Model Adaptation with Synthetic and Real Data for Semantic Dense Foggy Scene Understanding

Author(s):
Sakaridis, Christos; Dai, Dengxin; Hecker, Simon; Van Gool, Luc

Publication date:
2018

Permanent link:
https://doi.org/10.3929/ethz-b-000305722

Originally published in:
Lecture Notes in Computer Science 11217, https://doi.org/10.1007/978-3-030-01261-8_42

Model Adaptation with Synthetic and Real Data for Semantic Dense Foggy Scene Understanding

Christos Sakaridis[1(✉)], Dengxin Dai[1], Simon Hecker[1], and Luc Van Gool[1,2]

[1] ETH Zürich, Zürich, Switzerland
{csakarid,dai,heckers,vangool}@vision.ee.ethz.ch
[2] KU Leuven, Leuven, Belgium

Abstract. This work addresses the problem of semantic scene understanding under dense fog. Although considerable progress has been made in semantic scene understanding, it is mainly related to clear-weather scenes. Extending recognition methods to adverse weather conditions such as fog is crucial for outdoor applications. In this paper, we propose a novel method, named Curriculum Model Adaptation (CMAda), which *gradually* adapts a semantic segmentation model from light synthetic fog to dense real fog in multiple steps, using both synthetic and real foggy data. In addition, we present three other main stand-alone contributions: 1) a novel method to add synthetic fog to real, clear-weather scenes using semantic input; 2) a new fog density estimator; 3) the *Foggy Zurich* dataset comprising 3808 real foggy images, with pixel-level semantic annotations for 16 images with dense fog. Our experiments show that 1) our fog simulation slightly outperforms a state-of-the-art competing simulation with respect to the task of semantic foggy scene understanding (SFSU); 2) CMAda improves the performance of state-of-the-art models for SFSU significantly by leveraging unlabeled real foggy data. The datasets and code will be made publicly available.

Keywords: Semantic foggy scene understanding, fog simulation, synthetic data, curriculum model adaptation, curriculum learning

1 Introduction

Adverse weather conditions create visibility problems for both people and the sensors that power automated systems [25,37,48]. While sensors and the downstreaming vision algorithms are constantly getting better, their performance is mainly benchmarked with clear-weather images. Many outdoor applications, however, can hardly escape from bad weather. One typical example of adverse weather conditions is fog, which degrades the visibility of a scene significantly [36, 52]. The denser the fog is, the more severe this problem becomes.

During the past years, the community has made a tremendous progress on image dehazing (defogging) to increase the visibility of foggy images [24,40,56]. The last few years have also witnessed a leap in object recognition. The semantic understanding of foggy scenes, however, has received little attention, despite its

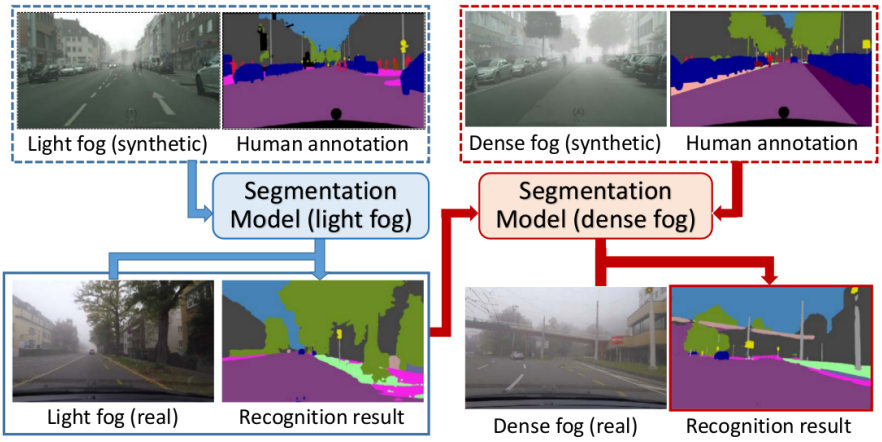


Fig. 1. The illustrative pipeline of our approach for semantic scene understanding under dense fog

importance in outdoor applications. For example, an automated car still needs to detect other traffic agents and traffic control devices in the presence of fog. This work investigates the problem of semantic foggy scene understanding (SFSU).

The current "standard" policy for addressing semantic scene understanding is to train a neural network with many annotations of real images [11,47]. Applying the same protocol to diverse weather conditions seems to be problematic, as the manual annotation part is hard to scale. The difficulty of data collection and annotation increases even more for adverse weather conditions. To overcome this problem, two streams of research have gained extensive attention: 1) transfer learning [9] and 2) learning with synthetic data [46,48].

Our method falls into the middle ground, and aims to combine the strength of these two kinds of methods. In particular, our method is developed to learn from 1) a dataset with high-quality synthetic fog and corresponding human annotations, and 2) a dataset with a large number of images with real fog. The goal of our method is to improve the performance of SFSU without requiring extra human annotations.

To this aim, this work proposes a novel fog simulator to generate high-quality synthetic fog into real images that contain clear-weather outdoor scenes, and then leverage these partially synthetic foggy images for SFSU. The new fog simulator builds on the recent work in [48], by introducing a semantic-aware filter to exploit the structures of object instances. We show that learning with our synthetic data improves the performance for SFSU. Furthermore, we present a novel method, dubbed Curriculum Model Adaptation (CMAda), which is able to *gradually* adapt a segmentation model from light synthetic fog to dense real fog in multiple steps, by using both synthetic and real foggy data. CMAda improves upon direct adaptation significantly on two datasets with dense real fog.

The main contributions of the paper are: 1) a new automatic and scalable pipeline to generate high-quality synthetic fog, with which new datasets are generated; 2) a novel curriculum model adaptation method to learn from both synthetic and (unlabeled) real foggy images; 3) a new real foggy dataset with 3808 images, including 16 finely annotated images with dense fog. A visual overview of our approach is presented in Fig. 1.

