Semantic Foggy Scene Understanding with Synthetic Data

Originally published in:
International Journal of Computer Vision 126(9), https://doi.org/10.1007/s11263-018-1072-8

International Journal of Computer Vision (2018) 126:973–992
https://doi.org/10.1007/s11263-018-1072-8

Semantic Foggy Scene Understanding with Synthetic Data

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

Received: 25 July 2017 / Accepted: 26 February 2018 / Published online: 23 March 2018

Abstract
This work addresses the problem of semantic foggy scene understanding (SFSU). Although extensive research has been performed on image dehazing and on semantic scene understanding with clear-weather images, little attention has been paid to SFSU. Due to the difficulty of collecting and annotating foggy images, we choose to generate synthetic fog on real images that depict clear-weather outdoor scenes, and then leverage these partially synthetic data for SFSU by employing state-of-the-art convolutional neural networks (CNN). In particular, a complete pipeline to add synthetic fog to real, clear-weather images using incomplete depth information is developed. We apply our fog synthesis on the Cityscapes dataset and generate *Foggy Cityscapes* with 20,550 images. SFSU is tackled in two ways: (1) with typical supervised learning, and (2) with a novel type of semi-supervised learning, which combines (1) with an unsupervised supervision transfer from clear-weather images to their synthetic foggy counterparts. In addition, we carefully study the usefulness of image dehazing for SFSU. For evaluation, we present *Foggy Driving*, a dataset with 101 real-world images depicting foggy driving scenes, which come with ground truth annotations for semantic segmentation and object detection. Extensive experiments show that (1) supervised learning with our synthetic data significantly improves the performance of state-of-the-art CNN for SFSU on *Foggy Driving*; (2) our semi-supervised learning strategy further improves performance; and (3) image dehazing marginally advances SFSU with our learning strategy. The datasets, models and code are made publicly available.

Keywords Foggy scene understanding · Semantic segmentation · Object detection · Depth denoising and completion · Dehazing · Transfer learning

1 Introduction

Cameras and the accompanying vision algorithms are widely used for applications such as surveillance (Buch et al. 2011), remote sensing (Dai and Yang 2011), and automated cars (Janai et al. 2017), and their deployment keeps expanding. While these sensors and algorithms are constantly getting better, they are mainly designed to operate on clear-weather images and videos (Narasimhan and Nayar 2002). Yet, outdoor applications can hardly escape from "bad" weather. Thus, such computer vision systems should also function under adverse weather conditions. Here we focus on the presence of fog.

Fog degrades the visibility of a scene significantly (Narasimhan and Nayar 2003; Tan 2008). This causes problems not only to human observers, but also to computer vision algorithms. During the past years, a large body of research has been conducted on image defogging (dehazing) to increase scene visibility (Nishino et al. 2012; He et al. 2011; Wang and Fan 2014). Meanwhile, marked progress has been made in semantic scene understanding with clear-weather images and videos (Ren et al. 2015; Cordts et al. 2016; Yu and Koltun 2016). In contrast, the semantic understanding of foggy scenes has received little attention, despite its importance in outdoor applications. For instance, an automated car still requires a robust detection of road lanes, traffic lights, and other traffic agents in the presence of fog. This work investigates semantic foggy scene understanding (SFSU).

High-level semantic scene understanding is usually tackled by learning from many annotations of real images (Russakovsky et al. 2015; Cordts et al. 2016). Yet, the difficulty of collecting and annotating images for unusual weather conditions such as fog renders this standard protocol problematic.

Communicated by Adrien Gaidon, Florent Perronnin and Antonio Lopez.

✉ Christos Sakaridis
csakarid@vision.ee.ethz.ch

[1] ETH Zürich, Zurich, Switzerland

[2] KU Leuven, Leuven, Belgium

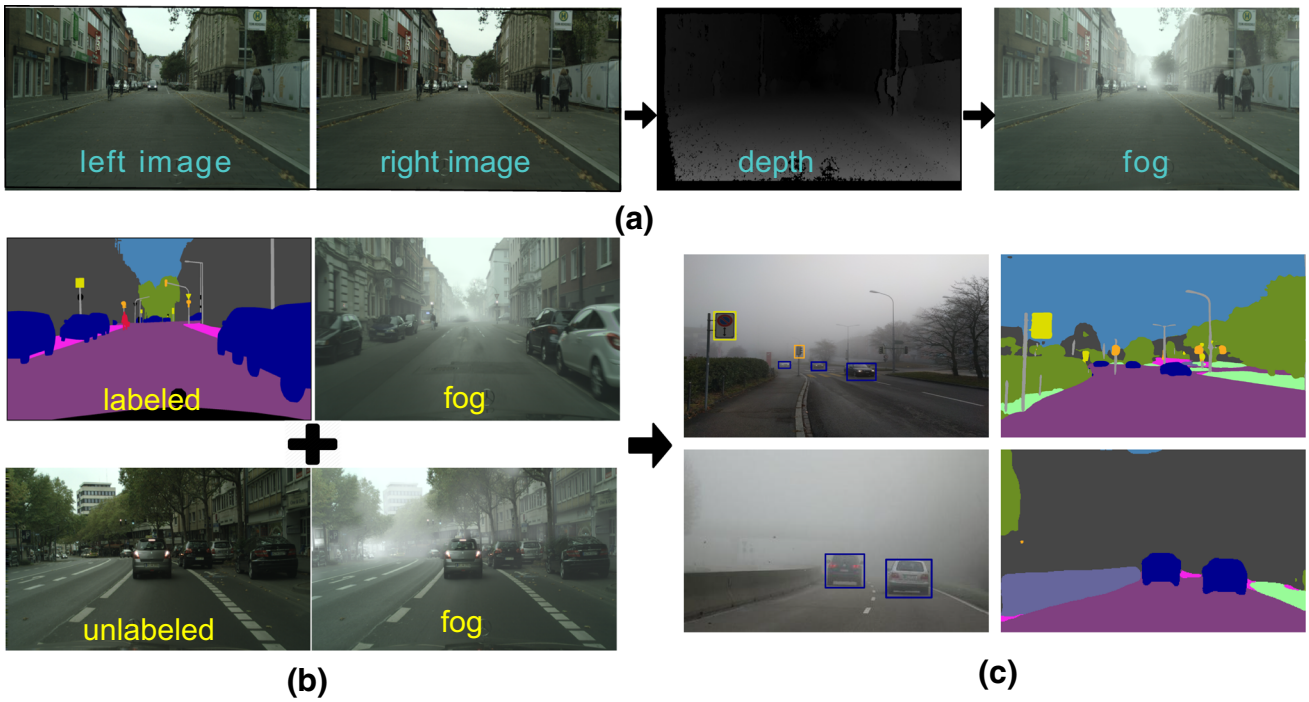

Fig. 1 The pipeline of semantic foggy scene understanding with partially synthetic data: from **a** fog simulation on real outdoor scenes, to **b** training with pairs of such partially synthetic foggy images and semantic annotations as well as pairs of foggy images and clear-weather images, and **c** scene understanding of real foggy scenes. This figure is seen better on a screen

To overcome this problem, we depart from this traditional paradigm and propose another route, also different from moving to fully synthetic scenes. Instead, we choose to generate synthetic fog into real images that contain clear-weather outdoor scenes, and then leverage these partially synthetic foggy images for SFSU.

Given the fact that large-scale annotated data are available for clear-weather images (Everingham et al. 2010; Geiger et al. 2012; Cordts et al. 2016; Russakovsky et al. 2015), we present an automatic pipeline to add synthetic yet highly realistic fog to such datasets. Our fog simulation uses the standard optical model for daytime fog (Koschmieder 1924) (which has already been used extensively in image dehazing) to overlay existing clear-weather images with synthetic fog in a physically sound way, simulating the underlying mechanism of foggy image formation. We leverage our fog simulation pipeline to create our *Foggy Cityscapes* dataset, by adding fog to urban scenes from the Cityscapes dataset (Cordts et al. 2016). This has led to 550 carefully refined high-quality synthetic foggy images with fine semantic annotations inherited directly from Cityscapes, plus an additional 20,000 synthetic foggy images without fine annotations. The resulting "synthetic-fog" images are used to adapt two semantic segmentation models (Yu and Koltun 2016; Lin et al. 2017) and an object detector (Girshick 2015) to foggy scenes. The models are trained in two fashions: (1) by the typical supervised learning scheme, using the 550 high-quality annotated foggy images, and (2) by a novel semi-supervised learning approach, which augments the dataset that is used in (1) with the additional 20,000 foggy images and draws the missing supervision for these images from the predictions of the source, clear-weather model on their clear-weather counterparts. For evaluation purposes, we collect and annotate a new dataset, *Foggy Driving*, with 101 images of driving scenes in the presence of fog. See Fig. 1 for the whole pipeline of our work. In addition, this work studies the utility of three state-of-the-art image dehazing methods for SFSU as well as human understanding of foggy scenes.

The main contributions of the paper are: (1) an automatic and scalable pipeline to impose high-quality synthetic fog on real clear-weather images; (2) two new datasets, one synthetic and one real, to facilitate training and evaluation of models used in SFSU; (3) a new semi-supervised learning approach for SFSU; and (4) a detailed study of the benefit of image dehazing for SFSU and human perception of foggy scenes.

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 is devoted to our fog simulation pipeline, followed by Sect. 4 that introduces our two foggy datasets. Section 5 describes supervised learning with our synthetic foggy data and studies the usefulness of image dehazing for SFSU in this context. Finally, Sect. 6

extends the learning to a semi-supervised paradigm, where supervision is transferred from clear-weather images to their synthetic foggy counterparts, and Sect. 7 concludes the paper.

2 Related Work

Our work is relevant to image defogging (dehazing), depth denoising and completion, foggy scene understanding, synthetic visual data, and transfer learning.

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. This ill-posed problem has been tackled from different perspectives. For instance, in contrast enhancement (Narasimhan and Nayar 2003; Tan 2008) the rationale is that clear-weather images have higher contrast than images degraded by fog. Depth and statistics of natural images are exploited as priors as well (Nishino et al. 2012; Fattal 2008; Berman et al. 2016; Fattal 2014). Another line of work is based on the dark channel prior (He et al. 2011), with the empirically validated assumption that pixels of clear-weather images are very likely to have low values in some of the three color channels. Certain works focus particularly on enhancing foggy road scenes (Tarel et al. 2012; Negru et al. 2015). Methods have also been developed for nighttime (Li et al. 2015), given its importance in outdoor applications. Fast dehazing approaches have been developed in Tarel and Hautière (2009) and Wang et al. (2017) towards real-time applications. Recent approaches also rely on trainable architectures (Tang et al. 2014), which have evolved to end-to-end models (Ren et al. 2016; 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) and Li et al. (2016). All these approaches can greatly increase visibility. Our work is complementary and focuses on the semantic understanding of foggy scenes.

