International Journal of Computer Vision (2020) 128:1182–1204
https://doi.org/10.1007/s11263-019-01182-4

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

Dengxin Dai[1] · Christos Sakaridis[1] · Simon Hecker[1] · Luc Van Gool[1,2]

Received: 19 July 2018 / Accepted: 27 April 2019 / Published online: 6 May 2019

Abstract

This work addresses the problem of semantic scene understanding under fog. Although marked progress has been made in semantic scene understanding, it is mainly concentrated on clear-weather scenes. Extending semantic segmentation 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 labeled synthetic foggy data and unlabeled real foggy data. The method is based on the fact that the results of semantic segmentation in moderately adverse conditions (light fog) can be bootstrapped to solve the same problem in highly adverse conditions (dense fog). CMAda is extensible to other adverse conditions and provides a new paradigm for learning with synthetic data and unlabeled real data. In addition, we present four 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) a novel fog densification method for real foggy scenes without known depth; and (4) the *Foggy Zurich* dataset comprising 3808 real foggy images, with pixel-level semantic annotations for 40 images with dense fog. Our experiments show that (1) our fog simulation and fog density estimator outperform their state-of-the-art counterparts 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, benefiting both from our synthetic and real foggy data. The foggy datasets and code are publicly available.

Keywords Semantic foggy scene understanding · Fog simulation · Learning with synthetic and real data · Curriculum model adaptation · Network distillation · Adverse weather conditions

1 Introduction

Adverse weather or illumination conditions create visibility problems for both people and the sensors that power automated systems (Narasimhan and Nayar 2002; Garg and Nayar 2007; Sakaridis et al. 2018; Dai and Van Gool 2018). While sensors and the downstream vision algorithms are constantly getting better, their performance is mainly benchmarked on clear-weather images (Cordts 2016; Hecker et al. 2018). Many outdoor applications, however, cannot escape from "bad" weather (Narasimhan and Nayar 2002). One typical example of adverse weather conditions is fog, which degrades the visibility of a scene significantly (Narasimhan and Nayar 2003; Tan 2008). The denser the fog is, the more severe this problem becomes.

During the past years, the community has made a tremendous progress in image dehazing (defogging) to increase the visibility in foggy images (Nishino et al. 2012; He et al. 2011; Wang and Fan 2014). The last few years have also witnessed a leap in object recognition. A great deal of effort is made specifically in semantic road scene understanding (Alvarez et al. 2012; Cordts 2016; Dhall et al. 2019). However, the extension of these techniques to other weather/illumination conditions has not received due attention, despite its 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 or rain. 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

Communicated by Anelia Angelova, Gustavo Carneiro, Niko Sünderhauf, Jürgen Leitner.

✉ Dengxin Dai
dai@vision.ee.ethz.ch

[1] ETH Zürich, Zurich, Switzerland

[2] KU Leuven, Leuven, Belgium

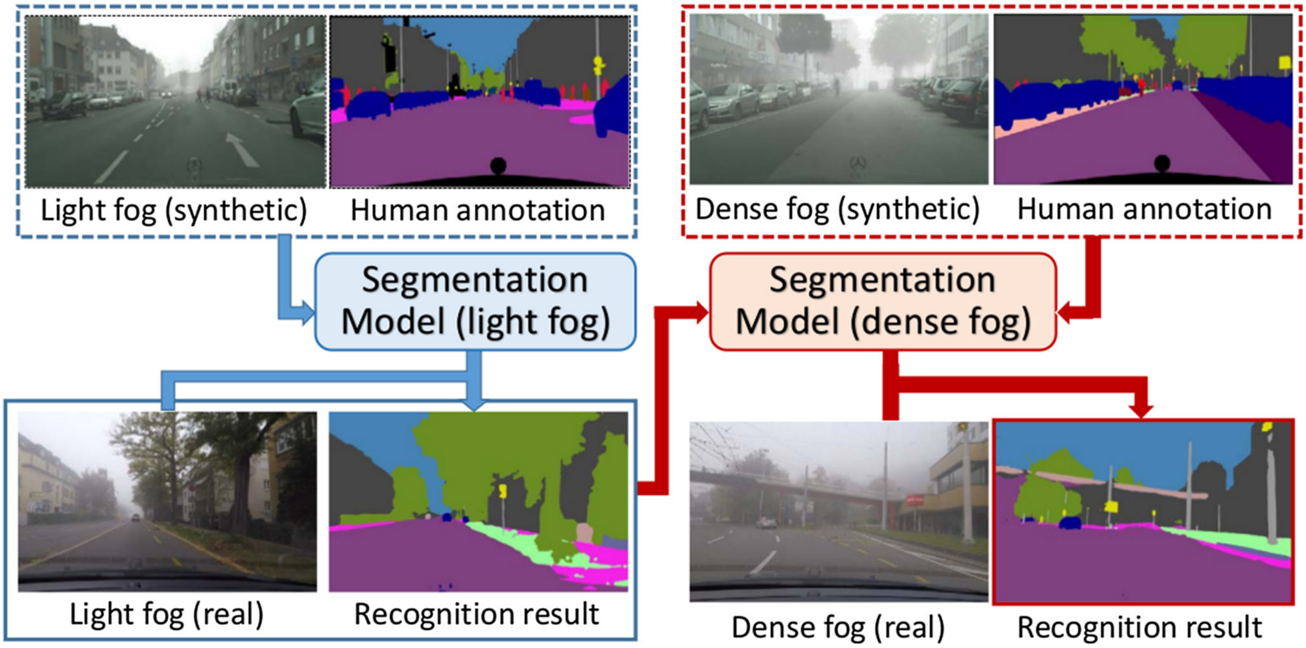


Fig. 1 The illustrative pipeline of a two-stage instantiation of CMAda for semantic scene understanding under dense fog

numerous annotated real images (Everingham et al. 2010; Russakovsky et al. 2015; Cordts 2016). While this trend of creating and using more human annotations may still continue, extending the same protocol to all conditions seems to be problematic, as the manual annotation part is hard to scale. The problem is more pronounced for adverse weather conditions, as the difficulty of data collection and annotation increases significantly. To overcome this problem, a few streams of research have gained extensive attention: learning with limited, weak supervision (Dai and Van Gool 2013; Misra et al. 2015), transfer learning (Hoffman 2014; Chen et al. 2018), and learning with synthetic data (Ros et al. 2016; Sakaridis et al. 2018).

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 the corresponding human annotations, and (2) a dataset with a large number of unlabeled images with real fog. The goal of our method is to improve the performance of SFSU without requiring extra human annotations for foggy images.

To this end, this work proposes a novel fog simulator to add high-quality synthetic fog to real images of clear-weather outdoor scenes, and then leverage these partially synthetic foggy images for SFSU. Our fog simulator builds on the recent work of Sakaridis et al. (2018), by introducing a semantic-aware filter to exploit the structures of object instances. We show that learning with our synthetic foggy data improves the performance for SFSU. Furthermore, we learn a fog density estimator from synthetic images of varying fog density, and order unlabeled real images by increasing fog density. This ordering forms the foundation of our novel learning method Curriculum Model Adaptation (CMAda) to *gradually* adapt a semantic segmentation model from *clear weather* to *dense fog*, through *light fog*. CMAda is based on the fact that recognition in moderately adverse conditions (light fog) is easier and its results can be re-used via *knowledge distillation* to solve a harder problem, i.e. recognition in highly adverse conditions (dense fog).

CMAda is iterative by nature and can be implemented for different numbers of steps. The pipeline of a two-step implementation of CMAda is shown in Fig. 1. CMAda has the potential to be used for other adverse weather conditions, and opens a new avenue for learning with synthetic data and unlabeled real data in general. Experiments show that CMAda yields the best results on two datasets with dense real fog as well as a dataset with real fog of varying density.

A shorter version of this work has been published to European Conference on Computer Vision (Sakaridis et al. 2018). Compared to the conference version, this paper makes the following six additional contributions:

1. An extension of the formulation of CMAda to accommodate multiple adaptation steps instead of only two steps, leading to improved performance over the conference paper as well.
2. A novel fog densification method for real foggy scenes. The fog densification method can close the domain gap between light real fog and dense real fog; using it

in CMAda significantly increases the performance for SFSU.
3. A method named Model Selection for the task of semantic scene understanding *in multiple weather conditions* where test images are a mixture of clear-weather images and foggy images. This extension is important for real world applications, as weather conditions change constantly. Semantic scene understanding methods need to be robust to such changes.
4. An enlarged annotated dense foggy set for our *Foggy Zurich* dataset, increasing its size from 16 to 40 images.[1]
5. More extensive experiments to diagnose the contribution of each component of the CMAda pipeline, to compare with more competing methods, and to comprehensively study the usefulness of image dehazing for SFSU.
6. Other sections are also enhanced, including related work as well as dataset collection and annotation.

The paper is structured as follows. Section 2 presents the related work. Section 3 is devoted to our method for simulating synthetic fog, which is followed by Sect. 4 for our learning approach. Section 5 summarizes our data collection and annotation. Finally, Sect. 6 presents our experimental results and Sect. 7 concludes this paper. Our foggy datasets and fog simulation code are publicly available at https://www.vision.ee.ethz.ch/~csakarid/Model_adaptation_SFSU_dense/.

2 Related Work

Our work is relevant to image defogging, joint image filtering, 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 (Narasimhan and Nayar 2003; Tan 2008; Nishino et al. 2012; Fattal 2008; Berman et al. 2016; Fattal 2014; He et al. 2011). Certain works focus particularly on enhancing foggy road scenes (Tarel et al. 2012; Negru et al. 2015). Recent approaches also rely on trainable architectures (Tang et al. 2014), which have evolved to end-to-end models (Zhang et al. 2017; Ling et al. 2016). For a comprehensive overview of defogging/dehazing algorithms, we point the reader to Xu et al. (2016), Li et al. (2016). Our work is complementary and mainly focuses on SFSU, while it also investigates the usefulness of image dehazing in the context of SFSU.

2.2 Joint Image Filtering

Using additional images as input for filtering a target image has been originally studied in settings where the target image has low photometric quality (Eisemann and Durand 2004; Petschnigg et al. 2004) or low resolution (Kopf et al. 2007). Compared to the bilateral filtering formulation of these approaches, subsequent works propose alternative formulations, such as the guided filter (He et al. 2013) and mutual structure filtering (Shen et al. 2015), for better incorporating the reference image into the filtering process. In comparison, we extend the classical cross-bilateral filter to a dual-reference cross-bilateral filter by accepting *two* reference images, one of which is a discrete label image that helps our filter adhere to the semantics of the scene.

2.3 Foggy Scene Understanding

Typical examples in this line include road and lane detection (Bar Hillel et al. 2014), traffic light detection (Jensen et al. 2016), car and pedestrian detection (Geiger et al. 2012), and a dense, pixel-level segmentation of road scenes into most of the relevant semantic classes (Brostow et al. 2008; Cordts 2016). While deep recognition networks have been developed (Yu and Koltun 2016; Lin et al. 2017; Zhao et al. 2017; Girshick 2015; Ren et al. 2015) and large-scale datasets have been presented (Geiger et al. 2012; Cordts 2016), that research mainly focused on clear weather. There is also a large body of work on fog detection (Bronte et al. 2009; Pavlić et al. 2012; Gallen et al. 2011; Spinneker et al. 2014). Classification of scenes into foggy and fog-free has been tackled as well (Pavlić et al. 2013). In addition, visibility estimation has been extensively studied for both daytime (Tarel et al. 2010; Miclea and Silea 2015; Hautière et al. 2006) and nighttime (Gallen et al. 2015), in the context of assisted and autonomous driving. The closest of these works to ours is Tarel et al. (2010), 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 scene understanding task is more complex and we tackle the problem from a different route by learning jointly from synthetic fog and real fog.

2.4 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 Levinkov and Fritz (2013). 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 (Shrivastava et al. 2017; Sankaranarayanan et al. 2018; Hoffman et al. 2018;

[1] Creating fine pixel-level annotations for dense foggy scenes is very difficult.

