. 4.3), respectively.

4.2. Phase-Amplitude Interaction

Learning both semantic and fog representations from only the image appearance is ill-posed, making it still difficult to precisely represent the semantic information and the fog information from $\mathbf{X}_l^S$ and $\mathbf{X}_l^F$. To more precisely describe the semantic and fog representation, we follow our analysis in the frequency space (in Sec. 3.3), and disentangle both $\mathbf{X}_l^S$ and $\mathbf{X}_l^F$ in the frequency space. To this end, the phase-amplitude interaction step is proposed to separate and refine both semantic and fog representation.

Take the mask queries (for semantics) $\mathbf{X}_l^S$ as an example. The Fast Fourier Transformation (FFT) transforms it from the spatial space to frequency space, given by

$$\mathbf{F}_l^s(u,v) = \sum_{w=1}^{W}\sum_{h=1}^{H} \mathbf{X}_l^S(w,h) e^{-j2\pi(wu/W + hv/H)}. \quad (6)$$

This frequency representation $F_l^s(u,v)$ (in Eq. 6) is further converted into the polar coordinate form, given by

$$\mathbf{F}_l^s(u,v) = |\mathbf{F}_l^s(u,v)| e^{-j\phi(u,v)}. \quad (7)$$

Consequently, the phase $\mathbf{P}_l^S$ and amplitude component $\mathbf{A}_l^S$ is computed as

$$\mathbf{P}_l^s(u,v) = \phi(u,v), \quad (8)$$
$$\mathbf{A}_l^s(u,v) = |\mathbf{F}_l^s(u,v)|. \quad (9)$$

Similarly, we can get the phase component and amplitude component from $\mathbf{X}_l^F$, which we denote as $\mathbf{P}_l^F$ and $\mathbf{A}_l^F$, respectively.

As explained in Sec. 3.3, fog primarily affects the amplitude component, whereas semantics are more prominent in the phase component. Consequently, our goal is to enable the semantic encoder E_S to exclusively attend to the phase component, while simultaneously enabling the fog encoder E_F to exclusively attend to the amplitude component.

Specifically, after interaction, the representation $\mathbf{P}_l$ from the semantic encoder E_S is computed as

$$\mathbf{P}_l = \mathbb{E}[\mathbf{P}_l^S, \mathbf{P}_l^F], \quad (10)$$

and the representation $\mathbf{A}_l$ from the fog encoder E_F is computed as

$$\mathbf{A}_l = \mathbb{E}[\mathbf{A}_l^S, \mathbf{A}_l^F], \quad (11)$$

where $\mathbb{E}[\cdot,\cdot]$ denotes the average operation.

4.3. Deep Amplitude Decoupling

The primary challenge in learning domain-generalized FSSS is the variation in fog density across different domains. After obtaining the amplitude component $\mathbf{A}_l$ which represents the presence of fog, it is natural to consider decoupling its impact from the semantics. To realize this objective, we propose the deep amplitude decoupling.

Instance normalization [59] has been widely used for style transfer. It applies a channel-wise normalization for a feature representation, and makes the feature representation less sensitive to the style variation [36]. More recently, it has been widely used for domain generalized semantic segmentation [23, 46, 48], aiming to decouple the impact of style. To this end, we implement the instance normalization function on amplitude component $\mathbf{A}_l$, making it robust to the variation of fog densities. For simplicity, here we describe the instance normalization on $\mathbf{A}_l$ from the perspective of spatial space $\mathbb{R}^{N\times C}$, which is computed as

$$\hat{\mathbf{A}}_l^c = \frac{\mathbf{A}_l^c - \boldsymbol{\mu}}{\boldsymbol{\sigma} + \epsilon} \cdot \gamma + \beta, \quad (12)$$

$$\boldsymbol{\mu} = \frac{1}{C}\sum_{c=1}^{C} \mathbf{A}_l^c, \boldsymbol{\sigma} = \sqrt{\frac{1}{C}\sum_{i=1}^{C}(\mathbf{A}_l^c - \boldsymbol{\mu})^2}, \quad (13)$$

where $c = 1, 2, \cdots, C$ represent the number of each individual channel.