2.2 Depth Denoising and Completion

Synthesizing a foggy image from its real, clear counterpart generally requires an accurate depth map. In previous works, the colorization approach of Levin et al. (2004) has been used to inpaint depth maps of the *indoor* NYU Depth dataset (Silberman et al. 2012). Such inpainted depth maps have been used in state-of-the-art dehazing approaches such as Ren et al. (2016) to generate training data in the form of synthetic indoor foggy images. In contrast, our work considers real *outdoor* urban scenes from the Cityscapes dataset (Cordts et al. 2016), which contains significantly more complex depth configurations than NYU Depth. Furthermore, the available depth information in Cityscapes is not provided by a depth sensor, but it is rather an estimate of the depth resulting from the application of a semiglobal matching stereo algorithm based on Hirschmüller (2008). This depth estimate usually contains a large amount of severe artifacts and large holes (cf. Fig. 1), which render it inappropriate for direct use in fog simulation. There are several recent approaches that handle highly noisy and incomplete depth maps, including stereoscopic inpainting (Wang et al. 2008), spatio-temporal hole filling (Camplani and Salgado 2012) and layer depth denoising and completion (Shen and Cheung 2013). Our method builds on the framework of stereoscopic inpainting (Wang et al. 2008) which performs depth completion at the level of superpixels, and introduces a novel, theoretically grounded objective for the superpixel-matching optimization that is involved.

2.3 Foggy Scene Understanding

Semantic understanding of outdoor scenes is a crucial enabler for applications such as assisted or autonomous driving. Typical examples 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 et al. 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 et al. 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: (1) our semantic understanding task is more complex, with 19 semantic classes that are commonly involved in driving scenarios, 8 of which occur as distinct objects; (2) we tackle the problem with modern deep CNN for semantic segmentation (Yu and Koltun 2016; Lin et al. 2017) and object detection (Girshick 2015), taking full advantage of the most recent advances in this field; and (3) we compile and release a large-scale dataset of synthetic foggy images based on real scenes plus a dataset of real-world foggy scenes, featuring both dense pixel-level semantic annotations and annotations for object detection.

2.4 Synthetic Visual Data

The leap of computer vision in recent years can to an important extent be attributed to the availability of large, labeled datasets (Everingham et al. 2010; Russakovsky et al. 2015; Cordts et al. 2016). However, acquiring and annotating such a dataset for each new problem is not (yet) doable. Thus, learning with synthetic data is gaining attention. We give some notable examples. Dosovitskiy et al. (2015) use the renderings of a floating chair to train dense optical flow regression networks. Gupta et al. (2016a) impose text onto natural images to learn an end-to-end text detection system. Vázquez et al. (2014) train pedestrian detectors with virtual data. In Ros et al. (2016) and Richter et al. (2016) the authors leverage video game engines to render images along with dense semantic annotations that are subsequently used in combination with real data to improve the semantic segmentation performance of modern CNN architectures on real scenes. Going one step further, Johnson-Roberson et al. (2017) shows that for the task of vehicle detection, training a CNN model *only* on massive amounts of synthetic images can outperform the same model trained on large-scale real datasets like Cityscapes. By contrast, our work tackles semantic segmentation and object detection for real *foggy* urban scenes, by adding synthetic fog to *real* images taken under clear weather. Hence, our approach is based on only partially synthetic data. In the same vein, Abu Alhaija et al. (2017) is based on real urban scenes, augmented with virtual cars. A very interesting project is "FOG" (Colomb et al. 2008). Its team developed a prototype of a small-scale fog chamber, able to produce stable visibility levels and homogeneous fog to test the reaction of drivers.

2.5 Transfer Learning

Our work bears resemblance to works from the broad field of transfer learning. Model adaptation across weather conditions to semantically segment simple road scenes is studied in Levinkov and Fritz (2013). More recently, a domain adversarial based approach was proposed to adapt semantic segmentation models both at pixel level and feature level from simulated to real environments (Hoffman et al. 2017). Our work generates synthetic fog from clear-weather data to close the domain gap. Combining our method and the aforementioned transfer learning methods is a promising direction for future work. The supervision transfer from clear weather to foggy weather in this paper is inspired by the stream of work on model distillation/imitation (Hinton et al. 2015; Gupta et al. 2016b; Dai et al. 2015). Our approach is similar in that knowledge is transferred from one domain (model) to another by using paired data samples as a bridge.

3 Fog Simulation on Real Outdoor Scenes

To simulate fog on input images that depict real scenes with clear weather, the standard approach is to model the effect of fog as a function that maps the radiance of the clear scene to the radiance observed at the camera sensor. Critically, this space-variant function is usually parameterized by the distance ℓ of the scene from the camera, which equals the length of the path along which light has traveled and is closely related to scene depth. As a result, the pair of the clear image and its depth map forms the basis of our foggy image synthesis. In this section, we first detail the optical model which we use for fog and then present our complete pipeline for fog simulation, with emphasis on our denoising and completion of the input depth. Finally, we present some criteria for selecting suitable images to generate high-quality synthetic fog.

3.1 Optical Model of Choice for Fog

In the image dehazing literature, various optical models have been used to model the effect of haze on the appearance of a scene. For instance, optical models tailored for nighttime haze removal have been proposed in Zhang et al. (2014) and Li et al. (2015), taking into account the space-variant lighting that characterizes most nighttime scenes. This variety of models is directly applicable to the case of fog as well, since the physical process for image formation in the presence of either haze or fog is essentially similar. For our synthesis of foggy images, we consider the standard optical model of Koschmieder (1924), which is used extensively in the literature (He et al. 2011; Fattal 2008; Tang et al. 2014; Tarel and Hautière 2009; Ren et al. 2016) and is formulated as

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

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. This model assumes the atmospheric light to be globally constant, which is generally valid only for *daytime* images. The transmission $t(\mathbf{x})$ determines the amount of scene radiance that reaches the camera. In case of a *homogeneous* medium, transmission depends on the distance $\ell(\mathbf{x})$ of the scene from the camera through

$$t(\mathbf{x}) = \exp\left(-\beta\ell(\mathbf{x})\right). \tag{2}$$

The parameter β is named attenuation coefficient and it effectively controls the thickness of the fog: larger values of β mean thicker fog. The meteorological optical range (MOR), also known as visibility, is defined as the maximum distance from the camera for which $t(\mathbf{x}) \geq 0.05$, which implies that if (2) is valid, then MOR $= 2.996/\beta$. Fog decreases the

MOR to less than 1 km by definition (Federal Meteorological Handbook No. 1: Surface Weather Observations and Reports 2005). Therefore, the attenuation coefficient in homogeneous fog is by definition

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

where the lower bound corresponds to the lightest fog configuration. In our fog simulation, the value that is used for β always obeys (3).

Model (1) provides a powerful basis for simulating fog on outdoor scenes with clear weather. Even though its assumption of homogeneous atmosphere is strong, it generates synthetic foggy images that can act as good proxies for real world foggy images where this assumption might not hold exactly, as long as it is provided with an *accurate* transmission map t. Straightforward extensions of (1) are used in Tarel et al. (2012) to simulate heterogeneous fog on synthetic scenes.

To sum up, the necessary inputs for fog simulation using (1) are a color image $\mathbf{R}$ of the original clear scene, atmospheric light $\mathbf{L}$ and a dense transmission map t defined at each pixel of $\mathbf{R}$. Our task is thus twofold:

1. estimation of t, and
2. estimation of $\mathbf{L}$ from $\mathbf{R}$.

Step 2 is simple: we use the method proposed in He et al. (2011) with the improvement of Tang et al. (2014). In the following, we focus on step 1 for the case of outdoor scenes with a noisy, incomplete estimate of depth serving as input.

3.2 Depth Denoising and Completion for Outdoor Scenes

The inputs that our method requires for generating an accurate transmission map t are:

- the original, clear-weather color image $\mathbf{R}$ to add synthetic fog on, which constitutes the *left* image of a stereo pair,
- the *right* image $\mathbf{Q}$ of the stereo pair,
- the intrinsic calibration parameters of the two cameras of the stereo pair as well as the length of the baseline,
- a dense, raw disparity estimate D for $\mathbf{R}$ of the same resolution as $\mathbf{R}$, and
- a set M comprising the pixels where the value of D is missing.

These requirements can be easily fulfilled with a stereo camera and a standard stereo matching algorithm (Hirschmüller 2008).

The main steps of our pipeline are the following:

1. calculation of a raw depth map d in meters,
2. *denoising and completion* of d to produce a refined depth map d' in meters,
3. calculation of a scene distance map ℓ in meters from d',
4. application of (2) to obtain an initial transmission map $\hat{t}$, and
5. guided filtering (He et al. 2013) of $\hat{t}$ using $\mathbf{R}$ as guidance to compute the final transmission map t.

The central idea is to leverage the accurate structure that is present in the color images of the stereo pair in order to improve the quality of depth, before using the latter as input for computing transmission. We now proceed in explaining each step in detail, except step 4 which is straightforward. In step 1, we use the input disparity D in combination with the values of the focal length and the baseline to obtain d. The missing values for D, indicated by M, are also missing in d.

Step 2 follows a segmentation-based depth filling approach, which builds on the stereoscopic inpainting method presented in Wang et al. (2008). More specifically, we use a superpixel segmentation of the clear image $\mathbf{R}$ to guide depth denoising and completion at the level of superpixels, making the assumption that each individual superpixel corresponds roughly to a plane in the 3D scene.

First, we apply a photo-consistency check between $\mathbf{R}$ and $\mathbf{Q}$, using the input disparity D to establish pixel correspondences between the two images of the stereo pair, similar to Eq. (12) in Wang et al. (2008). All pixels in $\mathbf{R}$ for which the color deviation (measured as difference in the RGB color space) from the corresponding pixel in $\mathbf{Q}$ has greater magnitude than $\epsilon = 12/255$ are deemed invalid regarding depth and hence are added to M.

We then segment $\mathbf{R}$ into superpixels with SLIC (Achanta et al. 2012), denoting the target number of superpixels as $\hat{K}$ and the relevant range domain scale parameter as $m = 10$. For depth denoising and completion on Cityscapes, we use $\hat{K} = 2048$. The final number of superpixels that are output by SLIC is denoted by K. These superpixels are classified into reliable and unreliable ones with respect to depth information, based on the number of pixels with missing or invalid depth that they contain. More formally, we use the criterion of Eq. (2) in Wang et al. (2008), which states that a superpixel T is reliable if and only if

$$\mathrm{card}(T \setminus M) \geq \max\{P,\ \lambda\, \mathrm{card}(T)\}, \tag{4}$$

setting $P = 20$ and $\lambda = 0.6$.