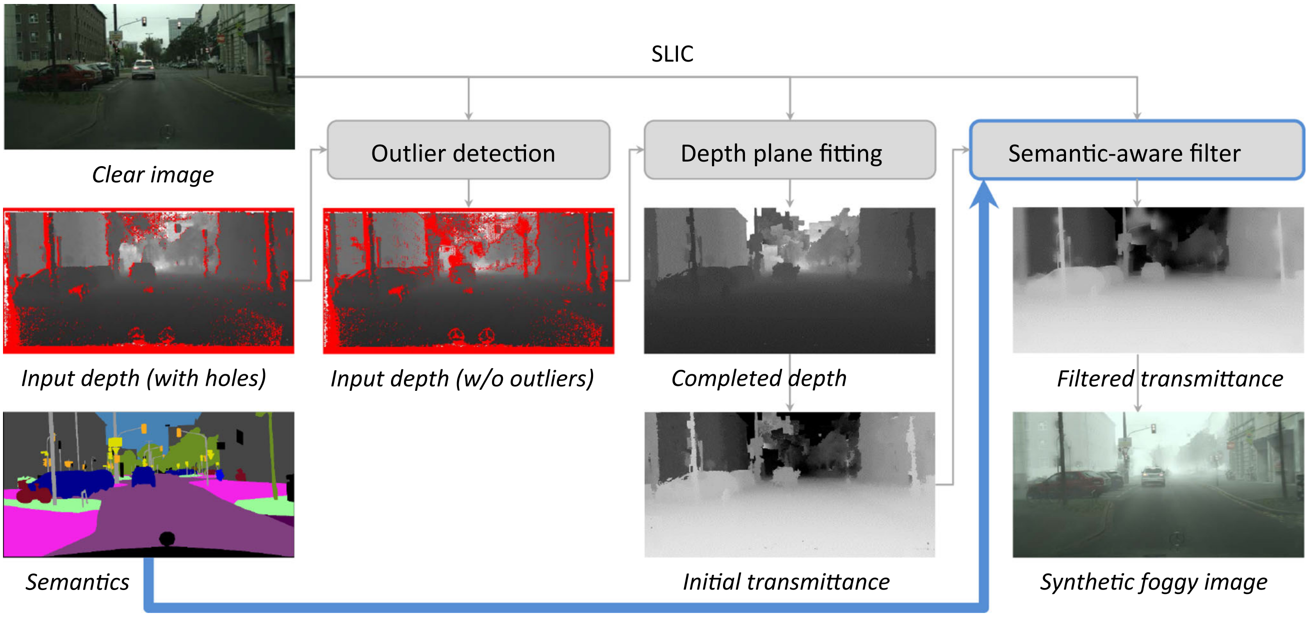


Fig. 2 The pipeline of our fog simulation using semantics

Wulfmeier et al. 2018). Most of these works are based on adversarial domain adaptation. Our work is complementary to methods in this vein; we adapt the model parameters with carefully generated data, leading to an algorithm whose behavior is easy to understand and whose performance is more predictable. Combining our method and adversarial domain adaptation is a promising direction. Our work also shares similarity to Zhang et al. (2017) in applying the general idea of curriculum learning to domain adaptation.

The concurrent work in Dai and Van Gool (2018) on adaptation of semantic segmentation models from daytime to nighttime using solely real data, which was preceded by the conference version of this paper, shows that real images captured at twilight are helpful for supervision transfer from daytime to nighttime. CMAda constitutes a more complex framework, since it leverages both synthetic foggy data and real foggy data *jointly* for adapting semantic segmentation models to fog, whereas the method in Dai and Van Gool (2018) uses solely real data for the adaptation. Moreover, the assignment of real foggy images to the correct target foggy domain through fog density estimation is another crucial and nontrivial component of CMAda and it is a prerequisite for using these real images as training data in the method. By contrast, the partition of the real dataset in Dai and Van Gool (2018) into subsets that correspond to different times of day from daytime to nighttime is trivially performed by using the time of capture of the images.

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 Sakaridis et al. (2018) 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 (Achanta et al. 2012) using RANSAC, and postprocessing of the completed depth map with guided image filtering (He et al. 2013). 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 Sakaridis et al. (2018) uses the clear-weather color image as guidance to filter the depth map. However, as previous works on image filtering (Shen et al. 2015) have shown, guided filtering and similar joint filtering methods such as cross-bilateral filtering (Eisemann and Durand 2004; Petschnigg et al. 2004) transfer every 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 transferred to the latter erroneously.

Whereas previous approaches such as mutual-structure filtering (Shen et al. 2015) 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 when creating semantic annotations are not pixel-accurate, and using the color image as additional reference helps to capture the precise location and orientation 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.

Before delving into the formulation of our filter, we briefly argue against alternative usage cases of semantic annotations in our fog simulation pipeline which might seem attractive at first sight. First, replacing SLIC superpixels with superpixels induced by the semantic labeling for the depth plane fitting step is not viable, because it induces very large superpixels, for which the planarity assumption breaks completely. Second, we have experimented with omitting the robust depth plane fitting step altogether and applying our dual-reference cross-bilateral filter directly on the incomplete depth map which is output from the outlier detection step. This approach, however, is highly sensitive to outliers that have not been detected and invalidated in the preceding step. By contrast, these remaining outliers are handled successfully by robust RANSAC-based depth plane fitting.

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 (Paris and Durand 2009). 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 Sect. 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

$$\begin{aligned} t(\mathbf{p}) = & \left\{ \sum_{\mathbf{q}\in\mathcal{N}(\mathbf{p})} G_{\sigma_s}(\|\mathbf{q}-\mathbf{p}\|) \left[\delta(h(\mathbf{q})-h(\mathbf{p})) \right.\right. \\ & \left.\left. + \; \mu G_{\sigma_c}(\|\mathbf{J}(\mathbf{q})-\mathbf{J}(\mathbf{p})\|)\right] \hat{t}(\mathbf{q}) \right\} \\ & \Bigg/ \left\{ \sum_{\mathbf{q}\in\mathcal{N}(\mathbf{p})} G_{\sigma_s}(\|\mathbf{q}-\mathbf{p}\|) \left[\delta(h(\mathbf{q})-h(\mathbf{p})) \right.\right. \\ & \left.\left. + \; \mu G_{\sigma_c}(\|\mathbf{J}(\mathbf{q})-\mathbf{J}(\mathbf{p})\|)\right] \right\}, \end{aligned} \tag{2}$$

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 $\mathbf{J}$, e.g. due to self-occlusion of an object.

The formulation of (2) enables an efficient implementation of our filter based on the bilateral grid (Paris and Durand 2009). More specifically, we construct two separate bilateral grids that correspond to the semantic and color domains respectively 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 (Paris and Durand 2009).

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 Sakaridis et al. (2018), 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 Koschmieder (1924) and is expressed as

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

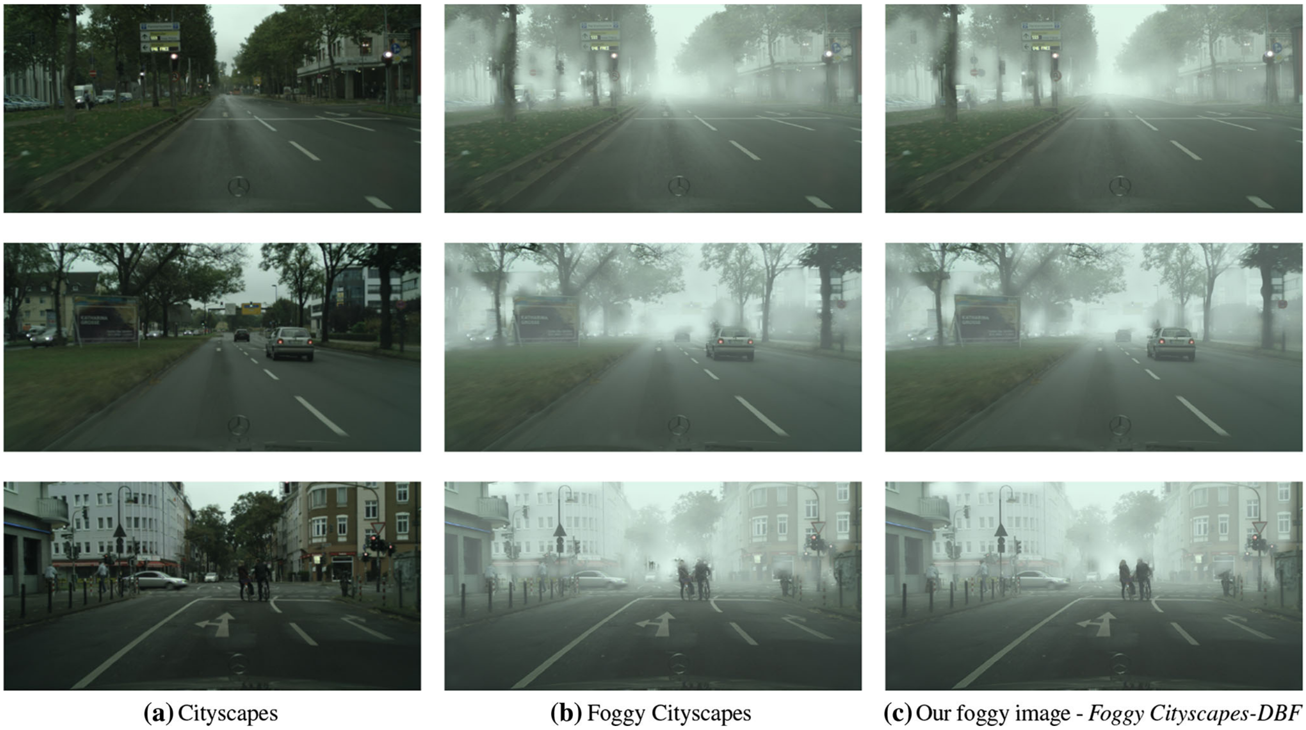


Fig. 3 Comparison of our synthetic foggy images against Foggy Cityscapes (Sakaridis et al. 2018). This figure is better seen on a screen and zoomed in

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 (Federal Meteorological Handbook 2005). For homogeneous fog 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 Sakaridis et al. (2018). Moreover, we adopt the stereoscopic inpainting method of Sakaridis et al. (2018) 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 (Cordts 2016) 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 Sakaridis et al. (2018). Our synthetic 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.

We have applied our fog simulation using semantics to the entire Cityscapes dataset. The resulting foggy dataset is named *Foggy Cityscapes-DBF* (**D**ual-reference cross-**B**ilateral **F**ilter). *Foggy Cityscapes-DBF* is publicly available at the Cityscapes website https://www.cityscapes-dataset.com.

4 Semantic Foggy Scene Understanding

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 Sect. 3, and then elaborate on our novel CMAda approach which uses 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 (Cordts 2016). 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 primary 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 Sakaridis et al. (2018). We adopt this policy and fine-tune the RefineNet model (Lin et al. 2017) on synthetic foggy images from our *Foggy Cityscapes-DBF* dataset. The performance of our adapted models on real fog is compared to that of the original clear-weather model as well as the models that are adapted on Foggy Cityscapes (Sakaridis et al. 2018), providing an objective comparison of our simulation method against (Sakaridis et al. 2018).

The learned model can be used as a standalone approach for semantic foggy scene understanding as shown in Sakaridis et al. (2018), or it can be used as an initialization step for our CMAda method, which is described next and learns both from synthetic and real data.

4.2 Curriculum Model Adaptation (CMAda)

In the previous section, the proposed method learns to adapt semantic segmentation models from the domain of clear weather to the domain of foggy weather in a single step. While considerable improvement can be achieved (as shown in Sect. 6.1.1), the method falls short when it is presented with dense fog. This is because domain discrepancies become more accentuated for denser fog: (1) the domain discrepancy between synthetic foggy images and real foggy images increases with fog density; and (2) the domain discrepancy between real clear-weather images and real foggy images increases with fog density. This section presents a method to gradually adapt the semantic segmentation model which was originally trained with clear-weather images to images with dense fog by using both labeled synthetic foggy images and unlabeled real foggy images. The method, which we term Curriculum Model Adaptation (CMAda), uses synthetic fog with a range of varying fog density—from light fog to dense fog—and a large dataset of unlabeled real foggy scenes with variable, unknown fog density. The goal is to improve the performance of state-of-the-art semantic segmentation models on dense foggy scenes without using any human annotations of foggy scenes. Below, we first present our fog density estimator and our method for densification of fog in real foggy images without depth information, and then proceed to the complete learning approach.

4.2.1 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 (Choi et al. 2015). 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 (Choi et al. 2015). 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 (Cordts 2016) for fog density estimation. Since simulated fog density is directly controlled through β, we generate several versions of *Foggy Cityscapes-DBF* with varying $\beta \in \{0, 0.005, 0.01, 0.02\}$ and train AlexNet (Krizhevsky et al. 2012) to regress the value of β for each image, lifting the need to handcraft features relevant to fog and to collect human annotations as Choi et al. (2015) did. The predicted fog density with our method on real images correlates well with human judgments of fog density, based on a user study conducted on our large real *Foggy Zurich* dataset via Amazon Mechanical Turk (cf. Sect. 6.1.2 for results). The fog density estimator is used to order images in *Foggy Zurich* according to fog density, paving the way for our curriculum adaptation which learns from images with progressively denser fog. We denote the estimator by $f : \mathbf{x} \rightarrow \mathbb{R}^+$, where $\mathbf{x}$ is an image.