Finally, let $\mathrm{IFFF}(\cdot,\cdot)$ denote the inverse fast Fourier transform (IFFT), which has a pairwise phase and amplitude as input. Then, the decoupled amplitude component $\hat{\mathbf{A}}_l$ and the phase component $\mathbf{P}$ is merged to get the refined mask queries $\mathbf{X}_l^{final}$, given by

$$\mathbf{X}_l^{final} = \mathrm{IFFF}(\mathbf{P}, \hat{\mathbf{A}}_l). \quad (14)$$

4.4. Implementation Details

The proposed frequency decoupling scheme is integrated into each layer of a Transformer encoder, which is the

key component of Mask2Former [11, 12] segmentation paradigm. For the image encoder, by default we use the Swin-base Transformer [41] as backbone, with a pre-trained weight on ImageNet. Notably, the proposed frequency decoupling based Transformer is also versatile to CNN based backbones such as ResNet-50, ResNet-101 and *etc*. For the image decoder, following the original Mask2Former [12] pipeline, the image feature is progressively upsampled to $\times 32$, $\times 16$, $\times 8$ and $\times 4$ resolution, respectively.

The Transformer decoder consists of nine frequency decoupling components. Following the original Mask2Former [12], for every three components, the image features of the $\times 32$, $\times 16$ and $\times 8$ resolution are inputted one-by-one. The final output from the ninth component is fused with the $\times 4$ resolution image features for prediction.

Without whistles and bells, all the loss functions and hyper-parameters keep the default settings of the original Mask2Former [12] without any additional fine-tuning. The model is trained by 50 epochs, with an initial learning rate 1×10^{-4} and the Adam optimizer.

5. Experiments

5.1. Datasets

CityScapes (C) [14] is one of the widely-used scene semantic segmentation datasets for driving scenes with 19 semantic categories. In our experiments, its training set with 2965 well-annotated samples are used as the source domain for domain generalized methods.

Clear CityScapes (CC) [53] is a subset of CityScapes [14]. It consists of 498 deliberately-selected samples from CityScapes, and is only used as the source domain for domain adaptation methods.

Foggy-CityScapes (FC) [53] is one of the most commonly-used dataset for FSSS. It has 498 images for training and 52 images for testing. The fog in this dataset is synthetic, and has three levels of fog densities, namely, light, medium, and dense.

Foggy Zurich (FZ) [52] is a FSSS dataset that captures real-fog scenes in the Zurich City. It has 3,808 real-world samples, but most of them are unlabelled. In our experiments, only its 40 images with annotation are used as foggy target domain.

Foggy Driving (FD) [53] is FSSS dataset only used for testing. It has 101 images that are captured under real fog conditions. 33 of 101 images are finely-annotated and the rest 68 of them are coarsely-annotated.

Adverse Conditions Dataset with Correspondences (ACDC) [54] is a recently published driving scene semantic segmentation dataset under adverse conditions. It has 4,006 samples in total, and 1000 of them are captured under real fog conditions. Following its original setting, the dataset split for training set, validation set and test set is 4:1:5.

