# Semantic-Aware Dehazing Network With Adaptive Feature Fusion

Shengdong Zhang, Wenqi Ren, *Member, IEEE*, Xin Tan, *Graduate Student Member, IEEE*, Zhi-Jie Wang, *Member, IEEE*, Yong Liu, Jingang Zhang, Xiaoqin Zhang, and Xiaochun Cao, *Senior Member, IEEE*

***Abstract*—Despite that convolutional neural networks (CNNs) have shown high-quality reconstruction for single image dehazing, recovering natural and realistic dehazed results remains a challenging problem due to semantic confusion in the hazy scene. In this article, we show that it is possible to recover textures faithfully by incorporating semantic prior into dehazing network since objects in haze-free images tend to show certain shapes, textures, and colors. We propose a semantic-aware dehazing network (SDNet) in which the semantic prior is taken as a color constraint for dehazing, benefiting the acquisition of a reasonable scene configuration. In addition, we design a densely connected block to capture global and local information for dehazing and semantic prior estimation. To eliminate the unnatural appearance of some objects, we propose to fuse the features from shallow and deep layers adaptively. Experimental results demonstrate that our proposed model performs favorably against the state-of-the-art single image dehazing approaches.**

***Index Terms*—Adaptive feature fusion, dehazing, image restoration, semantic aware.**

Manuscript received 31 December 2020; revised 31 July 2021; accepted 23 October 2021. Date of publication 19 November 2021; date of current version 21 December 2022. This work was supported by the National Natural Science Foundation of China under Grant 62076234, Grant 62172409, Grant 61972425, Grant U1811264, Grant 61922064, Grant U2033210, and Grant 62101387. This article was recommended by Associate Editor S. Chen. *(Corresponding authors: Jingang Zhang; Xiaoqin Zhang.)*

Shengdong Zhang is with the College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou 325035, China, and also with the Computer Science Engineering Department, Shaoxing University, Shaoxing 312000, China (e-mail: shengdong1986@gmail.com).

Wenqi Ren and Xiaochun Cao are with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: renwenqi@iie.ac.cn; caoxiaochun@iie.ac.cn).

Xin Tan is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: tanxin2017@sjtu.edu.cn).

Zhi-Jie Wang is with the College of Computer Science and the Ministry of Education Key Laboratory of Dependable Service Computing in Cyber Physical Society, Chongqing University, Chongqing 400044, China (e-mail: cszjwang@cqu.edu.cn).

Yong Liu is with the Beijing Key Laboratory of Big Data Management and Analysis Methods, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100093, China.

Jingang Zhang is with the School of Future Technology, The University of Chinese Academy of Sciences, Beijing 100039, China (e-mail: zhangjg@ucas.ac.cn).

Xiaoqin Zhang is with the College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou 325035, China (e-mail: zhangxiaoqinnan@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2021.3124231.

Digital Object Identifier 10.1109/TCYB.2021.3124231

## I. Introduction

SINGLE image dehazing aims to regain a haze-free image from a hazy input directly, which is a fundamental problem in the image processing field since dehazing can greatly facilitate related high-level tasks [1], for example, image recognition and scene understanding. In the literature, the image degradation process caused by air particles is mathematically formulated as [2]

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{1}$$

where $I(x)$ denotes the degenerative hazy image, the scene radiance needs to be recovered is represented by $J(x)$, $A$ is the environment lighting, and the transmittance map is represented by $t(x)$ that depends on scattering coefficient $\beta$ and scene depth $d(x)$.

Early methods [2], [3] employ multiple images or additional depth information to remove haze. However, it is hard to obtain the multiple images for the same scene or additional depth information in real cases. To overcome this problem, single image dehazing methods are proposed [4], [5] capitalized on sharp image priors. He *et al.* [6] discovered a dark channel prior (DCP) to predict the transmission map. However, DCP may be ineffective for the scene objects that are similar to the atmospheric light. Fattal [7] observed that pixels in a haze-free patch form a line in the RGB color space and recover transmission maps based on this prior. Berman *et al.* [8] introduced a haze-line prior, based on the fact that hundred color clusters can be used to represent a haze-free image well [8].

Recently, deep neural networks provide significantly improved performance in terms of peak signal-to-noise ratio (PSNR) in the single image dehazing task [9], [10]. Cai *et al.* [11] employed convolutional neural network (CNN) to extract more effective low-level features to predict the transmission map. Ren *et al.* [12] introduced a multiscale deep model to predict the transmission map, in which a large network is employed to predict a coarse transmission map, and then a small network is used to refine the coarse transmission map. However, such networks exhibit limitations in terms of faithful texture recovery.

In this work, we propose an efficient algorithm to predict semantic segmentation for single image dehazing. Suppose the semantic segmentation of the scene is known, this prior can characterize the semantic class of an object region (e.g., sky, building, and grass) and constrains the reasonable solution

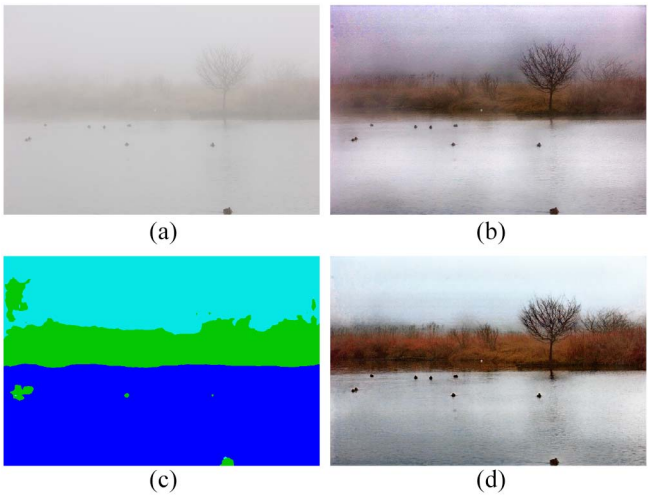


Fig. 1. Visual comparisons on a challenging real-world dense hazy example. Without using the semantic information, the network cannot recover a reasonable scene, for example, the sky region tends to be dark. In contrast, with the guidance from semantic information, our algorithm generates a visual faithful result. (a) Hazy input. (b) Without semantic prior. (c) Semantic segmentation. (d) With semantic prior.

space for dehazing. On the other hand, we note that a hazy image and the corresponding haze-free image share the same semantic information, which can be used to bridge the gap between synthetic training data and real-world hazy images.

Motivated by the above insights, we attempt to develop a method that can fully utilize the correlation between color space and semantic segmentation probability maps. To this end, we design a semantic-aware dehazing network (SDNet), which generates haze-relevant features and semantic prior based on the fusion of high-level and low-level features. Low-level features help recover texture details, and high-level features contain much semantic information, which also benefits the dehazing and semantic prior estimation. As shown in Fig. 1, the proposed SDNet generates faithful colors when considering the semantic prior.

Another advantage of using semantic prior is that semantic information could reduce the gap between synthetic training data and real-world test images. We note that most existing learning-based dehazing approaches ignore the gap between synthetic and real hazy images. This article employs semantic prior to bridge the gap between synthetic and real images since both share the same classifications and categories. Therefore, semantic labels of objects in training data can be generalized to the same ones in real photographs. To this end, we develop a benchmark dataset consisting of outdoor hazy images, semantic segmentations, and ground-truth depth maps from the SYSU-Scene dataset [13]. In addition, we note that directly fusing the shallow layers and deep layers results in artifacts in the final dehazed image. To address this issue, we propose an adaptive fusion module. The adaptive fusion module can pass the most representative features and rescale the weight of high-level and low-level features, which are helpful to generate a natural dehazed result.

The main contributions of this article are as follows.

1) We propose an SDNet to solve semantic segmentation and image dehazing simultaneously in a unified framework. Reconstruction of a dehazed image with rich semantic regions can be achieved by the learned semantic priors.
2) We propose an adaptive fusion module to fuse the features from shallow layers and deep layers adaptively. The adaptive fusion module is helpful to remove the unnatural artifact in the final dehazed image.
3) We develop a benchmark dataset consisting of outdoor hazy images and semantic segmentations to train the proposed network. We show the learned SDNet is able to dehaze real-world hazy images well with the semantic prior.
4) We evaluate the proposed dehazing method through extensive experiments on both synthetic datasets and real-world images. In addition, ablation studies are conducted to demonstrate the effectiveness of different modules in the proposed SDNet.

## II. Related Work

In this section, we review the most related work of single image dehazing and semantic knowledge learning.

*Single Image Dehazing:* The presence of haze reduces the color saturation and contrast of haze-free images, which degrades the performance of most high-level computer vision tasks. Dehazing methods can be mainly grouped into two categories: 1) image restoration methods based on sharp image priors and 2) deep-learning-based dehazing networks.

There existed many image restoration methods via handcrafted features [8], [14]–[16]. Based on the observation that haze-free image patch has at least one pixel with one color channel tends to be zero, He *et al.* employed DCP to predict the transmission map effectively in general cases. However, DCP cannot be applied to white scenes and sky regions. To improve the generalization ability of DCP, Meng *et al.* proposed a boundary-constraint prior (BCCR). Zhu *et al.* [15] proposed a linear model to predict the depth and solved the parameters of the model with a supervised machine-learning method. Chen and Huang [17] introduced an edge collapse-based dehazing algorithm, which dynamically repairs the transmission map and obtains satisfactory visibility dehazing results. Kim *et al.* [18] introduced a fast dehazing method based on transmission map estimation.