4.2.2 CMAda with Synthetic and Real Fog

The CMAda algorithm has a *source domain* denoted by S, an *ultimate target domain* denoted by T, and an ordered sequence of *intermediate target domains* indicated by $(\dot{T}_1, \ldots, \dot{T}_K)$ with K being the number of intermediate domains. In this work, S is clear weather, T is dense fog, and $\dot{T}_k$'s correspond to fog density that increases with k, ranging between the density of S (zero) and T. Our method adapts semantic segmentation models through the sequence of domains $(S, \dot{T}_1, \dot{T}_2, \ldots, \dot{T}_K, T)$. The *intermediate target domains* $\dot{T}_k$'s are optional; when $K = 0$, the method reduces to a single-stage adaptation as presented in Sect. 4.1. Similarly, $K = 1$ leads to a two-stage adaptation approach as presented in the conference version of this paper (Sakaridis et al. 2018), $K = 2$ to a three-stage adaptation approach, and so on. We abbreviate these instantiations of CMAda as CMAda1 ($K = 0$), CMAda2 ($K = 1$), CMAda3 ($K = 2$), and so on.

Let us denote by $z \in \{1, \ldots, Z\}$ the domain index in the above ordered sequence $(S, \dot{T}_1, \dot{T}_2, \ldots, \dot{T}_K, T)$, with $Z = K + 2$. In this work, the sequence of domains is sorted in ascending order with respect to fog density. For instance, it could be (*clear weather*, *light fog*, *dense fog*), with *clear weather* being the source domain, *dense fog* the ultimate target domain and *light fog* the intermediate target domain. The approach proceeds progressively and adapts the semantic segmentation model from the current domain (fog density) to the subsequent one by learning from the corresponding synthetic foggy dataset and the corresponding real foggy dataset. Once the model for the subsequent domain has been trained, its knowledge is distilled on unlabeled real foggy images from that domain, and then used along with a denser version of synthetic foggy data to adapt this model to the next domain (i.e. the immediately higher fog density).

Since the method proceeds in an iterative manner, we only present the algorithmic details for model adaptation from $z-1$ to z. Let us use β_z to indicate the fog density for domain z, represented as the attenuation coefficient. In order to adapt the semantic segmentation model ϕ^{z-1} from the previous domain $z-1$ to the current domain z, we generate synthetic fog of the exact fog density β_z and inherit the human annotations of the original clear-weather images. Thus, the synthetic foggy dataset for adapting to z is

$$\mathcal{D}_{\text{syn}}^{z} = \{(\bar{\mathbf{x}}_m^{\beta_z}, \mathbf{y}_m^1)\}_{m=1}^{M}, \tag{6}$$

where M is the total number of synthetic foggy images, $\mathbf{y}_m^1(i, j) \in \{1, \ldots, C\}$ is the label of pixel (i, j) of the clear-weather image $\mathbf{x}_m^{\beta_1}$ ($\beta_1 = 0$), and C is the total number of classes.

For real foggy images, since no human annotations are available, we rely on a strategy of self-learning or curriculum learning. Objects in lighter fog are easier to recognize than in denser fog, hence models trained for lighter fog are more generalizable to real data. The model ϕ^{z-1} for the previous domain $z-1$ can be applied to all real foggy images with fog density less than β_{z-1} in order to generate supervisory labels for training model ϕ^z for domain z. Specifically, the real foggy dataset for adapting to z is

$$\mathcal{D}_{\text{real}}^{z} = \{(\mathbf{x}_n, \hat{\mathbf{y}}_n^{z-1}) \mid f(\mathbf{x}_n) \leq \beta_{z-1}\}_{n=1}^{N}, \tag{7}$$

where $\hat{\mathbf{y}}_n^{z-1} = \phi^{z-1}(\mathbf{x}_n)$ denotes the predicted labels of image $\mathbf{x}_n$ using the model ϕ^{z-1}.

Once the two training sets are formed, the aim is to learn ϕ^z from $\mathcal{D}_{\text{syn}}^{z}$ and $\mathcal{D}_{\text{real}}^{z}$. The proposed scheme balances the contributions of both the synthetic foggy dataset $\mathcal{D}_{\text{syn}}^{z}$ from domain z with human annotations and the real foggy dataset $\mathcal{D}_{\text{real}}^{z}$ from domain $z-1$ with labels inferred using model ϕ^{z-1}:

$$\min_{\phi^z} \Bigg(\sum_{\substack{(\mathbf{x}', \mathbf{y}') \\ \in \mathcal{D}_{\text{syn}}^{z}}} L(\phi^z(\mathbf{x}'), \mathbf{y}') + \lambda \sum_{\substack{(\mathbf{x}'', \mathbf{y}'') \\ \in \mathcal{D}_{\text{real}}^{z}}} L(\phi^z(\mathbf{x}''), \mathbf{y}'') \Bigg), \tag{8}$$

where $L(.,.)$ is the cross entropy loss function and $\lambda = w\frac{R}{M}$ is a hyper-parameter balancing the weights of the two datasets, with w serving as the relative weight of each real noisily labeled image compared to each synthetic labeled one and R being the number of images in $\mathcal{D}_{\text{real}}^{z}$. We empirically set $w = 1$ in our experiments, but an optimal value can be obtained via cross-validation if needed. The optimization of (8) is implemented by generating a hybrid data stream and feeding it to a CNN for standard supervised training. More specifically, during training, training images are fetched from the randomly shuffled $\mathcal{D}_{\text{syn}}^{z}$ and $\mathcal{D}_{\text{real}}^{z}$ with a ratio of $1 : w$.

We now describe the initialization stage of our method, which is also a variant of our method when no *intermediate target domains* are used. When $z = 1$, we are in the clear-weather domain and the model ϕ^1 is directly trained on a labeled real dataset, so no adaptation is required. For the case $z = 2$, there are no real foggy images falling into the domain $z - 1 = 1$ which is the clear-weather domain. In this case, the model ϕ^2 is trained with the synthetic dataset $\mathcal{D}_{\text{syn}}^{2}$ only, as specified in Sect. 4.1. For the remaining steps from $z = 3$ on, we iteratively apply the adaptation approach introduced above to adapt to domain Z, which constitutes the *ultimate target domain* T. In this work, we have experimented with three instantiations of our method for $Z = \{2, 3, 4\}$, which we name CMAda1, CMAda2 and CMAda3 respectively. The sequences of attenuation coefficients (fog densities) for the three versions are $(0, 0.01)$, $(0, 0.005, 0.01)$ and $(0, 0.0025, 0.005, 0.01)$ respectively.

Figure 1 provides an overview of CMAda2. Below, we summarize the complete operations of CMAda2 to further help understand the method. With the chosen sequence of attenuation coefficients $(0, 0.005, 0.01)$, the whole pipeline of CMAda2 is 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 ($\beta = 0.005$), 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 2) to obtain noisy semantic labels;
6. generate a dataset with dense synthetic fog ($\beta = 0.01$);
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 according to (8).

4.2.3 Discussion

CMAda adapts segmentation models from clear weather to dense fog and is inspired by curriculum learning (Bengio et al. 2009), 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. semantic scene understanding under dense real fog. CMAda also exploits the direct control of fog density for synthetic foggy images.

This learning approach also bears resemblance to model distillation (Hinton et al. 2015; Gupta 2016) or imitation (Buciluǎ et al. 2006; Dai et al. 2015). 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; and (2) models trained on the source domain can be successfully applied to the target domain when the domain gap is small, hence incremental (curriculum) domain adaptation can better propagate semantic knowledge from the source domain to the ultimate target domain than single-step domain adaptation approaches.

The goal of CMAda is to train a semantic segmentation model for the ultimate target domain z. The standard recipe is to record foggy images $\mathbf{x}^{\beta_z}$'s and then to manually create semantic labels $\mathbf{y}^{\beta_z}$'s for those foggy images so that the standard supervised learning can be applied. As discussed in Sect. 1, there is difficulty to apply this recipe to all adverse weather conditions because manual creation of $\mathbf{y}^{\beta_z}$'s is very time-consuming and expensive. To address this problem, this work develops methods to automatically create two proxy datasets for $(\mathbf{x}^{\beta_z}, \mathbf{y}^{\beta_z})$. The two proxies are defined in (6) and in (7). These two proxies reflect different and complementary characteristics of $(\mathbf{x}^{\beta_z}, \mathbf{y}^{\beta_z})$. 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. Learning jointly from both proxy datasets in CMAda reduces the influence of their individual drawbacks.

The CMAda pipeline presented in Sect. 4.2.2 is an extension of the original method proposed in the conference version (Sakaridis et al. 2018) of this paper from a two-stage approach to a general multiple-stage approach. CMAda is a stand-alone approach and already outperforms competing methods for SFSU, as discussed in Sect. 6. In the next section, we present an extension of CMAda, CMAda+, that further boosts performance.

4.3 CMAda+ with Synthetic and Densified Real Fog

As defined in (6), images in the synthetic training set $\mathcal{D}^z_{\text{syn}}$ have exactly the same fog density β_z as images in the target domain z. Images in the real dataset $\mathcal{D}^z_{\text{real}}$, however, have lower fog density than the target fog density β_z, as defined in (7). While the lower fog density of the real training images facilitates the self-learning stream of CMAda with real foggy images, the remaining domain gap due to the disparity in fog density hampers finding a better solution. In Sect. 4.3.1, we present a method to densify fog in real foggy images so that it matches the desired fog density. The fog densification method is general and can be applied beyond CMAda. In Sect. 4.3.2, we use our fog densification method to upgrade the dataset defined in (7) to a densified foggy dataset, which is used in CMAda+ along with the synthetic dataset to train the model ϕ^z.

4.3.1 Fog Densification of a Real Foggy Scene

We aim at synthesizing images with increased fog density compared to *already foggy* real input images for which no depth information is available. In this way, we can generate multiple synthetic versions of each split of our real *Foggy Zurich* dataset, where each synthetic version is characterized by a different, controlled range of fog densities, so that these densified foggy images can be leveraged in our curriculum adaptation. To this end, we utilize our fog density estimator and propose a simple yet effective approach for increasing fog density when no depth information is available for the input foggy image, by using the assumption of constant transmittance in the scene.

More formally, we denote the input real foggy image with $\mathbf{I}_l$ and assume that it can be expressed through the optical model (3). Contrary to our fog simulation on clear-weather scenes in Sect. 3, the clear scene radiance $\mathbf{R}$ is unknown and the input foggy image $\mathbf{I}_l$ cannot be directly used as its substitute for synthesizing a foggy image $\mathbf{I}_d$ with increased fog density, as $\mathbf{I}_l$ does not correspond to clear weather. Since the scene distance ℓ which determines the transmittance through (4) is also unknown, we make the simplifying assumption that the transmittance map for $\mathbf{I}_l$ is globally constant, i.e.

$$t(\mathbf{x}) = t_l, \tag{9}$$

and use the statistics for scene distance ℓ computed on Cityscapes, which features depth maps, to estimate t_l. By using the distance statistics from Cityscapes, we implicitly assume that the distribution of distances of Cityscapes roughly matches that of our *Foggy Zurich* dataset, which is supported by the fact that both datasets contain similar, road scenes. In particular, we apply our fog density estimator on $\mathbf{I}_l$ to get an estimate β_l of the input attenuation coefficient. The values for scene distance ℓ of all pixels in Cityscapes are collected into a histogram $\mathcal{H} = \{(\ell_i, p_i) : i = 1, \ldots, N\}$ with N distance bins, where ℓ_i are the bin centers and p_i are the relative frequencies of the bins. We use each bin center as representative of all samples in the bin and compute t_l as a weighted average of the transmittance values that correspond to the different bins through (4):

$$t_l = \sum_{i=1}^{N} p_i \exp\left(-\beta_l \ell_i\right). \tag{10}$$

The calculation of t_l via (10) enables the estimation of the clear scene radiance $\mathbf{R}$ by re-expressing (3) for $\mathbf{I}_l$ when (9) holds as

$$\mathbf{R}(\mathbf{x}) = \frac{\mathbf{I}_l(\mathbf{x}) - \mathbf{L}}{t_l} + \mathbf{L}. \tag{11}$$

The globally constant atmospheric light $\mathbf{L}$ which is involved in (11) is estimated in the same way as in Sect. 3.3.