Method	Backbone	Target Domains			
		$\rightarrow$ AF	$\rightarrow$ FZ	$\rightarrow$ FD	$\rightarrow$ FC
CNN Encoder:					
IBNet [46]	Res-50	63.8	33.4	45.5	66.5
Iternorm [23]	Res-50	63.3	35.2	44.6	66.9
SW [47]	Res-50	62.4	34.1	45.8	66.4
ISW [13]	Res-50	64.3	36.1	46.2	66.6
SHADE [71]	Res-50	61.4	39.5	42.0	65.8
SAW [48]	Res-50	64.0	37.3	47.0	67.8
WildNet [28]	Res-50	64.7	39.2	42.6	64.4
SPC [24]	Res-50	68.0	39.3	43.5	64.7
FreD (Ours)	Res-50	**69.3**	**36.9**	**49.1**	**71.9**
		$\uparrow$1.3	$\downarrow$2.4	$\uparrow$2.1	$\uparrow$4.1
ViT Encoder:					
SegFormer* [64]	MiT-B4	59.2	43.9	46.6	75.5
ISSA* [38]	MiT-B4	67.5	-	-	-
HGFormer* [16]	Swin-L	69.9	-	-	-
Mask2Former [12]	Swin-B	73.3	49.4	51.1	73.8
FreD (Ours)	Swin-B	**75.1**	**50.9**	**53.2**	**76.6**
		$\uparrow$1.8	$\uparrow$1.5	$\uparrow$2.1	$\uparrow$2.8

Table 1. Comparison with existing domain generalized segmentation methods. Evaluation metric mIoU is in %. '-': no official performance report. ACDC-Fog: AF; Foggy Zurich: FZ; Foggy-driving: FD; Foggy-CityScapes: FC. By default all the results are implemented on the official source code and default parameter settings; *: directly cite from the corresponding papers.

5.2. State-of-the-art Comparison

Comparison with Generalization Methods Notice that we are the first generalization methods for FSSS. Here we compare with existing domain generalization methods which can be applied to FSSS. Our method is generally applicable to both CNN backbones and ViT backbones. As shown in Table 1, the first group compares with methods using CNN backbones, which are IBNet [46], Iternorm [23], SW [47], ISW [13], SHADE [71], SAW [48], WildNet [28] and SPC [24].

The second group compares with methods using ViT backbones, which are HGFormer [16] and ISSA [38], SegFormer [64] and Mask2Former [12]. ISSA [38] and HGFormer [16] do not have available code and the results are directly cited from their papers. Our method uses Swin-Base as backbone, which is same as Mask2Former [12].

In accordance with existing domain generalized segmentation setting, we use CityScapes as the source domain. Four FSSS datasets are used as unseen target domains, namely, FC, FZ, FD and AF.

Table 1 reports the performance. For CNN backbones, the proposed method outperforms second best by 1.3%, 2.1% and 4.1% mIoU on AF, FD and FC, respectively. On FZ, there is a 2.4% mIoU drop compared with SPC [24], which involves large-scale additional training data. For CNN backbones, the proposed method outperforms second-best by 1.8%, 1.5%, 2.1% and 2.8% mIoU improvement on AF, FZ, FD and FC, respectively. In conclusion, the proposed method is able to significantly outperform existing domain generalized segmentation methods when deployed

Method	Backbone	Intermediate Domain			Target Domains		
		CZ	FZ	Other	→ AF	→ FZ	→ FD
CNN Encoder:							
AdSegNet* [58]	Res-101	✓	✓	✗	31.8	26.1	37.6
ADVENT* [61]	Res-101	✓	✓	✗	32.9	24.5	36.1
DISE* [8]	Res-101	✓	✓	✗	42.4	40.7	45.2
CCM* [31]	Res-101	✓	✓	✗	-	35.8	42.6
SAC* [1]	Res-101	✓	✓	✗	-	37.0	43.4
ProDA* [70]	Res-101	✓	✓	✗	38.4	37.8	41.2
DMLC* [17]	Res-101	✓	✓	✗	-	33.5	32.6
DACS* [57]	Res-101	✓	✓	✗	-	28.7	35.0
CMAda3+* [15]	RefineNet	✓	✓	✓	-	46.8	49.8
FIFO* [29]	RefineNet	✓	✓	✓	54.1	48.4	50.7
CuDA-Net* [42]	Res-101	✓	✓	✗	55.6	48.2	**52.7**
ViT Encoder:							
DAFormer* [21]	MiT-B5	✓	✓	✗	48.9	44.4	-
CumFormer* [62]	MiT-B5	✓	✓	✗	60.7	-	56.2
FreD (w. CC)	Swin-B	✗	✗	✗	71.1	46.3	48.7
FreD (w. C)		✗	✗	✗	**75.1**	**50.9**	53.2
					↑14.4	↑2.5	↓3.0