Recently, thanks to the development of CNNs, researchers introduce numerous deep models for image dehazing, such as DehazeNet [11], DCPDN [19], DDN [20], HDDNet [21], and MSCNN [12]. DehazeNet [11] and MSCNN [12] are designed to predict transmission maps by stacking some CNN layers. DCPDN [19] recovers the final dehazed result via embedding the atmospheric scattering model into the network. EPDN [22] models the dehazing as an image-to-image translation problem. Chen *et al.* [23] introduced a deep-learning-based method to improve the generalization of DCP. Li *et al.* [24] introduced a progressive dehazing network with a haze-level aware. Liu *et al.* [25] introduced a deep prior for single image dehazing. By considering the nonlocal similarity [26], Zhang *et al.* [27] proposed a nonlocal dehazing network. Deng *et al.* [28] proposed to obtain different dehazed results and then fuse the intermediate results to obtain a high-quality dehazed result. A perception-inspired method [29]

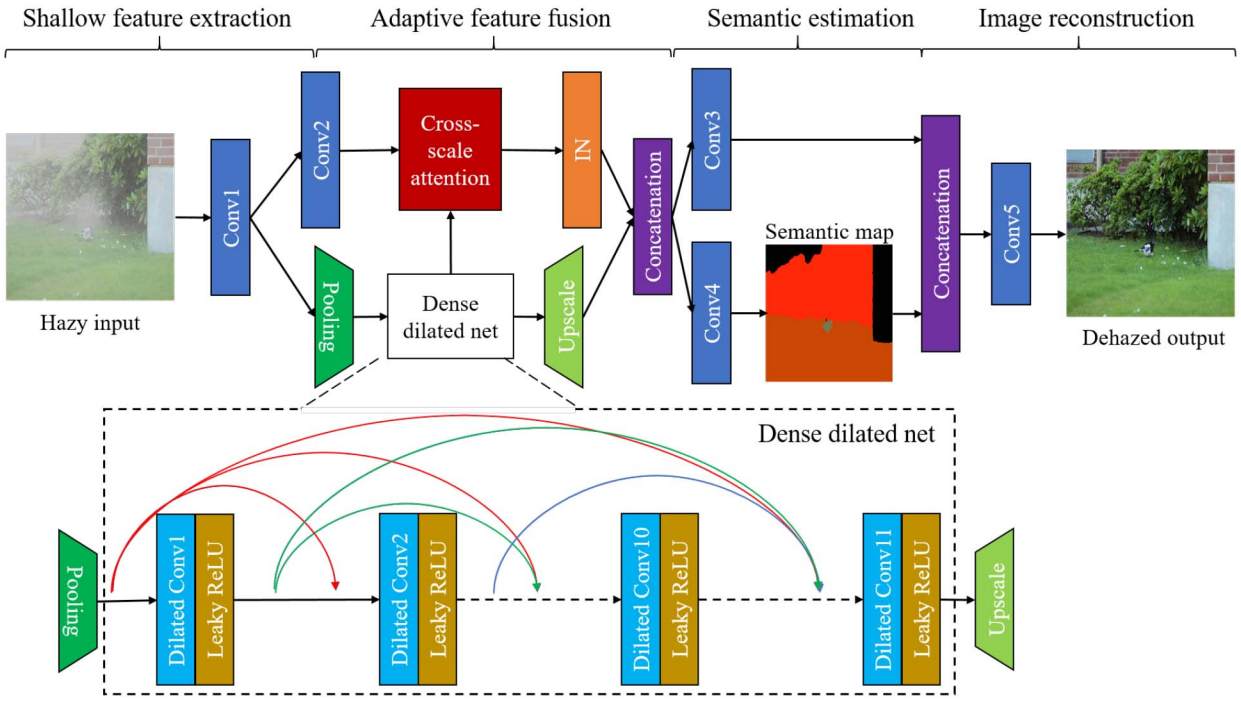


Fig. 2. Architecture of the proposed SDNet. The proposed network consists of four main parts: shallow feature extraction, dense dilated net for capturing global information and semantic prior, adaptive multiscale feature fusion module, and image reconstruction block. The first and last convolutional layers are shallow feature extractor and reconstruction layer, respectively.

was proposed to improve the dehazing quality. A radial basis function (RBF)-based dehazing method [30] was proposed through artificial neural networks dedicated to effectively removing haze effect while retaining the visible edges and the brightness of restored images. Huang *et al.* [31] introduced joint semantic learning for object detection, which improves the object detection performance in inclement weather conditions. Wang *et al.* [32] modeled the dehazing task as an image-to-image transform problem. Zhu *et al.* [33] designed a dehazing generator, which outputs the transmission map, air-light, and final dehazed result from the hazy input image. Zhang *et al.* [21] proposed a hierarchical density-aware dehazing network, which employs haze density to improve dehazing quality. Hong *et al.* [34] employed knowledge distillation for dehazing, which transfers the knowledge of clean image to student network. Dong *et al.* [35] designed a U-Net architecture to improve the dehazing performance by employing boosting and error feedback. Pang *et al.* [36] studied the binocular image dehazing problem and proposed a binocular image dehazing network (BidNet), which can remove haze both the right and left hazy images of binocular images.

The closest to ours is the research of [37] and [38], which employ semantic information to improve the dehazing performance. There exist four main differences between the proposed method and these two semantic-based dehazing approaches. First, the semantic segmentation of [37] and [38] is extracted by the pretrained VGGNet [39] and refineNet [40] on sharp images, respectively. In contrast, our semantic prior is trained on a hazy dataset, which can boost the accuracy of estimated semantic prior on real hazy images. Second, compared with SSD [37], our model can capture the relations between low-level and high-level features, which is critical for identifying the category of a pixel. Third, SSD captures image-level semantic information, while the proposed model captures pixel-level semantic information which keeps the spatial information. Finally, the method of [38] employs the semantic to estimate the transmission map. However, inaccurate transmission estimation would result in undesirable results. In contrast, our proposed model directly exploits the semantic to reconstruct the clean image from the hazy image.

*Semantic Knowledge Learning:* CNNs have been demonstrated to be effective in a lot of high-level tasks [41]–[46] and low-level tasks [47]–[49], which benefits from semantic knowledge learning. For example, VGGNet [39] increases the depth of layers for better feature extraction and becomes the basic model of many high-level works [50], [51]. To overcome the training difficulty of VGGNet, ResNet [41] was proposed by adding the residual connection, which achieves strong performance in many semantic knowledge learning tasks [52], [53]. Moreover, to overcome the heavy shrink of deep-learning model, the dilation convolution [54] was presented and shows effective in semantic segmentation. To better utilize the feature information in different layers, DenseNet [43] was developed by using the dense connection strategy, which receives the encouraging results in image classifications. In our work, we integrate the dilation and dense strategy to capture the global semantic features to guide the haze removal.

## III. Proposed Method

### A. Network

The architecture of the proposed SDNet is shown in Fig. 2. Let $I$ and $J$ denote as the hazy input and haze-free ground truth, respectively. The reconstructed image can be obtained

by $\mathcal{F}(I)$, where $\mathcal{F}(\cdot)$ denotes the function of our proposed SDNet.

As shown in Fig. 2, the first and last convolutional layers are shallow feature extractor and reconstruction layer, respectively. We propose a dense dilated net to extract hierarchical features, which can capture the global structure of the hazy input and preserve the main structure well. In addition, an adaptive feature fusion module is employed to fuse low- and high-level features for semantic estimation. Finally, we reconstruct the dehazed result capitalized on the intermediate features and segmentation probability map. We next introduce the details of the proposed SDNet.

*Dense Dilated Net:* Our dense dilated net is constructed by stacking several dilated convolutional layers shown in Fig. 2. To detect or identify the semantic of a pixel without contextual information is a challenging task since little information about the scene structure is available. To consider more contextual information, it essentially fuses the merits of dilated convolution and dense connections. The key point in the dense dilated net is how to grasp larger scope information (i.e., larger receptive field size) so that it is possible to obtain more sophisticated high-level semantic knowledge. One possible solution is to perform max pooling several times as used in conventional semantic segmentation methods. However, our dehazing task needs more pixel-level accurate results. Max pooling would lose lots of details of the image. To alleviate such drawbacks, we propose to use stacked dilated convolutions to enlarge receptive field size. In addition, low-level details in the image are important to recover the boundary. To this end, we employ dense connections to preserve the main structure of the input. Such connection learning tends to preserve more low-level features and allows us to form deep networks for high-quality image dehazing with stronger representation ability. Inspired by the above demands, the densely connected dilated network is naturally presented. In this way, it allows us to not only incorporate the semantic information into dehazing but also capture the global structure and local details well.