For the output densified foggy image $\mathbf{I}_d$, we select a target attenuation coefficient $\beta_d > \beta_l$ and again estimate the corresponding global transmittance value t_d similarly to (10), this time plugging β_d into the formula. The output image $\mathbf{I}_d$ is finally computed via (3) as

$$\mathbf{I}_d(\mathbf{x}) = \mathbf{R}(\mathbf{x}) t_d + \mathbf{L}\left(1 - t_d\right). \tag{12}$$

If we substitute $\mathbf{R}$ in (12) using (11), the output image is expressed only through t_l, t_d, the input image $\mathbf{I}_l$ and atmospheric light $\mathbf{L}$ as

$$\begin{aligned}\mathbf{I}_d(\mathbf{x}) &= \mathbf{I}_l(\mathbf{x}) + \frac{t_d - t_l}{t_l}\left(\mathbf{I}_l(\mathbf{x}) - \mathbf{L}\right)\\ &= \frac{t_d}{t_l}\mathbf{I}_l(\mathbf{x}) + \left(1 - \frac{t_d}{t_l}\right)\mathbf{L}.\end{aligned} \tag{13}$$

Equation (13) implies that our fog densification method can bypass the explicit calculation of the clear scene radiance $\mathbf{R}$ in (11), as the output image does not depend on $\mathbf{R}$. In this way, we completely avoid dehazing our input foggy image as an intermediate step, which would pose challenges as it constitutes an inverse problem, and reduce the inference problem just to the estimation of the attenuation coefficient by assuming a globally constant transmittance. Moreover, (13) implies that the change in the value of a pixel $\mathbf{I}_d(\mathbf{x})$ with respect to $\mathbf{I}_l(\mathbf{x})$ is linear in the difference $\mathbf{I}_l(\mathbf{x}) - \mathbf{L}$. This means that distant parts of the scene, where $\mathbf{I}_l(\mathbf{x}) \approx \mathbf{L}$, are not modified significantly in the output, i.e. $\mathbf{I}_d(\mathbf{x}) \approx \mathbf{I}_l(\mathbf{x})$. On the contrary, our fog densification modifies the appearance of those parts of the scene which are closer to the camera and shifts their color closer to that of the estimated atmospheric light irrespective of their exact distance from the camera. This can be observed in the example of Fig. 4, where the closer parts of the input scene such as the red car on the left and the vegetation on the right have brighter colors in the synthesized output. The overall shift to brighter colors is verified by the accompanying RGB histograms of the input and output images in Fig. 4.

4.3.2 Fog Densification of a Real Foggy Dataset

When applying our fog densification to an entire dataset in the context of CMAda+, a simple choice is to specify the same target fog density β_z for all images in the dataset. This may completely close the domain gap due to different fog density, but it ignores the variability of the true fog density across different images in the dataset and introduces other domain discrepancies, as our fog densification makes simplifying assumptions. Thus, we propose to define the target fog density independently for each input image.

Given the dataset $\mathcal{D}^z_{\text{real}}$ defined in (7), instead of mapping all $\beta_l \in [0, \beta_{z-1}]$ to $\beta_d = \beta_z$, we choose to perform a linear mapping from $[0, \beta_{z-1}]$ to $[\beta_{z-1}, \beta_z]$. In particular, given a real foggy image with its estimated attenuation coefficient $\beta_l \in [0, \beta_{z-1}]$, the target attenuation coefficient is determined as

$$\beta_d = \beta_{z-1} + \frac{\beta_l(\beta_z - \beta_{z-1})}{\beta_{z-1}}. \tag{14}$$

Using $\psi_{\beta_l \to \beta_d}(\mathbf{x}_n)$ to indicate the densified image for $\mathbf{x}_n$, the densified real foggy dataset for CMAda+ at step z is

$$\mathcal{D}^z_{\text{real}} = \{(\psi_{\beta_l \to \beta_d}(\mathbf{x}_n), \hat{\mathbf{y}}^{z-1}_n) \mid f(\mathbf{x}_n) \leq \beta_{z-1}\}_{n=1}^{N}. \tag{15}$$

This densified dataset is then used in CMAda+ for training, along with the synthetic dataset defined in (6), based on the same formulation (8) as CMAda.

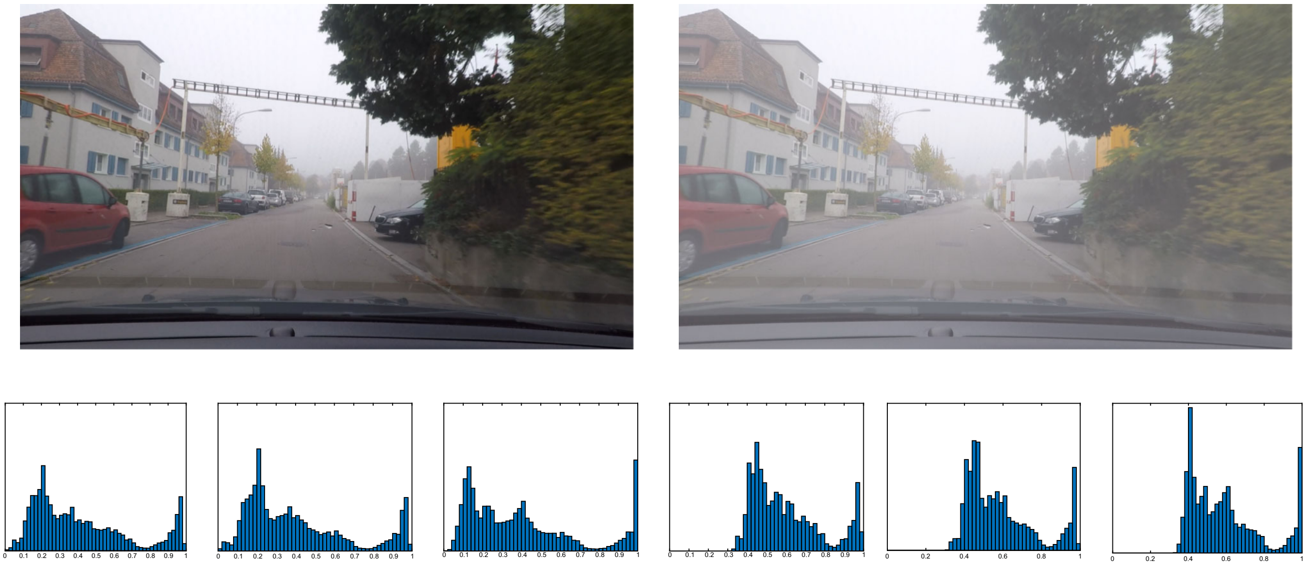


Fig. 4 Top row, left to right: example input image from *Foggy Zurich* and synthesized output image with our fog densification. Bottom row, left to right: R, G, and B histogram of the input image, R, G, and B histogram of the output image

4.4 Semantic Scene Understanding in Multiple Weather Conditions

In Sects. 4.2.2 and 4.3, specialized approaches have been developed for semantic scene understanding under fog. However, in real world applications weather conditions change constantly, e.g. the weather can change from foggy to sunny or vice versa at any time. We argue that semantic scene understanding methods need to be robust and adaptive to these changes. With this aim, we propose Model Selection, a method for selecting the appropriate model depending on the encountered weather condition.

4.4.1 Model Selection

Our method uses two expert models, one specialized for clear weather and the other for fog. In particular, a two-class classifier is trained to distinguish *clear weather* from *fog*, with images from the Cityscapes dataset used as samples of the former class and images from three versions of our *Foggy Cityscapes-DBF* dataset with attenuation coefficients 0.005, 0.01, and 0.02 as samples of the latter class. We select AlexNet (Krizhevsky et al. 2012) as the architecture of this classifier.

Denoting the semantic segmentation model specialized for fog by ϕ^Z, the respective model for clear weather by ϕ^1, and the aforementioned classifier by g, the semantic labels of a test image $\mathbf{x}$ are obtained through

$$\hat{\mathbf{y}} = \begin{cases} \phi^1(\mathbf{x}), & \text{if } g(\mathbf{x}) = 1, \\ \phi^Z(\mathbf{x}) & \text{otherwise}, \end{cases} \tag{16}$$

where label 1 indicates the *clear weather* class and label 0 indicates *fog*.

The method is not limited to these two conditions and can be directly generalized to handle multiple adverse conditions, such as *rain* or *snow*.

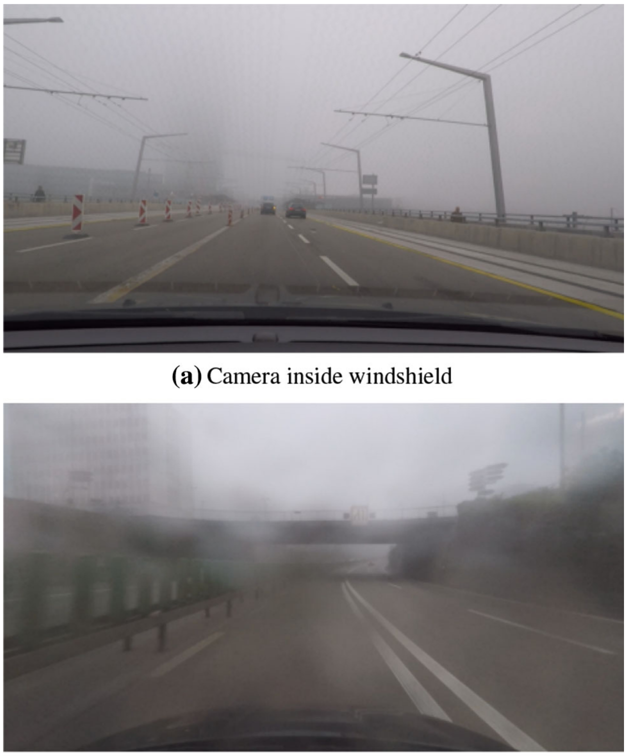
5 The Foggy Zurich Dataset

We present the *Foggy Zurich* dataset, which comprises 3808 images depicting foggy road scenes in the city of Zurich and its suburbs. We provide annotations for semantic segmentation for 40 of these scenes that contain dense fog.

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.

In particular, the small water droplets that compose fog condense and form dew on the surface of the lens very shortly after the vehicle starts moving, which causes severe blurring artifacts and contrast degradation in the image, as shown



(a) Camera inside windshield

(b) Camera outside windshield

Fig. 5 Comparison of images taken in fog with the camera mounted **a** inside and **b** outside the front windshield of the vehicle. We opt for the former configuration for collecting *Foggy Zurich*

in Fig. 5b. On the contrary, mounting the camera inside the windshield, as we did when collecting *Foggy Zurich*, prevents these blurring artifacts and affords much sharper images, to which the windshield surface incurs minimal artifacts, as shown in Fig. 5a.

5.2 Annotation of Images with Dense Fog

We use our fog density estimator presented in Sect. 4.2.1 to order all images in *Foggy Zurich* according to fog density. Based on this ordering, we manually select 40 images with *dense* fog and diverse visual scenes, and construct the test set of *Foggy Zurich* therefrom, which we term *Foggy Zurich-test*. The aforementioned selection is performed manually in order to guarantee that the test set has high diversity, which compensates for its relatively small size in terms of statistical significance of evaluation results. We annotate these images with fine pixel-level semantic annotations using the 19 evaluation classes of the Cityscapes dataset (Cordts 2016): *road*, *sidewalk*, *building*, *wall*, *fence*, *pole*, *traffic light*, *traffic sign*, *vegetation*, *terrain*, *sky*, *person*, *rider*, *car*, *truck*, *bus*, *train*, *motorcycle* and *bicycle*. 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. 6. Furthermore, we note that individual instances of *person*, *rider*, *car*, *truck*, *bus*, *train*, *motorcycle* and *bicycle* are annotated separately, which additionally induces bounding box annotations for object detection for these 8 classes, although we focus solely on semantic segmentation in this paper.

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 (Geiger et al. 2012), Cityscapes (Cordts 2016) and Mapillary Vistas (Neuhold et al. 2017), 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. Through manual inspection, we found that even Mapillary Vistas, which was

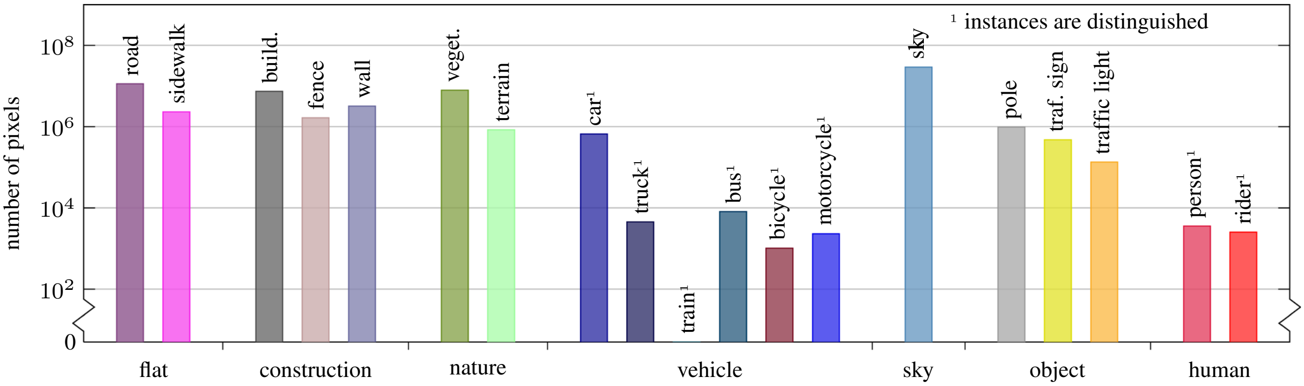


Fig. 6 Number of annotated pixels per class for *Foggy Zurich-test*

Table 1 Absolute and average number of annotated pixels, humans and vehicles for *Foggy Zurich-test*, *Foggy Driving*, KITTI and Cityscapes

	Pixels	Humans	Vehicles	h/im	v/im
Foggy Zurich	66.1M	27	135	0.7	3.4
Foggy Driving	72.8M	290	509	2.9	5.0
KITTI	0.23G	6.1k	30.3k	0.8	4.1
Cityscapes	9.43G	24.0k	41.0k	7.0	11.8

Only the training and validation sets of KITTI and Cityscapes are considered. “h/im” stands for humans per image, “v/im” for vehicles per image and “Foggy Zurich” for *Foggy Zurich-test*

specifically designed to also include scenes with adverse conditions such as snow, rain or nighttime, in fact contains very few images with fog, i.e. in the order of 10 images out of 25000, with relatively more images depicting *misty* scenes, which have MOR $\geq$ 1km, i.e. significantly better visibility than foggy scenes [1].

To the best of our knowledge, the only previous dataset for semantic foggy scene understanding whose scale exceeds that of *Foggy Zurich-test* is *Foggy Driving* (Sakaridis et al. 2018), with 101 annotated images. However, 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 Sect. 4.2.2, 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 Table 1, we compare the overall annotation statistics of *Foggy Zurich-test* to some of the aforementioned existing datasets; we note that the comparison involves a test set (*Foggy Zurich-test*) and unions of training plus validation sets (KITTI and Cityscapes), which are much larger than the respective test sets. The comparatively lower number of humans and vehicles per image in *Foggy Zurich-test* is not a surprise, as the condition of dense fog that characterizes the dataset discourages road transportation and reduces traffic.

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 that are used in CMAda those images which bear resemblance to any image in *Foggy Zurich-test* with respect to the depicted scene.

6 Experiments

Our model of choice for experiments on semantic segmentation with our CMAda pipeline is the state-of-the-art RefineNet (Lin et al. 2017). 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 *Foggy Cityscapes-DBF* by applying our fog simulation (which is denoted by “SDBF” in the following for short) on the same *refined* set of Cityscapes images that was used in Sakaridis et al. (2018) to compile Foggy Cityscapes-refined. This set comprises 498 training and 52 validation images; we use the former for training. In our experiments, we use the values 0.005 and 0.01 for attenuation coefficient β both in SDBF and the fog simulation of Sakaridis et al. (2018) (denoted by “SGF”) to generate different versions of *Foggy Cityscapes-DBF* and Foggy Cityscapes respectively with varying fog density.

6.1 Performance on Foggy Scenes

For evaluation, we use (1) *Foggy Zurich-test*, (2) a subset of *Foggy Driving* (Sakaridis et al. 2018) containing 21 images with dense fog, which we term *Foggy Driving-dense*, and (3) the entire *Foggy Driving* (Sakaridis et al. 2018).

We summarize our main experimental results in Table 2. Overall, our method significantly improves the performance of semantic segmentation under dense fog compared to the original RefineNet model which has been trained on clear-weather images of Cityscapes. More specifically, we improve the performance (mIoU) from 34.6 to **46.8**% on *Foggy Zurich-test* and from 35.8 to **43.0**% on *Foggy Driving-dense*. With the new extensions, our fully-fledged CMAda3+ method significantly outperforms CMAda2, which was originally presented in the conference version of this paper (Sakaridis et al. 2018).

It is worthwhile to mention that these improvements are achieved without using any extra human annotations on top of the original Cityscapes. Also, images in *Foggy Driving* were taken by different cameras than the GoPro Hero 5 camera used for *Foggy Zurich*, showing that CMAda also generalizes well to different sensors from that corresponding to the real training set of the method.

In the rest of Sect. 6.1, we analyze the effect of the individual components of our approach. This analysis demonstrates the benefit for semantic segmentation of real foggy scenes of: (1) our fog simulation for generating synthetic training data, (2) our fog density estimator against a state-of-the-art competing method, 3) combining our synthetic foggy data from *Foggy Cityscapes-DBF* with unlabeled real data from *Foggy Zurich* through our CMAda pipeline to adapt *gradually* to dense real fog in multiple steps, and 4) using our fog densification method to further close the gap between light real fog and dense real fog. Finally, we provide some qualitative results.

6.1.1 Benefit of Adaptation with Our Synthetic Fog

Our first segmentation experiment shows that our semantic-aware fog simulation (SDBF) performs competitively compared to the fog simulation of Sakaridis et al. (2018) (SGF) for generating synthetic data to adapt RefineNet to real dense fog. *RefineNet-res101-Cityscapes* is fine-tuned on *Foggy Cityscapes-DBF* and alternatively Foggy Cityscapes, both with attenuation coefficient $\beta = 0.01$, for 8 epochs. The corresponding results in Table 2 are presented in the top two rows under the group "CMAda1". Training on synthetic fog with either type of fog simulation helps to beat the baseline clear-weather RefineNet model on all three test sets, the improvement being more significant on *Foggy Zurich-test* and *Foggy Driving*. In addition, SDBF slightly outperforms SGF consistently.

Moreover, in all cases that both synthetic and real foggy data are used in the two-stage CMAda pipeline, corresponding to the rows of Table 2 grouped under "CMAda2", SDBF yields significantly higher segmentation performance on *Foggy Zurich-test* compared to SGF, while the two methods are on a par on the other two sets.

6.1.2 Benefit of Our Fog Density Estimator on Real Data

The second component of the CMAda pipeline that we ablate is the fog density estimator. In particular, Table 2 includes results for the single-stage pipeline with adaptation on real images from the unlabeled part of *Foggy Zurich* and the two-stage pipeline with adaptation on synthetic and real images from Foggy Cityscapes and *Foggy Zurich* respectively, where the ranking of real images according to fog density is performed either with the method of Choi et al. (2015) or with our AlexNet-based fog density estimator described in Sect. 4.2.1. In all experimental settings, our fog density estimator outperforms (Choi et al. 2015) significantly in terms of mIoU on all datasets. This fully lifts the need of manually designing features and labeling images for fog density estimation, as was done in Choi et al. (2015).

For further verification of our fog density estimator, we conduct a user study on Amazon Mechanical Turk (AMT). 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 with images from the real *Foggy Zurich* dataset, and 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 estimator. In order to avoid confusing cases, i.e. two images of similar fog densities, the two images of each pair need to be ranked at least 20 percentiles apart from each other by our estimator.

We have collected answers for 12,000 pairs in 4000 HITs. The HITs are considered for evaluation only when both validation questions are correctly answered. 87% of all HITs are valid for evaluation. On these 10,400 pairs, the agreement between our fog density estimator and human judgment is 89.3%. This high agreement confirms that fog density estimation is a relatively easier task which can be solved by using synthetic data, and the acquired knowledge can be further exploited for solving high-level tasks on foggy scenes. Figure 7 shows foggy images in ascending order of estimated fog density using our estimator.

6.1.3 Benefit of Adaptation with Synthetic and Real Fog

The main segmentation experiment showcases the effectiveness of our CMAda pipeline. *Foggy Cityscapes-DBF* and Foggy Cityscapes (Sakaridis et al. 2018) are the two alternatives for the synthetic foggy training sets in steps 4 and 6 of the pipeline, corresponding to the two alternatives for fog simulation (SDBF and SGF respectively). *Foggy Zurich* serves as the real foggy training set. We use the results of our fog density estimation to select 1556 images from *Foggy Zurich* with light fog and name this set *Foggy Zurich-light*. We implement CMAda2 by first fine-tuning RefineNet on *Foggy Cityscapes-DBF* (alternatively Foggy Cityscapes) with $\beta = 0.005$ for 6k iterations and then further fine-tuning it on the union of *Foggy Cityscapes-DBF* (alternatively Foggy Cityscapes) with $\beta = 0.01$ and *Foggy Zurich-light*, where the latter set is labeled by the aforementioned initially adapted model. Two-stage curriculum adaptation to dense fog with synthetic and real data, which corresponds to the results in the rows that are grouped under "CMAda2" in Table 2, consistently outperforms single-stage adaptation with either only synthetic or only real training data ("CMAda1"), irrespective of the selected fog simulation and fog density estimation methods. The combination of our fog simulation SDBF and our fog density estimator delivers the best result on all three test sets among all variants of CMAda2, improving upon the baseline RefineNet model on *Foggy Zurich-test* by 8.3%. The same combination also provides a clear generalization benefit of 4.2% against the baseline on *Foggy Driving*, even though this dataset involves different camera sensors and scenes than *Foggy Zurich*, which is the sole real-world dataset used in our training.

We note that the significant performance benefit delivered by CMAda both on *Foggy Zurich-test* and *Foggy Driving* is not matched by the state-of-the-art domain-adversarial approach of Tsai et al. (2018) for adaptation of semantic segmentation models, which we also trained both on our synthetic *Foggy Cityscapes-DBF* set and our unlabeled real *Foggy Zurich-light* set. This can be attributed to the fact that

Table 2 Performance comparison on *Foggy Zurich-test* (*FZ*), *Foggy Driving-dense* (*FDD*) and *Foggy Driving* (*FD*) of different variants of our CMAda pipeline as well as competing approaches, using with the mean intersection-over-union (mIoU) metric over all classes

	Clear-weather	Synthetic fog		Real fog	Density Estimator		*FZ*	*FDD*	*FD*
	Cityscapes Cordts (2016)	SGF Sakaridis et al. (2018)	SDBF (ours)	GoPro	FADE Choi et al. (2015)	Ours	mIoU (%)	mIoU (%)	mIoU (%)
Comparison									
RefineNet Lin et al. (2017)	✓						34.6	35.8	44.3
SFSU Sakaridis et al. (2018)	✓	✓					35.7	35.9	46.3
AdSegNet Tsai et al. (2018)	✓		✓	✓			25.0	15.8	29.7
CMAda2 Sakaridis et al. (2018)	✓		✓	✓		✓	42.9	37.3	48.5
CMAda3+	✓		✓	✓		✓	**46.8**	**43.0**	**49.8**
Ablation study									
Baseline Lin et al. (2017)	✓						34.6	35.8	44.3
CMAda1	✓	✓					35.7	35.9	46.3
	✓		✓				36.3	36.1	46.3
	✓			✓	✓		37.5	36.4	45.7
	✓			✓		✓	38.9	36.6	46.0
CMAda2	✓	✓		✓	✓		39.8	35.7	47.5
	✓	✓		✓		✓	41.5	37.0	48.5
	✓		✓	✓	✓		40.6	35.5	47.7
	✓		✓	✓		✓	42.9	37.3	48.5
CMAda3	✓		✓	✓		✓	43.7	40.6	48.9
CMAda2+	✓		✓	✓		✓	43.4	40.1	**49.9**
CMAda3+	✓		✓	✓		✓	**46.8**	**43.0**	49.8

Bold values indicate the best performance on each dataset by all segmentation methods

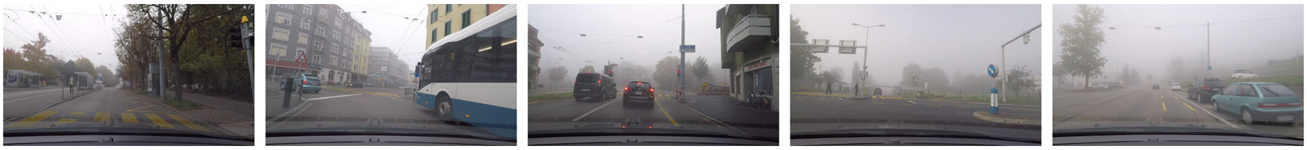


Fig. 7 Foggy images from *Foggy Zurich*, sorted from left to right in ascending order with respect to estimated fog density using our estimator

images captured under adverse conditions such as fog have large intra-domain variance as a result of poor visibility, effects of artificial lighting sources and motion blur. However, we believe that domain-adversarial approaches have the potential to be used for transferring knowledge to adverse weather domains.

6.1.4 Benefit of Adaptation at Finer Scales

We also experiment with the three-stage instantiation of CMAda, CMAda3, using the optimal configuration of all components of the pipeline based on the previous comparisons. Compared to CMAda2, CMAda3 adapts the semantic segmentation model at a finer scale, i.e. 1) from clear-weather to mist with synthetic misty data; 2) then to light fog with synthetic light foggy data and real misty data; and 3) finally to dense fog with synthetic dense foggy data and real light foggy data. The exact fog densities at each stage are defined in Sect. 4.2.2. In particular, the extra stage compared to CMAda2 consists in labeling a split of *Foggy Zurich* with very light estimated fog, which we term *Foggy Zurich-light+*, via the clear-weather RefineNet model and using it in conjunction with *Foggy Cityscapes-DBF* with $\beta = 0.005$ to form the training set for the first stage of CMAda.

Including this extra stage affords higher segmentation performance on all three test sets as reported in row "CMAda3" of Table 2, outperforming the respective best CMAda2 instance by 3.3% on *Foggy Driving-dense*. The improvement of CMAda3 over CMAda2 shows that our approach benefits from adaptation at finer scales, which is in line with the rationale of curriculum learning (Bengio et al. 2009). However, training for a large number of stages increases the computational cost significantly. Thus, selecting the "optimal" number of stages and the exact fog densities that correspond to the intermediate target domains needs further investigation and could be solved to some extent by cross-validation.

6.1.5 Benefit of Fog Densification

The final component of our proposed pipeline that we evaluate is our fog densification method, introduced in Sect. 4.3. Table 2 shows the results of CMAda2+ and CMAda3+ on the three test datasets, along with the results of their counterparts CMAda2 and CMAda3. CMAda2+ and CMAda2 use the same training parameters. The same holds for CMAda3+ and CMAda3. Applying our fog densification to the real foggy training sets used in CMAda significantly improves performance for both numbers of adaptation stages that are examined. For instance, CMAda3+ outperforms CMAda3 by 3.1%, 2.4% and 0.9% on *Foggy Zurich-test*, *Foggy Driving-dense* and *Foggy Driving* respectively. This is because without fog densification, the images in the synthetic dataset $\mathcal{D}^z_{\text{syn}}$ of each adaptation stage (defined in (6)) have the exact same fog density β_z as images in the target domain of that stage, whereas the images in the real dataset $\mathcal{D}^z_{\text{real}}$ have *lower* fog density than β_z (cf. 7). This lower fog density of the real training images facilitates the self-learning, bootstrapping strategy. However, it also creates a domain gap between training and test images due to the difference in their fog density. On the contrary, the dataset with densified fog defined in (15) matches the target fog density of the test images, which helps close this domain gap and significantly boosts the performance of CMAda.

6.1.6 Qualitative Results and Discussion

In Fig. 8, we show segmentation results on *Foggy Zurich-test* generated with our best-performing method CMAda3+, our conference paper method CMAda2 and the single-stage version CMAda1 using only synthetic training data from *Foggy Cityscapes-DBF*, compared to the method of Sakaridis et al. (2018) that only uses synthetic data from Foggy Cityscapes (Sakaridis et al. 2018) and the clear-weather RefineNet model (Lin et al. 2017). This visual comparison demonstrates that our multiple-stage methods CMAda3+ and CMAda2 yield significantly better results and generally capture the road layout more accurately than the two competing approaches and our single-stage method CMAda1. Moreover, the more stages CMAda involves, the more accurate the segmentation result is in general. For instance, on the leftmost image of Fig. 8, CMAda3+ segments the wall and the vegetation on the right side much better than the other methods and only misclassifies some parts of them as *building*, which is a much less detrimental error from a driving perspective than confusing these classes with *road*, as is the case for the other methods. Similarly, the buildings and the tree trunk in the third image are better segmented by CMAda3+.

To further demonstrate the behavior of CMAda, we also show semantic segmentation results of the clear-weather RefineNet model (Lin et al. 2017) and the three aforementioned variants of our method for variable fog density in Fig. 9. In particular, we have applied our fog density estimator to *Foggy Driving* and use four images therefrom for which the estimated fog density ranges from very low to very high. First, we observe that the clear-weather baseline performs comparably well for very light fog due to the small domain shift from clear weather, but for higher fog densities CMAda variants outperform this baseline. The advantage gets more pronounced as fog density increases. Second, comparing the different CMAda variants, we conclude that having more adaptation stages leads to increasing returns as fog density increases. For instance, the bus in the highly foggy rightmost image is correctly recognized only after all three adaptation stages have been applied.

While we observe a significant improvement with CMAda, semantic segmentation performance on foggy scenes is still much worse than the reported performance by existing papers on clear-weather scenes. Foggy scenes are indeed more challenging than clear-weather scenes with respect to understanding their semantics. There are more underlying causal factors of variation that generated foggy data, which requires either more training data or more intelligent learning approaches to disentangle the increased degrees of freedom. While our method shows considerable improvement by transferring semantic knowledge from clear-weather to fog, the models are adapted in an "unsupervised" manner, i.e. without using human annotations of real foggy data. Incorporating a moderate amount of human annotations of real foggy scenes into our learning approach is a promising research direction, if significantly better results are desired.

Our method involves two data streams: partially synthetic data with annotations and real data without annotations. Learning from the real data stream is based on a "self-learning" mechanism, which creates a risk of entering a negative reinforcement loop by adapting to mistakes made at previous stages. In practice, we find that our training process is stable. In order to further investigate this, we follow the literature (Radosavovic et al. 2018) to identify and exclude the

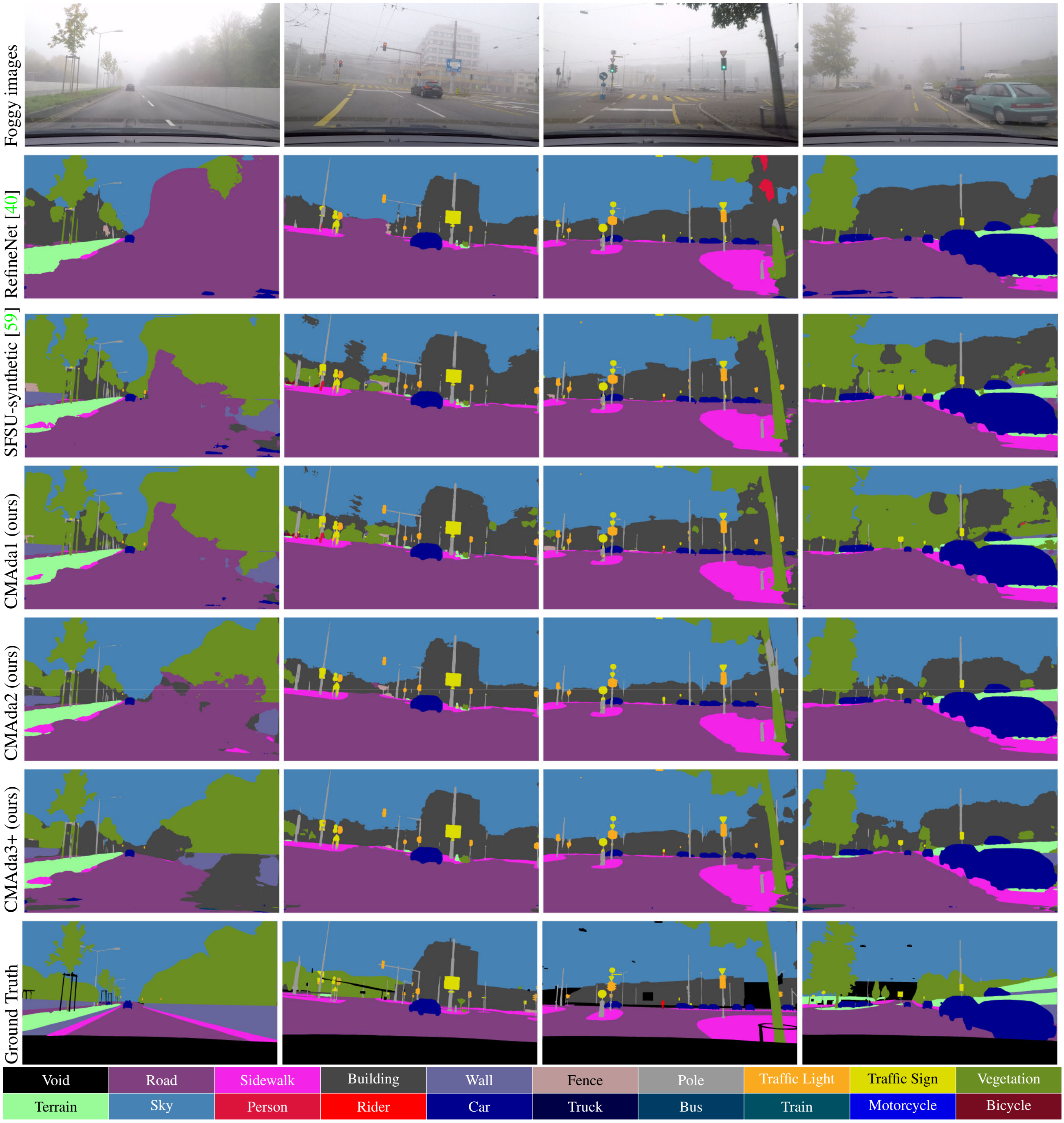


Fig. 8 Qualitative results for semantic segmentation on *Foggy Zurich-test*

erroneous predictions from training. In particular, the confidence scores of the predictions are used as a proxy for prediction quality and we generate pseudo-labels only for pixels where this confidence is higher than a defined threshold. This prediction selection step, however, does not provide clear benefit and thus is not included in our approach.

We believe that the low risk of entering the negative reinforcement loop and the steady improvement of our method can be ascribed to two factors: (1) the accurate human annotations of the partially synthetic data stream restrict the space of adapted models, ruling out solutions that would create severe errors in the inferred labels of the real data; and (2) each adaptation stage is initialized with the solution of the previous stage, which helps smoothly traverse the model space from the initial clear-weather model to the target foggy model.

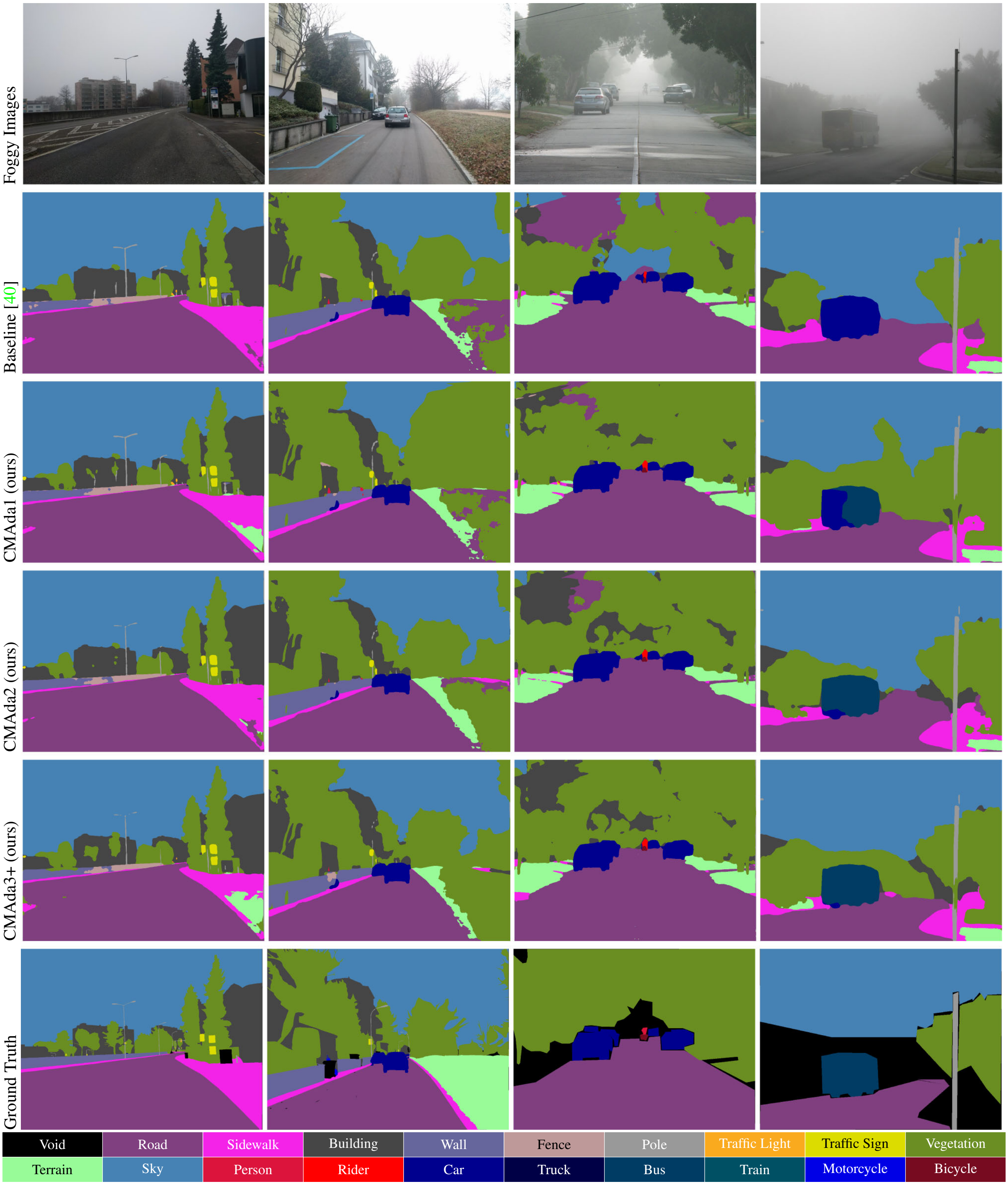


Fig. 9 Qualitative semantic segmentation results on images from *Foggy Driving* with varying fog density. Foggy images in the top row are sorted from left to right in ascending order of estimated fog density using our estimator

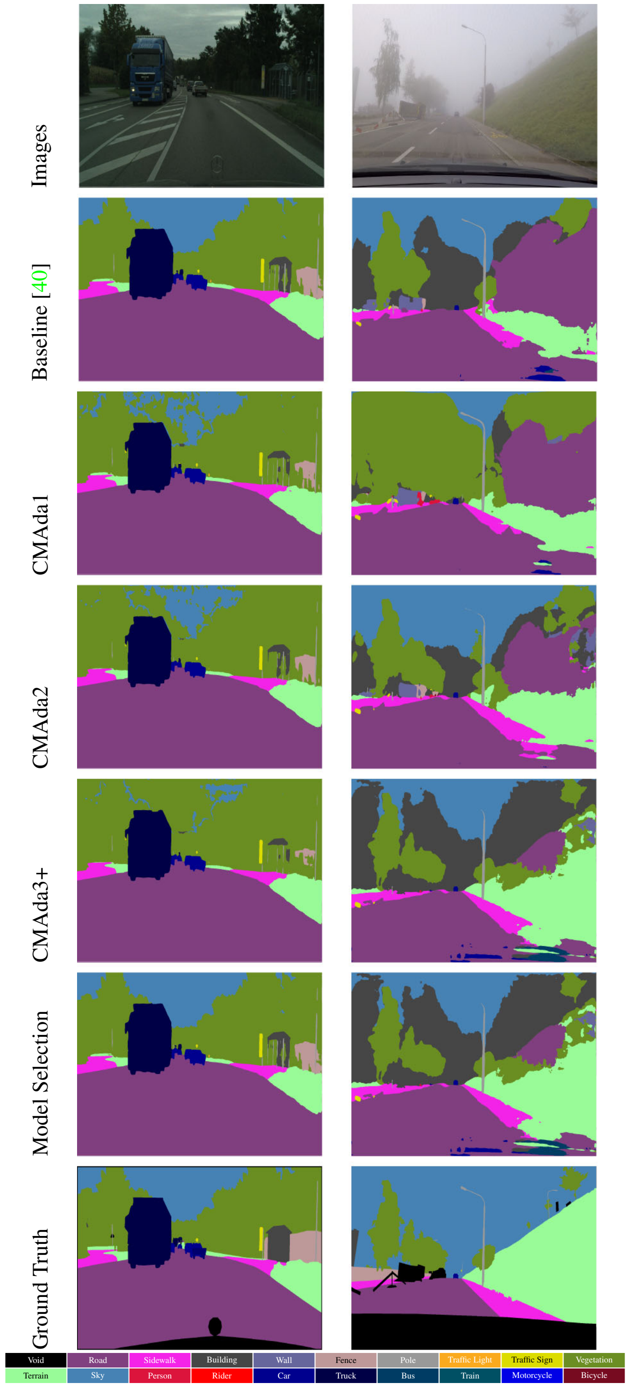


Fig. 10 Qualitative semantic segmentation results under two weather conditions: clear weather (left) and foggy weather (right)

6.2 Performance in Multiple Weather Conditions

We first note that the results which have been presented in Table 2 on the *Foggy Driving* dataset (Sakaridis et al. 2018), which contains images of varying fog densities from very low to high, show that adaptation with CMAda to dense fog also brings a significant benefit for *lower* fog densities.

In the following, we turn to evaluation of our Model Selection method presented in Sect. 4.4.1 for the task of *semantic scene understanding in multiple weather conditions*. We consider two conditions: foggy weather and clear weather. This means that the test set comprises a mixture of images captured either in clear weather or under fog. In particular, we report the performance of three domain-specific methods and two variants of our Model Selection on three datasets. The three domain-specific methods are: (1) RefineNet, which is trained on Cityscapes dataset (Lin et al. 2017) for clear weather, (2) CMAda2, which is trained for foggy weather, and (3) CMAda3+, which is also trained for foggy weather. The first variant of Model Selection uses RefineNet and CMAda2 as its two expert models and the second one uses RefineNet and CMAda3+ respectively. The three test datasets are *Cityscapes-lindau-40*, *Foggy Zurich-test*, and *Clear-Foggy-80*, which is the union of the two previous sets. *Cityscapes-lindau-40* contains the first 40 images (in alphabetical order) from the city of Lindau in the validation set of Cityscapes.

The performance of all five methods on the three datasets is reported in Table 3. We share a few observations. First, as discussed in previous sections, our adapted models significantly improve the recognition performance on foggy scenes. Second, it seems that some knowledge initially learned for recognition in clear-weather scenes is forgotten by our models during the adaptation process. This is also evidenced by the visual comparison in Fig. 10, where the sky in the first image is misclassified after the adaptation. This is because during the adaptation stages, we aim for the best expert model for (dense) foggy scenes and have not included any clear weather images. Adding some clear-weather images into the training data will alleviate this problem, but at a cost of lower performance on foggy scenes. Last but not least, both variants of our Model Selection method demonstrate higher performance than their constituent expert models. The second variant of Model Selection with RefineNet and CMAda3+ yields the best performance. It works especially well on the *Clear-Foggy-80* dataset which contains 40 foggy images and 40 clear weather images, due to the good performance of the two expert models in their own domains. The improved performance with Model Selection implies that training multiple expert models—each for a different condition—and adaptively selecting the best one at testing time based on the input is a promising direction for semantic scene understanding in adverse conditions. We also demonstrate the improvement with Model Selection in Fig. 10 when both clear weather and fog are considered.

Table 3 Performance comparison of RefineNet (trained for clear weather), CMAda2 (trained for foggy weather), CMAda3+ (trained for foggy weather), and our Model Selection method on three datasets: *Cityscapes-lindau-40* (clear weather), *Foggy Zurich-test* (foggy weather) and the union of the two *Clear-Foggy-80* (clear + foggy weather)

Weather	RefineNet	CMAda2	CMAda3+	MS_R2	MS_R3+
Mean IoU over *all* classes (%)					
Clear	**67.2**	65.1	59.6	**67.2**	**67.2**
Foggy	34.6	42.9	**46.8**	42.9	**46.8**
Clear + Foggy	54.3	59.1	58.1	59.3	**62.2**

Bold values indicate the best performance on each dataset by all segmentation methods

"MS_R2" stands for Model Selection with RefineNet and CMAda2 as the two expert models and "MS_R3+" for Model Selection with RefineNet and CMAda3+ as the two expert models

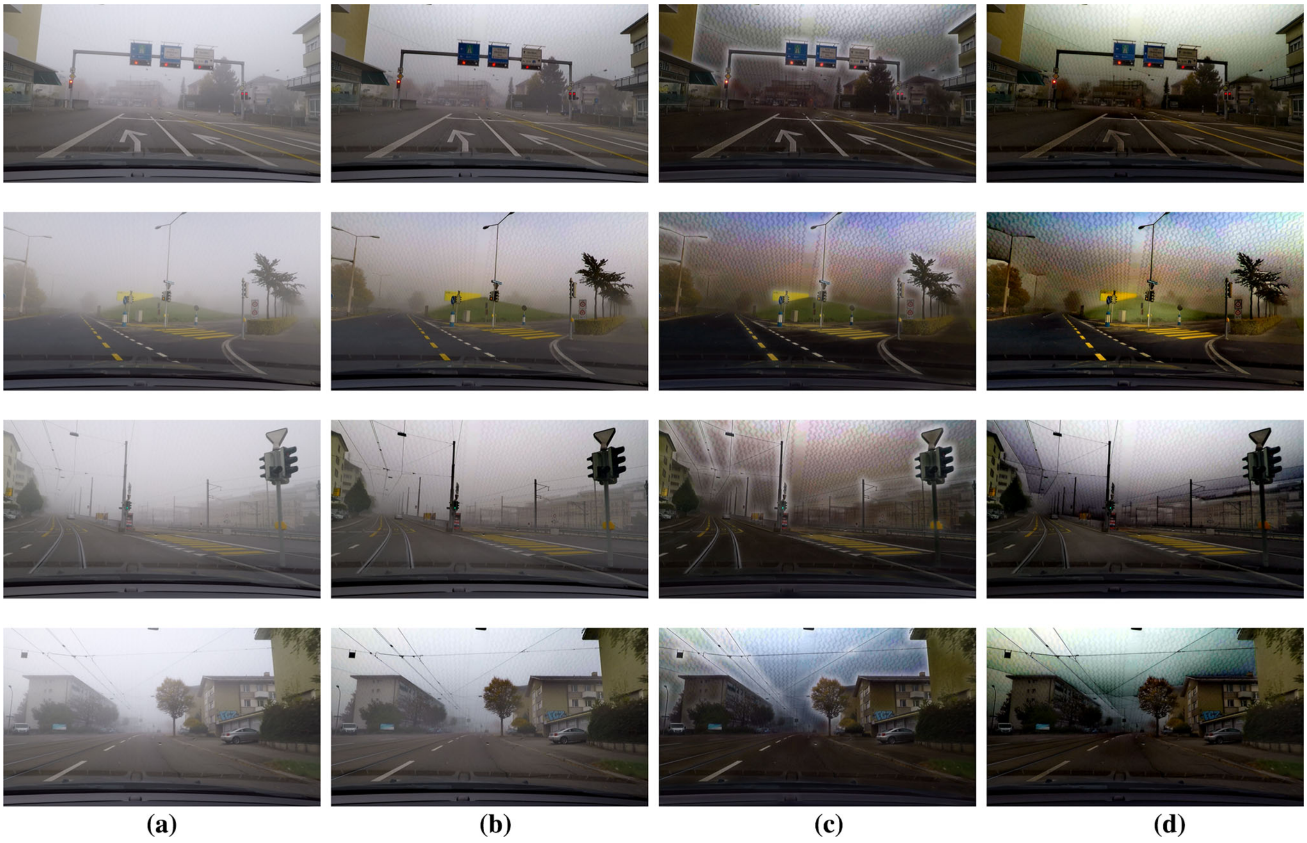


Fig. 11 Representative images from *Foggy Zurich-test* and dehazed versions of them obtained with the three dehazing methods that we consider in our experiments on utility of dehazing preprocessing. **a** *Foggy Zurich-test image*. **b** MSCNN (Ren et al. 2016). **c** DCP (He et al. 2011). **d** Non-local (Berman et al. 2016). This figure is better seen on an screen and zoomed in

6.3 Investigating the Utility of Dehazing Preprocessing

For completeness, we conduct an experimental comparison of the baseline RefineNet model of Table 2 and our single-stage CMAda pipeline using only synthetic training data against a dehazing preprocessing baseline, and report the results on *Foggy Zurich-test* and *Foggy Driving-dense* in Tables 4 and 5 respectively. In particular, we consider dehazing as an optional preprocessing step before feeding the input foggy images to the segmentation model, and experiment with four options with respect to this dehazing preprocessing: no dehazing at all (already examined in Sect. 6.1.1), multi-scale convolutional neural networks (MSCNN) (Ren et al. 2016), dark channel prior (DCP) (He et al. 2011), and non-local dehazing (Berman et al. 2016). Apart from directly applying the original clear-weather RefineNet model on the dehazed test images, the results of which are included in the

Table 4 Performance comparison on *Foggy Zurich-test* of RefineNet ("w/o FT") versus fine-tuned versions of it ("FT") trained on *Foggy Cityscapes-DBF* with attenuation coefficient $\beta = 0.005$, for four options regarding dehazing: no dehazing, MSCNN (Ren et al. 2016), DCP (He et al. 2011), and Non-local (Berman et al. 2016)

	No dehazing	MSCNN	DCP	Non-local
Mean IoU over *all* classes (%)				
w/o FT	**34.6**	34.4	31.2	27.6
FT	**36.7**	36.1	34.2	29.1
Mean IoU over *frequent* classes (%)				
w/o FT	**51.8**	48.6	42.9	41.1
FT	**51.7**	49.8	46.6	44.2

Bold values indicate the best performance among all pre-processing (dehazing) methods

Table 5 Performance comparison on Foggy Driving-dense of RefineNet ("w/o FT") versus fine-tuned versions of it ("FT") trained on *Foggy Cityscapes-DBF* with attenuation coefficient $\beta = 0.005$, for four options regarding dehazing: no dehazing, MSCNN (Ren et al. 2016), DCP (He et al. 2011), and Non-local (Berman et al. 2016)

	No dehazing	MSCNN	DCP	Non-local
Mean IoU over *all* classes (%)				
w/o FT	35.8	**38.3**	33.2	32.8
FT	36.6	**40.0**	35.8	37.5
Mean IoU over *frequent* classes (%)				
w/o FT	**57.6**	55.5	47.4	50.7
FT	**60.8**	60.6	54.6	58.9

Bold values indicate the best performance among all pre-processing (dehazing) methods

"w/o FT" rows of Tables 4 and 5, we also fine-tune this model on the dehazed versions of our synthetic *Foggy Cityscapes-DBF* dataset, and compare against fine-tuning directly on the synthetic foggy images (already examined in Sect. 6.1.1). Our experimental protocol is consistent: the same dehazing option is used both before fine-tuning and at testing time. The attenuation coefficient for *Foggy Cityscapes-DBF* is $\beta = 0.005$. The rest details are the same as in Sect. 6.1.1. *Not* applying dehazing generally leads to the best results irrespective of using the original model or fine-tuned versions of it. Fine-tuning without dehazing performs best in all cases but one (*Foggy Driving-dense* and evaluation on all classes), which confirms the merit of our approach. This lack of significant improvement with dehazing preprocessing is in congruence with the findings of (Sakaridis et al. 2018), which has dissuaded us from including dehazing preprocessing in our default CMAda pipeline.