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 domain generalization setting. '*': directly cited from the corresponding references. ACDC-Fog: AF; Foggy Zurich: FZ; Foggy-driving: FD; Clear Zurich: CZ; Clear CityScapes: CC; CityScapes: C.

on either CNN or ViT backbone.

Comparison with Domain Adaptation Methods While our proposed method has not seen the target domain during training and does not need any intermediate domains, nevertheless, the proposed method demonstrates superior performance compared with domain adaption methods on FSSS.

For comparison, we choose AdSegNet [58], DISE [8], CCM [31], ADVENT [61], CMAda3+ [15], DACS [57], SAC [1], DMLC [17], ProDA [70], DAFormer [21], FIFO [29] and CuDA-Net [42]. They use Clear-Cityscapes as source domain and a variety of intermediate domains. Details are listed in Table 2.

As shown in Table 2, when using Clear-Cityscapes as source domain, our method outperforms the second-best by 10.4% mIoU on AF. The performance on foggy zurich (FZ) is slightly slower because the compared methods use it as both intermediate domain and target domain. The performance on FD is also lower because FD is a very small dataset and can be easily trained if visible. Our method can be easily improved when use a larger source domain. Here we replace clear-Cityscapes to Cityscapes. As can be seen, the performance on AF and FZ are both the best, and on FD is close to the best. Especially, we achieve a performance gain by 14.4% mIoU on AF.

Comparison with De-fog Methods De-fog paradigm is also compared. For each defog based method, following the implementation in [42], the first stage is to implement de-fog algorithms on each target domain. Then, the second stage is to do inference on each target domain by the pre-trained segmentation model on the source domain. The de-fog methods include DCP [18], MSCNN [49], Non-local [2], DCPDN [69], GFN [50] and DISE [8], and TransWeather [60], For stage 2, RefineNet [39], Res-101 [19] and SegFormer [64] are used. All the source domains are CC. The results are reported in Table 3. Our proposed methods outperforms all de-fog methods greatly.

Method	Target Domains		
	→ AF	→ FZ	→ FD
DCP [18] + RefineNet [39]	34.7	31.2	33.2
MSCNN [49] + RefineNet [39]	38.5	34.4	38.3
Non-local [2] + RefineNet [39]	31.9	27.6	32.8
DCPDN [69] + Res-101 [19]	33.4	28.7	37.9
GFN [50] + RefineNet [39]	33.6	28.7	37.2
DISE [8] + Res-101 [19] †	-	38.6	37.1
TransWeather [60] + SegFormer [64]	39.4	37.3	-
FreD	71.1	46.3	48.7
	↑31.7	↑7.7	↑10.4

Table 3. Comparison with existing de-fog based methods. ClearCityScapes as source domain. Evaluation metric mIoU is in %. '-': either no official code or no performance report. N/A: the result is not meaningful under domain generalization setting. '*': directly cited from the corresponding references. †: needs domain adaptation, uses CZ and FZ as intermediate domain. Foggy Zurich: FZ; Clear Zurich: CZ.

Encoders		Target Domains		
E_S	E_F	→ AF	→ FZ	→ FD
✗	✗	73.3	49.4	51.1
✓	✗	74.3	50.1	52.4
✗	✓	74.2	49.9	52.3
✓	✓	**75.1**	**50.9**	**53.2**

Table 4. Ablation studies on the semantic encoder E_S and the fog encoder E_F in the proposed method. ACDC-Fog: AF; Foggy Zurich: FZ; Foggy-driving: FD. Evaluation metric mIoU is in %.