Specifically, in our model, the dense connections contain 11 dilated convolutional layers, and each layer receives different types of feature maps from previous layers. This can be formulated as

$$Fd_l = \mathrm{DConv}(C(F_0, Fd_1, \ldots, Fd_{l-1})),\ \ l = 1, 2, \ldots, 11 \quad (2)$$

where $Fd_l$ denotes the intermediate feature maps learned by the dense dilated net at the $l$th dilated convolutional layer, for example, $Fd_1$ is the output of Dilated Conv1, $F_0$ means the shallow features extracted from the first convolutional layer and downsampled by the average pooling layer (it makes the resolution of features reduce to 1/2 size of the input). In addition, DConv and $C$ are the dilated convolution and the concatenation operations, respectively. In this case, our network is trained on the high-resolution features to keep the mid- and high-frequency for the realistic appearance.

Compared with the typical dilated residual network [54], we use a more light network to keep the capacity to recover the detailed photographic appearance of scene objects. Furthermore, our model can capture multiscale objects well. For example, each dilated layer is equivalent to a kernel in different scales, for example, different receptive fields. Consequently, our model can obtain a feature map with many more scales, which helps the model capture multiscale objects well.

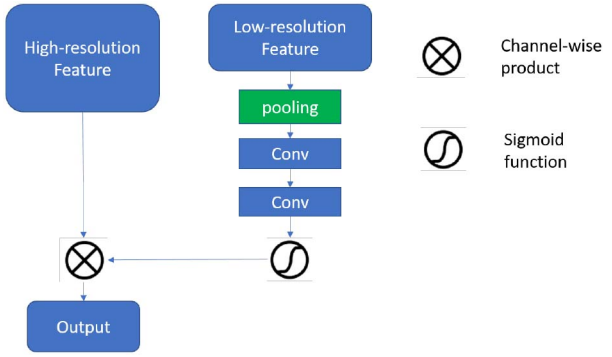


Fig. 3. Architecture of the proposed cross-scale attention module, which receives the input from high-resolution feature maps and low-resolution feature maps as inputs and outputs the most informative feature maps.

*Adaptive Feature Fusion Module:* After obtaining the high-level features $Fd_{11}$ from the dense dilated net in (2), we further estimate the segmentation probability map and dehazed result based on the extracted features. However, we note that only using the high-level feature maps from the dense dilated net would result in some artifacts in the estimated segmentation and the final dehazed result. Therefore, we propose an adaptive feature fusion module to fuse both low-level features (from the second convolutional layers, i.e., $F_2$) and high-level features (from the dense dilated network, i.e., $Fd_{11}$) to improve the performance of semantic segmentation. The proposed adaptive feature fusion module consists of three main parts: 1) a cross-scale attention module; 2) an instance normalization layer; and 3) a concat layer. We note that the semantic properties of deep layers are more abstract, while the shadow layers have more low-level features. Directly fusing these two types of features may cause feature incompatibility. To circumvent this issue, we present an adaptive feature fusion method to alleviate the gap between the high-level and low-level features by using instance normalization [55], [56]. This can be formulated as

$$F_{\mathrm{fuse}} = C(\sigma(CA(F_2, Fd_{11})), u \uparrow (Fd_{11})) \quad (3)$$

where $\sigma(\cdot)$ is the instance normalization operation, $u \uparrow$ is the upsampling layer, $CA$ denotes as cross-scale attention, and $F_2$ is the output of Conv2. There are several choices to serve as upscale modules, such as a transposed layer, bilinear upsampling, and nearest-neighbor upsampling. In this work, we directly use a simple bilinear upsampling layer to resize the output features from the dense dilated network.

The proposed cross-scale attention is designed to choose the most representative features and pass them to the instance normalization layer. As shown in Fig. 3, cross-scale attention receives features from shallow and deep layers and then uses the deep layer features to activate the most informative features from shallow layers. Conventional deep dehazing methods handle channelwise features equally, which is not suitable for our network. Our model is designed to use shallow features to compensate for the deep features. Especially,

we design cross-scale attention to determine which channels are helpful to restore the semantic segmentation and final dehazed image. The proposed cross-scale attention is different from the conventional channel attention [57]. First, channel attention employs the global pooling of the input features to determine the informative channelwise features, while our cross-scale attention employs the global pooling of the high-level features to determine the informative low-level features. Second, cross-scale attention is designed to choose the features which can compensate for high-level features, while channel attention only chooses the most informative features. The proposed cross-scale attention is different from the squeeze-and-excitation (SE) module [58] and nonlocal module [59]. First, the SE module and nonlocal module obtain new features from one feature map. In comparison, the proposed cross-scale attention performs between two feature maps that are far away from each other. Second, NL blocks compute the response at each position by attending to all other positions and computing a weighted average of the features in all positions, which incurs a large computation burden. In comparison, the proposed cross-scale attention extracts high-level features using a densely connected network and collects useful low-level features by cross-scale attention, which can capture global and local features and avoid a large computation burden.

The instance normalization in our work has two advantages in our work. First, we note that hazy inputs often lack of contrast due to attenuation, while we aim to enhance the contrast in the dehazed result. Therefore, the reconstructed result should not, in general, depend on the contrast of the hazy image. Fortunately, instance normalization is able to learn a highly nonlinear contrast normalization function for enhancing the contrast in the dehazed result. Second, instance normalization could address the problem of feature incompatibility caused by low- and high-level features [60]. With the fused features, our model could effectively take advantage of both low-level structural and high-level semantic information. As a result, the semantic segmentation map $M$ can be generated by

$$M = \text{Conv}(F_{\text{fuse}}) = \text{Conv}(C(\sigma(F_{ca}), u \uparrow (Fd_{11}))) \tag{4}$$

where $F_{ca} = CA(F_2, Fd_{11})$, which extracts most representative features for the next stage.

*Image Reconstruction:* In this block, we seek to use semantic prior to improve the dehazing quality. Objects in the scene tend to have limited color appearance for a given semantic prior. For example, trees and grasses tend to show a green appearance. Therefore, providing a category (based on the scene context) for a pixel may make the network generate a reasonable color appearance easily. Specifically, our basic idea is to use the constrain between color space and semantic segmentation probability maps. To achieve this, we design a module that models the dehazing as a posterior problem. The module can allow us to generate a clean image conditional on the segmentation probability maps by identifying which category the pixel belongs to. In this way, it shall provide us the additional information to remove haze. The generated dehazing feature maps and the segmentation probability maps are fused to restore the final dehazed result as

$$\mathcal{F}(I) = \text{Conv}(C(\text{Conv}(F_{\text{fuse}}), M)) \tag{5}$$

where $\mathcal{F}(I)$ means the reconstructed image by the proposed SDNet conditioned on the segmentation map.

### B. Loss Functions

To train our semantic-aware model, we use a pixelwise softmax classifier to predict a class label for each pixel. The class label will be used to generate the segmentation probability maps, defined as follows:

$$\mathcal{L}_{\text{sem}}(s, s^*) = -\frac{1}{P}\sum_i s_i^* \log(s_i) \tag{6}$$

where $P$ is the number of pixels in an image, $s_i = exp(z_i)/\sum_s exp(z_{i,s})$ is the class prediction at pixel $i$ given the output $z$ of the semantic module, and $s^*$ is the ground-truth semantic label. Moreover, for the dehazed result, we also define a reconstruction loss between the recovered image and the ground truth based on the $L_1$ norm

$$\mathcal{L}_{\text{rec}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathcal{F}(I_i, \Theta) - J_i\|_1 \tag{7}$$

where $N$ is the number of images in the training dataset, $\|\cdot\|_1$ is the $L_1$ norm, $J$ is the ground-truth haze-free image, and $\Theta$ keeps the weights of the learned filters.

In particular, to further improve the dehazing quality, we propose to exploit a smooth loss by restricting the predict results having the same gradient with ground truths, which is formulated as

$$\mathcal{L}_g = \frac{1}{N}\sum_{i=1}^{N} \|\nabla(\mathcal{F}(I_i, \Theta) - \nabla J_i\|_1 \tag{8}$$

where $\nabla$ denotes the gradient extraction operation.

Finally, by combining the semantic loss and reconstruction losses for dehazing, our final loss function is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_1 \mathcal{L}_{\text{sem}} + \lambda_2 \mathcal{L}_g \tag{9}$$

where $\lambda_1$ and $\lambda_2$ are the positive weights, which are used to control the importance degree of the corresponding loss.

### C. Training Dataset

There is no existing dataset that contains a hazy image, ground-truth haze-free image, and the semantic segmentation for training our network. To solve this issue, we make the following efforts. First, we collect 1200 images from the public segmentation dataset of SYSU-Scene [13], then we estimate the depth map for each image using the depth estimation method [67] and synthesize hazy images by following the protocol of learning-based dehazing methods [12], [19], [64]. Specifically, we choose ten random $\beta \in [0, 0.5]$ for $t(x) = e^{-\beta d(x)}$. We do not use a big $\beta \in [0.5, \infty]$, since such a setting generates a very small transmission, which may be not plausible for real case. As a result, we synthesize 12 000 hazy images and corresponding clean ones as well as ground-truth

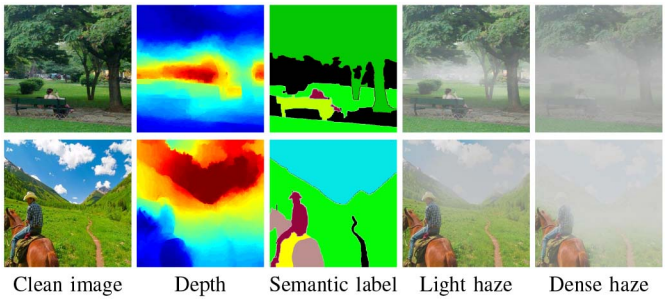


Fig. 4. Visual examples of the proposed dataset. The depth can represent the distance between the objects and camera, and the semantic label can represent the object class.

semantic segmentations in the training set. For the segmentation labels, we directly use the 37 labels defined in [13], including building, grass, car, person, sea, airplane, bag, ball, bench, bicycle, bird, boat, bottle, bus, camera, cat, cellphone, chair, cow, cup, dog, glasses, horse, laptop, motorbike, racket, rail, sheep, sky, sofa, street/road, suitcase, table, train, tree, tv/monitor, and umbrella. We show the generated dense and light hazy images accompanied with semantic label and depth map in Fig. 4.