For each superpixel that fulfills (4), we fit a depth plane by running RANSAC on its pixels that have a valid value for depth. We use an adaptive inlier threshold to account for differences in the range of depth values between distinct superpixels. For a superpixel T, the inlier threshold is set as

$$\theta = 0.01 \underset{\mathbf{x} \in T \setminus M}{\text{median}} \{d(\mathbf{x})\}. \tag{5}$$

We use adaptive RANSAC and set the maximum number of iterations to 2000 and the bound on the probability of having obtained a pure inlier sample to $p = 0.99$.

The greedy approach of Wang et al. (2008) is used subsequently to match unreliable superpixels to reliable ones pairwise and assign the fitted depth planes of the latter to the former. Different than Wang et al. (2008), we propose a novel objective function for matching pairs of superpixels. For a superpixel pair (s, t), our proposed objective is formulated as

$$E(s,t) = \|\mathbf{C}_s - \mathbf{C}_t\|^2 + \alpha \|\mathbf{x}_s - \mathbf{x}_t\|^2. \tag{6}$$

The first term measures the proximity of the two superpixels in the range domain, where we denote the average CIELAB color of superpixel s with $\mathbf{C}_s$. In other words, we penalize the squared Euclidean distance between the average colors of the superpixels in the CIELAB color space, which has been designed to increase perceptual uniformity (Comaniciu and Meer 2002). On the contrary, the objective of Wang et al. (2008) uses the cosine similarity of average superpixel colors to form the range domain cost:

$$1 - \frac{\mathbf{C}_s}{\|\mathbf{C}_s\|} \cdot \frac{\mathbf{C}_t}{\|\mathbf{C}_t\|}. \tag{7}$$

The disadvantage of (7) is that it assigns zero matching cost to dissimilar colors in certain cases. For instance, in the RGB color space, the pair of colors (δ, δ, δ) and $(1-\delta, 1-\delta, 1-\delta)$, where δ is a small positive constant, is assigned zero penalty, even though the former color is very dark gray and the latter is very light gray.

The second term on the right-hand side of (6) measures the proximity of the two superpixels in the spatial domain as the squared Euclidean distance between their centroids $\mathbf{x}_s$ and $\mathbf{x}_t$. By contrast, the spatial proximity term of Wang et al. (2008) assigns zero cost to pairs of adjacent superpixels and unit cost to non-adjacent pairs. This implies that close yet non-adjacent superpixels are penalized equally to very distant superpixels by Wang et al. (2008). As a result, a certain superpixel s can be erroneously matched to a very distant superpixel t which is highly unlikely to share the same depth plane as s, as long as the range domain term for this pair is minimal and all superpixels adjacent to s are dissimilar to it with respect to appearance. Our proposed spatial cost handles these cases successfully: t is assigned a very large spatial cost for being matched to s, and other superpixels that have less similar appearance yet smaller distance to s are preferred.

In (6), $\alpha > 0$ is a parameter that weights the relative importance of the spatial domain term versus the range domain term. Similarly to Achanta et al. (2012), we set $\alpha = m^2/S^2$, where $S = \sqrt{N/K}$, N denotes the total number of pixels in the image, and $m = 10$ and K are the same as for SLIC. Our matching objective (6) is similar to the distance that is defined in SLIC (Achanta et al. 2012) and other superpixel segmentation methods for assigning *an individual pixel to a superpixel*. In our case though, this distance is rather used to measure similarity between *pairs of superpixels*.

After all superpixels have been assigned a depth plane, we use these planes to complete the missing depth values for pixels belonging to M. In addition, we replace the depth values of pixels which do not belong to M but constitute large-margin outliers with respect to their corresponding plane (deviation larger than $\hat{\theta} = 50$ m) with the values imputed by the plane. This results in a complete, denoised depth map d', and concludes step 2.

In step 3, we compute the distance $\ell(\mathbf{x})$ of the scene from the camera at each pixel $\mathbf{x}$ based on $d'(\mathbf{x})$, using the coordinates of the principal point plus the focal length of the camera.

Finally, in step 5 we post-process the initial transmission map $\hat{t}$ with guided filtering (He et al. 2013), in order to smooth transmission while respecting the boundaries of the clear image $\mathbf{R}$. We fix the radius of the guided filter window to $r = 20$ and the regularization parameter to $\mu = 10^{-3}$, i.e. we use the same values as in the haze removal experiments of He et al. (2013).

Results of the presented pipeline for fog simulation on example images from Cityscapes are provided in Fig. 2 for $\beta = 0.01$, which corresponds to visibility of ca. 300 m. We compare our fog simulation to an alternative implementation, which employs nearest-neighbor interpolation to complete the missing values of the depth map before computing the transmission and does not involve guided filtering as a post-processing step.

3.3 Input Selection for High-Quality Fog Simulation

Applying the presented pipeline to simulate fog on large datasets with real outdoor scenes such as Cityscapes with the aim of producing synthetic foggy images of high quality calls for careful refinement of the input.

To be more precise, the sky is clear in the majority of scenes in Cityscapes, with intense direct or indirect sunlight, as shown in Fig. 3a. These images usually contain sharp shadows and have high contrast compared to images that depict foggy scenes. This causes our fog simulation to generate synthetic images which do not resemble real fog very well, e.g. Fig. 3b. Therefore, our first refinement criterion is whether the sky is overcast, ensuring that the light in the input real scene is not strongly directional.

Secondly, we observe that atmospheric light estimation in step 2 of our fog simulation sometimes fails to select a

Fig. 2 Comparison of our fog simulation to nearest-neighbor interpolation for depth completion on images from Cityscapes. This figure is better seen on a screen and zoomed in. **a** Input from Cityscapes. **b** Nearest-neighbor depth completion. **c** Our fog simulation

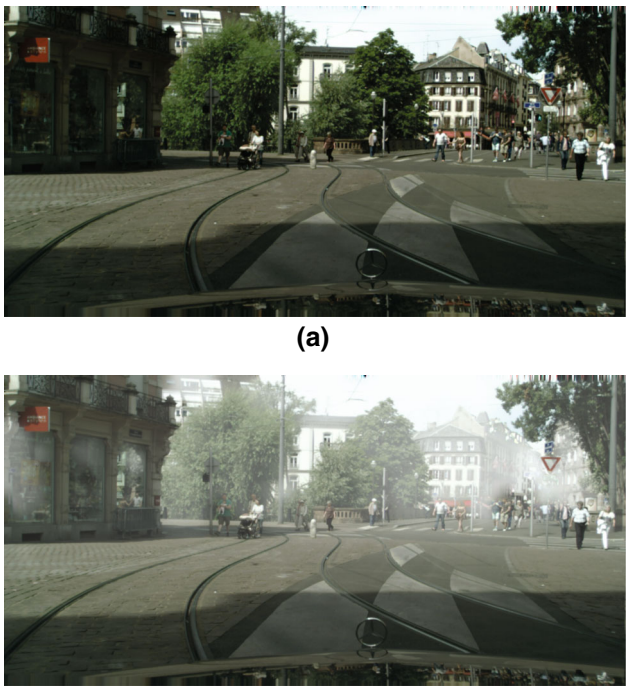

Fig. 3 Sunny scene from Cityscapes and the result of our fog simulation. **a** Input image from Cityscapes. **b** Output of our fog simulation

pixel with ground truth semantic label *sky* as the representative of the value of atmospheric light. In rare cases, it even happens that the sky is not visible at all in an image. This results in an erroneous, physically invalid value of atmospheric light being used in (1) to synthesize the foggy image. Consequently, our second refinement criterion is whether the pixel that is selected as atmospheric light is labeled as *sky*, and affords an automatic implementation.

4 Foggy Datasets

We present two distinct datasets for semantic understanding of foggy scenes: *Foggy Cityscapes* and *Foggy Driving*. The former derives from the Cityscapes dataset (Cordts et al. 2016) and constitutes a collection of synthetic foggy images generated with our proposed fog simulation that automatically inherit the semantic annotations of their real, clear counterparts. On the other hand, *Foggy Driving* is a collection of 101 real-world foggy road scenes with annotations for semantic segmentation and object detection, used as a benchmark for the domain of foggy weather.

4.1 Foggy Cityscapes

We apply the fog simulation pipeline that is presented in Sect. 3 to the complete set of images provided in the Cityscapes dataset. More specifically, we first obtain 20,000 synthetic foggy images from the larger, coarsely annotated part of the dataset, and keep all of them, without applying the refinement criteria of Sect. 3.3. In this way, we trade the high visual quality of the synthetic images for a very large scale and variability of the synthetic dataset. We do not make use of the original coarse annotations of these images for semantic segmentation; rather, we produce labellings with state-of-the-art semantic segmentation models on the original, clear images and use them to transfer knowledge from clear weather to foggy weather, as will be discussed in Sect. 6. We name this set *Foggy Cityscapes-coarse*.

In addition, we use the two criteria of Sect. 3.3 in conjunction to filter the finely annotated part of Cityscapes that originally comprises 2975 training and 500 validation images, and obtain a refined set of 550 images, 498 from the training set and 52 from the validation set, which fulfill

Fig. 4 Different versions of an exemplar scene from *Foggy Cityscapes* for varying visibility. **a** clear-weather. **b** $\beta = 0.005$. **c** $\beta = 0.01$. **d** $\beta = 0.02$

both criteria. Running our fog simulation on this refined set provides us with a moderate-scale collection of high-quality synthetic foggy images. This collection automatically inherits the original fine annotations for *semantic segmentation*, as well as bounding box annotations for *object detection* which we generate by leveraging the instance-level semantic annotations that are provided in Cityscapes for the 8 classes *person*, *rider*, *car*, *truck*, *bus*, *train*, *motorcycle* and *bicycle*. We term this collection *Foggy Cityscapes-refined*.

Since MOR can vary significantly in reality for different instances of fog, we generate five distinct versions of *Foggy Cityscapes*, each of which is characterized by a constant simulated attenuation coefficient β in (2), hence a constant MOR. In particular, we use $\beta \in \{0.005, 0.01, 0.02, 0.03, 0.06\}$, which correspond approximately to MOR of 600, 300, 150, 100 and 50 m respectively. Figure 4 shows three of the five synthesized foggy versions of a clear scene in *Foggy Cityscapes*.

4.2 Foggy Driving

Foggy Driving consists of 101 color images depicting real-world foggy driving scenes. We captured 51 of these images with a cell phone camera in foggy conditions at various areas of Zurich, and the rest 50 images were carefully collected from the web. We note that all images have been preprocessed so that they have a maximum resolution of 960×1280 pixels.