$P_F \rightarrow P_S$	$A_S \rightarrow A_F$	$P \rightarrow A$	→ AF	→ FD
✗	✗	✗	73.3	51.1
✓	✗	✗	74.5	52.6
✓	✓	✗	74.7	52.8
✓	✓	✓	**75.1**	**53.2**

Table 5. Ablation studies on each step of phase-amplitude interaction. ACDC-Fog: AF; Foggy-driving: FD. Metric mIoU is in %.

5.3. Ablation Studies

On Each Encoder The proposed method consists of two self-attention based encoders to represent the scene semantic and the fog, which we denote as E_S and E_F, respectively. We compare the proposed method with the scenarios when there is only one encoder. CityScapes is used as the source domain. ACDC-fog (AF), Foggy Zurich (FZ) and Foggy Driving (FD) are used as the unseen target domains.

The results are reported in Table 4. When only using E_S, the mIoU improvement on AF, FZ and FD is 1.0%, 0.7% and 1.3%. When only using E_F, the improvement on AF, FZ and FD, in contrast, is 0.9%, 0.5% and 1.2%. Clearly, jointly using both E_S and E_F leads to a better FSSS performance, by 1.9%, 1.5% and 2.1% on AF, FZ and FD.

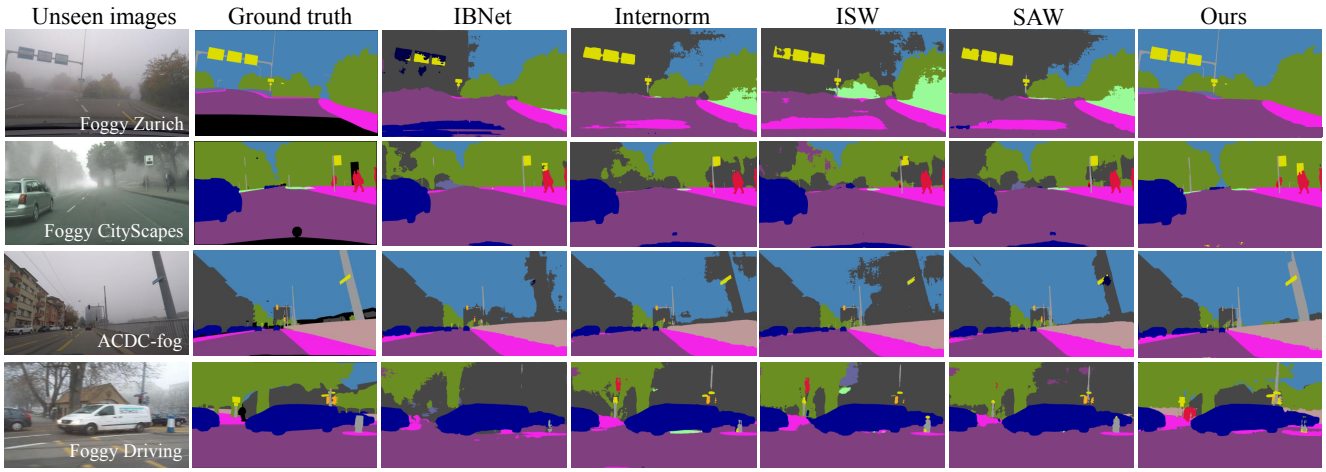

Figure 6. Visualized foggy-scene semantic segmentation results when using CityScapes as the source domain. From the first to forth row, Foggy Zurich, Foggy CityScapes, ACDC-fog and Foggy Driving is the unseen target domains. The proposed method is compared with existing domain generalized semantic segmentation methods, namely, IBNet [46], Internorm [23], ISW [13] and SAW [48].