Figure 11 illustrates the results of the examined dehazing methods on sample images from *Foggy Zurich-test* and reveals the issues these methods face on real-world outdoor images with dense fog. Only MSCNN are able to slightly enhance the image contrast while introducing only minor artifacts. This correlates with the superior performance of the segmentation model that uses MSCNN for dehazing preprocessing compared to the models that use the other two methods, as reported in Table 4. Still, directly using the original foggy images generally outperforms all dehazing preprocessing alternatives.

7 Conclusion

In this article, 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 additional input to a novel dual-reference cross-bilateral filter, and applied it to the Cityscapes dataset (Cordts 2016) to obtain *Foggy Cityscapes-DBF*. In addition, we have introduced a simple CNN-based fog density estimator which can benefit from large synthetic datasets such as *Foggy Cityscapes-DBF* that provide straightforward ground truth for this task. On the real data side, we have presented *Foggy Zurich*, a large-scale real-world dataset of foggy scenes, including pixel-level semantic annotations for 40 scenes with dense fog. Through extensive evaluation, we have showcased that: (1) our Curriculum Model Adaptation exploits both our synthetic and our real data in a synergistic manner and significantly boosts performance on real fog without using any labeled real foggy image, and (2) our fog simulation and fog density estimation methods outperform their state-of-the-art counterparts.

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

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(11), 2274–2282.

Alvarez, J. M., Gevers, T., LeCun, Y., & Lopez, A. M .(2012). Road scene segmentation from a single image. In *European Conference on Computer Vision*.

Bar Hillel, A., Lerner, R., Levi, D., & Raz, G. (2014). Recent progress in road and lane detection: A survey. *Machine Vision and Applications*, *25*(3), 727–745.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *International conference on machine learning* (pp. 41–48).

Berman, D., Treibitz, T., & Avidan, S. (2016). Non-local image dehazing. In *IEEE conference on computer vision and pattern recognition (CVPR).*

Bronte, S., Bergasa, L. M., & Alcantarilla, P. F. (2009). Fog detection system based on computer vision techniques. In *International IEEE conference on intelligent transportation systems*.

Brostow, G. J., Shotton, J., Fauqueur, J., & Cipolla, R. (2008). Segmentation and recognition using structure from motion point clouds. In *European conference on computer vision*.

Buciluă, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. In *International conference on knowledge discovery and data mining (SIGKDD)*.

Chen, Y., Li, W., Sakaridis, C., Dai, D., & Van Gool, L. (2018). Domain adaptive faster r-CNN for object detection in the wild. In *Conference on computer vision and pattern recognition (CVPR)*.

Choi, L. K., You, J., & Bovik, A. C. (2015). Referenceless prediction of perceptual fog density and perceptual image defogging. *IEEE Transactions on Image Processing*, *24*(11), 3888–3901.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes dataset for semantic urban scene understanding. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Dai, D., Kroeger, T., Timofte, R., & Van Gool, L. (2015). Metric imitation by manifold transfer for efficient vision applications. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Dai, D., & Van Gool, L. (2013). Ensemble projection for semi-supervised image classification. In *International conference on computer vision (ICCV)*.

Dai, D., & Van Gool, L. (2018). Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *IEEE international conference on intelligent transportation systems*.

Dhall, A., Dai, D., & Van Gool, L. (2019). Real-time 3D traffic cone detection for autonomous driving. In *IEEE intelligent vehicles symposium (IV)*.

Eisemann, E., & Durand, F. (2004). Flash photography enhancement via intrinsic relighting. In *ACM SIGGRAPH*.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The PASCAL visual object classes (VOC) challenge. *IJCV*, *88*(2), 303–338.

Fattal, R. (2008). Single image dehazing. *ACM transactions on graphics (TOG)*, *27*(3), 1.

Fattal, R. (2014). Dehazing using color-lines. *ACM transactions on graphics (TOG)*, *34*(1), 13.

Federal Meteorological Handbook No. 1: Surface Weather Observations and Reports. (2005). U.S. Department of Commerce, National Oceanic and Atmospheric Administration.

Gallen, R., Cord, A., Hautière, N., & Aubert, D. (2011). Towards night fog detection through use of in-vehicle multipurpose cameras. In *IEEE intelligent vehicles symposium (IV)*.

Gallen, R., Cord, A., Hautière, N., Dumont, É., & Aubert, D. (2015). Nighttime visibility analysis and estimation method in the presence of dense fog. *IEEE Transactions on Intelligent Transportation Systems*, *16*(1), 310–320.

Garg, K., & Nayar, S. K. (2007). Vision and rain. *International Journal of Computer Vision*, *75*(1), 3–27.

Geiger, A., Lenz, P., & Urtasun, R. (2012) Are we ready for autonomous driving? The KITTI vision benchmark suite. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Girshick, R. (2015). Fast R-CNN. In *International conference on computer vision (ICCV)*.

Gupta, S., Hoffman, J., & Malik, J. (2016). Cross modal distillation for supervision transfer. In *The IEEE conference on computer vision and pattern recognition (CVPR)*

Hautière, N., Tarel, J. P., Lavenant, J., & Aubert, D. (2006). Automatic fog detection and estimation of visibility distance through use of an onboard camera. *Machine Vision and Applications*, *17*(1), 8–20.

He, K., Sun, J., & Tang, X. (2011). Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(12), 2341–2353.

He, K., Sun, J., & Tang, X. (2013). Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(6), 1397–1409.

Hecker, S., Dai, D., & Van Gool, L. (2018). End-to-end learning of driving models with surround-view cameras and route planners. In *European conference on computer vision (ECCV)*.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hoffman, J., Guadarrama, S., Tzeng, E., Hu, R., Donahue, J., Girshick, R., Darrell, T., & Saenko, K. (2014). LSDA: Large scale detection through adaptation. In *Neural information processing systems (NIPS)*.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., & Darrell, T. (2018). CyCADA: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*.

Jensen, M. B., Philipsen, M. P., Møgelmose, A., Moeslund, T. B., & Trivedi, M. M. (2016). Vision for looking at traffic lights: Issues, survey, and perspectives. *IEEE Transactions on Intelligent Transportation Systems*, *17*(7), 1800–1815.

Kopf, J., Cohen, M. F., Lischinski, D., & Uyttendaele, M. (2007). Joint bilateral upsampling. *ACM transactions on graphics*, *26*, 3.

Koschmieder, H. (1924). Theorie der horizontalen Sichtweite. Beitrage zur Physik der freien Atmosphäre.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *NIPS*.

Levinkov, E., & Fritz, M. (2013). Sequential bayesian model update under structured scene prior for semantic road scenes labeling. In *IEEE international conference on computer vision*.

Li, Y., You, S., Brown, M. S., & Tan, R. T. (2016). Haze visibility enhancement: A survey and quantitative benchmarking. In *CoRR*. arXiv:1607.06235.

Lin, G., Milan, A., Shen, C., & Reid, I. (2017). Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Ling, Z., Fan, G., Wang, Y., & Lu, X. (2016). Learning deep transmission network for single image dehazing. In *IEEE international conference on image processing (ICIP)*.

Miclea, R. C., & Silea, I. (2015). Visibility detection in foggy environment. In *International Conference on Control Systems and Computer Science*.

Misra, I., Shrivastava, A., & Hebert, M. (2015). Watch and learn: Semi-supervised learning for object detectors from video. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Narasimhan, S. G., & Nayar, S. K. (2002). Vision and the atmosphere. *International Journal of Computer Vision*, *48*(3), 233–254.

Narasimhan, S. G., & Nayar, S. K. (2003). Contrast restoration of weather degraded images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(6), 713–724.

Negru, M., Nedevschi, S., & Peter, R. I. (2015). Exponential contrast restoration in fog conditions for driving assistance. *IEEE Transactions on Intelligent Transportation Systems*, *16*(4), 2257–2268.

Neuhold, G., Ollmann, T., Rota Bulò, S., & Kontschieder, P. (2017). The Mapillary Vistas dataset for semantic understanding of street scenes. In *The IEEE international conference on computer vision (ICCV)*.

Nishino, K., Kratz, L., & Lombardi, S. (2012). Bayesian defogging. *International Journal of Computer Vision*, *98*(3), 263–278.

Paris, S., & Durand, F. (2009). A fast approximation of the bilateral filter using a signal processing approach. *International Journal of Computer Vision*, *81*, 24.

Pavlić, M., Belzner, H., Rigoll, G., & Ilić, S. (2012). Image based fog detection in vehicles. In *IEEE intelligent vehicles symposium*.

Pavlić, M., Rigoll, G., & Ilić, S. (2013). Classification of images in fog and fog-free scenes for use in vehicles. In *IEEE intelligent vehicles symposium (IV)*.

Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M., Hoppe, H., & Toyama, K. (2004). Digital photography with flash and no-flash image pairs. In *ACM SIGGRAPH*.

Radosavovic, I., Dollár, P., Girshick, R., Gkioxari, G., & He, K. (2018). Data distillation: Towards omni-supervised learning. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *4*, 91–99.

Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., & Yang, M.H. (2016). Single image dehazing via multi-scale convolutional neural networks. In *European conference on computer vision*.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016). 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)*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Sakaridis, C., Dai, D., Hecker, S., & Van Gool, L. (2018). Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *European conference on computer vision (ECCV)*.

Sakaridis, C., Dai, D., & Van Gool, L. (2018). Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, *126*(9), 973–992.

Sankaranarayanan, S., Balaji, Y., Jain, A., Nam Lim, S., & Chellappa, R. (2018). Learning from synthetic data: Addressing domain shift for semantic segmentation. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Shen, X., Zhou, C., Xu, L., & Jia, J. (2015). Mutual-structure for joint filtering. In *The IEEE international conference on computer vision (ICCV)*.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Spinneker, R., Koch, C., Park, S. B., & Yoon, J. J. (2014). Fast fog detection for camera based advanced driver assistance systems. In *International IEEE conference on intelligent transportation systems (ITSC)*.

Tan, R. T. (2008). Visibility in bad weather from a single image. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Tang, K., Yang, J., & Wang, J. (2014). Investigating haze-relevant features in a learning framework for image dehazing. In *IEEE conference on computer vision and pattern recognition*.

Tarel, J. P., Hautière, N., Caraffa, L., Cord, A., Halmaoui, H., & Gruyer, D. (2012). Vision enhancement in homogeneous and heterogeneous fog. *IEEE Intelligent Transportation Systems Magazine*, *4*(2), 6–20.

Tarel, J. P., Hautière, N., Cord, A., Gruyer, D., & Halmaoui, H. (2010) Improved visibility of road scene images under heterogeneous fog. In *IEEE intelligent vehicles symposium* (pp. 478–485).

Tsai, Y. H., Hung, W. C., Schulter, S., Sohn, K., Yang, M. H., & Chandraker, M. (2018). Learning to adapt structured output space for semantic segmentation. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Wang, Y. K., & Fan, C. T. (2014). Single image defogging by multiscale depth fusion. *IEEE Transactions on Image Processing*, *23*(11), 4826–4837.

Wulfmeier, M., Bewley, A., & Posner, I. (2018). Incremental adversarial domain adaptation for continually changing environments. In *International conference on robotics and automation (ICRA)*.

Xu, Y., Wen, J., Fei, L., & Zhang, Z. (2016). Review of video and image defogging algorithms and related studies on image restoration and enhancement. *IEEE Access*, *4*, 165–188.

Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In *International conference on learning representations*.

Zhang, H., Sindagi, V. A., & Patel, V. M. (2017). Joint transmission map estimation and dehazing using deep networks. In *CoRR*. arXiv:1708.00581.

Zhang, Y., David, P., & Gong, B. (2017). Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.