We provide dense, pixel-level semantic annotations for all images of *Foggy Driving*. In particular, we use the 19 evaluation classes of Cityscapes: *road*, *sidewalk*, *building*, *wall*, *fence*, *pole*, *traffic light*, *traffic sign*, *vegetation*, *terrain*, *sky*, *person*, *rider*, *car*, *truck*, *bus*, *train*, *motorcycle* and *bicycle*. Pixels that do not belong to any of the above classes or are not labeled are assigned the *void* label, and they are ignored for semantic segmentation evaluation. At annotation time, we label individual instances of *person*, *rider*, *car*, *truck*, *bus*, *train*, *motorcycle* and *bicycle* separately following the Cityscapes annotation protocol, which directly affords bounding box annotations for these 8 classes.

In total, 33 images have been finely annotated (cf. the last three rows of Fig. 13) in the aforementioned procedure, and the rest 68 images have been coarsely annotated (cf. the top three rows of Fig. 13). We provide per-class statistics for the pixel-level semantic annotations of *Foggy Driving* in Fig. 5. Furthermore, statistics for the number of objects in the bounding box annotations are shown in Fig. 6. Because of the coarse annotation that is created for one part of *Foggy Driving*, we do not use this part in evaluation of object detection approaches, as difficult objects that are not included in the annotations may be detected by a good method and missed by a comparatively worse method, resulting in incorrect comparisons with respect to precision. On the contrary, the coarsely annotated images are used without such issues in evaluation of semantic segmentation approaches, since predictions at unlabeled pixels are simply ignored and thus do not affect the measured performance.

Foggy Driving may have a smaller size than other recent datasets for semantic scene understanding, however, it features challenging foggy scenes with comparatively high complexity. As Table 1 shows, the subset of 33 images with fine annotations is roughly on par with Cityscapes regarding the average number of humans and vehicles per image. In total, *Foggy Driving* contains more than 500 vehicles and almost 300 humans. We also underline the fact that

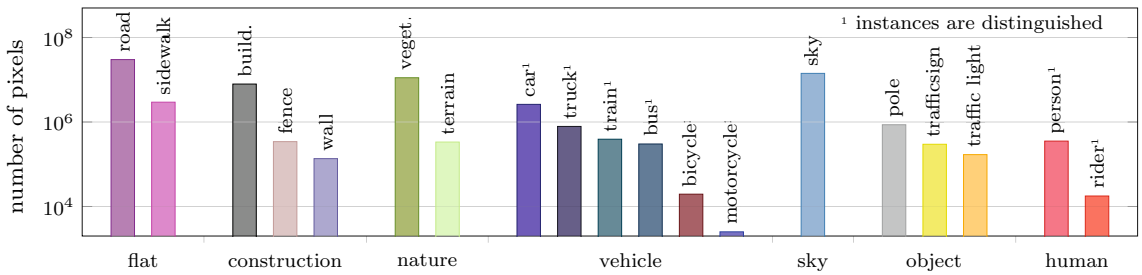

Fig. 5 Number of annotated pixels per class for *Foggy Driving*

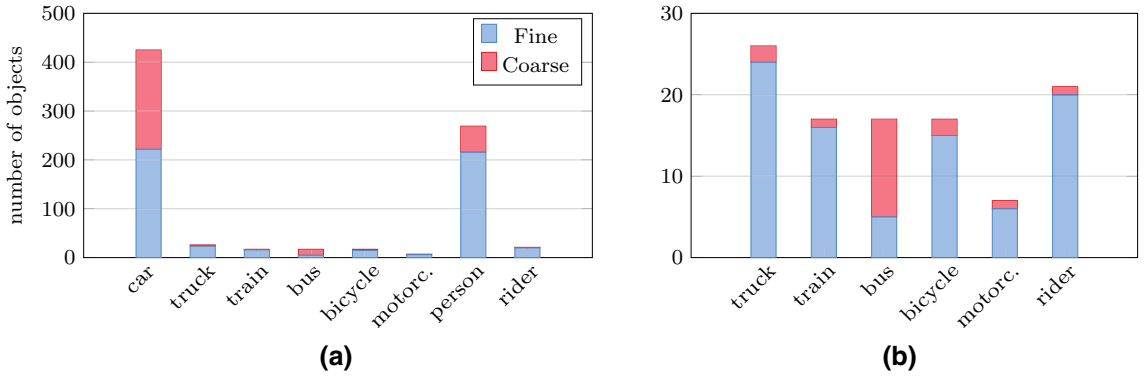

Fig. 6 Number of objects per class in *Foggy Driving*. **a** includes statistics for the complete set of eight classes for which instances are distinguished, whereas **b** presents a zoomed version of (**a**) for six of these classes

Table 1 Absolute and average number of annotated pixels, humans and vehicles for *Foggy Driving* ("Ours"), KITTI and Cityscapes

	Pixels	Humans	Vehicles	h/im	v/im
Ours (fine)	38.3*M*	236	288	**7.2**	8.7
Ours (coarse)	34.6*M*	54	221	0.8	3.3
KITTI	0.23*G*	6.1k	30.3k	0.8	4.1
Cityscapes	**9.43G**	**24.0k**	**41.0k**	7.0	**11.8**

"h/im" stands for humans per image and "v/im" for vehicles per image. Only the training and validation sets of KITTI and Cityscapes are considered
Maximum entries are given in bold

Table 1 compares *Foggy Driving*—a dataset used purely for testing—against the unions of training and validation sets of KITTI (Geiger et al. 2012) and Cityscapes, which are much larger than their respective testing sets that would provide a better comparison.

As a final note, we identify the subset of the 19 annotated classes that occur frequently in *Foggy Driving*. These "frequent" classes either have a larger number of total annotated pixels, e.g. *road*, or a larger number of total annotated polygons or instances, e.g. *pole* and *person*, compared to the rest of the classes. They are: *road*, *sidewalk*, *building*, *pole*, *traffic light*, *traffic sign*, *vegetation*, *sky*, *person*, and *car*. In the experiments that follow in Sect. 5.1, we occasionally use this set of frequent semantic classes as an alternative to the complete set of semantic classes for averaging per-class scores, in order to further verify results based only on classes with plenty of examples.

5 Supervised Learning with Synthetic Fog

We first show that our synthetic *Foggy Cityscapes-refined* dataset can be used per se for successfully adapting modern CNN models to the condition of fog with the usual supervised learning paradigm. Our experiments focus primarily on the task of semantic segmentation and additionally include comparisons on the task of object detection, evidencing clearly the usefulness of our synthetic foggy data in understanding the semantics of *real* foggy scenes such as those in *Foggy Driving*.

More specifically, the general outline of our main experiments can be summarized in two steps:

1. fine-tuning a model that has been trained on the original Cityscapes dataset for clear weather by using only synthetic images of *Foggy Cityscapes-refined*, and
2. evaluating the fine-tuned model on *Foggy Driving* and showing that its performance is improved compared to the original, clear-weather model. Thus, the reported results pertain to *Foggy Driving* unless otherwise mentioned.

In other words, all models are ultimately evaluated on data from a different domain than that of the data on which they have been fitted, revealing their true generalization potential on previously unseen foggy scenes.

We also consider dehazing as an optional preprocessing step before feeding the input images to semantic segmentation models for training and testing, and examine the effect of this dehazing preprocessing on the performance of such a model using state-of-the-art dehazing methods. The effect of dehazing on semantic segmentation performance is additionally correlated with its utility for human understanding of foggy scenes by conducting a user study on Amazon Mechanical Turk.

5.1 Semantic Segmentation

Our model of choice for conducting experiments on semantic segmentation with the supervised pipeline is the modern dilated convolutions network (DCN) (Yu and Koltun 2016). In particular, we make use of the publicly available *Dila-*

tion10 model, which has been trained on the 2975 images of the training set of Cityscapes. We wish to note that this model was originally trained and tested on 1396×1396 image crops by Yu and Koltun (2016), but due to GPU memory limitations we train it on 756×756 crops and test it on 700×700 crops. Still, *Dilation10* enjoys a fair mean intersection over union (IoU) score of 34.9% on *Foggy Driving*.

In the following experiments of Sect. 5.1, we fine-tune *Dilation10* on the training set of *Foggy Cityscapes-refined* which consists of 498 images, and reserve the 52 images of the respective validation set for additional evaluation. In particular, we fine-tune all layers of the original model for 3k iterations (ca. 6 epochs) using mini-batches of size 1. Unless otherwise mentioned, the attenuation coefficient β used in *Foggy Cityscapes* is equal to 0.01.

Overall, we consider four different options with respect to dehazing preprocessing: applying no dehazing at all, dehazing with multi-scale convolutional neural networks (MSCNN) (Ren et al. 2016), dehazing using the dark channel prior (DCP) (He et al. 2011), and non-local image dehazing (Berman et al. 2016). Unless otherwise specified, no dehazing is applied. Our experimental protocol is consistent with respect to dehazing preprocessing: the same option for dehazing preprocessing is used both at training time and test time. More specifically, at training time we first process the synthetic foggy images of *Foggy Cityscapes-refined* according to the specified option for dehazing preprocessing and then use the processed images as input for fine-tuning *Dilation10*. At evaluation time, we process the images in *Foggy Driving* with the same dehazing preprocessing that was used at training time (if any was), and use the processed images to test the fine-tuned model.

Benefit of Fine-tuning on Synthetic Fog Our first experiment evidences the benefit of fine-tuning on *Foggy Cityscapes-refined* for improving semantic segmentation performance on *Foggy Driving*. Table 2 presents comparative performance of the original *Dilation10* model against its fine-tuned counterparts in terms of mean IoU over all annotated classes in *Foggy Driving* as well as over frequent classes only. All four options regarding dehazing preprocessing are considered. Note that we also evaluate the original *Dilation10* model for all dehazing preprocessing alternatives (only relevant at test time in this case) in the first row of each part of Table 2. Indeed, all fine-tuned models outperform *Dilation10* irrespective of the type of dehazing preprocessing that is applied, both for mean IoU over all classes and over frequent classes only. The best-performing fine-tuned model, which we refer to as *FT-0.01*, involves no dehazing and outperforms *Dilation10* significantly, i.e. by 3% for mean IoU over all classes and 5% for mean IoU over frequent classes. Note additionally that *FT-0.01* has been fine-tuned on only 498 training images of *Foggy Cityscapes-refined*, compared to the 2975 training images of Cityscapes for *Dilation10*.