### D. Implementation Details

It has been shown that identity initialization [68] is better than Gaussian random variables for dilation layers. In our implementation, we initialize the weights of dilation layers using the identity initializer. We utilize Leaky ReLU after each convolutional and dilated convolutional layer as an activation function. During training, all modules are trained end to end using the loss function defined in (9). ADAM is employed to train the proposed model as the optimization solver and set the initialized learning rate as 0.0001 to train our network. We decrease the learning rate by 0.5 every 25 epochs. We train our model for 150 epochs, which takes about 48 h on an RTX 2080Ti. We use a batch size of 2 with a patch size of $300 \times 300$ cropped randomly from hazy images. All experiments were conducted using Python 3.6 and PyTorch 1.1. We set $\lambda_1 = 0.5$ and $\lambda_2 = 0.001$ in our experiments. In addition, we use dilation rates of 1, 2, 4, 8, 16, 32, 16, 8, 4, 2, and 1 for the 11 dilated convolutional layers of the densely connected dilated network, respectively, which help the net to leverage more context to capture structure information well.

## IV. Experiments

In this section, an ablation study is conducted to show the improvements obtained by the proposed modules in the model, and then we compare our method against state-of-the-art dehazing methods, including AOD-Net [62], GFN [69], DCPDN [19], BCCR [16], CAP [15], GRM [61], DCP [14], NLD [8], DehazeNet [11], MSCNN [12], PDN [63], EPDN [22], GridDehazeNet [25], BPPNet [70], PhysicsGan [71], FFA-Net [72], and MSBDN [35] on hazy images. (We will release the source code, trained model, as well as the dataset on our project website.)

TABLE I
Quantitative Comparison Results on the HazeRD Dataset Using Different Proposed Module

| | w/o DCB | w/o CSA | w/o AFF | w/o semantic | SDNet |
|---|---|---|---|---|---|
| PSNR | 20.10 | 22.31 | 22.20 | 21.51 | 22.50 |
| SSIM | 0.854 | 0.909 | 0.901 | 0.879 | 0.917 |

TABLE II
Average PSNR and SSIM of Hazy Images From Outdoor Images From RESIDE

| | O-HAZE | Dense-Haze | HazeRD | Our dataset |
|---|---|---|---|---|
| PSNR | 19.35 | 18.50 | 20.20 | 24.56 |
| SSIM | 0.832 | 0.812 | 0.850 | 0.914 |

TABLE III
Average PSNR and SSIM of Dehazed Results on the HazeRD Dataset With Different Batch Sizes

| | 1 | 2 |
|---|---|---|
| PSNR | 21.71 | 22.50 |
| SSIM | 0.866 | 0.917 |

### A. Ablation Study

To demonstrate the improvements of each component introduced in our network, we conduct an ablation study on the HazeRD dataset using four variant methods: 1) full model without adaptive feature fusion module (w/o AFF); 2) full model without semantic prior (w/o semantic); 3) full model without cross-scale attention (w/o CSA) module; and 4) full model without densely connected block (w/o DCB). The compared results are listed in Table I. It is observed that densely connected block is critical for improving the dehazing performance, which enlarges the receptive field size of the model. Semantic information restricts the solve space and improves the dehazing performance. The adaptive feature fusion module and cross-scale attention module contribute to the performance improvements. In addition, from Fig. 5, we can observe the adaptive feature fusion module can be used to improve the contribution of low-level features and make the dehazing result smoother and remove the artifacts and color distortion [see the sky region in (c)]; while the semantic prior is beneficial to recover a semantic reasonable result and help the model generate a clearer result [observed on the building in (d)].

To reveal the influence of training data, we investigate the performance of models trained on different datasets. As suggested, we obtain a semantic segmentation by EncNet [53] for O-HAZE, Dense-Haze, and Hazerd. Then, we trained the proposed model on these datasets. Finally, we evaluate the performance of the trained models on outdoor haze images from RESIDE. The performance of models is listed in Table II. As shown in Table II, the model trained on the proposed dataset has a good generalization ability. We also show one example for each dataset in Fig. 6.

We experiment with the influence of batch size. We use batch sizes 1 and 2 to train dehazing network. We list the performance of different batch sizes in Table III. When the

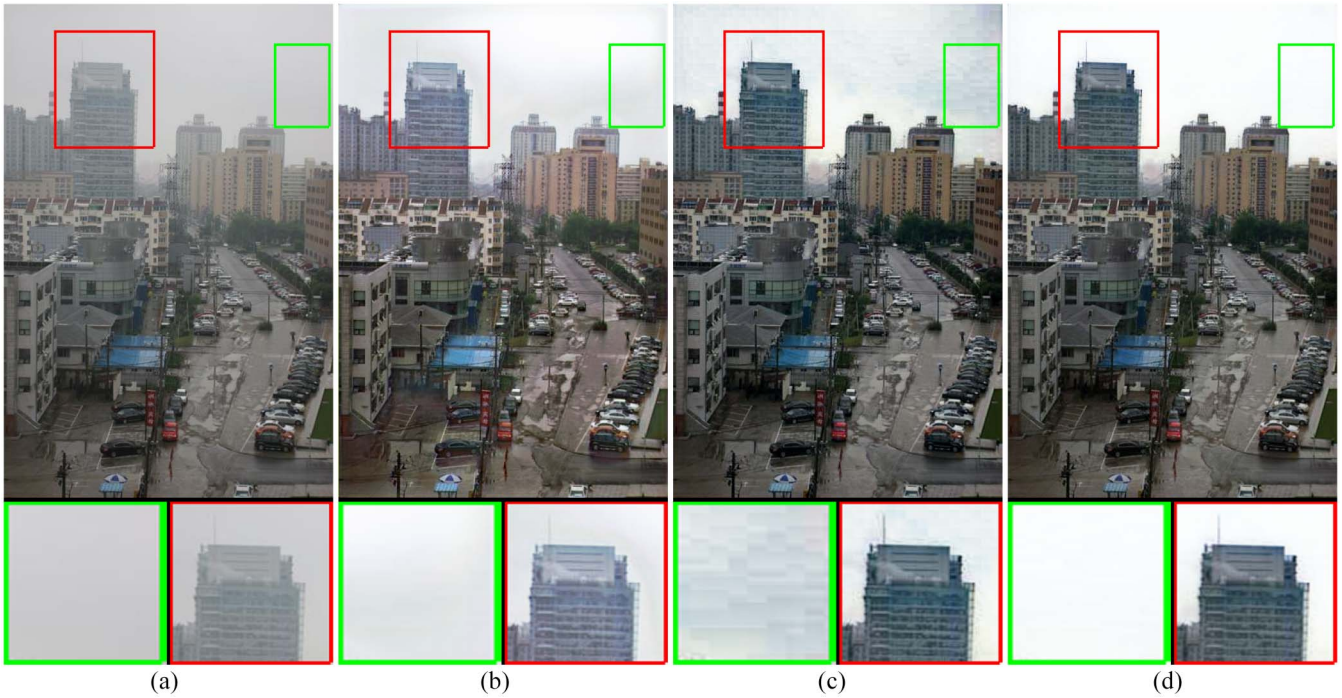


Fig. 5. Visual comparisons of dehazed results with different modules on a challenging real-world hazy example. Semantic information helps the network recover faithful color information and adaptive feature fusion module could generate smooth result and avoid color distortion and artifacts. (a) Input. (b) w/o semantic. (c) w/o AFF. (d) SDNet.