2 Related Work

Our work is relevant to image defogging (dehazing), foggy scene understanding, and domain adaptation.

2.1 Image Defogging/Dehazing

Fog fades the color of observed objects and reduces their contrast. Extensive research has been conducted on image defogging (dehazing) to increase the visibility of foggy scenes [5,15,16,24,36,40,52]. Certain works focus particularly on enhancing foggy road scenes [38,54]. Recent approaches also rely on trainable architectures [53], which have evolved to end-to-end models [34,59]. For a comprehensive overview of dehazing algorithms, we point the reader to [32,57]. Our work is complementary and focuses on semantic foggy scene understanding.

2.2 Foggy Scene Understanding

Typical examples in this line include road and lane detection [3], traffic light detection [28], car and pedestrian detection [19], and a dense, pixel-level segmentation of road scenes into most of the relevant semantic classes [7,11]. While deep recognition networks have been developed [20,33,45,58,60] and large-scale datasets have been presented [11,19], that research mainly focused on clear weather. There is also a large body of work on fog detection [6,17,42,51]. Classification of scenes into foggy and fog-free has been tackled as well [43]. In addition, visibility estimation has been extensively studied for both daytime [22,35,55] and nighttime [18], in the context of assisted and autonomous driving. The closest of these works to ours is [55], in which synthetic fog is generated and foggy images are segmented to *free-space area* and *vertical objects*. Our work differs in that our semantic understanding task is more complex and we tackle the problem from a different route by learning jointly from synthetic fog and real fog.

2.3 Domain Adaptation

Our work bears resemblance to transfer learning and model adaptation. Model adaptation across weather conditions to semantically segment simple road scenes is studied in [31]. More recently, domain adversarial based approaches were proposed to adapt semantic segmentation models both at pixel level and feature level from simulated to real environments [27,49]. Our work closes the domain

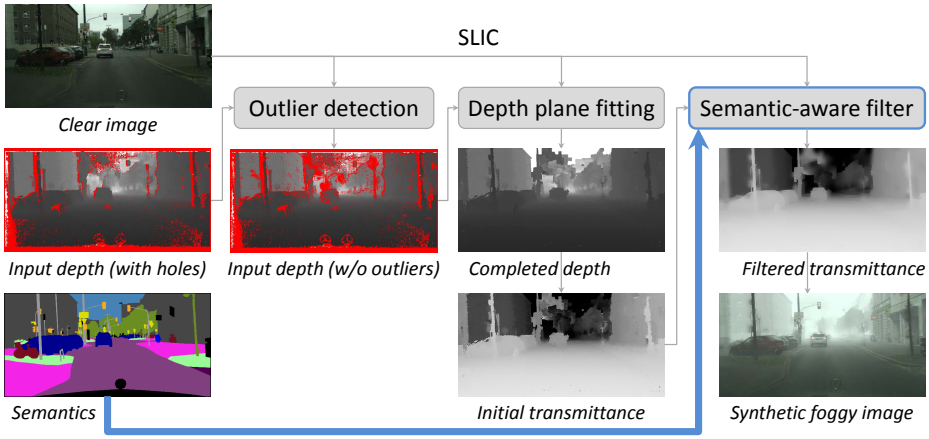


Fig. 2. The pipeline of our fog simulation using semantics

gap by generating synthetic fog and by using the policy of *gradual* adaptation. Combining our method and the aforementioned transfer learning methods is a promising direction. The concurrent work in [13] on adaptation of semantic models from daytime to nighttime solely with real data is closely related to ours.

3 Fog Simulation on Real Scenes Using Semantics

3.1 Motivation

We drive our motivation for fog simulation on real scenes using semantic input from the pipeline that was used in [48] to generate the Foggy Cityscapes dataset, which primarily focuses on depth denoising and completion. This pipeline is denoted in Fig. 2 with thin gray arrows and consists of three main steps: depth outlier detection, robust depth plane fitting at the level of SLIC superpixels [2] using RANSAC, and postprocessing of the completed depth map with guided image filtering [23]. Our approach adopts the general configuration of this pipeline, but aims to improve its postprocessing step by leveraging the semantic annotation of the scene as additional reference for filtering, which is indicated in Fig. 2 with the thick blue arrow.

The guided filtering step in [48] uses the clear-weather color image as guidance to filter depth. However, as previous works on image filtering [50] have shown, guided filtering and similar joint filtering methods such as cross-bilateral filtering [14, 44] transfer the structure that is present in the guidance/reference image to the output target image. Thus, any structure that is specific to the reference image but irrelevant for the target image is also transferred to the latter erroneously.

Whereas previous approaches such as mutual-structure filtering [50] attempt to estimate the common structure between reference and target images, we identify this common structure with the structure that is present in the ground-truth

semantic labeling of the image. In other words, we assume that edges which are shared by the color image and the depth map generally coincide with *semantic edges*, *i.e.* locations in the image where the semantic classes of adjacent pixels are different. Under this assumption, the semantic labeling can be used directly as the reference image in a classical cross-bilateral filtering setting, since it contains exactly the mutual structure between the color image and the depth map. In practice, however, the boundaries drawn by humans in the semantic annotation are not pixel-accurate, and using the color image as additional reference helps to capture the precise shape of edges better. As a result, we formulate the postprocessing step of the completed depth map in our fog simulation as a *dual-reference* cross-bilateral filter, with color and semantic reference.