Table 2 Performance comparison on *Foggy Driving* of *Dilation10* versus fine-tuned versions of it using *Foggy Cityscapes-refined*, for four options regarding dehazing preprocessing

	No dehazing	MSCNN	DCP	Non-local
Mean IoU over *all* classes (%)				
W/o FT	34.9	34.7	29.9	29.3
FT	**37.8**	**37.1**	**37.4**	**36.6**
Mean IoU over *frequent* classes in *Foggy Driving* (%)				
W/o FT	52.4	52.4	45.5	46.2
FT	**57.4**	**56.2**	**56.7**	**55.1**

"FT" stands for using fine-tuning and "W/o FT" for not using fine-tuning
Best results are given in bold

Table 3 Performance comparison on *Foggy Driving* of various fine-tuned versions of *Dilation10* that correspond to different fog simulation methods for generating the training dataset *Foggy Cityscapes-refined* that is used for fine-tuning, and different learning rate policies during fine-tuning

	Constant l.r.	"Poly" l.r.
Nearest neighbor	32.9	36.2
Ours w/o guided filtering	33.0	36.8
Ours	**34.4**	**37.8**

Mean IoU (%) over *all* classes is used to report results
Best results are given in bold

Comparison of Fog Simulation Approaches Next, we compare in Table 3 the utility of our proposed fog simulation method for generating useful synthetic training data in terms of semantic segmentation performance on *Foggy Driving*, against two alternative approaches: the baseline that we considered in Fig. 2 and a truncated version of our method, where we omit the guided filtering step. We consider two different policies for the learning rate when fine-tuning on *Foggy Cityscapes-refined*: a constant learning rate of 10^{-5} and a polynomially decaying learning rate, commonly referred to as "poly" (Chen et al. 2018), with a base learning rate of 10^{-5} and a power parameter of 0.9. Our method for fog simulation consistently outperforms the two baselines and the "poly" learning rate policy allows the model to be fine-tuned more effectively than the constant policy. In all other experiments with DCN, we use the "poly" learning rate policy with the parameters specified above for fine-tuning.

Increasing Returns at Larger Distance As can easily be deduced from (2), fog has a growing effect on the appearance of the scene as distance from the camera increases. Ideally, a model that is dedicated to foggy scenes must deliver a greater benefit for distant parts of the scene. In order to examine this aspect of semantic segmentation of foggy scenes, we use the

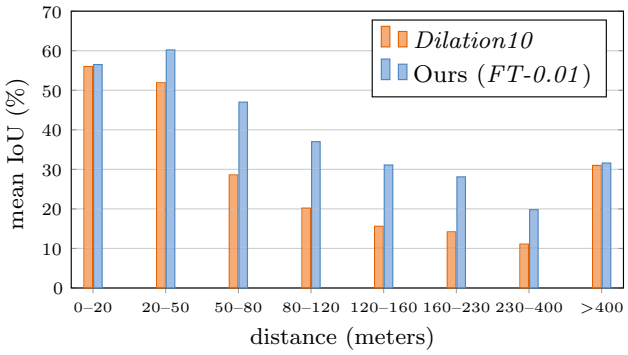

Fig. 7 Performance of semantic segmentation models on *Foggy Cityscapes-refined* at distinct ranges of scene distance from the camera

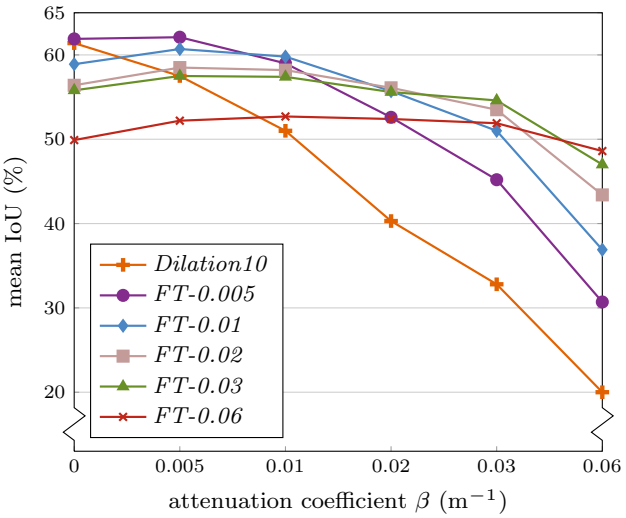

Fig. 8 Performance of semantic segmentation models on various versions of the validation set of *Foggy Cityscapes-refined* corresponding to different values of attenuation coefficient β

completed, dense distance maps of Cityscapes images that have been computed as an intermediate output of our fog simulation, given that *Foggy Driving* does not include depth information. In more detail, we consider the validation set of *Foggy Cityscapes-refined*, the images of which are unseen both for *Dilation10* and our fine-tuned models, and bin the pixels according to their value in the corresponding distance map. Each distance range is considered separately for evaluation by ignoring all pixels that do not belong to it. In Fig. 7, we compare mean IoU of *Dilation10* and *FT-0.01* individually for each distance range. *FT-0.01* brings a consistent gain in performance across all distance ranges. What is more, this gain is larger in both absolute and relative terms for pixels that are more than 50 m away from the camera, implying that our model is able to handle better the most challenging parts of a foggy scene. Note that most pixels in the very last distance range (more than 400 m away from the camera) belong to the *sky* class and their appearance does not change much between the clear and the synthetic foggy images.

Generalization in Synthetic Fog across Densities In order to verify the ability of a model that has been fine-tuned on *Foggy Cityscapes-refined* for a fixed value $\beta^{(t)}$ of the attenuation coefficient, hence fixed fog density, to generalize well to new, unseen fog densities, we evaluate the model on multiple versions of the validation set of *Foggy Cityscapes-refined*, each rendered using a different value for β which is in general not equal to $\beta^{(t)}$. In particular, we use the five different versions of *Foggy Cityscapes-refined* as described in Sect. 4.1 and obtain five models by fine-tuning *Dilation10* on the training set of each version. In congruence with notation in previous experiments, we denote such a fine-tuned model by *FT-*$\beta^{(t)}$, e.g. *FT-0.02*. Afterwards, we evaluate each of these models plus *Dilation10* on the validation set of each of the five foggy versions plus the original, clear-weather version where $\beta = 0$. The mean IoU performance of the six models is presented in Fig. 8. Whereas the performance of *Dilation10* drops rapidly as β increases, all five fine-tuned "foggy" models are more robust to changes in β across the examined range. Analyzing the performance of each fine-tuned model individually, we observe that performance is high and fairly stable in the range $[0, \beta^{(t)}]$ and drops for $\beta > \beta^{(t)}$. This implies that a "foggy" model is able to generalize well to lighter synthetic fog than what was used to fine-tune it. Moreover, all "foggy" models compare favorably to *Dilation10* across the largest part of the range of β, with most "foggy" models being beaten by *Dilation10* only for clear weather. Note also that the performance gain with "foggy" models under foggy conditions is much larger than the corresponding performance loss for clear weather.

Effect of Synthetic Fog Density on Real-world Performance Our final experiment on semantic segmentation serves two purposes: to examine the effect of varying the fog density of the synthetic training data as well as that of dehazing preprocessing on the performance of the fine-tuned model on real foggy data. To this end, we use three of the versions of *Foggy Cityscapes-refined* corresponding to the values $\{0.005,\ 0.01,\ 0.02\}$ for β and consider all four options regarding dehazing preprocessing for fine-tuning *Dilation10*. The performance of the 12 resulting fine-tuned models on *Foggy Driving* in terms of mean IoU over all annotated classes as well as over frequent classes only is reported in Table 4. We first discuss the effect of varying fog density for each dehazing option individually and defer a general comparison of the various dehazing preprocessing options to the next paragraph.

The two conditions that must be met in order for the examined models to achieve better performance are:

1. a good matching of the distributions of the synthetic training data and the real, testing data, and

Table 4 Performance comparison on *Foggy Driving* of fine-tuned versions of *Dilation10* using *Foggy Cityscapes-refined*, for three different values of attenuation coefficient β in fog simulation and four options regarding dehazing preprocessing

	$\beta = 0.005$	$\beta = 0.01$	$\beta = 0.02$
Mean IoU over *all* classes (%)			
No dehazing	37.6	**37.8**	36.1
MSCNN	**38.3**	37.1	36.9
DCP	36.6	37.4	36.1
Non-local	36.2	36.6	35.3
Mean IoU over *frequent* classes in *Foggy Driving* (%)			
No dehazing	57.0	**57.4**	56.2
MSCNN	**57.3**	56.2	56.3
DCP	56.0	56.7	55.2
Non-local	55.1	55.1	54.5

Best results are given in bold

2. a clear appearance of both sets of data, in the sense that the segmentation model should have an easy job in mining discriminative features from the data.

Focusing on the case that does not involve dehazing, we observe that the models with $\beta = 0.005$ and $\beta = 0.01$ perform significantly better than that with $\beta = 0.02$, implying that according to point 1 *Foggy Driving* is dominated by scenes with light or medium fog. On the other hand, each of the three dehazing methods that are used for preprocessing has its own particularities in enhancing the appearance and contrast of foggy scenes while also introducing artifacts to the output. More specifically, MSCNN is slightly conservative in removing fog, as was found for other learning-based dehazing methods in Li et al. (2016), and operates best under lighter fog, providing a significant improvement in this setting with regard to point 2. In conjunction with the light-fog character of *Foggy Driving*, this explains why fine-tuning on light fog ($\beta = 0.005$) combined with MSCNN preprocessing delivers one of the two best overall results. By contrast, the more aggressive DCP is known to operate better at high levels of fog, as its estimated transmission is biased towards lower values (Tang et al. 2014). The performance of models with DCP preprocessing thus peaks at medium rather than low simulated-fog density, which signifies a trade-off between removing fog to the proper extent and minimal introduction of artifacts. Non-local dehazing has also been found to operate best at medium levels of fog (Li et al. 2016), which results in a similar performance trend to DCP.

Effect of Dehazing Preprocessing on Real-world Performance and Discussion Comparing the four options regarding dehazing preprocessing via Table 4, we observe that applying no dehazing is the best or second best option for both measures and across all three values of β. Only MSCNN marginally beats the no-dehazing option in some cases, while overall these two options are roughly on a par. The absence of a significant performance gain on *Foggy Driving* when performing dehazing preprocessing can be ascribed to generic as well as method-specific reasons.

First, in the real-world setting of *Foggy Driving*, the homogeneity and uniformity assumptions of the optical model (1) that is used by all examined dehazing methods may not hold exactly. Of course, this model is also used in our fog simulation, however, foggy image synthesis is a *forward* problem, whereas image defogging/dehazing is an *inverse* problem, hence inherently more difficult. Thus, the artifacts that are introduced by our fog simulation are likely to be less prominent than those introduced by dehazing. This fact appears to outweigh the potential increase in visibility for dehazed images as far as point 2 above is concerned. An interesting insight that follows is the use of forward techniques to generate training data for hard target domains based on data from the source domain as an alternative to the application of inverse techniques to transform such target domains into the easier source domain.