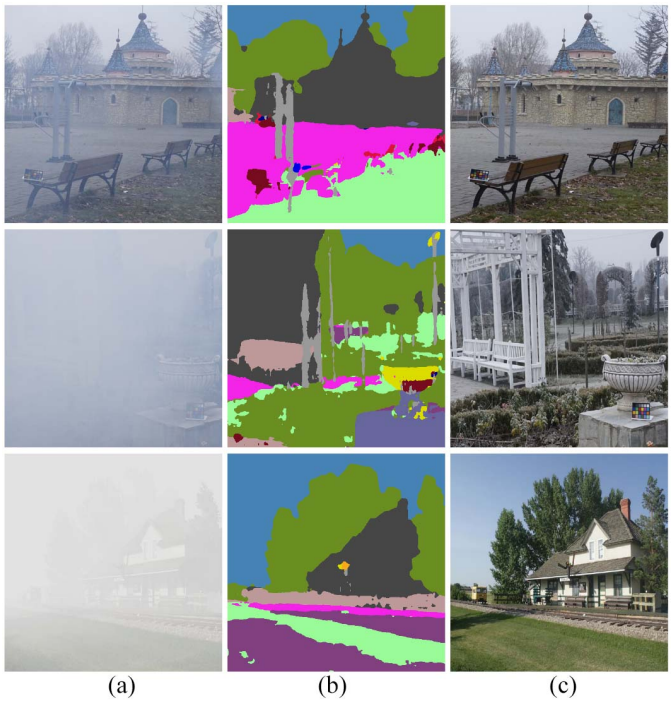


Fig. 6. Visual results of different training dataset. The first row is from O-HAZE. The second row is from Dense-Haze. The third row is from Hazerd. (a) Hazy image. (b) Semantic label. (c) GTs.

batch size is set to 1, the dehazing result tends to show overenhancement. When the batch size is set to 2, the result tends to show a normal dehazing result. Based on the experiment on real hazy image and simulated hazy images, we trained the proposed model with batch size 2.

TABLE IV
QUANTITATIVE COMPARISON RESULTS ON THE TEST PART OF THE PROPOSED DATASET WITH DIFFERENT ACCURATE OF ESTIMATED SEMANTIC MAP

| Accurate | 40.51 | 60.64 | 69.46 |
|---|---|---|---|
| PSNR | 22.51 | 24.62 | 25.13 |

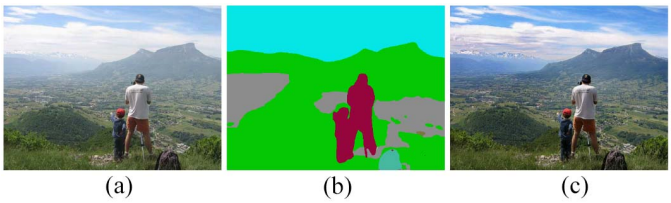


Fig. 7. Visual results of dehazed result and semantic label. (a) Input. (b) Semantic label. (c) Dehazed result.

The proposed method is depending on the information provided by the estimated semantic map. We do an experiment to show the relation between the accuracy of the estimated semantic map and dehazing performance. We test the segmentation and dehazing performance on the test part of the proposed dataset, the pixelwise accurate of the proposed method is 69.46% and the PSNR is 25.13, which shows that the dehazed result can be boosted from the coarse estimated semantic map. As shown in Table IV, we can observe that the dehazing PSNR increases with the increase of the accuracy of the estimated semantic map. The estimated semantic map is a rough structure representing the scene, which is helpful for dehazing. As shown in Fig. 7, some areas (marked as gray) are labeled as road, however, these areas still achieve a reasonable dehazed result.

TABLE V
QUANTITATIVE COMPARISON RESULTS OF VARIOUS DEHAZING METHODS AND PROPOSED METHOD ON OUTDOOR IMAGES FROM THE RESIDE DATASET [64], HAZERD DATASET [65], AND O-HAZE DATASET [66]

| | | NLD | MSCNN | DehazeNet | DCP | FVR | CAP | AOD-Net | DCPDN | GFN | PDNet | EPDN | GridDehazeNet | SDNet |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RESIZED | PSNR | 16.85 | 18.64 | 22.46 | 19.13 | 15.39 | 22.27 | 19.06 | 19.93 | 21.55 | 20.89 | 23.82 | 23.49 | **24.56** |
| | SSIM | 0.78 | 0.78 | 0.85 | 0.81 | 0.72 | 0.90 | 0.90 | 0.84 | 0.84 | 0.85 | 0.89 | 0.87 | **0.91** |
| HazeRD | PSNR | 18.82 | 19.10 | 19.53 | 17.66 | 16.17 | 18.56 | 18.13 | 18.82 | 19.18 | 20.14 | 18.92 | 14.89 | **22.50** |
| | SSIM | 0.84 | 0.85 | 0.85 | 0.84 | 0.85 | 0.83 | 0.83 | 0.89 | 0.86 | 0.89 | 0.86 | 0.74 | **0.92** |
| O-HAZE | PSNR | 16.61 | 19.07 | 16.21 | 16.59 | 14.92 | 17.62 | 16.72 | 15.62 | 18.30 | 18.52 | 16.44 | 12.95 | **19.38** |
| | SSIM | 0.75 | 0.77 | 0.67 | 0.74 | 0.70 | 0.71 | 0.68 | 0.62 | 0.72 | 0.75 | 0.71 | 0.31 | **0.77** |

TABLE VI
QUANTITATIVE COMPARISON RESULTS OF VARIOUS DEHAZING METHODS AND PROPOSED METHOD ON THE INDOOR HAZY IMAGES OF SOTS TEST DATA FROM THE RESIDE DATASET [64]

| | | DCP | FVR | BCCR | CAP | NLD | GRM | MSCNN | DehazeNet | AOD-Net | DCPDN | GFN | PDNet | SDNet |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| w/o noise | PSNR | 16.62 | 15.72 | 16.88 | 19.05 | 17.29 | 18.86 | 17.57 | 21.14 | 19.06 | 15.86 | 22.30 | 22.83 | **24.91** |
| | SSIM | 0.82 | 0.75 | 0.79 | 0.84 | 0.86 | 0.81 | 0.75 | 0.85 | 0.85 | 0.82 | 0.88 | 0.91 | **0.94** |
| 1% noise | PSNR | 17.13 | 15.68 | 17.89 | 18.69 | 16.44 | 19.00 | 17.16 | 20.75 | 18.96 | 15.82 | 22.21 | 21.80 | **22.85** |
| | SSIM | 0.65 | 0.53 | 0.67 | 0.71 | 0.6211 | 0.79 | 0.71 | 0.70 | 0.73 | 0.80 | 0.82 | 0.78 | **0.85** |
| 3% noise | PSNR | 17.03 | 15.53 | 16.78 | 17.39 | 16.14 | 18.33 | 17.06 | 19.94 | 18.06 | 15.82 | 20.97 | 20.80 | **21.94** |
| | SSIM | 0.64 | 0.51 | 0.65 | 0.70 | 0.60 | 0.75 | 0.69 | 0.68 | 0.70 | 0.75 | 0.76 | 0.75 | **0.82** |

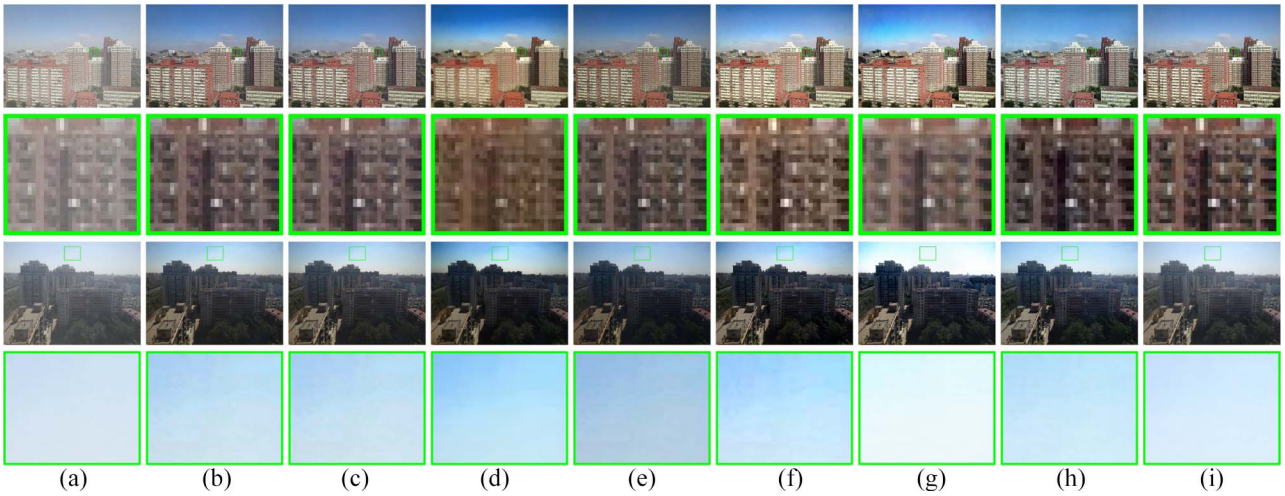


Fig. 8. Visual comparisons of dehazed results of various dehazing methods and proposed method on the RESIDE dataset. (a) Input. (b) DCP [12]. (c) CAP [13]. (d) GRM [58]. (e) AOD-Net [59]. (f) PDNet [60]. (g) DCPDN [17]. (h) SDNet. (i) GTs.