3.2 Dual-reference Cross-bilateral Filter Using Color and Semantics

Let us denote the RGB image of the clear-weather scene by $\mathbf{R}$ and its CIELAB counterpart by $\mathbf{J}$. We consider CIELAB, as it has been designed to increase perceptual uniformity and gives better results for bilateral filtering of color images [41]. The input image to be filtered in the postprocessing step of our pipeline constitutes a scalar-valued transmittance map $\hat{t}$. We provide more details on this transmittance map in Sec. 3.3. Last, we are given a labeling function

$$h : \mathcal{P} \to \{1, \ldots, C\} \tag{1}$$

which maps pixels to semantic labels, where $\mathcal{P}$ is the discrete domain of pixel positions and C is the total number of semantic classes in the scene. We define our dual-reference cross-bilateral filter with color and semantic reference as

$$t(\mathbf{p}) = \frac{\sum\limits_{q \in \mathcal{N}(\mathbf{p})} G_{\sigma_s}(\|\mathbf{q}-\mathbf{p}\|)\left[\delta(h(\mathbf{q})-h(\mathbf{p})) + \mu G_{\sigma_c}(\|\mathbf{J}(\mathbf{q})-\mathbf{J}(\mathbf{p})\|)\right]\hat{t}(\mathbf{q})}{\sum\limits_{q \in \mathcal{N}(\mathbf{p})} G_{\sigma_s}(\|\mathbf{q}-\mathbf{p}\|)\left[\delta(h(\mathbf{q})-h(\mathbf{p})) + \mu G_{\sigma_c}(\|\mathbf{J}(\mathbf{q})-\mathbf{J}(\mathbf{p})\|)\right]}, \tag{2}$$

where $\mathbf{p}$ and $\mathbf{q}$ denote pixel positions, $\mathcal{N}(\mathbf{p})$ is the neighborhood of $\mathbf{p}$, δ denotes the Kronecker delta, G_{σ_s} is the spatial Gaussian kernel, G_{σ_c} is the color-domain Gaussian kernel and μ is a positive constant. The novel dual reference is demonstrated in the second factor of the filter weights, which constitutes a sum of the terms $\delta(h(\mathbf{q}) - h(\mathbf{p}))$ for semantic reference and $G_{\sigma_c}(\|\mathbf{J}(\mathbf{q}) - \mathbf{J}(\mathbf{p})\|)$ for color reference, weighted by μ. The formulation of the semantic term implies that only pixels $\mathbf{q}$ with the same semantic label as the examined pixel $\mathbf{p}$ contribute to the output at $\mathbf{p}$ through this term, which prevents blurring of semantic edges. At the same time, the color term helps to better preserve true depth edges that do not coincide with any semantic boundary but are present in $\mathbf{J}$.

The formulation of (2) enables an efficient implementation of our filter based on the bilateral grid [41]. More specifically, we construct two separate bilateral grids that correspond to the semantic and color domains and operate separately

on each grid to perform filtering, combining the results in the end. In this way, we handle a 3D bilateral grid for the semantic domain and a 5D grid for the color domain instead of a single joint 6D grid that would dramatically increase computation time [41].

In our experiments, we set $\mu = 5$, $\sigma_s = 20$, and $\sigma_c = 10$.

3.3 Remaining Steps

Here we outline the rest parts of our fog simulation pipeline of Fig. 2. For more details, we refer the reader to [48], with which most parts of the pipeline are common. The standard optical model for fog that forms the basis of our fog simulation was introduced in [29] and is expressed as

$$\mathbf{I}(\mathbf{x}) = \mathbf{R}(\mathbf{x})t(\mathbf{x}) + \mathbf{L}(1 - t(\mathbf{x})), \tag{3}$$

where $\mathbf{I}(\mathbf{x})$ is the observed foggy image at pixel $\mathbf{x}$, $\mathbf{R}(\mathbf{x})$ is the clear scene radiance and $\mathbf{L}$ is the atmospheric light, which is assumed to be globally constant. The transmittance $t(\mathbf{x})$ determines the amount of scene radiance that reaches the camera. For homogeneous fog, transmittance depends on the distance $\ell(\mathbf{x})$ of the scene from the camera through

$$t(\mathbf{x}) = \exp\left(-\beta \ell(\mathbf{x})\right). \tag{4}$$

The attenuation coefficient β controls the density of the fog: larger values of β mean denser fog. Fog decreases the meteorological optical range (MOR), also known as visibility, to less than 1 km by definition [1]. For homogeneous fog $\mathrm{MOR} = 2.996/\beta$, which implies

$$\beta \geq 2.996 \times 10^{-3}\ \mathrm{m}^{-1}, \tag{5}$$

where the lower bound corresponds to the lightest fog configuration. In our fog simulation, the value that is used for β always obeys (5).

The required inputs for fog simulation with (3) are the image $\mathbf{R}$ of the original clear scene, atmospheric light $\mathbf{L}$ and a complete transmittance map t. We use the same approach for atmospheric light estimation as that in [48]. Moreover, we adopt the stereoscopic inpainting method of [48] for depth denoising and completion to obtain an initial complete transmittance map $\hat{t}$ from a noisy and incomplete input disparity map D, using the recommended parameters. We filter $\hat{t}$ with our dual-reference cross-bilateral filter (2) to compute the final transmittance map t, which is used in (3) to synthesize the foggy image $\mathbf{I}$.

Results of the presented pipeline for fog simulation on example images from Cityscapes [11] are provided in Fig. 3 for $\beta = 0.02$, which corresponds to visibility of ca. 150m. We specifically leverage the instance-level semantic annotations that are provided in Cityscapes and set the labeling h of (1) to a different value for each distinct instance of the same semantic class in order to distinguish adjacent instances. We compare our synthetic foggy images against the respective images of Foggy Cityscapes that were generated with the approach of [48]. Our synthetic

(a) Cityscapes (b) Foggy Cityscapes (c) Our foggy image

Fig. 3. Comparison of our synthetic foggy images against Foggy Cityscapes [48]. This figure is better seen on a screen and zoomed in

foggy images generally preserve the edges between adjacent objects with large discrepancy in depth better than the images in Foggy Cityscapes, because our approach utilizes semantic boundaries, which usually encompass these edges. The incorrect structure transfer of color textures to the transmittance map, which deteriorates the quality of Foggy Cityscapes, is also reduced with our method.

4 Semantic Segmentation of Scenes with Dense Fog

In this section, we first present a standard supervised learning approach for semantic segmentation under dense fog using our synthetic foggy data with the novel fog simulation of Sec. 3, and then elaborate on our novel curriculum model adaptation approach using both synthetic and real foggy data.

4.1 Learning with Synthetic Fog

Generating synthetic fog from real clear-weather scenes grants the potential of inheriting the existing human annotations of these scenes, such as those from the Cityscapes dataset [11]. This is a significant asset that enables training of standard segmentation models. Therefore, an effective way of evaluating the merit of a fog simulator is to adapt a segmentation model originally trained on clear weather to the synthesized foggy images and then evaluate the adapted model against the original one on real foggy images. The goal is to verify that the standard learning methods for semantic segmentation can benefit from our simulated fog in the challenging scenario of real fog. This evaluation policy has been

proposed in [48]. We adopt this policy and fine-tune the RefineNet model [33] on synthetic foggy images generated with our simulation. The performance of our adapted models on dense real fog is compared to that of the original clear-weather model as well as the models that are adapted on Foggy Cityscapes [48], providing an objective comparison of our simulation method against [48].

4.2 Curriculum Model Adaptation with Synthetic and Real Fog

While adapting a standard segmentation model to our synthetic fog improves its performance as shown in Sec. 6.2, the paradigm still suffers from the domain discrepancy between synthetic and real foggy images. This discrepancy becomes more accentuated for denser fog. We present a method which can learn from our synthetic fog plus unlabeled real foggy data.

The method, which we term Curriculum Model Adaptation (CMAda), uses two versions of synthetic fog—one with light fog and another with dense fog—and a large dataset of unlabeled real foggy scenes with variable, unknown fog density, and works as follows:

1. generate a synthetic foggy dataset with multiple versions of varying fog density;
2. train a model for fog density estimation on the dataset of step 1;
3. rank the images in the real foggy dataset with the model of step 2 according to fog density;
4. generate a dataset with light synthetic fog, and train a segmentation model on it;
5. apply the segmentation model from step 4 to the light-fog images of the real dataset (ranked lower in step 3) to obtain "noisy" semantic labels;
6. generate a dataset with dense synthetic fog;
7. adapt the segmentation model from step 4 to the union of the dense synthetic foggy dataset from step 6 and the light real foggy one from step 5.

CMAda adapts segmentation models from light synthetic fog to dense real fog and is inspired by curriculum learning [4], in the sense that we first solve easier tasks with our synthetic data, *i.e.* fog density estimation and semantic scene understanding under light fog, and then acquire new knowledge from the already "solved" tasks in order to better tackle the harder task, *i.e.* scene understanding under dense real fog. CMAda also exploits the direct control of fog density for synthetic foggy images. Fig. 1 provides an overview of our method. Below we present details on our fog density estimation, *i.e.* step 2, and the training of the model, *i.e.* step 7.

Fog Density Estimation. Fog density is usually determined by the visibility of the foggy scene. An accurate estimate of fog density can benefit many applications, such as image defogging [10]. Since annotating images in a fine-grained manner regarding fog density is very challenging, previous methods are trained on a few hundreds of images divided into only two classes: foggy and

fog-free [10]. The performance of the system, however, is affected by the small amount of training data and the coarse class granularity.

In this paper, we leverage our fog simulation applied to Cityscapes [11] for fog density estimation. Since simulated fog density is directly controlled through β, we generate several versions of Foggy Cityscapes with varying $\beta \in \{0, 0.005, 0.01, 0.02\}$ and train AlexNet [30] to regress the value of β for each image, lifting the need to handcraft features relevant to fog as [10] did. The predicted fog density using our method correlates well with human judgments of fog density taken in a subjective study on a large foggy image database on Amazon Mechanical Turk (cf. Sec. 6.1 for results). The fog density estimator is used to rank our new *Foggy Zurich* dataset, to select light foggy images for usage in CMAda, and to select dense foggy images for manual annotation.

Curriculum Model Adaptation. We formulate CMAda for semantic segmentation as follows. Let us denote a clear-weather image by $\mathbf{x}$, the corresponding image under light synthetic fog by $\mathbf{x}'$, the corresponding image under dense synthetic fog by $\mathbf{x}''$, and the corresponding human annotation by $\mathbf{y}$. Then, the training data consist of labeled data with light synthetic fog $\mathcal{D}'_l = \{(\mathbf{x}'_i, \mathbf{y}_i)\}_{i=1}^{l}$, labeled data with dense synthetic fog $\mathcal{D}''_l = \{(\mathbf{x}''_i, \mathbf{y}_i)\}_{i=1}^{l}$ and unlabeled images with light real fog $\bar{\mathcal{D}}'_u = \{\bar{\mathbf{x}}'_j\}_{j=l+1}^{l+u}$, where $\mathbf{y}_i^{m,n} \in \{1, ..., C\}$ is the label of pixel (m, n), and C is the total number of classes. l is the number of labeled training images with synthetic fog, and u is the number of unlabeled images with light real fog. The aim is to learn a mapping function $\phi'' : \mathcal{X}'' \mapsto \mathcal{Y}$ from $\mathcal{D}'_l$, $\mathcal{D}''_l$ and $\bar{\mathcal{D}}'_u$, and evaluate it on images with dense real fog $\bar{\mathcal{D}}'' = \{\bar{\mathbf{x}}''_1, \ldots, \bar{\mathbf{x}}''_k\}$, where k is the number of images with dense real fog.

Since $\bar{\mathcal{D}}'_u$ does not have human annotations, we generate the supervisory labels as previously described in step 5. In particular, we first learn a mapping function $\phi' : \mathcal{X}' \mapsto \mathcal{Y}$ with $\mathcal{D}'_l$ and then obtain the labels $\bar{\mathbf{y}}'_j = \phi'(\bar{\mathbf{x}}'_j)$ for $\bar{\mathbf{x}}'_j$, $\forall j \in \{l+1, \ldots, l+u\}$. $\bar{\mathcal{D}}'_u$ is then upgraded to $\bar{\mathcal{D}}'_u = \{(\bar{\mathbf{x}}'_j, \bar{\mathbf{y}}'_j)\}_{j=l+1}^{l+u}$. The proposed scheme for training semantic segmentation models for dense foggy image $\bar{\mathbf{x}}''$ is to learn a mapping function ϕ'' so that human annotations for dense synthetic fog and the generated labels for light real fog are both taken into account:

$$\min_{\phi''} \frac{1}{l} \sum_{i=1}^{l} L(\phi''(\mathbf{x}''_i), \mathbf{y}_i) + \lambda \frac{1}{u} \sum_{j=l+1}^{l+u} L(\phi''(\bar{\mathbf{x}}'_j), \bar{\mathbf{y}}'_j), \tag{6}$$

where $L(.,.)$ is the cross entropy loss function and $\lambda = \frac{u}{l} \times w$ is a hyper-parameter balancing the weights of the two data sources, with w serving as the relative weight of each real weakly labeled image compared to each synthetic labeled one. We empirically set $w = 1/3$ in our experiment, but an optimal value can be obtained via cross-validation if needed. The optimization of (6) is implemented by mixing images from $\mathcal{D}''_l$ and $\bar{\mathcal{D}}'_u$ in a proportion of $1 : w$ and feeding the stream of hybrid data to a CNN for standard supervised training.

This learning approach bears resemblance to model distillation [21, 26] or imitation [8, 12]. The underpinnings of our proposed approach are the following:

1) in light fog objects are easier to recognize than in dense fog, hence models trained on synthetic data are more generalizable to real data in case both data sources contain light rather than dense fog; 2) dense synthetic fog and light real fog reflect different and complementary characteristics of the target domain of dense real fog. On the one hand, dense synthetic fog features a similar overall visibility obstruction to dense real fog, but includes artifacts. On the other hand, light real fog captures the true nonuniform and spatially varying structure of fog, but at a different density than dense fog.

5 The Foggy Zurich Dataset

5.1 Data Collection

Foggy Zurich was collected during multiple rides with a car inside the city of Zurich and its suburbs using a GoPro Hero 5 camera. We recorded four large video sequences, and extracted video frames corresponding to those parts of the sequences where fog is (almost) ubiquitous in the scene at a rate of one frame per second. The extracted images are manually cleaned by removing the duplicates (if any), resulting in 3808 foggy images in total. The resolution of the frames is 1920×1080 pixels. We mounted the camera inside the front windshield, since we found that mounting it outside the vehicle resulted in significant deterioration in image quality due to blurring artifacts caused by dew.

5.2 Annotation of Images with Dense Fog

We use our fog density estimator presented in Sec. 4.2 to rank all images in *Foggy Zurich* according to fog density. Based on the ordering, we manually select 16 images with *dense* fog and diverse visual scenes, and construct the test set of *Foggy Zurich* therefrom, which we term *Foggy Zurich-test*. We annotate these images with fine pixel-level semantic annotations using the 19 evaluation classes of the Cityscapes dataset [11]. In addition, we assign the *void* label to pixels which do not belong to any of the above 19 classes, or the class of which is uncertain due to the presence of fog. Every such pixel is ignored for semantic segmentation evaluation. Comprehensive statistics for the semantic annotations of *Foggy Zurich-test* are presented in Fig. 4. We also distinguish the semantic classes that occur frequently in *Foggy Zurich-test*. These "frequent" classes are: *road*, *sidewalk*, *building*, *wall*, *fence*, *pole*, *traffic light*, *traffic sign*, *vegetation*, *sky*, and *car*. When performing evaluation on *Foggy Zurich-test*, we occasionally report the average score over this set of frequent classes, which feature plenty of examples, as a second metric to support the corresponding results.

Despite the fact that there exists a number of prominent large-scale datasets for semantic road scene understanding, such as KITTI [19], Cityscapes [11] and Mapillary Vistas [39], most of these datasets contain few or even no foggy scenes, which can be attributed partly to the rarity of the condition of fog and the difficulty of annotating foggy images. To the best of our knowledge, the only

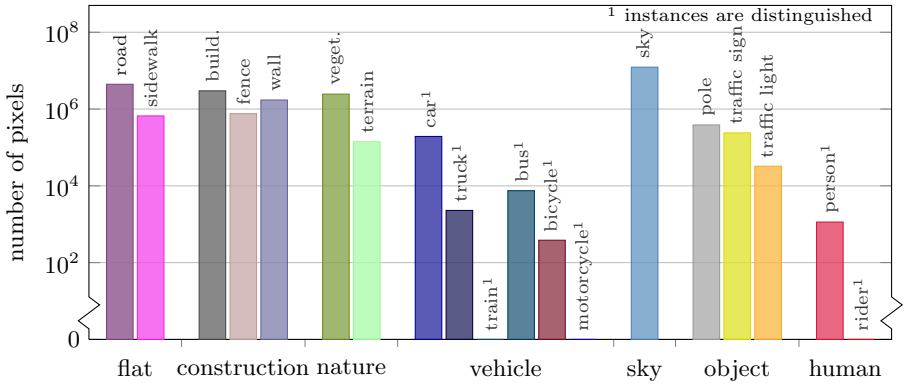


Fig. 4. Number of annotated pixels per class for *Foggy Zurich-test*

previous dataset for semantic foggy scene understanding whose scale exceeds that of *Foggy Zurich-test* is Foggy Driving [48], with 101 annotated images. However, we found that most images in Foggy Driving contain relatively light fog and most images with dense fog are annotated *coarsely*. Compared to Foggy Driving, *Foggy Zurich* comprises a much greater number of high-resolution foggy images. Its larger, unlabeled part is highly relevant for unsupervised or semi-supervised approaches such as the one we have presented in Sec. 4.2, while the smaller, labeled *Foggy Zurich-test* set features *fine* semantic annotations for the particularly challenging setting of dense fog, making a significant step towards evaluation of semantic segmentation models in this setting.