Second, the optical model (1), on which most of the popular dehazing approaches rely, assumes a *linear* relation between the irradiance at a pixel and the actual value of the pixel in the processed hazy image. Therefore, these approaches require that an initial gamma correction step be applied before dehazing, otherwise their performance may deteriorate significantly. This in turn implies that the value of gamma must be known for each image, which is *not* the case for Cityscapes and *Foggy Driving*. Manually searching for "best" per-image values is also infeasible for these large datasets. In the absence of any further information, we have used a constant value of 1 for gamma as Berman et al. (2016) recommend, which is probably suboptimal for most of the images. We thus wish to point out that future work on outdoor datasets, whether considering fog/haze or not, should ideally record the value of gamma for each image, so that dehazing methods can show their full potential on such datasets.

Specifically for DCP, performance decreases compared to MSCNN partly due to the light-fog character of *Foggy Driving* which does not match the optimal operating point of DCP. On the other hand, non-local dehazing uses a different model for estimating atmospheric light than the one that is shared by our fog simulation, MSCNN, and DCP, and thus already faces greater difficulty in dehazing images from *Foggy Cityscapes*.

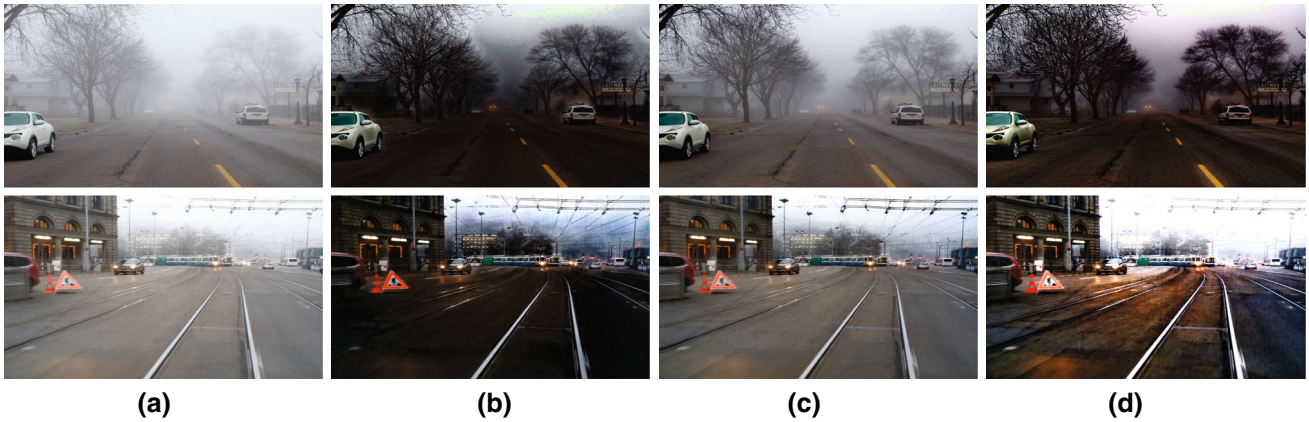

Fig. 9 Example images from *Foggy Driving* and their dehazed versions using three state-of-the-art dehazing methods that are examined in our experiments. **a** Foggy. **b** DCP. **c** MSCNN. **d** Non-local

5.2 Linking the Objective and Subjective Utility of Dehazing Preprocessing in Foggy Scene Understanding

Our experiments in Sect. 5.1 indicate that using any of the three examined state-of-the-art dehazing methods to preprocess foggy images before feeding them to a CNN for semantic segmentation does not provide a clear benefit over feeding the foggy images directly in the objective terms of mean IoU performance of the trained model. In this section, we complement this objective evaluation with a study of the utility of dehazing preprocessing for human understanding of foggy scenes and show that the comparative results of the objective evaluation generally agree with the comparative results of the human-based evaluation.

Both for the objective semantic-segmentation-based and the subjective human-based evaluation, we compare the four aforementioned options with regard to dehazing preprocessing *individually* on each image of our datasets. Figure 9 presents examples of the tetrads of images that we consider: the foggy image, which either belongs to the validation set of *Foggy Cityscapes-refined* with $\beta = 0.01$ or to *Foggy Driving* and corresponds to no usage of dehazing, and its dehazed versions using DCP, MSCNN and non-local dehazing. For comparative objective evaluation of the four alternatives on each image, we use the mean IoU scores of the respective fine-tuned DCN models that are considered in the experiment of Table 2, measured on that image. The classes that do not occur in an image are not considered for computing mean IoU on this image. The four alternatives are ranked for each image according to their mean IoU scores on it. Comparative evaluation based on human subjects considers the same tetrads of images but employs a more composite protocol, which is detailed below.

User Study via Amazon Mechanical Turk Humans are subjective and are not good at giving scores to individual images in a linear scale (Kendall and Smith 1940). We thus follow the literature (Rubinstein et al. 2010) and choose the paired comparisons technique to let human subjects compare the four options regarding dehazing preprocessing. The participants are shown two images at a time that both pertain to the same scene, side by side, and are simply asked to choose the one which is more suitable for safe driving (i.e. easier to interpret). Thus, six comparisons need to be performed per scene, corresponding to all possible pairs.

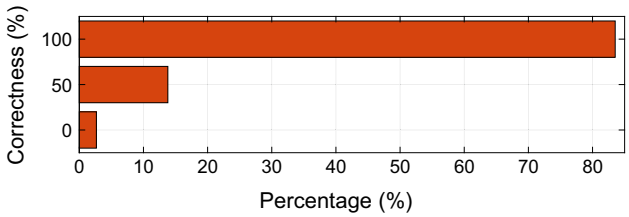

Fig. 10 Quality of our user survey on AMT, computed using known-answer questions

We use Amazon Mechanical Turk (AMT) to perform these comparisons. In order to guarantee high quality, we only employ AMT Masters in our study and verify the answers via a Known Answer Review Policy. Masters are an elite group of subjects, who have consistently demonstrated superior performance on AMT. Each individual task completed by the participants, referred to as Human Intelligence Task (HIT), comprises five image pairs to be compared, out of which three pairs are the true query pairs and the rest two pairs have a known correct answer and are only used for validation. In particular, each known-answer pair consists of two versions of a scene from *Foggy Cityscapes-refined* with different levels of fog, choosing from three versions of the dataset corresponding to clear weather, $\beta = 0.005$ and $\beta = 0.01$. The version with less fog is considered the correct answer. In order to avoid answers based on memorized patterns, the five image pairs in each HIT are randomly shuffled

Table 5 Agreement coefficients for all pairwise comparisons of the four dehazing options

Foggy versus DCP	0.155
Foggy versus MSCNN	0.115
Foggy versus non-local	0.010
DCP versus MSCNN	0.182
DCP versus non-local	0.036
MSCNN versus non-local	0.182
Mean	0.113

and the left-right order of the images in each pair is randomly swapped. In addition, each HIT is completed by three different subjects to increase reliability. The overall quality of the user survey is shown in Fig. 10, which demonstrates that the subjects have done a decent job: for 83% of the HITs, both known-answer questions are answered correctly. We only use results from these HITs in our following analysis.

Consistency of Subjects' Answers We first study the consistency of choices among subjects; all subjects are in high agreement if the advantage of one option over the other is obvious and consistent. To measure this, we employ the coefficient of agreement (Kendall and Smith 1940):

$$\mu = \frac{2\sigma}{\binom{m}{2}\binom{t}{2}} - 1, \text{ with } \sigma = \sum_{i=1}^{t}\sum_{j=1}^{t}\binom{a_{ij}}{2}, \tag{8}$$

where a_{ij} is the number of times that option i is chosen over option j, $m = 3$ is the number of subjects, and $t = 4$ is the number of dehazing options. The maximum of μ is 1 for complete agreement and its minimum is $-1/3$ for complete disagreement. The values of μ for all pairs of options are shown in Table 5. The small positive numbers in the table suggest that subjects tend to agree when comparing options pairwise but no single option has dominant advantage over another one.

Ranking and Correlation with Objective Evaluation We finally compute the overall ranking of all four options for each image based on the number of times each option is chosen in all relevant pairwise comparisons. The correlation of these rankings with those induced by mean IoU performance is measured with *Kendall's τ coefficient* (Kendall 1938) with $-1 \leq \tau \leq 1$, where a value of 1 implies perfect agreement, -1 implies perfect disagreement, and 0 implies zero correlation. Figure 11 provides a complete overview of the comparative results both for our user study and the semantic-segmentation-based evaluation on *Foggy Cityscapes-refined* and *Foggy Driving*, including rank correlation results for the two types of evaluation.

The results in the top row of Fig. 11 indicate that none of the three examined methods for dehazing preprocessing improves reliably the human understanding of synthetic foggy scenes from *Foggy Cityscapes* or real foggy scenes from *Foggy Driving*. In particular, the no-dehazing option beats all other three options in pairwise comparisons on *Foggy Cityscapes-refined* and loses only to DCP marginally on *Foggy Driving*, while it is also ranked first on more images than any other option for both datasets.

In addition, the rankings obtained with the two types of evaluation are generally in congruence for the real-world case of *Foggy Driving*. The no-dehazing and DCP options are ranked higher than MSCNN and non-local dehazing both in the user study and in the objective evaluation. The high performance of DCP compared to MSCNN is due to the usage of $\beta = 0.01$ for *Foggy Cityscapes-refined* (cf. the discussion in Sect. 5.1). What is more, the two rankings exhibit a positive correlation on average for *Foggy Driving* based on the respective distribution of τ in the bottom right chart of Fig. 11, which supports our conclusion in Sect. 5.1 about the marginal benefit of dehazing preprocessing for foggy scene understanding.

5.3 Object Detection

For our experiment on object detection in foggy scenes, we select the modern Fast R-CNN (Girshick 2015) as the architecture of the evaluated models. We prefer Fast R-CNN over more recent approaches such as Faster R-CNN (Ren et al. 2015) because the former involves a simpler training pipeline, making fine-tuning to foggy conditions straightforward. Consequently, we do not learn the front-end of the object detection pipeline which involves generation of object proposals; rather, we use multiscale combinatorial grouping (Arbeláez et al. 2014) for this task.

In order to ensure a fair comparison, we first obtain a baseline Fast R-CNN model for the original Cityscapes dataset, similarly to the preceding semantic segmentation experiments. Since no such model is publicly available, we begin with the model released by Girshick (2015) which has been trained on PASCAL VOC 2007 (Everingham et al. 2010) and fine-tune it on the union of the training and validation sets of Cityscapes which comprises 3475 images. Fine-tuning through all layers is run with the same configurations as in Girshick (2015), except that we use the "poly" learning rate policy with a base learning rate of 2×10^{-4} and a power parameter of 0.9, with 7k iterations (4 epochs).

This baseline model that has been trained on the real Cityscapes with clear weather serves as initialization for fine-tuning on our synthetic images from *Foggy Cityscapes-refined*. To this end, we use all 550 training and validation images of *Foggy Cityscapes-refined* and fine-tune with the same settings as before, only that the base learning rate is set to 10^{-4} and we run 1650 iterations (6 epochs).

We experiment with two values of the attenuation coefficient β for *Foggy Cityscapes-refined* and present comparative performance on the 33 finely annotated images of *Foggy Driving* in Table 6. No dehazing is involved in this experiment. We concentrate on the classes *car* and *person* for evaluation, since they constitute the intersection of the set of frequent classes in *Foggy Driving* and the set of annotated classes with distinct instances. Individual average precision (AP) scores for *car* and *person* are reported, as well as mean scores over these two classes ("mean frequent") and over the complete set of 8 classes occurring in instances ("mean all"). For completeness, we note that the original VOC 2007 model of Girshick (2015) exhibits an AP of 2.1% for *car* and 1.9% for *person*.

Both of our fine-tuned models outperform the baseline model by a significant margin for *car*. At the same time, they are on a par with the baseline model for *person*. The overall winner is the model that has been fine-tuned on light fog, which we refer to as *FT-0.005*: it outperforms the baseline model by 2.4% on average on the two frequent classes and it is also slightly better when taking all 8 classes into account.

We provide a visual comparison of *FT-0.005* and the baseline model for car detection on example images from *Foggy Driving* in Fig. 12. Note the ability of our model to detect distant cars, such as the two cars in the image of the second row which are moving on the left side of the road and are visible from their front part. These two cars are both missed by the baseline model.

Table 6 Performance comparison on *Foggy Driving* of baseline Fast R-CNN model trained on Cityscapes ("W/o FT") versus fine-tuned versions of it using *Foggy Cityscapes-refined*

	Mean all	Car	Person	Mean frequent
W/o FT	11.1	30.5	**10.3**	20.4
FT $\beta = 0.01$	11.1	34.6	10.0	22.3
FT $\beta = 0.005$	**11.7**	**35.3**	**10.3**	**22.8**

"FT" stands for using fine-tuning and "W/o FT" for not using fine-tuning. AP (%) is used to report results
Best results are given in bold

6 Semi-supervised Learning with Synthetic Fog

While standard supervised learning can improve the performance of SFSU using our synthetic fog, the paradigm still needs manual annotations for corresponding clear-weather images. In this section, we extend the learning to a new paradigm which is also able to acquire knowledge from unlabeled pairs of foggy images and clear-weather images. In particular, we train a semantic segmentation model on clear-weather images using the standard supervised learning

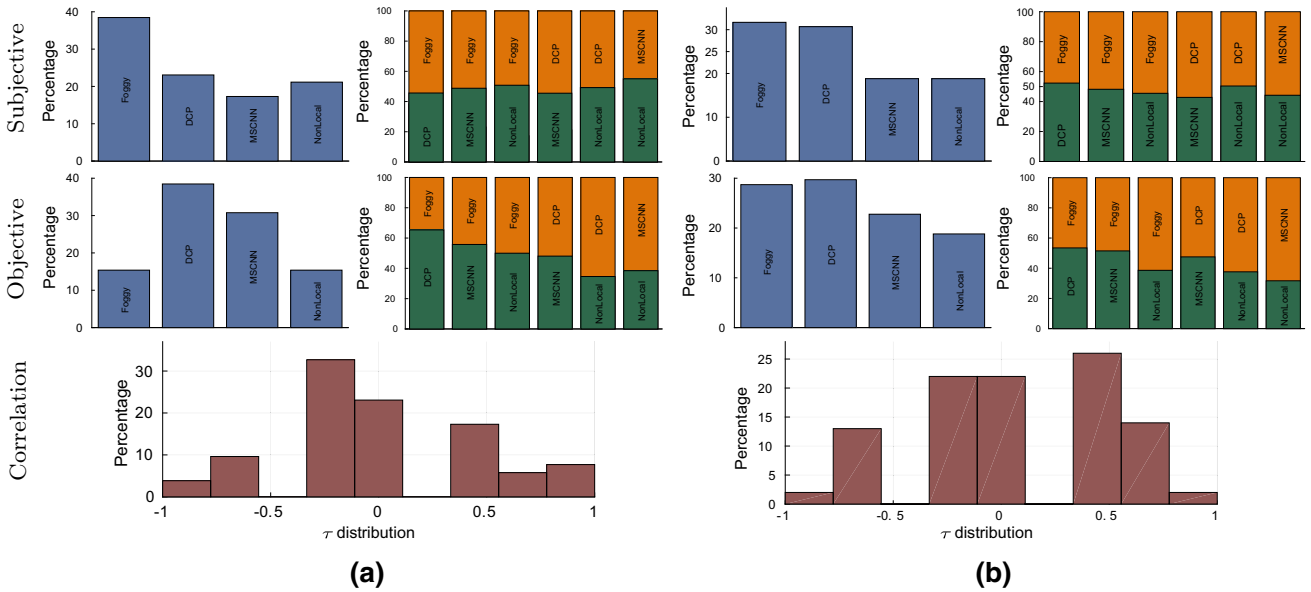

Fig. 11 Comparison of four options for dehazing preprocessing, i.e. no dehazing ("Foggy"), "DCP" (He et al. 2011), "MSCNN" (Ren et al. 2016), and "NonLocal" (Berman et al. 2016), on **a** the validation set of *Foggy Cityscapes-refined* for $\beta = 0.01$ and **b** *Foggy Driving*, in terms of subjective human understanding of the foggy scenes (top) and performance of the corresponding fine-tuned DCN models (middle). For each combination of dataset and evaluation setting, we show the percentage of scenes for which each option is ranked first overall on the left, and the respective percentages for pairwise comparisons of the options on the right. Bottom: Histograms of correlation of the rankings obtained for the two evaluation settings over the datasets, measured with Kendall's τ

Fig. 12 Qualitative results for detection of cars on *Foggy Driving*. From left to right: ground truth annotation, baseline Fast R-CNN model trained on original Cityscapes, and our model *FT-0.005* fine-tuned on *Foggy Cityscapes-refined* with light fog. This figure is seen better when zoomed in on a screen

paradigm, and apply the model to an even larger set of clear but "unlabeled" images (e.g. our 20,000 unlabeled images of *Foggy Cityscapes-coarse*) to generate the class responses. Since we have created a foggy version for the unlabeled dataset, these class responses can then be used to supervise the training of models for SFSU.

This learning approach is inspired by the stream of work on model distillation (Hinton et al. 2015; Gupta et al. 2016b) or imitation (Buciluǎ et al. 2006; Dai et al. 2015). Buciluǎ et al. (2006), Hinton et al. (2015) and Dai et al. (2015) transfer supervision from sophisticated models to simpler models for efficiency, and Gupta et al. (2016b) transfers supervision from the domain of images to other domains such as depth maps. In our case, supervision is transferred from clear weather to foggy weather. The underpinnings of our proposed approach are the following: (1) in clear weather, objects are easier to recognize than in foggy weather, thus models trained on images with clear weather in principle generalize better to new images of the same condition than those trained on foggy images; and (2) since the synthetic foggy images and their clear-weather counterparts depict exactly the same scene, recognition results should also be the same for both images ideally.

We formulate our semi-supervised learning (SSL) for semantic segmentation as follows. Let us denote a clear-weather image by $\mathbf{x}$, the corresponding foggy one by $\mathbf{x}'$, and the corresponding human annotation by $\mathbf{y}$. Then, the training data consist of both labeled data $\mathcal{D}_l = \{(\mathbf{x}_i, \mathbf{x}'_i, \mathbf{y}_i)\}_{i=1}^{l}$ and unlabeled data $\mathcal{D}_u = \{(\mathbf{x}_j, \mathbf{x}'_j)\}_{j=l+1}^{l+u}$, where $\mathbf{y}_i^{m,n} \in \{1, ..., K\}$ is the label of pixel (m, n), and K is the total number of classes. l is the number of labeled training images, and u is the number of unlabeled training images. The aim is to learn a mapping function $\phi' : \mathcal{X}' \mapsto \mathcal{Y}$ from $\mathcal{D}_l$ and $\mathcal{D}_u$. In our case, $\mathcal{D}_l$ consists of the 498 high-quality foggy images in the training set of *Foggy Cityscapes-refined* which have human annotations with fine details, and $\mathcal{D}_u$ consists of the

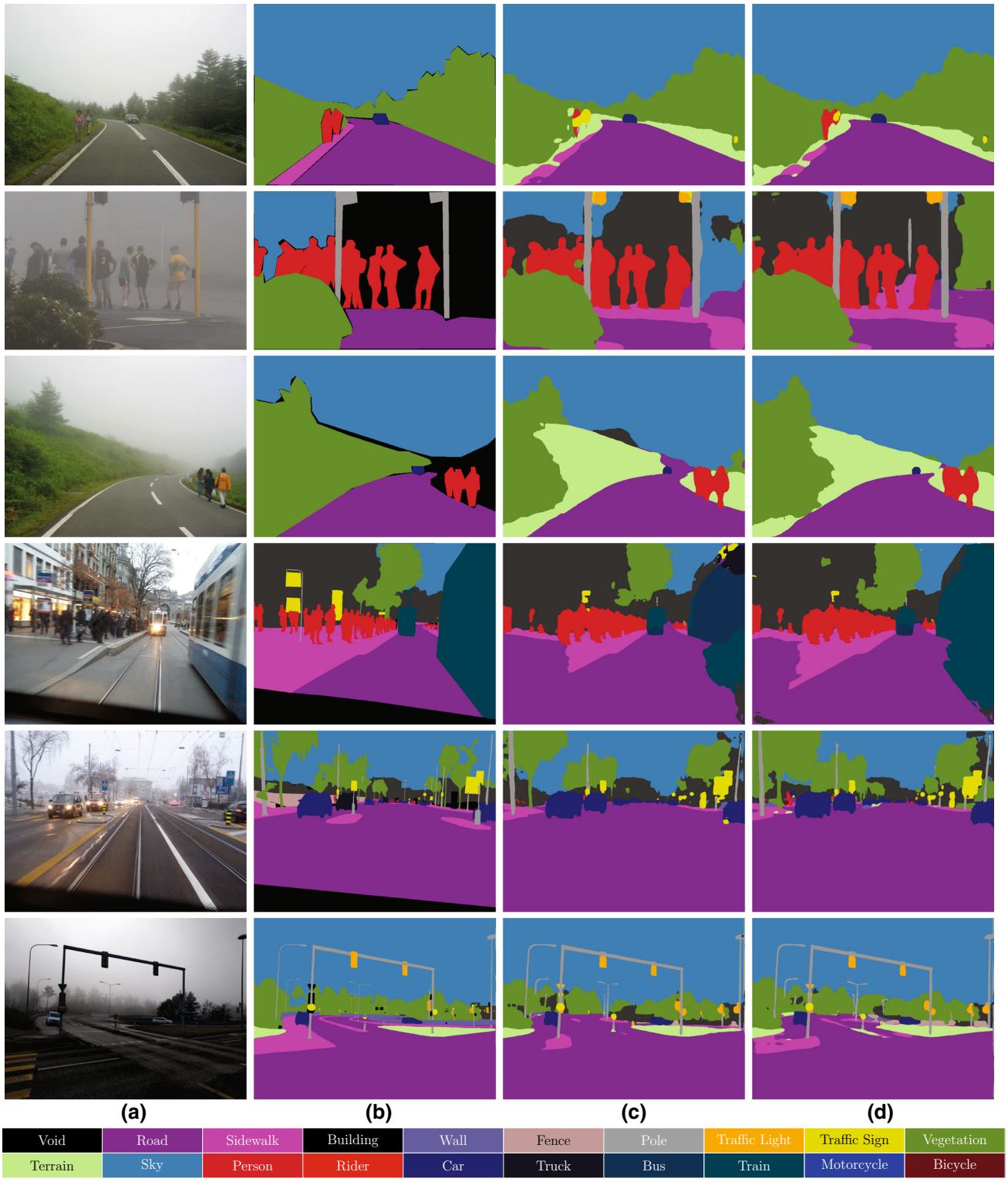

Fig. 13 Qualitative results for semantic segmentation on *Foggy Driving*, both for coarsely annotated images (top three rows) and finely annotated images (bottom three rows). **a** Foggy image. **b** Ground truth. **c** Lin et al. (2017). **d** Ours: Lin et al. (2017) fine-tuned with our SSL on *Foggy Cityscapes*

additional 20,000 foggy images in *Foggy Cityscapes-coarse* which do not have fine human annotations.

Since $\mathcal{D}_u$ does not have class labels, we use the idea of supervision transfer to generate the supervisory labels for all the images therein. To this end, we first learn a mapping function $\phi : \mathcal{X} \mapsto \mathcal{Y}$ with $\mathcal{D}_l$ and then obtain the labels $\hat{\mathbf{y}}_j = \phi(\mathbf{x}_j)$ for $\mathbf{x}_j$ and $\mathbf{x}'_j$, $\forall j \in \{l+1, ..., l+u\}$. $\mathcal{D}_u$ is then upgraded to $\hat{\mathcal{D}}_u = \{(\mathbf{x}_j, \mathbf{x}'_j, \hat{\mathbf{y}}_j)\}_{j=l+1}^{l+u}$. The proposed scheme for training semantic segmentation models for foggy images $\mathbf{x}'$ is to learn a mapping function ϕ' so that human annotations $\mathbf{y}$ and the transferred labels $\hat{\mathbf{y}}$ are both taken into account:

$$\min_{\phi'} \sum_{i=1}^{l} L(\phi'(\mathbf{x}'_i), \mathbf{y}_i) + \lambda \sum_{j=l+1}^{l+u} L(\phi'(\mathbf{x}'_j), \hat{\mathbf{y}}_j), \tag{9}$$

where $L(.,.)$ is the Categorical Cross Entropy Loss function for classification, and $\lambda = \frac{l}{u} \times w$ is a parameter for balancing the contribution of the two terms, serving as the relative weight of each unlabeled image compared to each labeled one. We empirically set $w = 5$ in our experiment, but an optimal value can be obtained via cross-validation if needed. In our implementation, we approximate the optimization of (9) by mixing images from $\mathcal{D}_l$ and $\hat{\mathcal{D}}_u$ in a proportion of $1 : w$ and feeding the stream of hybrid data to a CNN for standard supervised training.

We select RefineNet (Lin et al. 2017) as the CNN model for semantic segmentation, which is a more recent and better performing method than DCN (Yu and Koltun 2016) that is used in Sect. 5. The reason for using DCN in Sect. 5 is that RefineNet had not been published yet at the time that we were conducting the experiments of Sect. 5. We would like to note that the state-of-the-art PSPNet (Zhao et al. 2017), which has been trained on the Cityscapes dataset similarly to the original version of RefineNet that we use as our baseline, achieved a mean IoU of only 24.0% on *Foggy Driving* in our initial experiments.

We use mean IoU for evaluation, similarly to Sect. 5, and $\beta = 0.01$ for *Foggy Cityscapes*. We compare the performance of three trained models: (1) original RefineNet (Lin et al. 2017) trained on Cityscapes, (2) RefineNet fine-tuned on $\mathcal{D}_l$, and (3) RefineNet fine-tuned on $\mathcal{D}_l$ and $\hat{\mathcal{D}}_u$. The mean IoU scores of the three models on *Foggy Driving* are 44.3%, 46.3%, and **49.7**% respectively. The 2% improvement of (2) over (1) confirms the conclusion we draw in Sect. 5 that fine-tuning with our synthetic fog can indeed improve the performance of semantic foggy scene understanding. The 3.4% improvement of (3) over (2) validates the efficacy of the SSL paradigm. Figure 13 shows visual results of (1) and (3), along with the foggy images and human annotations. The re-trained model with our SSL paradigm can better segment certain parts of the images which are misclassified by the original RefineNet, e.g. the pedestrian in the first example, the tram in the fourth one, and the sidewalk in the last one.

Both the quantitative and qualitative results suggest that our approach is able to alleviate the need for collecting large-scale training data for semantic understanding of foggy scenes, by training with the annotations that are already available for clear-weather images and the generated foggy images directly and by transferring supervision from clear-weather images to foggy images of the same scenes.

7 Conclusion

In this paper, we have demonstrated the benefit of synthetic data that are based on real images for semantic understanding of foggy scenes. Two foggy datasets have been constructed to this end: the partially synthetic *Foggy Cityscapes* dataset which derives from Cityscapes, and the real-world *Foggy Driving* dataset, both with dense pixel-level semantic annotations for 19 classes and bounding box annotations for objects belonging to 8 classes. We have shown that *Foggy Cityscapes* can be used to boost performance of state-of-the-art CNN models for semantic segmentation and object detection on the challenging real foggy scenes of *Foggy Driving*, both in a usual supervised setting and in a novel, semi-supervised setting. Last but not least, we have exposed through detailed experiments the fact that image dehazing faces difficulties in working out of the box on real outdoor foggy data and thus is marginally helpful for SFSU. In the future, we would like to combine dehazing and semantic understanding of foggy scenes into a unified, end-to-end learned pipeline, which can also leverage the type of synthetic foggy data we have introduced. The datasets, models and code are available at http://www.vision.ee.ethz.ch/~csakarid/SFSU_synthetic.

Acknowledgements The authors would like to thank Kevis Maninis for useful discussions. This work is funded by Toyota Motor Europe via the research project TRACE-Zürich.

References

Abu Alhaija, H., Mustikovela, S. K., Mescheder, L., Geiger, A., & Rother, C. (2017). Augmented reality meets deep learning for car instance segmentation in urban scenes. In *Proceedings of the British machine vision conference (BMVC)*.

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.

Arbeláez, P., Pont-Tuset, J., Barron, J., Marques, F., & Malik, J. (2014). Multiscale combinatorial grouping. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

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.

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*.

Buch, N., Velastin, S. A., & Orwell, J. (2011). A review of computer vision techniques for the analysis of urban traffic. *IEEE Transactions on Intelligent Transportation Systems*, *12*(3), 920–939.

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

Camplani, M., & Salgado, L. (2012). Efficient spatio-temporal hole filling strategy for Kinect depth maps. In *SPIE/IS&T electronic imaging*.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848.

Colomb, M., Hirech, K., André, P., Boreux, J. J., Lacote, P., & Dufour, J. (2008). An innovative artificial fog production device improved in the European project FOG. *Atmospheric Research*, *87*(3), 242–251.

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 603–619.

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., & Yang, W. (2011). Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Geoscience and Remote Sensing Letters*, *8*(1), 173–176.

Dosovitskiy, A., Fischery, P., Ilg, E., Häusser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., & Brox, T. (2015). FlowNet: Learning optical flow with convolutional networks. In *IEEE international conference on computer vision (ICCV)*.

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), 72.

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.

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, A., Vedaldi, A., & Zisserman, A. (2016). Synthetic data for text localisation in natural images. In *IEEE conference on computer vision and pattern recognition*.

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.

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

Hirschmüller, H. (2008). Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(2), 328–341.

Hoffman, J., Tzeng, E., Park, T., Zhu, J. Y., Isola, P., Saenko, K., Efros, A. A., & Darrell, T. (2017). CyCADA: Cycle-consistent adversarial domain adaptation. ArXiv e-prints.

Janai, J., Güney, F., Behl, A., & Geiger, A. (2017). Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. arXiv preprint arXiv:1704.05519.

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.

Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S. N., Rosaen, K., & Vasudevan, R. (2017). Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *IEEE international conference on robotics and automation*.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93.

Kendall, M. G., & Smith, B. B. (1940). On the method of paired comparisons. *Biometrika*, *31*(3/4), 324–345.

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

Levin, A., Lischinski, D., & Weiss, Y. (2004). Colorization using optimization. In *ACM SIGGRAPH*.

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., Tan, R. T., & Brown, M. S. (2015). Nighttime haze removal with glow and multiple light colors. In *IEEE international conference on computer vision (ICCV)*.

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

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.

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

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)*.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 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*.

Richter, S. R., Vineet, V., Roth, S., & Koltun, V. (2016). Playing for data: Ground truth from computer games. In *European conference on computer vision*. Springer.

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)*.

Rubinstein, M., Gutierrez, D., Sorkine, O., & Shamir, A. (2010). A comparative study of image retargeting. *ACM Transactions on Graphics*, *29*(6), 160:1–160:10.

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.

Shen, J., & Cheung, S. C. S. (2013). Layer depth denoising and completion for structured-light RGB-D cameras. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from RGBD images. In *European conference on computer vision*.

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. (2009). Fast visibility restoration from a single color or gray level image. In *IEEE international conference on computer vision*.

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).

Vázquez, D., Lopez, A. M., Marin, J., Ponsa, D., & Geronimo, D. (2014). Virtual and real world adaptation for pedestrian detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(4), 797–809.

Wang, L., Jin, H., Yang, R., & Gong, M. (2008). Stereoscopic inpainting: Joint color and depth completion from stereo images. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Wang, W., Yuan, X., Wu, X., & Liu, Y. (2017). Fast image dehazing method based on linear transformation. *IEEE Transactions on Multimedia*, *19*(6), 1142–1155.

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

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. CoRR arXiv:1708.00581.

Zhang, J., Cao, Y., & Wang, Z. (2014). Nighttime haze removal based on a new imaging model. In *IEEE international conference on image processing (ICIP)* (pp. 4557–4561).

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)*.