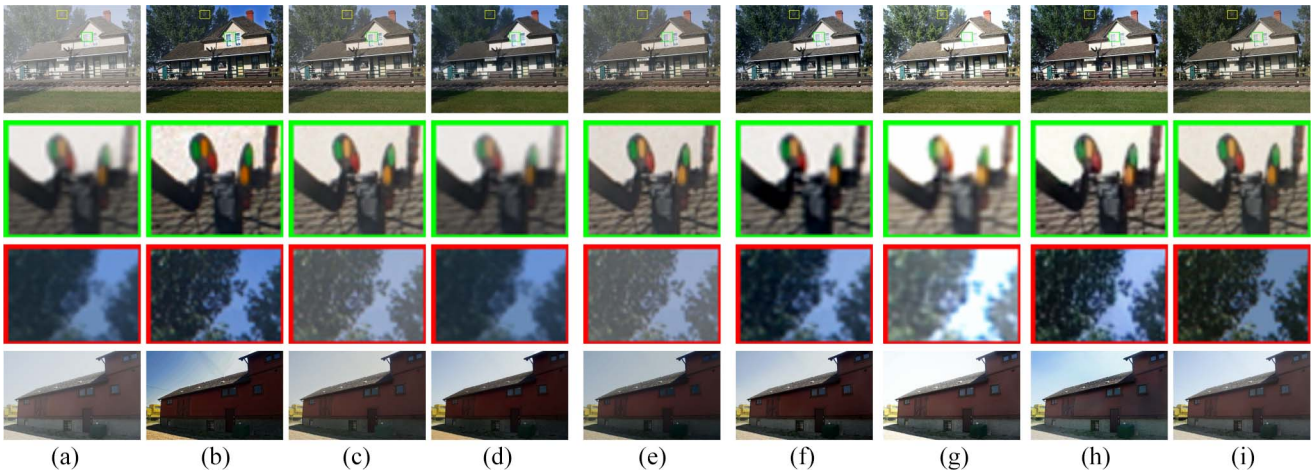


Fig. 9. Visual comparisons of dehazed results of various dehazing methods and proposed method on the HazeRD dataset. (a) Input. (b) DCP [12]. (c) CAP [13]. (d) GRM [58]. (e) AOD-Net [59]. (f) PDNet [60]. (g) DCPDN [17]. (h) SDNet. (i) GTs.

### *B. Quantitative Evaluations on Benchmarks*

We evaluate the proposed network on the public dahazing test dataset HazeRD [65], RESIDE [64], and O-HAZE [66]. All these datasets contain the ground-truth haze-free images, which can make us able to evaluate the performance of dehazing methods qualitatively. We compare the proposed method

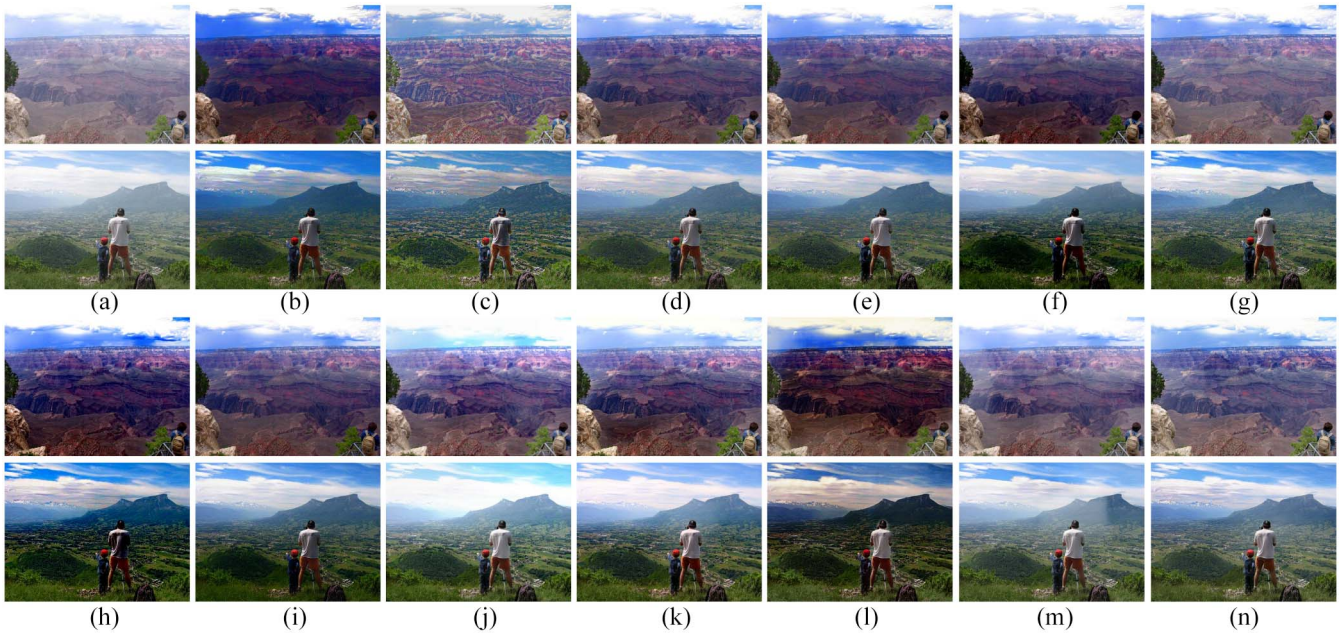


Fig. 10. Visual comparisons of dehazed results of various dehazing methods and proposed method on real-world hazy images. It can be observed that dehazed results of the proposed method are much clearer than the results of other state-of-the-art methods. (a) Hazy input. (b) DCP [12]. (c) FVR [70]. (d) RF [71]. (e) CAP [13]. (f) DehazeNet [9]. (g) MSCNN [10]. (h) NLD [8]. (i) AOD-Net [59]. (j) DCPCN [17]. (k) GFN [66]. (l) EPDN [19]. (m) GridDehazeNet [22]. (n) SDNet.

with the state-of-the-art methods [8], [11], [12], [14]–[16], [19], [62], [63], [69], [73] using PSNR and SSIM.

*Evaluations on the RESIDE Dataset [64]:* RESIDE is a large-scale hazy image dataset, which includes two simulated test datasets of indoor and outdoor scenes, respectively. We first quantitatively compare the proposed SDNet with other state-of-the-art methods in Table V. As indicated, the proposed model outperforms the competitor in terms of SSIM and PSNR metrics. For example, compared with PDNet, SDNet achieves better results by up to 3.67 dB and 0.06 in terms of PSNR and SSIM, respectively. The proposed algorithm also advances 0.76 dB in terms of PSNR than the very recent work of EPDN.

We also show visual comparisons of various methods on the RESIDE dataset in Fig. 8. As shown, most compared dehazing methods tend to remain some haze in the dehazed results. In contrast, the proposed SDNet obtains clearer details and vivid colors in outdoor scenes as shown in Fig. 8.

In order to conduct a fair comparison, we trained the proposed model on indoor hazy images from the RESIDE dataset as other learning-based methods and report the result on the tested dataset of RESIDE. As shown in Table VI, we can see that the proposed model achieved the highest dehazing performance.

*Evaluations on the HazeRD Dataset [65]:* The HazeRD dataset contains natural outdoor images and the corresponding high-accuracy depth maps, therefore, can simulate more realistic haze to evaluate the performance of dehazing methods. Note that the images in HazeRD are not used by all of the CNN-based methods as training data. Table V shows the compared results. It can be seen that the proposed SDNet obtains the highest SSIM and PNSR on the HazeRD testing data. In particular, our algorithm exceeds the second best method (PDNet [63]) by up to 2.36 dB and 0.03 in terms of PSNR and SSIM, respectively.

We further show two examples from the HazeRD dataset in Fig. 9. As shown, our method generates more close results to the ground-truth haze-free images than other state-of-the-art methods. From the zoomed-in area, we can see that the leaves generated by CAP, AOD-Net, and DCPDN seem to remain in some haze and the details are lost. The results by GRM seem to show some blurry artifacts. In addition, the window area in the dehazed result by DCP shows some color distortions. Compared with the state-of-the-art methods, our algorithm produces more visually pleasant results that are similar to ground truths.

*Evaluations on the O-HAZE Dataset [66]:* Although we have evaluated the performance of SDNet on the RESIDE and HazeRD datasets, we note that synthetic hazy images from these two benchmarks are different from real scenes. Different from RESIDE and HazeRD, hazy images in O-HAZE are captured in the presence of real haze produced by haze machines. We quantitatively compare our method against the state-of-the-art dehazing approaches in Table V. We can observe that our model achieves the best performance in terms of SSIM and PSNR.

### C. Evaluations on Real-World Hazy Images

We further qualitatively evaluate the proposed SDNet on the natural hazy images from [74]. Fig. 10 shows several real-world hazy images and the dehazed results generated by the proposed approach and state-of-the-art dehazing methods [14], [73], [74], [15], [11], [12], [8], [19], [69], [22], [25].