Method	Backbone	Target Domains			
		→ light	→ medium	→ dense	mean
IBNet [46]	Res-50	72.4	67.9	59.5	66.6
Iternorm [23]	Res-50	72.0	68.3	60.7	66.9
SW [47]	Res-50	73.3	69.4	61.7	66.5
ISW [13]	Res-50	72.1	67.9	60.1	66.7
DeepLabv3+ [10]	Res-101	67.1	65.2	61.6	63.4
SegFormer [64]	MiT-B4	70.5	66.1	62.3	65.3
Mask2Former [12]	Swin-B	76.2	74.5	70.8	73.7
Ours	Swin-B	**79.5**	**78.0**	**74.8**	**77.4**
		↑3.3	↑3.5	↑4.0	↑3.7

Table 6. Sensitivity analysis of the proposed method 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.

On Phase-Amplitude Interaction Under the above domain generalization setting, we further analyze how each step of the phase-amplitude interaction impacts the overall performance. As shown in Fig. 4, the phase component P_F is merged into the phase component P_S, the amplitude component A_S is merged into the amplitude component A_F, and finally the merged phase P is fused with the decoupled amplitude A. We denote these three steps as $P_F \rightarrow P_S$, $A_S \rightarrow A_F$ and $P \rightarrow A$, respectively.

The results are reported in Table 5. The impact of $P_F \rightarrow P_S$ is slightly higher than $A_S \rightarrow A_F$, as the phase component has more semantic information. On the other hand, fusing decoupled amplitude component can lead to another slight segmentation improvement.

On Foggy Density We further analyze the generalization ability of the proposed method and existing domain generalized segmentation on different levels of fog density. CityScapes is used as the source domain, and the Foggy-CityScapes is used as unseen target domains when posed on the light, middle and dense fog densities.

Table 6 reports the experimental outcomes. The proposed method shows the state-of-the-art performance when generalized to different levels of fog densities. Especially, when generalized to middle and dense fog densities, it outperforms the second-best by up to 3.5% and 4.0% mIoU.

5.4. Visualization

We show some visual segmentation results of the proposed method and other domain generalized semantic segmentation methods in Fig. 6. Each model is trained on CityScapes as the source domain, and does inference on the four unseen target domains, namely, ACDC-fog, Foggy Zurich, Foggy Driving and Foggy CityScapes. The proposed method shows a better visual prediction than existing methods. More visual segmentation results are provided in the supplementary material.

6. Conclusion

This paper addresses domain generalized foggy-scene semantic segmentation (FSSS), which can generalize to unseen foggy domains when training only on a clear source domain. It investigates the impact of fog formulation on FSSS methods and introduces a **Fre**quency **D**ecoupling (FreD) approach for domain-generalized FSSS. To address the domain gap among foggy scenes, this approach leverages the frequency space to fog-related effects (amplitude) from scene semantics (phase) in feature representations. Experimental results across multiple datasets demonstrate the superiority of the proposed method over existing domain-generalized, domain adaptation, and de-fog FSSS methods.

References

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021. 2, 7

[2] Dana Berman, Shai Avidan, et al. Non-local image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1674–1682, 2016. 2, 7

[3] Qi Bi, Shaodi You, and Theo Gevers. Interactive learning of intrinsic and extrinsic properties for all-day semantic segmentation. *IEEE Transactions on Image Processing*, 2023. 1, 2

[4] Qi Bi, Shaodi You, and Theo Gevers. Learning content-enhanced mask transformer for domain generalized urban-scene segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 819–827, 2024. 1

[5] Qi Bi, Shaodi You, and Theo Gevers. Learning generalized segmentation for foggy-scenes by bi-directional wavelet guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 801–809, 2024. 1, 2

[6] Qi Bi, Shaodi You, and Theo Gevers. Modeling weather uncertainty for multi-weather co-presence estimation. *arXiv preprint arXiv:2403.20092*, 2024. 1

[7] David Brüggemann, Christos Sakaridis, Prune Truong, and Luc Van Gool. 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*, pages 3174–3184, 2023. 2

[8] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. 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*, pages 1900–1909, 2019. 2, 7

[9] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023. 2

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pages 801–818, 2018. 3, 8

[11] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 4, 6

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 2, 3, 4, 6, 8

[13] S. Choi, S. Jung, H. Yun, J. Kim, S. Kim, and J. Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021. 1, 2, 3, 6, 8

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 3, 6

[15] Dengxin Dai, Christos Sakaridis, Simon Hecker, and Luc Van Gool. Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *International Journal of Computer Vision*, 128:1182–1204, 2020. 2, 7

[16] Jian Ding, Nan Xue, Gui-Song Xia, Bernt Schiele, and Dengxin Dai. Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15413–15423, 2023. 2, 6

[17] Xiaoqing Guo, Chen Yang, Baopu Li, and Yixuan Yuan. 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*, pages 3927–3936, 2021. 2, 7

[18] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010. 1, 2, 7

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[20] Yang Hong, Kaixuan Wei, Linwei Chen, and Ying Fu. Crafting object detection in very low light. In *BMVC*, page 3, 2021. 2

[21] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. 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*, pages 9924–9935, 2022. 2, 7

[22] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021. 1, 2, 4

[23] L. Huang, Y. Zhou, F. Zhu, L. Liu, and L. Shao. Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4874–4883, 2019. 2, 3, 5, 6, 8

[24] Wei Huang, Chang Chen, Yong Li, Jiacheng Li, Cheng Li, Fenglong Song, Youliang Yan, and Zhiwei Xiong. Style projected clustering for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3061–3071, 2023. 2, 6

[25] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 12341–12351, 2021. 2

[26] Wei Ji, Jingjing Li, Qi Bi, Tingwei Liu, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *arXiv preprint arXiv:2304.05750*, 2023.

[27] Wei Ji, Jingjing Li, Cheng Bian, Zongwei Zhou, Jiaying Zhao, Alan L Yuille, and Li Cheng. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1094–1104, 2023. 2

[28] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9946, 2022. 2, 6

[29] Sohyun Lee, Taeyoung Son, and Suha Kwak. Fifo: Learning fog-invariant features for foggy scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18911–18921, 2022. 2, 7

[30] Sangrok Lee, Jongseong Bae, and Ha Young Kim. Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11776–11785, 2023. 1, 2, 4

[31] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *European Conference on Computer Vision*, pages 440–456, 2020. 2, 7

[32] Jingjing Li, Wei Ji, Qi Bi, Cheng Yan, Miao Zhang, Yongri Piao, Huchuan Lu, et al. Joint semantic mining for weakly supervised rgb-d salient object detection. *Advances in Neural Information Processing Systems*, 34:11945–11959, 2021. 2

[33] Kunming Li, Yu Li, Shaodi You, and Nick Barnes. Photo-realistic simulation of road scene for data-driven methods in bad weather. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 491–500, 2017. 1

[34] Ruoteng Li, Loong-Fah Cheong, and Robby T Tan. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1633–1642, 2019. 2

[35] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3175–3185, 2020. 2

[36] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017. 5

[37] Yu Li, Shaodi You, Michael S Brown, and Robby T Tan. Haze visibility enhancement: A survey and quantitative benchmarking. *Computer Vision and Image Understanding*, 165:1–16, 2017. 2

[38] Yumeng Li, Dan Zhang, Margret Keuper, and Anna Khoreva. Intra-source style augmentation for improved domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 509–519, 2023. 2, 6

[39] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 3, 7

[40] Shiqi Lin, Zhizheng Zhang, Zhipeng Huang, Yan Lu, Cuiling Lan, Peng Chu, Quanzeng You, Jiang Wang, Zicheng Liu, Amey Parulkar, et al. Deep frequency filtering for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11797–11807, 2023. 1, 2, 4

[41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[42] Xianzheng Ma, Zhixiang Wang, Yacheng Zhan, Yinqiang Zheng, Zheng Wang, Dengxin Dai, and Chia-Wen Lin. 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*, pages 18922–18931, 2022. 1, 2, 7

[43] Srinivasa G Narasimhan and Shree K Nayar. Vision and the atmosphere. *International journal of computer vision*, 48: 233–254, 2002. 2

[44] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 3

[45] Junwen Pan, Qi Bi, Yanzhan Yang, Pengfei Zhu, and Cheng Bian. Label-efficient hybrid-supervised learning for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2026–2034, 2022. 2

[46] X. Pan, P. Luo, J. Shi, and X. Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *European Conference on Computer Vision*, pages 464–479, 2018. 2, 3, 5, 6, 8

[47] X. Pan, X. Zhan, J. Shi, X. Tang, and P. Luo. Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1863–1871, 2019. 2, 6, 8

[48] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2594–2605, 2022. 2, 3, 5, 6, 8

[49] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 154–169. Springer, 2016. 2, 7

[50] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3253–3261, 2018. 2, 7

[51] Wenqi Ren, Jingang Zhang, Xiangyu Xu, Lin Ma, Xiaochun Cao, Gaofeng Meng, and Wei Liu. Deep video dehazing with semantic segmentation. *IEEE transactions on image processing*, 28(4):1895–1908, 2018. 1, 2

[52] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *European Conference on Computer Vision*, pages 687–704, 2018. 1, 2, 6

[53] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 1, 2, 3, 6

[54] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021. 1, 2, 6

[55] Hao Shen, Zhong-Qiu Zhao, Yulun Zhang, and Zhao Zhang. Mutual information-driven triple interaction network for efficient image dehazing. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7–16, 2023. 2, 4

[56] Gabriel Tjio, Ping Liu, Joey Tianyi Zhou, and Rick Siow Mong Goh. Adversarial semantic hallucination for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 318–327, 2022. 2

[57] Thanh-Dat Truong, Chi Nhan Duong, Ngan Le, Son Lam Phung, Chase Rainwater, and Khoa Luu. Bimal: Bijective maximum likelihood approach to domain adaptation in semantic scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8548–8557, 2021. 2, 7

[58] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 2, 7

[59] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 5

[60] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2353–2363, 2022. 2, 7

[61] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 2, 7

[62] Ziquan Wang, Yongsheng Zhang, XianZheng Ma, Ying Yu, ZhenChao Zhang, Zhipeng Jiang, and Binbin Cheng. Semantic segmentation of foggy scenes based on progressive domain gap decoupling. *TechRxiv*, 2023. 7

[63] Yanyan Wei, Zhao Zhang, Huan Zheng, Richang Hong, Yi Yang, and Meng Wang. Sginet: Toward sufficient interaction between single image deraining and semantic segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6202–6210, 2022. 1, 2

[64] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2, 3, 6, 7, 8

[65] Zizheng Yang, Jie Huang, Jiahao Chang, Man Zhou, Hu Yu, Jinghao Zhang, and Feng Zhao. 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*, pages 14059–14070, 2023. 2

[66] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018. 3

[67] Hu Yu, Jie Huang, Yajing Liu, Qi Zhu, Man Zhou, and Feng Zhao. Source-free domain adaptation for real-world image dehazing. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6645–6654, 2022. 2, 4

[68] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao. Frequency and spatial dual guidance for image dehazing. In *European Conference on Computer Vision*, pages 181–198, 2022. 2, 4

[69] He Zhang and Vishal M Patel. Densely connected pyramid dehazing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3194–3203, 2018. 2, 7

[70] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. 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*, pages 12414–12424, 2021. 2, 7

[71] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *European Conference on Computer Vision*, pages 535–552, 2022. 2, 6

[72] Yurui Zhu, Tianyu Wang, Xueyang Fu, Xuanyu Yang, Xin Guo, Jifeng Dai, Yu Qiao, and Xiaowei Hu. Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21747–21758, 2023. 2