In order to ensure a sound training and evaluation, we manually filter the unlabeled part of *Foggy Zurich* and exclude from the resulting training sets those images which bear resemblance to any image in *Foggy Zurich-test* with respect to the depicted scene.

6 Experiments

6.1 Fog Density Estimation with Synthetic Data

We conduct a user study on Amazon Mechanical Turk (AMT) to evaluate the ranking results of our fog density estimator. In order to guarantee high quality, we only employ AMT Masters in our study and verify the answers via a Known Answer Review Policy. Each human intelligence task (HIT) comprises five image pairs to be compared: three pairs are the true query pairs; the rest two pairs contain synthetic fog of different densities and are used for validation. The participants are shown two images at a time, side by side, and are simply asked to choose the one which is more foggy. The query pairs are sampled based on the ranking results of our method. In order to avoid confusing cases, *i.e.* two images of similar fog densities, the two images of each pair need to be at least 20 percentiles apart based on the ranking results.

We have collected answers for 12000 pairs in 4000 HITs. The HITs are only considered for evaluation only when both the validation questions are correctly answered. 87% of all HITs are valid for evaluation. For these 10400 annotations, we find that the agreement between our ranking method and human judgment is 89.3%. The high accuracy confirms that fog density estimation is a relatively easier task, and the solution to it can be exploited for solving high-level tasks of foggy scenes.

6.2 Benefit of Adaptation with Our Synthetic Fog

Our model of choice for experiments on semantic segmentation is the state-of-the-art RefineNet [33]. We use the publicly available *RefineNet-res101-Cityscapes* model, which has been trained on the clear-weather training set of Cityscapes. In all experiments of this section, we use a constant learning rate of 5×10^{-5} and mini-batches of size 1. Moreover, we compile all versions of our synthetic foggy dataset by applying our fog simulation (which is denoted by "Stereo-DBF" in the following for short) on the same *refined* set of Cityscapes images that was used in [48] to compile Foggy Cityscapes-refined. This set comprises 498 training and 52 validation images; we use the former for training. We considered dehazing as a preprocessing step as in [48] but did not observe a gain against *no* dehazing and thus omit such comparisons from the following presentation.

Our first segmentation experiment shows that our semantic-aware fog simulation performs competitively compared to the fog simulation of [48] (denoted by "Stereo-GF") for generating synthetic data to adapt RefineNet to dense real fog. *RefineNet-res101-Cityscapes* is fine-tuned on the version of Foggy Cityscapes-refined that corresponds to each simulation method for 8 epochs. We experiment with two synthetic fog densities. For evaluation, we use *Foggy Zurich-test* as well as a subset of Foggy Driving [48] containing 21 images with dense fog, which we term Foggy Driving-dense, and report results in Tables 1 and 2 respectively. Training on lighter synthetic fog helps to beat the baseline clear-weather model in all cases and yields consistently better results than denser synthetic fog, which verifies the first motivating assumption of CMAda at the end of Sec. 4.2. In addition, Stereo-DBF beats Stereo-GF in most cases by a small margin and is consistently better at generating denser synthetic foggy data. On the other hand, Stereo-GF with light fog is slightly better for *Foggy Zurich-test*. This motivates us to consistently use the model that has been trained with Stereo-GF in steps 4 and 5 of CMAda for the experiments of Sec. 6.3, assuming that its merit for dense real fog extends to lighter fog. However, Stereo-DBF is still fully relevant for step 6 of CMAda based on its favorable comparison for denser synthetic fog.

6.3 Benefit of Curriculum Adaptation with Synthetic and Real Fog

Our second segmentation experiment showcases the effectiveness of our CMAda pipeline, using Stereo-DBF and Stereo-GF as alternatives for generating synthetic Foggy Cityscapes-refined in steps 4 and 6 of the pipeline. *Foggy Zurich* serves as the real foggy dataset in the pipeline. We use the results of our fog

Table 1. Performance comparison on *Foggy Zurich-test* of RefineNet and fine-tuned versions of it using Foggy Cityscapes-refined, rendered with different fog simulations and attenuation coefficients β

Mean IoU over *all* classes (%)			Mean IoU over *frequent* classes (%)		
RefineNet [33]	32.0		RefineNet [33]	48.8	
Fog simulation	$\beta = 0.005$	$\beta = 0.01$	Fog simulation	$\beta = 0.005$	$\beta = 0.01$
Stereo-GF [48]	**33.9**	30.2	Stereo-GF [48]	**49.3**	45.8
Stereo-DBF	33.4	**31.2**	Stereo-DBF	49.0	**46.6**

Table 2. Performance comparison on Foggy Driving-dense of RefineNet and fine-tuned versions of it using Foggy Cityscapes-refined, rendered with different fog simulations and attenuation coefficients β

Mean IoU over *all* classes (%)			Mean IoU over *frequent* classes (%)		
RefineNet [33]	30.4		RefineNet [33]	57.6	
Fog simulation	$\beta = 0.005$	$\beta = 0.01$	Fog simulation	$\beta = 0.005$	$\beta = 0.01$
Stereo-GF [48]	32.5	32.4	Stereo-GF [48]	60.4	58.7
Stereo-DBF	**32.8**	**32.8**	Stereo-DBF	**60.8**	**59.2**

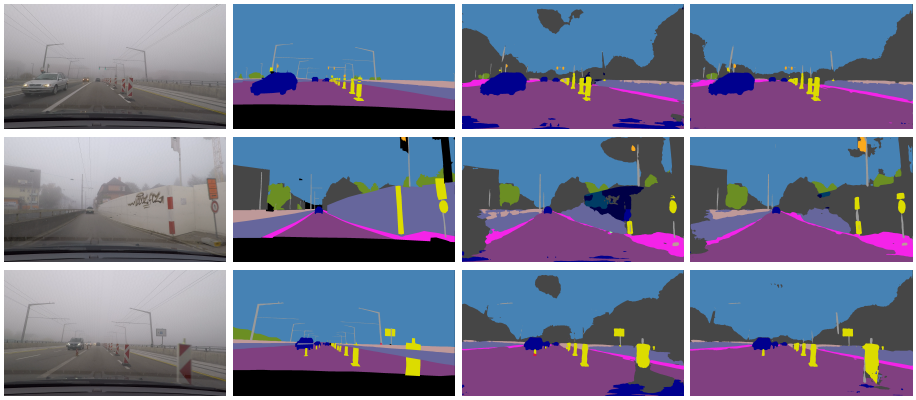
Table 3. Performance comparison on *Foggy Zurich-test* of the two adaptation steps of CMAda using Foggy Cityscapes-refined and *Foggy Zurich-light* for training

Mean IoU over *all* classes (%)			Mean IoU over *frequent* classes (%)		
Fog simulation	CMAda-4	CMAda-7	Fog simulation	CMAda-4	CMAda-7
Stereo-GF [48]	33.9	34.7	Stereo-GF [48]	49.3	53.3
Stereo-DBF	33.4	**37.9**	Stereo-DBF	49.0	**56.7**

density estimation to select 1556 images with light fog and name this set *Foggy Zurich-light*. The models which are obtained after the initial adaptation step that uses Foggy Cityscapes-refined with $\beta = 0.005$ are further fine-tuned for 6k iterations on the union of Foggy Cityscapes-refined with $\beta = 0.01$ and *Foggy Zurich-light* setting $w = 1/3$, where the latter set is noisily labeled by the aforementioned initially adapted models. Results for the two adaptation steps (denoted by "CMAda-4" and "CMAda-7") on *Foggy Zurich-test* and *Foggy Driving-dense* are reported in Tables 3 and 4 respectively. The second adaptation step CMAda-7, which involves dense synthetic fog and light real fog, consistently improves upon the first step CMAda-4. Moreover, using our fog simulation to simulate dense synthetic fog for CMAda-7 gives the best result on *Foggy Zurich-test*, improving the clear-weather baseline by 5.9% and 7.9% in terms of mean IoU over all classes and frequent classes respectively. Fig. 5 supports this result with visual comparisons. The real foggy images of *Foggy Zurich-light* used in CMAda-7 additionally provide a clear generalization benefit on Foggy Driving-dense, which involves different camera sensors than *Foggy Zurich*.

Table 4. Performance comparison on Foggy Driving-dense of the two adaptation steps of CMAda using Foggy Cityscapes-refined and *Foggy Zurich-light* for training

Mean IoU over *all* classes (%)			Mean IoU over *frequent* classes (%)		
Fog simulation	CMAda-4	CMAda-7	Fog simulation	CMAda-4	CMAda-7
Stereo-GF [48]	32.5	34.1	Stereo-GF [48]	60.4	**61.6**
Stereo-DBF	32.8	**34.3**	Stereo-DBF	60.8	61.5



(a) foggy image (b) ground truth (c) RefineNet [33] (d) CMAda

Fig. 5. Qualitative results for semantic segmentation on *Foggy Zurich-test*. "CMAda" stands for RefineNet [33] fine-tuned with our full CMAda pipeline on the union of Foggy Cityscapes-refined using our simulation and *Foggy Zurich-light*

7 Conclusion

In this paper, we have shown the benefit of using partially synthetic as well as unlabeled real foggy data in a curriculum adaptation framework to progressively improve performance of state-of-the-art semantic segmentation models in dense real fog. To this end, we have proposed a novel fog simulation approach on real scenes, which leverages the semantic annotation of the scene as input to a novel dual-reference cross-bilateral filter, and applied it to the Cityscapes dataset. We have presented *Foggy Zurich*, a large-scale dataset of real foggy scenes, including pixel-level semantic annotations for 16 scenes with dense fog. Through detailed evaluation, we have evidenced clearly that our curriculum adaptation method exploits both our synthetic and real data and significantly boosts performance on dense real fog without using any labeled real foggy image and that our fog simulation performs competitively to state-of-the-art counterparts.

Acknowledgements. This work is funded by Toyota Motor Europe via the research project TRACE-Zürich.

References

1. Federal Meteorological Handbook No. 1: Surface Weather Observations and Reports. U.S. Department of Commerce / National Oceanic and Atmospheric Administration (2005)
2. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2274–2282 (2012)
3. Bar Hillel, A., Lerner, R., Levi, D., Raz, G.: Recent progress in road and lane detection: A survey. Mach. Vision Appl. **25**(3), 727–745 (Apr 2014)
4. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: International Conference on Machine Learning. pp. 41–48 (2009)
5. Berman, D., Treibitz, T., Avidan, S.: Non-local image dehazing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
6. Bronte, S., Bergasa, L.M., Alcantarilla, P.F.: Fog detection system based on computer vision techniques. In: IEEE International Conference on Intelligent Transportation Systems (ITSC) (2009)
7. Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: European Conference on Computer Vision (2008)
8. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: International Conference on Knowledge Discovery and Data Mining (SIGKDD) (2006)
9. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
10. Choi, L.K., You, J., Bovik, A.C.: Referenceless prediction of perceptual fog density and perceptual image defogging. IEEE Transactions on Image Processing **24**(11), 3888–3901 (2015)
11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
12. Dai, D., Kroeger, T., Timofte, R., Van Gool, L.: Metric imitation by manifold transfer for efficient vision applications. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
13. Dai, D., Van Gool, L.: Progressive model adaptation and knowledge transfer from daytime to nighttime for semantic road scene understanding. In: IEEE International Conference on Intelligent Transportation Systems (2018)
14. Eisemann, E., Durand, F.: Flash photography enhancement via intrinsic relighting. In: ACM SIGGRAPH (2004)
15. Fattal, R.: Single image dehazing. ACM transactions on graphics (TOG) **27**(3) (2008)
16. Fattal, R.: Dehazing using color-lines. ACM Transactions on Graphics (TOG) **34**(1) (2014)
17. Gallen, R., Cord, A., Hautière, N., Aubert, D.: Towards night fog detection through use of in-vehicle multipurpose cameras. In: IEEE Intelligent Vehicles Symposium (IV) (2011)
18. Gallen, R., Cord, A., Hautière, N., Dumont, É., Aubert, D.: Nighttime visibility analysis and estimation method in the presence of dense fog. IEEE Transactions on Intelligent Transportation Systems **16**(1), 310–320 (2015)

19. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
20. Girshick, R.: Fast R-CNN. In: International Conference on Computer Vision (ICCV) (2015)
21. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
22. Hautière, N., Tarel, J.P., Lavenant, J., Aubert, D.: Automatic fog detection and estimation of visibility distance through use of an onboard camera. Machine Vision and Applications **17**(1), 8–20 (2006)
23. He, K., Sun, J., Tang, X.: Guided image filtering. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(6), 1397–1409 (2013)
24. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(12), 2341–2353 (2011)
25. Hecker, S., Dai, D., Van Gool, L.: Learning driving models with a surround-view camera system and a route planner. In: European Conference on Computer Vision (ECCV) (2018)
26. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
27. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: CyCADA: Cycle-consistent adversarial domain adaptation. In: International Conference on Machine Learning (2018)
28. Jensen, M.B., Philipsen, M.P., Møgelmose, A., Moeslund, T.B., Trivedi, M.M.: Vision for looking at traffic lights: Issues, survey, and perspectives. IEEE Transactions on Intelligent Transportation Systems **17**(7), 1800–1815 (July 2016)
29. Koschmieder, H.: Theorie der horizontalen Sichtweite. Beitrage zur Physik der freien Atmosphäre (1924)
30. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS (2012)
31. Levinkov, E., Fritz, M.: Sequential bayesian model update under structured scene prior for semantic road scenes labeling. In: IEEE International Conference on Computer Vision (2013)
32. Li, Y., You, S., Brown, M.S., Tan, R.T.: Haze visibility enhancement: A survey and quantitative benchmarking (2016), coRR abs/1607.06235
33. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
34. Ling, Z., Fan, G., Wang, Y., Lu, X.: Learning deep transmission network for single image dehazing. In: IEEE International Conference on Image Processing (ICIP) (2016)
35. Miclea, R.C., Silea, I.: Visibility detection in foggy environment. In: International Conference on Control Systems and Computer Science (2015)
36. Narasimhan, S.G., Nayar, S.K.: Contrast restoration of weather degraded images. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(6), 713–724 (Jun 2003)
37. Narasimhan, S.G., Nayar, S.K.: Vision and the atmosphere. Int. J. Comput. Vision **48**(3), 233–254 (Jul 2002)

38. Negru, M., Nedevschi, S., Peter, R.I.: Exponential contrast restoration in fog conditions for driving assistance. IEEE Transactions on Intelligent Transportation Systems **16**(4), 2257–2268 (2015)
39. Neuhold, G., Ollmann, T., Rota Bulò, S., Kontschieder, P.: The Mapillary Vistas dataset for semantic understanding of street scenes. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
40. Nishino, K., Kratz, L., Lombardi, S.: Bayesian defogging. International Journal of Computer Vision **98**(3), 263–278 (2012)
41. Paris, S., Durand, F.: A fast approximation of the bilateral filter using a signal processing approach. International Journal of Computer Vision (2009)
42. Pavlić, M., Belzner, H., Rigoll, G., Ilić, S.: Image based fog detection in vehicles. In: IEEE Intelligent Vehicles Symposium (2012)
43. Pavlić, M., Rigoll, G., Ilić, S.: Classification of images in fog and fog-free scenes for use in vehicles. In: IEEE Intelligent Vehicles Symposium (IV) (2013)
44. Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M., Hoppe, H., Toyama, K.: Digital photography with flash and no-flash image pairs. In: ACM SIGGRAPH (2004)
45. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. pp. 91–99 (2015)
46. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
47. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
48. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision (2018)
49. Sankaranarayanan, S., Balaji, Y., Jain, A., Nam Lim, S., Chellappa, R.: Learning from synthetic data: Addressing domain shift for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
50. Shen, X., Zhou, C., Xu, L., Jia, J.: Mutual-structure for joint filtering. In: The IEEE International Conference on Computer Vision (ICCV) (2015)
51. Spinneker, R., Koch, C., Park, S.B., Yoon, J.J.: Fast fog detection for camera based advanced driver assistance systems. In: IEEE International Conference on Intelligent Transportation Systems (ITSC) (2014)
52. Tan, R.T.: Visibility in bad weather from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
53. Tang, K., Yang, J., Wang, J.: Investigating haze-relevant features in a learning framework for image dehazing. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
54. Tarel, J.P., Hautière, N., Caraffa, L., Cord, A., Halmaoui, H., Gruyer, D.: Vision enhancement in homogeneous and heterogeneous fog. IEEE Intelligent Transportation Systems Magazine **4**(2), 6–20 (2012)
55. Tarel, J.P., Hautière, N., Cord, A., Gruyer, D., Halmaoui, H.: Improved visibility of road scene images under heterogeneous fog. In: IEEE Intelligent Vehicles Symposium. pp. 478–485 (2010)
56. Wang, Y.K., Fan, C.T.: Single image defogging by multiscale depth fusion. IEEE Transactions on Image Processing **23**(11), 4826–4837 (2014)

57. Xu, Y., Wen, J., Fei, L., Zhang, Z.: Review of video and image defogging algorithms and related studies on image restoration and enhancement. IEEE Access **4**, 165–188 (2016)
58. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: International Conference on Learning Representations (2016)
59. Zhang, H., Sindagi, V.A., Patel, V.M.: Joint transmission map estimation and dehazing using deep networks (2017), coRR abs/1708.00581
60. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)