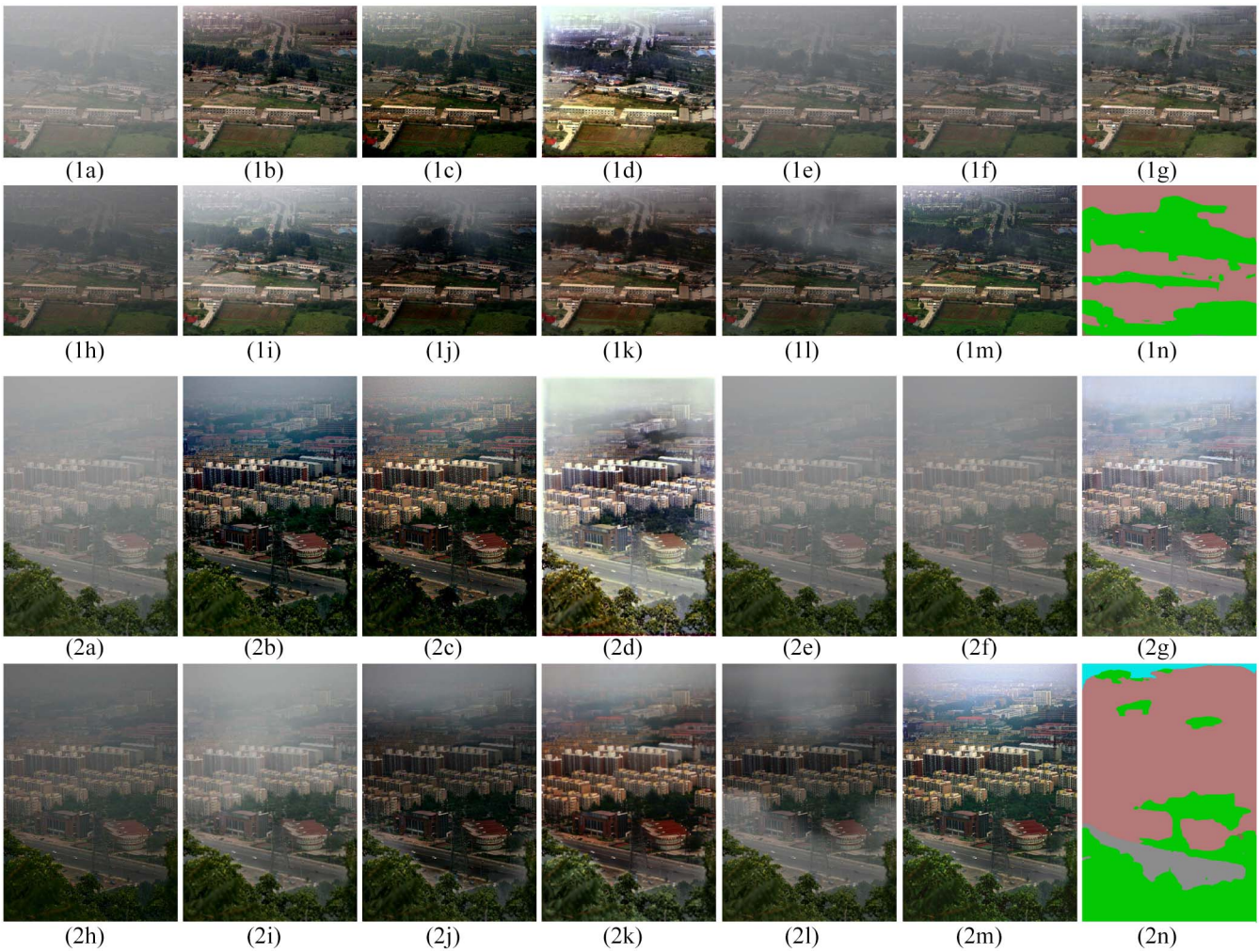


Fig. 11. Visual comparisons of dehazed results of various dehazing methods and proposed method on real-world hazy images. It can be observed that dehazed results of the proposed method are much clearer than the results of other state-of-the-art methods. (1a) Hazy input. (1b) NLD [8]. (1c) MSCNN [10]. (1d) BPPNet [67]. (1e) FFA-Net [69]. (1f) MSBDN [33]. (1g) PhysicsGan [68]. (1h) AOD-Net [59]. (1i) DCPCN [17]. (1j) GFN [66]. (1k) EPDN [19]. (1l) GridDehazeNet [22]. (1m) SDNet. (1n) Semantic. (2a) Hazy input. (2b) NLD [8]. (2c) MSCNN [10]. (2d) BPPNet [67]. (2e) FFA-Net [69]. (2f) MSBDN [33]. (2g) PhysicsGan [68]. (2h) AOD-Net [59]. (2i) DCPCN [17]. (2j) GFN [66]. (2k) EPDN [19]. (2l) GridDehazeNet [22]. (2m) SDNet. (2n) Semantic.

It can be observed that the traditional dehazing methods of DCP [14], FVR [73], and RF [74] fail to generate clear images and tend to introduce some color distortion as shown in Fig. 10(b)–(d). NLD [8] and EPDN [22] overestimate the haze density and obtain darker results than others such as the second and third images in Fig. 10(h) and (l). MSCNN [12] and GridDehazeNet [25] tend to leave haze in the results and methods of DehazeNet [11] tend to result in color distortion (which can be found in the area of the sky region of the first row).

In addition, we note that atmospheric model-based dehazing methods of [11], [12], and [19] use the conventional atmospheric model in (1) to recover clear images. However, due to the imperfect estimated transmission maps, the final recovered images contain some artifacts, as shown in Fig. 10(f), (g), and (j). Furthermore, the end-to-end deep-learning networks proposed in [62] and [69] use a CNN to directly predict haze-free images from hazy inputs. However, these methods fail to regain clean images as shown in Fig. 10(i) and (k).

In contrast, the proposed SDNet utilizes the semantic information and alleviates the traditional atmospheric constraint in (1). Thus, our model captures the global structure and reduces the gap between the synthetic and real-world hazy images, which facilitates haze removal and avoids artifacts. It can be observed that the results generated by our algorithm in Fig. 10(n) are much clearer than the ones generated by other algorithms.

We compare the proposed method with some recently state-of-the-dehazing methods [35], [70]–[72] in Fig. 11. As shown in Fig. 11(1d) and (2d), BPPNet [70] tend to show white appearance. As shown in Fig. 11(1e), (2e), (1f), and (2f), FFA-Net [72] and MSBDN [35] tend to retain haze in dehazed result. As shown in Fig. 11(1g) and (2g), PhysicsGan [71] tend to show a dark appearance. In contrast, the proposed method can obtain a visual pleased dehazed result as shown in 11(1m) and (2m). In addition, we also show the estimated semantic label in Fig. 11(1n) and (2n) for the input hazy images.

To further evaluate the proposed method on real-world images, we compare our method with recent

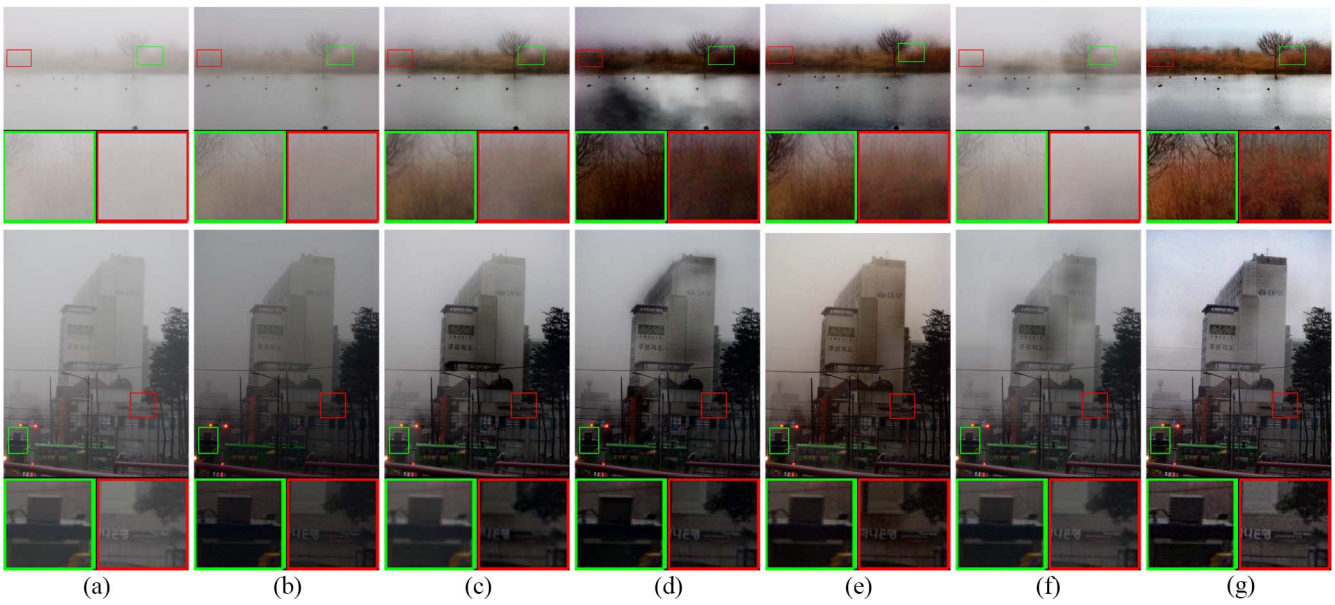


Fig. 12. Visual comparisons of dehazed results of various dehazing methods and proposed method on real-world dense hazy images. AOD-Net, DCPDN, and GridDehazeNet leave some haze, while GFN and EPDN generate some artifacts in the dehazed results. In contrast, our method removes haze moderately and preserves the image details well. (a) Dense hazy input. (b) AOD-Net [62]. (c) DCPDN [19]. (d) GFN [69]. (e) EPDN [22]. (f) GridDehazeNet [25]. (g) SDNet.

TABLE VII
AVERAGE RUNNING TIME ON THE RESIDE DATASET

| Method | DCP | BCCR | NLD | MSCNN | PDNet | DehazeNet | GFN | DCPDN | EPDN | Our |
|---|---|---|---|---|---|---|---|---|---|---|
| Language | Matlab | | | | | | | Python | | |
| Platform | | | | MatConvNet | | Caffe | | pytorch | | PyTorch |
| CPU (s) | 25.08 | 3.52 | 8.41 | 2.45 | 4.02 | 3.09 | 19.03 | 3.24 | 3.51 | 1.22 |

deep-learning-based dehazing methods on dense hazy images. As shown in Fig. 12, AOD-Net [62], DCPDN [19], and GridDehazeNet [25] cannot remove haze effectively in heavily hazy scenes. GFN [69] and EPDN [22] generate some color distortions in the dehazed results. In contrast, the proposed method yields visually pleasing results and removes haze as much as possible.

### D. Runtime

We show the runtime of state-of-the-art image dehazing methods and the proposed method on the same machine (8-GB memory and i5-6300HQ CPU@2.3 GHz) without using GPU implementation. We select 100 images from the RESIDE dataset [64]. Table VII shows the average running time of all the methods. The traditional algorithms of DCP [14], BCCR [16], and NLD [8] are time consuming due to a complex optimization process. Therefore, MSCNN [12], DehazeNet [11], GFN [69], PDNet [63], and DCPDN [19] utilize CNNs to estimate haze-free images. However, they are still time consuming since the traditional atmospheric model-based recovering method or the complicated networks. The results in Table VII show the high efficiency of our proposed method.

## V. CONCLUSION

In this article, we have proposed a novel SDNet to learn the semantic prior for a single image dehazing task. Our method models the dehazing problem as a maximizing the probability of color conditioned on the semantic information, which is achieved by obtaining a semantic prior with a dense dilated network. Thus, the proposed SDNet is capable of generating distinct and vivid colors by incorporating the categorical labels into the dehazing network. To efficiently estimate semantic prior, we present a densely connected dilated network, which can leverage more contextual information and capture scene structures. In addition, we propose an adaptive feature fusion module to fuse the multiscale features and adopt the instance normalization to remove artifacts and smooth the dehazed result. Extensive experiments on both synthetic and real-world datasets demonstrate that the proposed algorithm performs favorably against the state-of-the-art dehazing methods.

Our work currently relies on 37 semantic categories given in the SYSU-Scene dataset [13], and thus does not consider semantic priors of finer categories, such as valley, trucks, bridge, and river. In such a case, it puts forward challenging requirements for segmentation tasks from a hazy input. In future work, we will address this issue by considering more semantic categories.

## REFERENCES

[1] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 973–992, 2018.
[2] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," *Int. J. Comput. Vis.*, vol. 48, no. 3, pp. 233–254, 2002.

[3] J. Kopf *et al.*, "Deep photo: Model-based photograph enhancement and viewing," *ACM Trans. Graph.*, vol. 27, p. 116, Dec. 2008.

[4] R. T. Tan, "Visibility in bad weather from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[5] R. Fattal, "Single image dehazing," *ACM Trans. Graph.*, vol. 27, no. 3, p. 72, 2008.

[6] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1956–1963.

[7] R. Fattal, "Dehazing using color-lines," *ACM Trans. Graph.*, vol. 34, no. 1, p. 13, 2014.

[8] D. Berman, T. Treibitz, and S. Avidan, "Non-local image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1674–1682.

[9] X. Zhang, R. Jiang, T. Wang, and W. Luo, "Single image dehazing via dual-path recurrent network," *IEEE Trans. Image Process.*, vol. 30, pp. 5211–5222, 2021.

[10] X. Zhang, T. Wang, W. Luo, and P. Huang, "Multi-level fusion and attention-guided CNN for image dehazing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4162–4173, Nov. 2021.

[11] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, pp. 5187–5198, 2016.

[12] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 154–169.

[13] L. Lin, G. Wang, R. Zhang, R. Zhang, X. Liang, and W. Zuo, "Deep structured scene parsing by learning with image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 2276–2284.

[14] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[15] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Trans. Image Process.*, vol. 24, pp. 3522–3533, 2015.

[16] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 617–624.

[17] B.-H. Chen and S.-C. Huang, "Edge collapse-based dehazing algorithm for visibility restoration in real scenes," *J. Display Technol.*, vol. 12, no. 9, pp. 964–970, Sep. 2016.

[18] S. E. Kim, T. H. Park, and I. K. Eom, "Fast single image dehazing using saturation based transmission map estimation," *IEEE Trans. Image Process.*, vol. 29, pp. 1985–1998, 2019.

[19] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3194–3203.

[20] X. Yang, Z. Xu, and J. Luo, "Towards perceptual image dehazing by physics-based disentanglement and adversarial training," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7485–7492.

[21] J. Zhang *et al.*, "Hierarchical density-aware dehazing network," *IEEE Trans. Cybern.*, early access, May 7, 2021, doi: 10.1109/TCYB.2021.3070310.

[22] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 8160–8168.

[23] W.-T. Chen, J.-J. Ding, and S.-Y. Kuo, "PMS-net: Robust haze removal based on patch map for single images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11681–11689.

[24] Y. Li *et al.*, "LAP-net: Level-aware progressive network for image dehazing," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 3275–3284.

[25] X. Liu, Y. Ma, Z. Shi, and J. Chen, "GridDehazeNet: Attention-based multi-scale network for image dehazing," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7313–7322.

[26] L. Zhang, L. Song, B. Du, and Y. Zhang, "Nonlocal low-rank tensor completion for visual data," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 673–685, Feb. 2021.

[27] S. Zhang, F. He, and W. Ren, "NLDN: Non-local dehazing network for dense haze removal," *Neurocomputing*, vol. 410, pp. 363–373, Oct. 2020.

[28] Z. Deng *et al.*, "Deep multi-model fusion for single-image dehazing," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 2453–2462.

[29] C. Li, C. Guo, J. Guo, P. Han, H. Fu, and R. Cong, "PDR-net: Perception-inspired single image dehazing network with refinement," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 704–716, Mar. 2020.

[30] B.-H. Chen, S.-C. Huang, C.-Y. Li, and S.-Y. Kuo, "Haze removal using radial basis function networks for visibility restoration applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3828–3838, Aug. 2018.

[31] S.-C. Huang, T.-H. Le, and D.-W. Jaw, "DSNet: Joint semantic learning for object detection in inclement weather conditions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2623–2633, Aug. 2021.

[32] A. Wang, W. Wang, J. Liu, and N. Gu, "AIPNet: Image-to-image single image dehazing with atmospheric illumination prior," *IEEE Trans. Image Process.*, vol. 28, pp. 381–393, Jan. 2019.

[33] H. Zhu *et al.*, "Single-image dehazing via compositional adversarial network," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 829–838, Feb. 2021.

[34] M. Hong, Y. Xie, C. Li, and Y. Qu, "Distilling image dehazing with heterogeneous task imitation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3459–3468.

[35] H. Dong *et al.*, "Multi-scale boosted dehazing network with dense feature fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2154–2164.

[36] Y. Pang, J. Nie, J. Xie, J. Han, and X. Li, "BidNet: Binocular image dehazing without explicit disparity estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 5930–5939.

[37] Z. Cheng, S. You, V. Ila, and H. Li, "Semantic single-image dehazing," 2018, *arXiv:1804.05624*.

[38] W. Ren *et al.*, "Deep video dehazing with semantic segmentation," *IEEE Trans. Image Process.*, vol. 28, pp. 1895–1908, 2019.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2014. [Online]. Available: arXiv:1409.1556.

[40] G. Lin, F. Liu, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for dense prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1228–1242, May 2020.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[42] X. Tan, K. Xu, Y. Cao, Y. Zhang, L. Ma, and R. W. H. Lau, "Night-time scene parsing with a large real dataset," *IEEE Trans. Image Process.*, early access, Oct. 27, 2021, doi: 10.1109/TIP.2021.3122004.

[43] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[44] D. Lin, R. Zhang, Y. Ji, P. Li, and H. Huang, "SCN: Switchable context network for semantic segmentation of RGB-D images," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1120–1131, Mar. 2020.

[45] F. Zheng, H. Tang, and Y.-H. Liu, "Odometry-vision-based ground vehicle motion estimation with SE(2)-constrained SE(3) poses," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2652–2663, Jul. 2019.

[46] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran, 2015, pp. 91–99.

[47] Y. Chen, X. Cao, Q. Zhao, D. Meng, and Z. Xu, "Denoising hyperspectral image with non-i.i.d. noise structure," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 1054–1066, Mar. 2018.

[48] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo, "MADNet: A fast and lightweight network for single-image super resolution," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1443–1453, Mar. 2021.

[49] K. Zeng, J. Yu, R. Wang, C. Li, and D. Tao, "Coupled deep autoencoder for single image super-resolution," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 27–37, Jan. 2017.

[50] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3203–3212.

[51] X. Tan, H. Zhu, Z. Shao, X. Hou, Y. