FIFO: LEARNING FOG-INVARIANT FEATURES FOR FOGGY SCENE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust visual recognition under adverse weather conditions is of great importance in real-world applications. In this context, we propose a new method for learning semantic segmentation models robust against fog. Its key idea is to consider the fog condition of an image as its style and close the gap between images with different fog conditions in neural style spaces of a segmentation model. In particular, since the neural style of an image is in general affected by other factors as well as fog, we introduce a *fog-pass filter* module that learns to extract a fog-relevant factor from the style. Optimizing the fog-pass filter and the segmentation model alternately gradually closes the style gap between different fog conditions and allows to learn fog-invariant features in consequence. Our method substantially outperforms previous work on three real foggy image datasets. Moreover, it improves performance on both foggy and clear weather images, while existing methods often degrade performance on clear scenes.

1 INTRODUCTION

We have witnessed great advances in semantic segmentation for the last decade. However, most of existing models and datasets focus merely on improving accuracy under controlled environments, without considering image degradation caused by adverse weather conditions (*e.g.*, fog, rain, and snow), over- and under-exposure, motion blur, sensor noise, *etc*. The robustness of semantic segmentation models against these factors is of great importance in safety-critical applications and recently has gained increasing attention (Sakaridis et al., 2018a; Dai et al., 2020; Sakaridis et al., 2019; Zendel et al., 2018; Sakaridis et al., 2018b; Son et al., 2020; Choi et al., 2021).

Motivated by this, we study semantic segmentation of *foggy* scenes. The task is challenging since fog often damages visibility of images seriously, leading to substantial performance degradation. Attaching a fog removal network to the front of an existing model is not always useful for mitigating this issue (Pei et al., 2018; Sakaridis et al., 2018b) as well as being heavy in computation and memory. The other reason for the difficulty is the absence of fully annotated data for the task. Collecting a large set of foggy scenes is not straightforward since they can be captured under only a specific condition, and it is hard to label them due to their limited visibility.

Existing methods (Sakaridis et al., 2018b;a; Dai et al., 2020) tackle these issues through synthetic foggy image datasets, which are obtained by applying realistic fog effects to fully annotated clear weather images and are used for supervised learning of semantic segmentation. Furthermore, they introduce curriculum learning approaches (Sakaridis et al., 2018a; Dai et al., 2020) that gradually adapt a model from light synthetic fog to dense real fog using unlabeled real foggy images additionally. Although these methods have achieved impressive robustness, there remains room for further improvement in that their training strategies are limited to ordinary supervised learning. In addition, the curriculum adaptation demands external modules to control the fog density of real foggy images in training, and tends to make the final model biased to foggy scenes; it thus imposes extra computation cost and often degrades performance on clear images.

To resolve the above issues, we proposed a new method that learns Fog-Invariant features for FOggy scene segmentation, dubbed FIFO. Its overall pipeline is illustrated in Fig. 1. FIFO considers the fog condition of an image as its style, ideally independent of its content, and aims to learn a segmentation model insensitive to fog style variation of input image. To this end, we first define three different domains of training images, *i.e.*, clear weather (CW), synthetic fog (SF), and real fog (RF), where

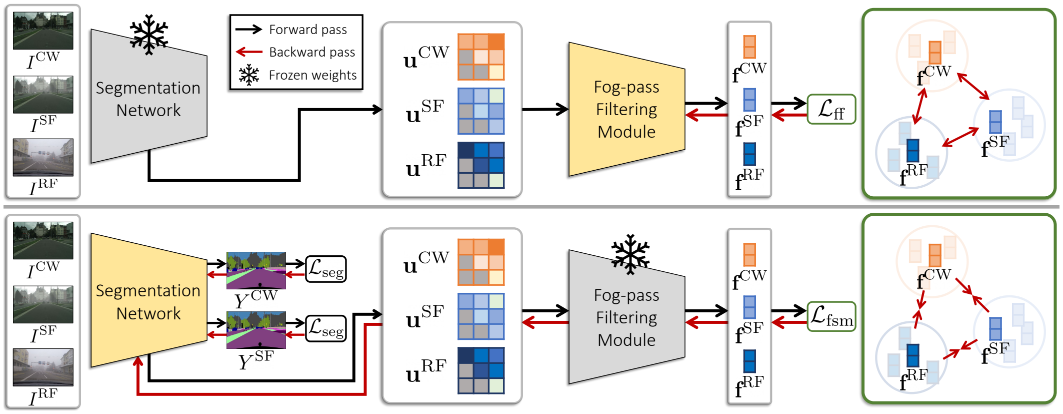
Forward pass
Backward pass
Frozen weights
I^{CW}
I^{SF}
I^{RF}
Segmentation Network
$\mathbf{u}^{\mathrm{CW}}$
$\mathbf{u}^{\mathrm{SF}}$
$\mathbf{u}^{\mathrm{RF}}$
Fog-pass Filtering Module
$\mathbf{f}^{\mathrm{CW}}$
$\mathbf{f}^{\mathrm{SF}}$
$\mathbf{f}^{\mathrm{RF}}$
$\mathcal{L}_{\mathrm{ff}}$
$\mathcal{L}_{\mathrm{seg}}$
Y^{CW}
Y^{SF}
$\mathcal{L}_{\mathrm{fsm}}$

Figure 1: Overall pipeline of FIFO. For each iteration of training, the fog-pass filtering module and the segmentation network are updated alternately. (*top*) Given Gram matrices of feature maps of the segmentation network as input, the fog-pass filtering module learns to extract fog factors, which are drawn together if they are from the same fog domain so that fog conditions of images are discriminated by their fog factors. (*bottom*) The segmentation network is trained by reducing the gap between fog factors of images with different fog conditions as well as by the segmentation loss.

images of the first two domains are labeled while those of the last one are not. FIFO then encourages the segmentation network to close the style discrepancy between different fog domains in feature spaces so that it learns fog-invariant features.

Then the success of FIFO depends heavily on the quality of the fog style representation. Unfortunately, existing style representation schemes (Gatys et al., 2016; Ulyanov et al., 2016b) are not desirable for our task since they are manually designed to capture the holistic style of an image that is affected also by factors other than fog (*e.g.*, when and where the image was taken) and even the content of the image (Choi et al., 2021); the direct use of these neural styles thus may introduce side-effects like content alteration and result in suboptimal solutions consequently.

To address this issue, we present *fog-pass filters*, learnable modules that take an ordinary neural style—the Gram matrix of a feature map (Gatys et al., 2016) as input and extract only a fog-relevant information from the style precisely in the form of embedding vectors, called *fog factors*. In particular, they learn to draw fog factors of the same domain together and hold those of different domains apart so that they discriminate fog conditions of input images through their fog factors. The segmentation model is in turn encouraged to reduce the gap between fog factors of images from different domains during training. The alternating optimization of the fog-pass filter and the segmentation network enables to gradually close the fog style gap between different domains and eventually to learn fog-invariant features.

In summary, we present a new perspective on the effect of fog in semantic segmentation and propose FIFO, a new method based on fog-invariant feature learning for semantic foggy scene segmentation. FIFO has advantages over the previous work (Sakaridis et al., 2018b;a; Dai et al., 2020) in terms of both efficiency and efficacy. It is more efficient since it allows end-to-end learning of a segmentation model and does not demand extra modules for controlling fog density of real foggy images. Also, it clearly outperforms existing records, and improves performance on both foggy and clear weather domains while existing methods often degrade performance on clear scenes.

2 Related Work

Semantic Foggy Scene Segmentation. Previous work (Sakaridis et al., 2018b;a; Dai et al., 2020) has developed fog simulators that are applied to clear images with full annotations to obtain labeled synthetic foggy images. Since supervised learning on the synthetic data limits performance due to the visual gap between synthetic and real foggy images, recent methods (Dai et al., 2020; Sakaridis et al., 2018a) further employ curriculum learning to gradually adapt a model from light synthetic fog to dense real fog. However, the curriculum adaptation often degrades performance on clear weather images and demands extra modules to control density levels of real foggy images during training.

FIFO also utilizes the synthetic dataset yet is free from the above limitations. It allows models to keep accurate on clear images while improving their performance on foggy scenes substantially. It also demands no extra module and enables end-to-end training.

Image Dehazing. Fog damages visibility of image, and accordingly, degrades visual recognition performance substantially. Numerous dehazing algorithms have been proposed so far to restore latent clean image from foggy input (Fattal, 2008; He et al., 2010; Fattal, 2014; Berman et al., 2016; Li et al., 2017a; Zhang & Patel, 2018; Chen et al., 2019; Liu et al., 2019). However, they are usually too heavy in computation to be attached to the front of recognition models. Further, recent studies (Pei et al., 2018; Sakaridis et al., 2018b) suggest that most dehazing models do not help improve recognition performance on foggy scenes. Our work instead learns a segmentation model whose features are invariant to fog. Such a model is more efficient as it does not require a separate dehazing model and is effective since its features are learned for both semantic segmentation and robustness against fog.

Robustness. Robustness of visual recognition networks against adverse conditions has been actively studied due to its importance in real-world applications (Goodfellow et al., 2015; Hendrycks & Dietterich, 2019; Zendel et al., 2018; Hendrycks et al., 2018; Sakaridis et al., 2021), and a variety of methods have been proposed to improve robustness (Son et al., 2020; Schneider et al., 2020; Shi et al., 2020; Wang et al., 2016; Lee et al., 2017; Choi et al., 2021). FIFO shares a similar idea with RobustNet (Choi et al., 2021) that regards adverse condition of input as its style. Specifically, RobustNet removes the effect of photometric transform of input from a neural style representation of a recognition model so that the model becomes invariant to the transform. Compared to RobustNet, FIFO more explicitly quantifies the effect of adverse condition through a learnable module (*i.e.*, fog-pass filter), so is able to more precisely manipulate the adverse effects during training.

Style Transfer. Neural style transfer has been studied to comprehend the style of an image apart from its content (Berger & Memisevic, 2016; Ulyanov et al., 2016a; Gatys et al., 2017; Huang & Belongie, 2017; Dumoulin et al., 2016; Li et al., 2017c; Song et al., 2019). In particular, Gatys et al. (2016) studied the Gram matrix of a feature map as a neural style representation and showed that the style of an image can be transferred to another by approximating its Gram matrix; the efficacy of Gram matrix has been proven further in later studies (Johnson et al., 2016; Luan et al., 2017). Also, Li et al. (2017b) proved that matching Gram matrices is useful for domain adaptation since it is equivalent to minimizing maximum mean discrepancy between domains. However, we found that Gram matrix is not appropriate as-is for quantifying the effect of fog in our task since it is affected other style factors or even content (Choi et al., 2021) as well as fog. We thus introduce the fog-pass filter to precisely capture only a fog-relevant factor from the Gram matrix of a feature map.

Unsupervised Domain Adaptation (UDA). Our work is also relevant to UDA since both adapt models to an unlabeled target domain. UDA methods for semantic segmentation can be categorized by the level at which adaptation is performed: Input-level (Murez et al., 2018; Hoffman et al., 2018; Pizzati et al., 2020; Kang et al., 2020), feature-level (Wang et al., 2020; Tsai et al., 2018; Luo et al., 2019a), and output-level (Zou et al., 2018; 2019; Luo et al., 2019b). FIFO is related in particular to the feature-level adaptation that learns domain-invariant features. Most of existing methods in this category (Wang et al., 2020; Tsai et al., 2018; Luo et al., 2019a) train a discriminator together with a segmentation model so that the discriminator maximizes a discrepancy between source and target domains while the segmentation model learns to minimize the discrepancy. FIFO shares a similar idea with these methods, but as will be demonstrated, closing the gap between fog factors in FIFO is more effective than fooling a fog domain classifier in fog-invariant feature learning.

3 Configuration of Training Data

Training images for FIFO are categorized into three different domains according to their fog types: clear weather (CW), synthetic fog (SF), and real fog (RF). For CW images, we adopt the Cityscapes dataset (Cordts et al., 2016), which is fully annotated for supervised learning of semantic segmentation. Meanwhile, as SF images, we utilize the Foggy Cityscapes-DBF dataset (Sakaridis et al., 2018a), which is constructed by simulating realistic fog effects on images of the Cityscapes dataset, thus also fully annotated. Finally, RF images are taken from the Foggy Zurich dataset (Sakaridis et al., 2018a), which is a collection of unlabeled foggy scenes captured in the real world.

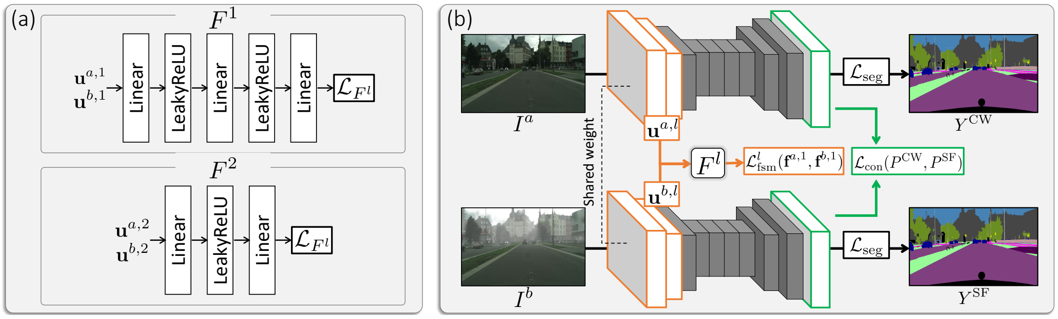

Figure 2: A schematic of FIFO. (a) The fog-pass filtering modules. Each of them takes as input the upper triangular part of a Gram matrix and returns a fog factor. The loss pulls or pushes a pair of fog factors according to the equivalence of their fog conditions. (b) Training of the segmentation network with the frozen fog-pass filters. Given a pair of images with different fog conditions as input, the network is trained by closing the gap between their fog factors and that between their segmentation predictions as well as by the ordinary segmentation loss.

Note that the way FIFO uses the two foggy image datasets is different from that of existing methods (Sakaridis et al., 2018a; Dai et al., 2020). First, in the Foggy Cityscapes-DBF dataset, FIFO fixes the density level of synthetic fog by a single value (*i.e.*, the attenuation coefficient $\beta = 0.005$) and utilizes the entire dataset. On the other hand, the previous work adopts only a refined, high-quality subset of the dataset and varies the fog level during training for the curriculum learning. Second, FIFO utilizes the Foggy Zurich dataset as a whole, whereas the previous work divides it into multiple sets of different density levels using extra modules that estimate the fog density of the images. These differences allow the pipeline of FIFO to be more concise and efficient.

4 Proposed Method

The training procedure for the fog-pass filters and the segmentation network is illustrated in Fig. 1 and their structures are presented in Fig. 2. The segmentation network is first pretrained on the Cityscapes dataset (Cordts et al., 2016), and fog-pass filters initialized randomly are attached to different feature maps of the network. Thereafter, on the dataset of the three fog domains introduced in Sec. 3, the two parts of FIFO are trained alternately per mini-batch, except for the first 5K iterations where the fog-pass filters are solely trained to avoid cold-start. Note that the fog-pass filters are used only in training for fog-invariant feature learning of the segmentation network.

To construct a mini-batch, we randomly sample the same number of images from CW and RF domains, and choose SF counterparts of the sampled CW images. Given such a mini-batch, the fog-pass filters are learned to draw fog factors of the same fog domain together and hold those of different domains apart so that they discriminate input according to its fog domain. On the other hand, the segmentation network is optimized for closing the distance between fog factors of different domains as well as for minimizing the ordinary segmentation loss. This alternating optimization closes the fog style gap between different domains precisely, leading to fog-invariant features.

The remaining part of this section first gives details of training the fog-pass filtering modules and the segmentation network, then empirically verify the key ideas of FIFO.

4.1 Fog-pass Filtering Modules

Instead of a raw feature map of the segmentation network, a fog-pass filtering module takes a holistic style representation of the feature map as input in order to focus more on the style of image by filtering out most of its content information. In this context, the style representation can be considered as a hardwired layer (Ji et al., 2010) that encodes our prior knowledge. We in particular adopt the Gram matrix of the feature map (Gatys et al., 2016) as the style representation as it provides richer style information than other methods, *e.g.*, channel-wise feature statistics (Ulyanov et al., 2016b). The Gram matrix, denoted by $\mathbf{G} \in \mathbb{R}^{c \times c}$, captures correlations between c channels of its input feature map. The (i, j) element of $\mathbf{G}$ indicates the correlation between i^{th} and j^{th} feature channels and is computed by $\mathbf{G}_{i,j} = \mathbf{a}_i^\top \mathbf{a}_j$, where $\mathbf{a}_i$ is the vector form of the i^{th} channel of the input feature map.

Specifically, since the Gram matrix is symmetric, the vector form of only the upper triangular part of the matrix is used as input to the fog-pass filtering module.

Let I^a and I^b be a pair of images from the mini-batch and F^l denote the fog-pass filter attached to the l^{th} layer of the segmentation network. Then the fog factors of the two images are computed by $\mathbf{f}^{a,l} = F^l(\mathbf{u}^{a,l})$ and $\mathbf{f}^{b,l} = F^l(\mathbf{u}^{b,l})$, respectively, where $\mathbf{u}^{a,l}$ and $\mathbf{u}^{b,l}$ denote the vectorized upper triangular parts of the Gram matrices computed from their l^{th} intermediate feature maps. The role of the fog-pass filter is to inform the segmentation network how I^a and I^b are different in terms of fog condition through $\mathbf{f}^{a,l}$ and $\mathbf{f}^{b,l}$. For this purpose, the fog-pass filter learns a space of fog factors where those of the same fog domain are grouped closely together and those of different domains are far from each other. The loss function for F^l is designed accordingly as follows:

$$\mathcal{L}_{F^l} = \sum_{(a,b)\in\mathcal{P}} \Big\{ \big(1 - \mathbb{I}(a,b)\big) \Big[m - d\big(\mathbf{f}^{a,l}, \mathbf{f}^{b,l}\big) \Big]_+^2 + \mathbb{I}(a,b) \Big[d\big(\mathbf{f}^{a,l}, \mathbf{f}^{b,l}\big) - m \Big]_+^2 \Big\}, \tag{1}$$

where $\mathcal{P}$ denotes the set of every image pair in the mini-batch and $\mathbb{I}(a,b)$ is the indicator function that returns 1 if I^a and I^b are of the same fog domain and 0 otherwise.

4.2 Segmentation Network

The segmentation network is trained using three different objectives, which are designed for semantic segmentation, fog-invariant feature learning, and consistent prediction regardless of fog condition of input, respectively. We elaborate on each of the loss functions below.

Segmentation Loss. For learning semantic segmentation, we apply the pixel-wise cross-entropy loss to individual images. To be specific, the loss is given by

$$\mathcal{L}_{\text{seg}}(\mathbf{P}, \mathbf{Y}) = -\frac{1}{n} \sum_i \sum_j \mathbf{Y}_{i,j} \log \mathbf{P}_{i,j}, \tag{2}$$

where $\mathbf{P}_{i,j} \in \mathbb{R}$ and $\mathbf{Y}_{i,j} \in \{0,1\}$ denote the predicted score and groundtruth label of class j at pixel i, respectively, while n is the number of pixels.

Fog Style Matching Loss. Given a pair of images from different fog domains, the segmentation network learns fog-invariant features that close the distance between their fog factors. To this end, the second loss matches the two fog factors given by the frozen fog-pass filters. Let $\mathbf{f}_i^{a,l}$ and $\mathbf{f}_i^{b,l}$ be the fog factors of the images computed by the fog-pass filter F^l. Then the loss is given by

$$\mathcal{L}_{\text{fsm}}^l(\mathbf{f}^{a,l}, \mathbf{f}^{b,l}) = \frac{1}{4d_l^2 n_l^2} \sum_{i=1}^{d_l} \Big(\mathbf{f}_i^{a,l} - \mathbf{f}_i^{b,l} \Big)^2, \tag{3}$$

where d_l and n_l denote the dimension of their fog factors and the spatial size of the l^{th} feature map, respectively.

Prediction Consistency Loss. A CW image and its SF counterpart have exactly the same semantic layout. By forcing predictions for these images being identical, we can align the CW and SF domains more aggressively in the learned representation. Hence, only for CW and SF images of the same origin, we encourage the model to predict the same segmentation map. Let $\mathbf{P}_i^{\mathsf{CW}} \in \mathbb{R}^c$ and $\mathbf{P}_i^{\mathsf{SF}} \in \mathbb{R}^c$ denote their class probability vectors predicted by the segmentation model for pixel i, where c is the number of classes. The third loss is designed to force the consistency between $\mathbf{P}_i^{\mathsf{CW}}$ and $\mathbf{P}_i^{\mathsf{SF}}$ for all pixels, and is given by

$$\mathcal{L}_{\text{con}}(\mathbf{P}^{\mathsf{CW}}, \mathbf{P}^{\mathsf{SF}}) = \sum_i \text{KLdiv}(\mathbf{P}_i^{\mathsf{CW}}, \mathbf{P}_i^{\mathsf{SF}}), \tag{4}$$

where $\text{KLdiv}(\cdot,\cdot)$ is the Kullback–Leibler divergence. This loss shares the same goal with the fog style matching loss in Eq. (3), but more strongly forces fog-invariance in the prediction level through a small number of CW–SF pairs. Also, it is complementary to the segmentation loss in Eq. (2) since the class probability distribution in Eq. (4) provides information beyond the categorical labels used by the segmentation loss.

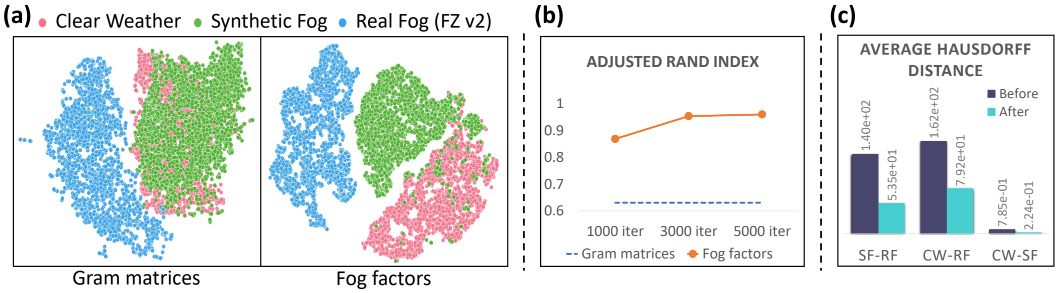

Figure 3: Empirical analysis on the impact of FIFO. (a) 2D visualization of distributions of Gram matrices and their fog factors. (b) Comparison between the quality of k-means clustering of the Gram matrices and that of the corresponding fog factors in adjusted Rand index. (c) The fog-style gap between different domains before and after training with FIFO, where the gap is measured by the average Hausdorff distance between two sets of fog factors.

Training Strategy. Given a mini-batch, the same number of image pairs are sampled from each of the three different domain pairs, *i.e.*, CW–SF, CW–RF, and SF–RF. Note that images of each CW–SF pair are of the same semantic layout so that the prediction consistency loss is applied to them. For CW–SF pairs, the segmentation network is optimized by

$$\mathcal{L}_S^{\text{CW-SF}} = \mathcal{L}_{\text{seg}}(\mathbf{P}^{\text{CW}}, \mathbf{Y}^{\text{CW}}) + \mathcal{L}_{\text{seg}}(\mathbf{P}^{\text{SF}}, \mathbf{Y}^{\text{SF}}) + \lambda_{\text{fsm}} \sum_l \mathcal{L}_{\text{fsm}}^l(\mathbf{f}^{\text{CW},l}, \mathbf{f}^{\text{SF},l}) + \lambda_{\text{con}} \mathcal{L}_{\text{con}}(\mathbf{P}^{\text{CW}}, \mathbf{P}^{\text{SF}}), \tag{5}$$

where λ_{fsm} and λ_{con} are balancing hyper-parameters and $Y^{\text{CW}} = Y^{\text{SF}}$. On the other hand, for the other pairs of input domains including RF, the loss consists of the segmentation and fog style matching terms only, and is given by

$$\mathcal{L}_S^{\text{D-RF}} = \mathcal{L}_{\text{seg}}(\mathbf{P}^{\text{D}}, \mathbf{Y}^{\text{D}}) + \lambda_{\text{fsm}} \sum_l \mathcal{L}_{\text{fsm}}^l(\mathbf{f}^{\text{D},l}, \mathbf{f}^{\text{RF},l}), \tag{6}$$

where $\text{D} \in \{\text{CW}, \text{SF}\}$. Note that $\mathcal{L}_{\text{seg}}$ is not applied to the prediction for real foggy image $\mathbf{P}^{\text{RF}}$ due to the absence of its segmentation label.

4.3 Empirical Verification

Impact of Fog-pass Filtering Modules. To justify the use of the fog-pass filters, we compare Gram matrices and their fog factors computed by a fog-pass filter in how well they disentangle the fog condition and the other aspects of an image. To this end, we examine distributions of Gram matrices and fog factors of CW, SF, and RF. To be specific, training images of the Cityscapes dataset (Cordts et al., 2016), their synthetic foggy counterparts given by the fog simulator of Dai et al. (2020) with the attenuation coefficient $\beta = 0.005$, and those of the Foggy Zurich-test v2 dataset (Sakaridis et al., 2018a) are adopted as CW, SF, and RF images, respectively; their Gram matrices are computed from ResBlock1 outputs of RefineNet-lw (Nekrasov et al., 2018) pre-trained on the Cityscapes, and the corresponding fog factors are computed from the Gram matrices through the fog-pass filtering module. Fig. 3(a) presents t-SNE visualization (van der Maaten & Hinton, 2008) of the distributions, in which Gram matrices of CW and SF largely overlap each other while fog factors are well separated according to their fog domains. This result suggests that Gram matrices are affected substantially by image content while fog factors represent only fog-relevant information as desired. The same trend is observed in Fig. 3(b) that quantitatively evaluates the quality of k-means clusters of the Gram matrices and fog factors via adjusted Rand index (Hubert & Arabie, 1985).

Fog-invariance Learned by FIFO. To investigate the impact of FIFO on fog-invariance learning, we first demonstrate that FIFO effectively reduces the gap between fog domains in the space of fog factors. To this end, each domain is represented as the set of fog factors of its images, and the gap between a pair of domains is measured by the average Hausdorff distance (Dubuisson & Jain, 1994) between such sets of the domains before and after training with FIFO. Fig. 3(c) shows that FIFO closes the fog-style gap in all three domain pairs. The impact is also verified qualitatively through images reconstructed from intermediate features of the segmentation network trained by FIFO. As a

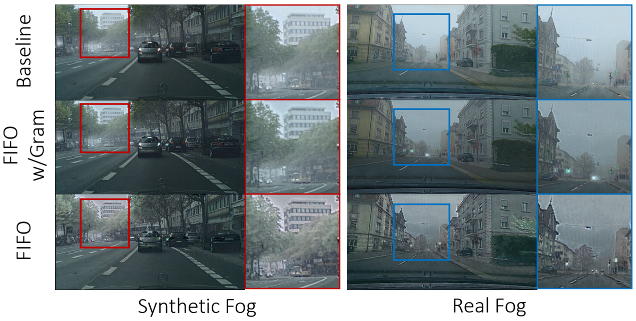

Figure 4: Images reconstructed by the baseline, a variant of FIFO closing the gap between Gram matrices, and FIFO.

reconstruction model, we adopt RefineNet-lw with two additional upsampling layers; its decoder is first trained to reconstruct images while freezing its encoder pretrained on the Cityscapes dataset, then the encoder is replaced with that of the segmentation network trained by FIFO. Also, for comparisons, we reconstruct images using a baseline (*i.e.*, training only on the Cityscapes) and a variant of FIFO with no fog-pass filter (*i.e.*, training by directly reducing the gap between Gram matrices) in the same manner. Fig. 4 presents examples of the reconstructed images, which demonstrate that FIFO allows to sharpen images effectively, well emphasize object boundaries in particular, and make fewer artifacts.

5 Experiments

5.1 Implementation Details

Network Architecture. We adopt RefineNet-lw (Nekrasov et al., 2018) with ResNet-101 (He et al., 2016) backbone as our segmentation network. Two fog-pass filtering modules are then respectively attached to the outputs of Conv1 and ResBlock1 layers of the segmentation network. As illustrated in Fig. 2(a), the two modules are implemented by multi-layer perceptrons with leaky ReLU activation functions (Maas et al., 2013).

Optimization and Hyper-parameters. The segmentation network is trained by SGD with a momentum of 0.9 and the initial learning rate of $6\mathrm{e}{-4}$ for the encoder and $6\mathrm{e}{-3}$ for the decoder; both learning rates are decreased by polynomial decay with a power of 0.5. The two fog-pass filtering modules are trained by Adamax (Kingma & Ba, 2015) with initial learning rates of $5\mathrm{e}{-4}$ (Conv1) and $1\mathrm{e}{-3}$ (ResBlock1), respectively; the dimensionality of fog factors is set to 64. Each mini-batch is constructed by sampling 4 images from each fog domain, thus its size is 12. During training, images are resized, cropped to 600×600, and flipped horizontally at random. Finally, the hyper-parameters λ_{fsm}, λ_{con}, and m are set to $5\mathrm{e}{-8}$, $1\mathrm{e}{-4}$ and 0.1, respectively.

5.2 Datasets for Evaluation

FIFO is evaluated and compared to previous work on three real foggy datasets: Foggy Zurich (FZ) test v2 (Sakaridis et al., 2018a), Foggy Driving (FD) (Sakaridis et al., 2018b), and Foggy Driving Dense (FDD) (Sakaridis et al., 2018a), where FDD is a subset of FD. Images of these datasets are foggy, captured in the real world, and fully annotated. Also, they share the same class set with the Cityscapes dataset, thus allow to validate models trained on the (Foggy) Cityscapes dataset. We further apply FIFO and previous work to an unseen clear dataset, Cityscapes lindau 40 introduced in (Dai et al., 2020), to evaluate their performance on clear weather scenes.

5.3 Quantitative Analysis

Quantitative results of FIFO and previous arts are summarized in Table 1. As shown in the table, FIFO largely outperforms CMAda3+ (Dai et al., 2020), the current best performing model based on RefineNet backbone (Nekrasov et al., 2018), on all the three foggy image datasets. These results indicate that our method of closing the fog style gap is superior to the curriculum adaptation.

We also validate the performance of the models on the clear weather dataset. In this experiment, the accuracy of CMAda3+ drops substantially; we suspect this is a side effect of the curriculum adaptation, which may lead to overfitting to foggy scenes due to catastrophic forgetting. On the other hand, FIFO enhances performance on clear weather, probably because of the data augmentation effects of using the three domains altogether during training.

Table 1: Quantitative results in mean intersection over union (mIoU) on three real foggy datasets—Foggy Zurich (FZ) test v2, Foggy Driving Dense (FDD), Foggy Driving (FD), and a clear weather dataset—Cityscapes lindau 40.

Method	Clear-weather	Synthetic&Real fog		FZ test v2	FDD	FD	C-lindau 40
	Cityscapes	SDBF	GoPro	mIoU (%)	mIoU (%)	mIoU (%)	mIoU (%)
RefineNet	✓			34.6	35.8	44.3	67.2
RefineNet-lw	✓			28.5	35.9	43.6	63.8
AdSegNet	✓	✓		25.0	15.8	29.7	-
AdvEnt	✓	✓		39.7	41.7	46.9	61.7
FDA	✓	✓		22.2	29.8	21.8	39.3
DANN	✓	✓		43.1	41.4	46.0	60.1
DANN-Gram	✓	✓		43.4	43.3	47.3	67.1
CMAda2+	✓	✓		43.4	40.1	49.9	-
CMAda3+	✓	✓		46.8	43.0	49.8	59.6
FIFO	✓	✓		**48.4**	**48.9**	**50.7**	64.8

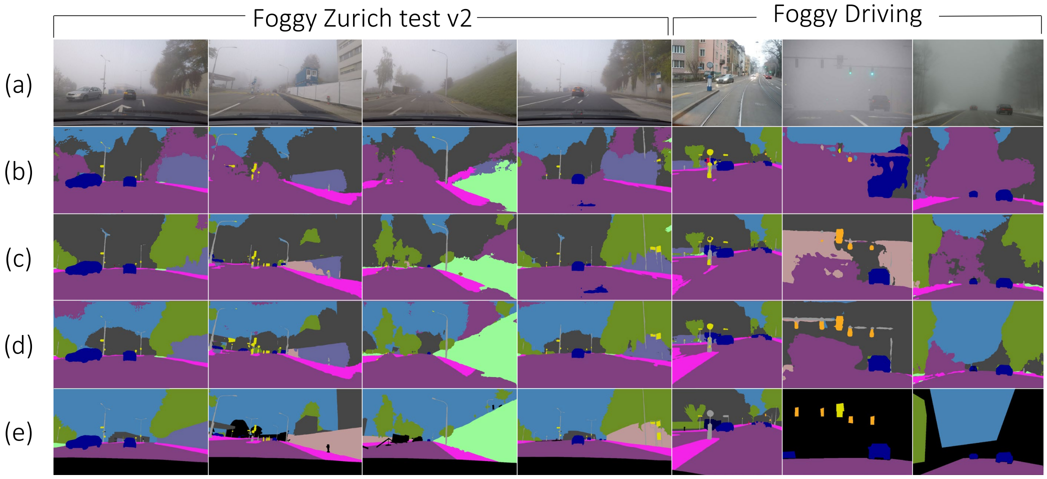

Figure 5: Qualitative results on the real foggy datasets. (a) Input images. (b) Baseline. (c) FIFO without the fog-pass filtering. (d) FIFO. (e) Groundtruth.

5.4 Qualitative Results

FIFO is qualitatively compared with two other methods. One is RefineNet-lw trained only on the Cityscapes dataset, which we call baseline. The other is a reduced version of FIFO using no fog-pass filtering module; this method learns the segmentation network while closing the gap between Gram matrices from different domains. Qualitative examples of their predictions are presented in Fig. 5. The baseline yields poor results in most cases. The reduced version of FIFO outperforms the baseline, especially for car, road, and vegetation classes, but FIFO using the fog-pass filtering modules clearly demonstrates the best segmentation results.

5.5 Comparison to Unsupervised Domain Adaptation

The task of FIFO is ostensibly identical to that of unsupervised domain adaptation (UDA) since both of them adapt models to an unlabeled target domain. Hence, one may wonder how well UDA models work for semantic foggy scene segmentation under the setting of FIFO. In this context, FIFO is compared to multiple UDA methods based on various levels of adaptation: FDA (Yang & Soatto, 2020) for input-level adaptation, AdSegNet (Tsai et al., 2018) and AdvEnt (Vu et al., 2019) for output-level adaptation, and DANN (Ganin et al., 2016) for feature-level adaptation. We also evaluate a variant of DANN whose domain classifier takes as input a Gram matrix, instead

Table 2: Analysis on the impact of domain pairs. C, S and R denote CW, SF and RF, respectively.

C–S	C–R	S–R	FZ	FDD	FD	CW
✓			43.7	38.6	46.1	67.6
	✓		37.7	40.3	47.2	66.0
		✓	39.3	42.8	49.7	61.6
✓	✓		49.7	46.0	49.9	65.8
✓		✓	46.0	47.6	50.0	62.3
✓	✓		47.4	38.2	47.0	64.3
✓	✓	✓	48.4	48.9	50.7	64.8

Table 3: Analysis on the impact of the fog style matching loss, the prediction consistency loss, and the fog-pass filtering.

Method	FZ	FDD	FD	CW
Baseline	28.5	35.9	43.6	63.8
FIFO w/o $\mathcal{L}_{\text{fsm}}$	31.7	38.5	45.1	62.4
FIFO w/o $\mathcal{L}_{\text{con}}$	41.6	45.4	48.9	63.5
FIFO w/ Gram	41.3	43.8	49.1	63.1
FIFO	**48.4**	**48.9**	**50.7**	**64.8**

of a raw feature map, like FIFO; this variant of DANN is denoted by DANN-Gram. The UDA models are trained using the same datasets, *i.e.*, CW, SF, and RF. To be specific, input- and output-level adaptation models are first pretrained on CW, then trained using SF while adapting to RF. Meanwhile, DANN and DANN-Gram are trained using CW, SF, and RF at once like FIFO: Their discriminators are optimized to maximize the discrepancy of fog domains like the fog-pass filtering modules in FIFO while their segmentation networks are learned to minimize the discrepancy.

The performance of these UDA methods is reported in Table 1. As shown in the table, the UDA models are all inferior to FIFO and CMAda3+, which suggests that, even though it is apparently similar to UDA for semantic segmentation, semantic foggy scene segmentation is of a different nature and has its own challenges. In the typical UDA setting, each of source and target domain has its own style, by which it can be defined. However, the style of a foggy scene is not determined only by its fog condition, rather is the result of the chemical combination of fog and other style factors of the scene. The UDA methods that consider SF and RF as domains with unique styles are thus not well suited to the task. Moreover, fog substantially damages visibility, and enlarges intra-domain variations since the effect of fog varies significantly according to the 3D configuration of the scene (Dai et al., 2020). Due to these differences and challenges, semantic foggy scene segmentation demands dedicated solutions like FIFO.

5.6 Ablation Study

We conduct extensive experiments while varying domain pairs of FIFO to investigate their effects. Table 2 summarizes the results. We found that the model trained using the CW pair has better performance than the model learned without the CW pair (*i.e.*, RF and SF). Also, using every domain pair contributes to performance on real foggy datasets as FIFO using all 3 pairs beats most variants using only 2 pairs.

We also investigate contributions of the fog-pass filtering modules and the prediction consistency loss $\mathcal{L}_{\text{con}}$ to the performance. Table 3 compares FIFO with its variants with and without the fog-pass filters and $\mathcal{L}_{\text{con}}$ in terms of segmentation quality on real foggy images. Note that, during training the segmentation network, FIFO w/o $\mathcal{L}_{\text{fsm}}$ completely drops fog style matching between different fog domains, while FIFO w/ Gram uses Gram matrices instead of fog factors when matching fog styles. The results in the tables suggest that all the losses and the fog-pass filtering contribute to the performance on all the three real foggy datasets, but the impact of the fog style matching is substantially larger than the others. Also, the gap between FIFO and FIFO w/ Gram demonstrates the superiority of fog factors over Gram matrices, which justifies the use of the fog-pass filters.

6 Conclusion

We have presented a new approach to learning fog-invariant features for foggy scene segmentation. It precisely quantifies the fog style of an image through the fog-pass filtering modules and learns a segmentation network for closing the gap between images of different fog conditions in the fog style space. Its efficacy has been demonstrated on public benchmarks for semantic foggy scene segmentation, where it beats every previous art without sacrificing performance on clear weather images. Moreover, unlike the current best-performing method, it enables end-to-end learning of segmentation models and demands no extra module nor human intervention for training.

7 REPRODUCIBILITY

In our paper, the detailed architectures of the fog-pass filtering modules and the segmentation network are shown in Fig. 2, the overall training pipeline is elaborated in Algorithm A.1, and implementation details are presented in Sec. 5.1.

8 CODE OF ETHICS

The semantic segmentation models dealing with robustness issues are related to safety-critical applications such as autonomous driving. FIFO can be also applied to the safety-critical situations, so there is a problem regarding the responsibility when applied to.

REFERENCES

Guillaume Berger and Roland Memisevic. Incorporating long-range consistency in cnn-based texture generation. In *Proc. International Conference on Learning Representations (ICLR)*, 2016.

Dana Berman, Tali Treibitz, and Shai Avidan. Non-local image dehazing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Wei-Ting Chen, Jian-Jiun Ding, and Sy-Yen Kuo. Pms-net: Robust haze removal based on patch map for single images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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 (IJCV)*, 2020.

M-P Dubuisson and Anil K Jain. A modified hausdorff distance for object matching. In *Proc. International Conference on Pattern Recognition (ICPR)*, 1994.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *Proc. International Conference on Learning Representations (ICLR)*, 2016.

Raanan Fattal. Single image dehazing. *ACM transactions on graphics (TOG)*, 2008.

Raanan Fattal. Dehazing using color-lines. *ACM transactions on graphics (TOG)*, 2014.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 17(1):2096–2030, 2016.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.

Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. International Conference on Learning Representations (ICLR)*, 2019.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proc. International Conference on Machine Learning (ICML)*, 2018.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 1985.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. In *Proc. International Conference on Machine Learning (ICML)*, 2010.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. European Conference on Computer Vision (ECCV)*, 2016.

Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander G Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.

S. Lee, J. Kim, J. S. Yoon, S. Shin, O. Bailo, N. Kim, T. H. Lee, H. S. Hong, S. H. Han, and I. S. Kweon. Vpgnet: Vanishing point guided network for lane and road marking detection and recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017a.

Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017b.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2017c.

Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019a.

Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019b.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. International Conference on Machine Learning (ICML)*, 2013.

Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Vladimir Nekrasov, Chunhua Shen, and Ian Reid. Light-weight refinenet for real-time semantic segmentation. *arXiv preprint arXiv:1810.03272*, 2018.

Yanting Pei, Yaping Huang, Qi Zou, Yuhang Lu, and Song Wang. Does haze removal help cnn-based image classification? In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

Fabio Pizzati, Raoul de Charette, Michela Zaccaria, and Pietro Cerri. Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proc. European Conference on Computer Vision (ECCV)*, 2018a.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision (IJCV)*, 2018b.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

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 (ICCV)*, 2021.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020.

Baifeng Shi, Dinghuai Zhang, Qi Dai, Zhanxing Zhu, Yadong Mu, and Jingdong Wang. Informative dropout for robust representation learning: A shape-bias perspective. In *Proc. International Conference on Machine Learning (ICML)*, 2020.

Taeyoung Son, Juwon Kang, Namyup Kim, Sunghyun Cho, and Suha Kwak. Urie: Universal image enhancement for visual recognition in the wild. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

Chunjin Song, Zhijie Wu, Yang Zhou, Minglun Gong, and Hui Huang. Etnet: Error transition network for arbitrary style transfer. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2019.

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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *Proc. International Conference on Machine Learning (ICML)*, 2016a.

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

L.J.P van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9: 2579–2605, 2008.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang. Studying very low resolution recognition using deep networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. Wilddash - creating hazard-aware benchmarks. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

He Zhang and Vishal M Patel. Densely connected pyramid dehazing network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

A Appendix

A.1 Algorithm of FIFO

We present detailed training procedure of FIFO in Algorithm 1.

Algorithm 1 : Training FIFO

Input: Pretrained fog-pass filtering module for the l^{th} layer: $F^l(\cdot)$, Segmentation network: $S(\cdot)$, Number of layers: L, Batch size per domain: m, Segmentation prediction: P, Segmentation label: Y, Input image set $\{I^{\mathsf{CW}}, I^{\mathsf{SF}}, I^{\mathsf{RF}}\}$: x, Subset of two elements from domain set $\{\mathsf{CW},\mathsf{SF},\mathsf{RF}\}$: $\{a, b\}$ and Segmentation label set $\{Y^{\mathsf{CW}}, Y^{\mathsf{RF}}\}$: y.
Output: Optimized segmentation network $S(\cdot)$.

```
for $\{1, \ldots,$ # of training iterations$\}$ do
    Sample mini-batch $\{x_i\}_{i=1}^m$
    for $\{l \leftarrow 1$ to $L\}$ do
        $\mathcal{L}_{\mathrm{F}^l} \leftarrow \mathcal{L}_{\mathrm{F}^l}(\{\mathbf{f}_i^l\}_{i=1}^m)$                    ▷ Eq. 1
        Update the fog-pass filtering module $F^l$
    end for
    Sample mini-batch $\{x_j\}_{j=1}^m$ and $\{y_j\}_{j=1}^m$
    for $\{l \leftarrow 1$ to $L\}$ do
        $\{\mathbf{f}_j^l\}_{j=1}^m, \leftarrow \{F^l(\mathbf{u}_j^l)\}_{j=1}^m$
        $\mathcal{L}_{\text{fsm}}^l \leftarrow \{\mathcal{L}_{\text{fsm}}^l(\mathbf{f}_j^{a,l}, \mathbf{f}_j^{b,l})\}_{j=1}^m$                    ▷ Eq. 3
    end for
    Sample the pair $\{I^a, I^b\} \in x_j$
    if $\{a, b\} == \{\mathsf{CW}, \mathsf{SF}\}$ then
        $\mathcal{L}_{\text{con}} \leftarrow \sum_i \text{KLdiv}(P_i^a, P_i^b)$                    ▷ Eq. 4
    end if
    if $\{a, b\} \cap \{\mathsf{CW}, \mathsf{SF}\} \neq \emptyset$ then
        $\mathcal{L}_{\text{seg}} \leftarrow -\frac{1}{n}\sum Y \log P$                    ▷ Eq. 2
    end if
    $\mathcal{L}_{\mathrm{S}} \leftarrow \sum_l \mathcal{L}_{\text{fsm}}^l + \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{seg}}$                    ▷ Eq. 5, 6
    Update the segmentation network $S$
end for
```

Consequently, the total objective of FIFO is following:

$$\sum_l \min_{F^l} \mathcal{L}_{F^l}^l + \min_S(\sum_l \mathcal{L}_{\text{fsm}}^l + \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{seg}}), \tag{7}$$

where l is the layer index.

A.2 Independence Analysis of Fog Factors

In this section, we quantitatively evaluate the independence of the fog factors and Gram matrices, with related to image content. To this end, we design content-pass filtering module which is optimized to extract content-relevant information; which we call content factors.

Training Content-pass Filtering Module Let I^a and I^b be a pair of images from the mini-batch and C^l denote the content-pass filtering module attached to the l^{th} layer of the segmentation network. Then the content factors of the two images are computed by $\mathbf{c}^{a,l} = C^l(\mathbf{u}^{a,l})$ and $\mathbf{c}^{b,l} = C^l(\mathbf{u}^{b,l})$. In contrast to the fog-pass filtering module, this module is optimized to learn an embedding space of content factors where the pairs having the same content, *i.e.*, CW–SF are grouped closely and else pairs are far from each other. The loss function for C^l is designed accordingly as follows:

$$\mathcal{L}_{C^l} = \sum_{(a,b)\in\mathcal{P}} \Big\{ \big(1 - \mathbb{I}(a,b)\big) \Big[m - d\big(\mathbf{f}^{a,l}, \mathbf{f}^{b,l}\big) \Big]_+^2 + \mathbb{I}(a,b) \Big[d\big(\mathbf{f}^{a,l}, \mathbf{f}^{b,l}\big) - m \Big]_+^2 \Big\}, \tag{8}$$

where $\mathcal{P}$ denotes the set of every image pair in the mini-batch and $\mathbb{I}(a, b)$ is the indicator function that returns 1 if the pair of I^a and I^b is a CW–SF pair and 0 otherwise.

Independence Analysis of Fog Factors We design an *independence score* to quantitatively evaluate the independence of fog factors with relation to content factors. The score is measured as an intersection of the similarity relationship of elements within each embedding space of fog factors and content factors. For each embedding space, we select one factor and search k th most similar factors having high cosine similarity with the selected factor. Then, for each space, we check which image the k factors were extracted from, and compute the proportion of the corresponding images overlapped between embedding spaces. Finally, we repeat the process for all of the fog factors in the fog embedding space and define the average proportion for all of the fog factors as an *independence score*.

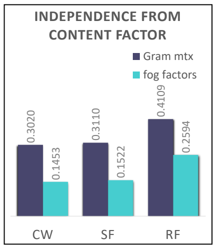

Figure 6: Independence score of fog factors and Gram matrices on content factors.

Let I, f and c be an image, a fog factor and a content factor. Consequently, the *independence score* is calculated as follows:

$$score(\mathcal{F}, \mathcal{C}) = \frac{1}{N}\sum_{i=1}^{N} \frac{|\{I_n | f_n \in \mathcal{F}, d(f_i, f_n) \leq d(f_i, f_k)\} \cap \{I_m | c_m \in \mathcal{C}, d(c_i, c_m) \leq d(c_i, c_k)\}|}{k}, \tag{9}$$

where k is a number of selecting similar factors and set to 200, f_k and c_k are the k th most similar fog factor from f_i and content factor from c_i, then $\mathcal{F}$ and $\mathcal{C}$ denote the set of fog factors and content factors, respectively.

Fig. 6 presents an *independence score* computed from fog factors and Gram matrices, with relation to content factors. Note that the experiment settings for Gram matrices, fog factors, and dataset configurations are all the same as in the main paper. The scores show fog factors are more independent of content factors compared to Gram matrices, as we desired. These indicate that the fog-pass filtering module extracts only fog-relevant information apart from the image content.

A.3 Class-wise Quantitative Results

We evaluate class-wise IoU for baseline, RefineNet-lw trained on Cityscapes, and FIFO as presented in Tab. 4 and Tab. 5, respectively. The '-' means the image of that class does not exist in the corresponding evaluation dataset. FIFO shows better performance on most classes compared to baseline, which indicates the superiority of our method.

Table 4: Classwise performance of baseline model, *i.e.*, RefineNet-lw pretrained Cityscapes dataset. We report IoU for each class and mIoU as the evaluation metric.

Baseline	roa	sid	bui	wal	fen	pol	tli	tsi	veg	ter	sky	per	rid	car	tru	bus	tra	mot	bic	mean
FZ	53.2	51.6	35.1	28.6	19.4	36.8	54.0	56.1	24.5	34.7	57.7	0.0	3.6	79.7	-	0.0	-	0.0	5.7	28.5
FDD	83.7	0.0	27.0	-	-	67.0	8.3	73.2	51.7	-	77.9	91.1	-	44.7	14.4	0.0	-	-	-	35.9
FD	88.0	26.6	68.1	28.6	14.6	42.5	44.3	54.5	63.0	9.1	86.9	64.5	46.7	65.0	6.8	13.0	27.5	28.6	49.2	43.6
C-Lindau	91.5	58.3	92.1	45.7	68.4	53.7	46.4	83.4	89.9	74.0	94.4	84.1	67.1	90.0	74.5	18.2	-	-	79.7	63.8

Table 5: Classwise performance of FIFO. We report IoU for each class and mIoU as the evaluation metric.

Ours	roa	sid	bui	wal	fen	pol	tli	tsi	veg	ter	sky	per	rid	car	tru	bus	tra	mot	bic	mean
FZ	65.5	57.7	54.3	49.9	25.8	44.1	60.2	61.6	73.1	63.6	70.4	6.3	45.5	84.9	-	46.7	-	46.7	15.1	48.4
FDD	89.3	2.48	46.7	-	-	62.0	29.3	80.8	67.2	-	88.6	91.1	-	67.0	34.3	75.3	-	-	-	48.9
FD	90.8	39.1	72.9	24.2	19.9	42.3	51.0	59.1	72.0	9.4	90.2	64.7	48.5	71.0	25.4	65.7	43.5	24.8	49.1	50.7
C-Lindau	90.2	55.7	91.5	43.7	57.7	50.1	55.2	80.5	89.0	61.9	93.7	83.0	70.6	86.5	63.6	12.0	-	-	80.9	64.8

Table 6: Performance of FIFO trained with different density levels of synthetic fog images. $\beta = 0.005$ denotes FIFO in the main paper, and $\beta = 0$ indicates the model trained only with clear weather and real foggy images. The reported scores are mIoU on the Foggy Zurich (FZ) test v2, Foggy Driving Dense (FDD), Foggy Driving (FD).

β	FZ test v2	FDD	FD	C-Lindau
0	37.7	40.3	47.2	66.0
0.0025	45.5	40.4	45.0	67.5
0.01	42.4	45.4	50.0	61.9
0.02	42.9	42.5	48.6	60.6
0.005	48.4	48.9	50.7	64.8

Table 7: Ablation study on layers where fog style matching loss is applied. FIFO utilizes the output of the first convolutional layer and first residual blocks to apply the fog style matching loss. We report mIoU as the evaluation metric.

	Conv1	Res Block1	FZ test v2	FDD	FD	C-Lindau
C1	✓		45.3	41.0	48.3	60.6
R1		✓	45.1	39.1	47.0	61.8
Ours (C1:R2)	✓	✓	48.4	48.9	50.7	64.8

A.4 Effect of Synthetic Fog Density

This section demonstrates the effect of our chosen value of β, which is the attenuation coefficient used for generating synthetic fog Sakaridis et al. (2018b). Note that, unlike the previous work Sakaridis et al. (2018a); Dai et al. (2020), our method fixes the thickness of synthetic fog by a single value to resolves issues of the curriculum adaptation. The value of β thus has to be well defined in our model; overly dense fog blinds images too much, while overly light fog is not enough to simulate real foggy scenes. To investigate the impact of β on the semantic segmentation performance on real foggy images, we design and evaluate variants of our model trained using different values of β. In all experiments, we train FIFO as proposed in the main paper, but with different synthetic foggy images generated from various values of β. Note that the larger the value of β, the thicker the generated fog, and vice versa.

Table. 6 show that 0.005 is the most optimal value for FIFO. For light synthetic fog (β=0.0025), the performance is lower since the synthetic fog does not simulate real fog well enough due to the fog density discrepancy between them. Especially, the performance is poor on FDD and FD datasets since they are dense foggy datasets. The performance obtained using dense synthetic foggy images (β=0.02) is also poorer because they are hard to be recognized due to limited visibility from the dense fog and artifacts.

A.5 Effect of Layer Selection

This section presents an ablation study regarding layers of the segmentation network. To this end, we present the segmentation performance of FIFO, changing the layer to apply fog style matching loss ($\mathcal{L}_{fsm}$). Specifically, we start by applying fog style matching loss to the output of the first convolutional layer of ResNet and denote it as C1. Also, applying fog style matching loss to the output of the first residual block layer of ResNet and denote it as R1. Then we apply fog style matching loss to the output of all of them and denote it as C1:R1. FIFO and its variants are evaluated on real foggy images in Table. 7

As summarized in Table. 7, the semantic segmentation performance improves as more feature map outputs are involved in FIFO. Overall, C1:R2, which is our final model, shows the best performance compared to other alternatives.

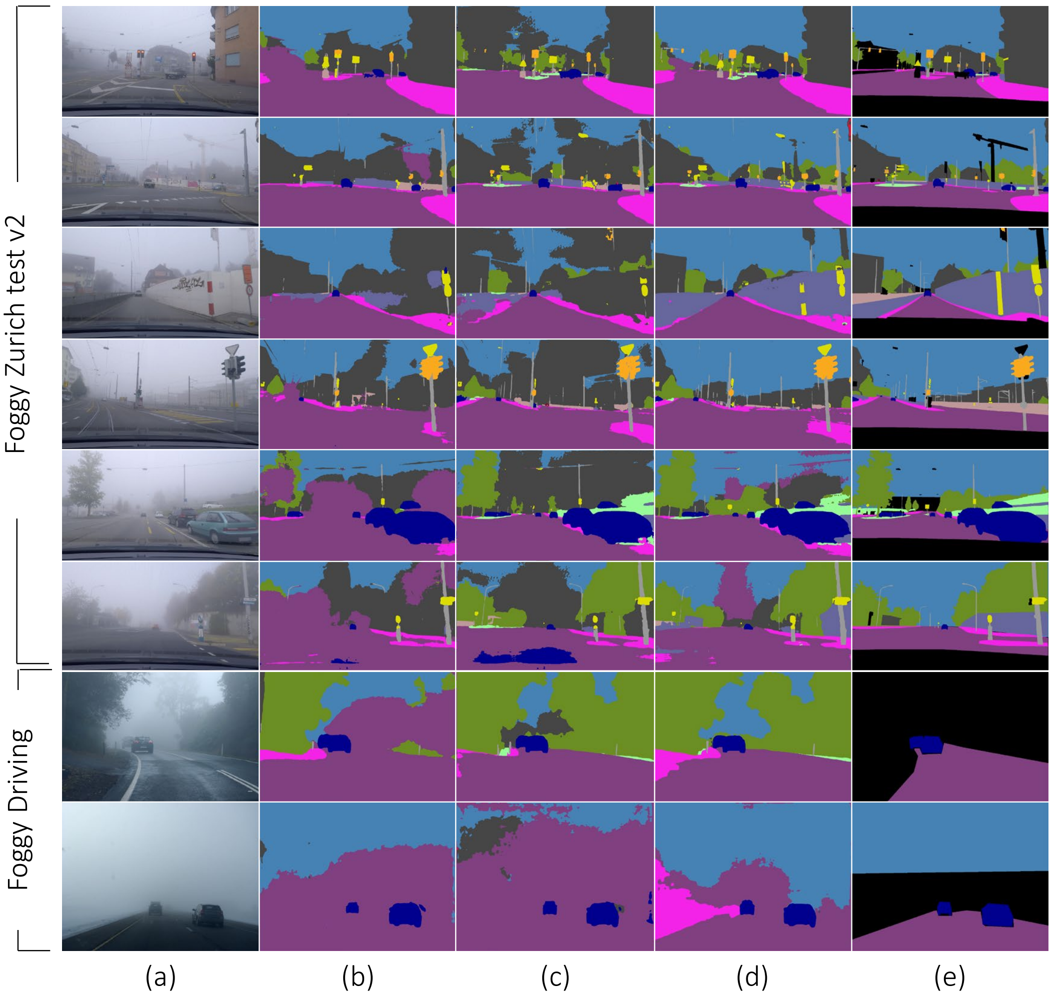

Figure 7: Additional qualitative results on the real foggy datasets. (a) Real foggy images. (b) Baseline. (c) FIFO closing the gap between gram matrices. (d) FIFO. (e) Groundtruth.

A.6 Additional Qualitative Results

This section presents additional qualitative results omitted in the main sections due to the space limit. More segmentation results of FIFO are illustrated in Fig. 7. We compare the results between FIFO, a variant of FIFO by reducing directly the gap between gram matrices, and baseline. Overall, FIFO offers higher quality segmentation results than the baseline regardless of fog density and datasets. Specifically, FIFO seems best performing on parts where dense fog is laid while other models fail, which indicates FIFO working as desired. Fig 8 exihibits additional qualitative results on image reconstruction. Likewise, the image quality where dense fog is laid is improved, which implies FIFO extract fog-invariant features. In addition, clear weather images, as well as foggy images, become more clear when the features are trained by FIFO.

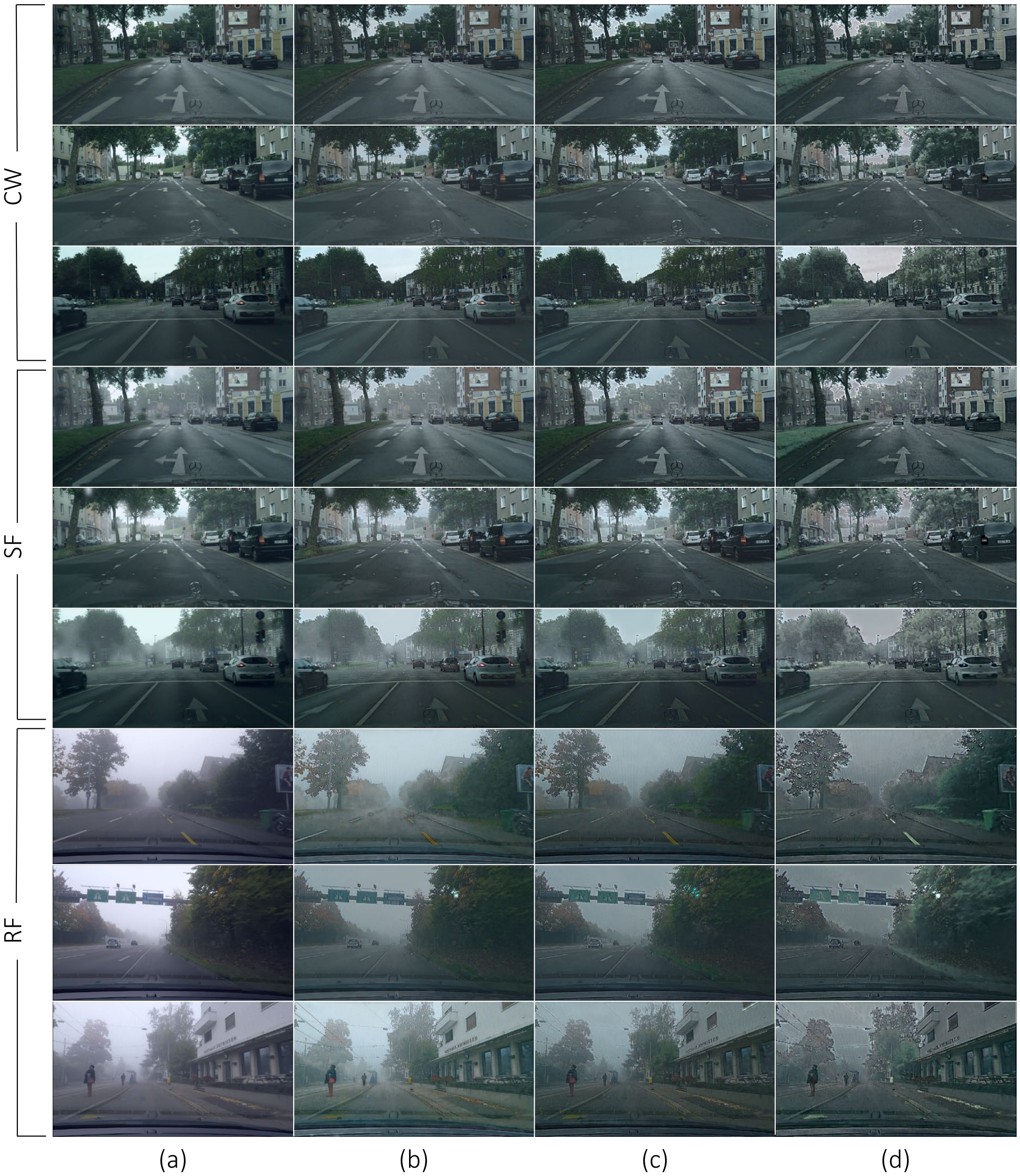

Figure 8: Additional qualitative results on image reconstruction. (a) Real foggy images. (b) Baseline. (c) FIFO closing the gap between gram matrices. (d) FIFO.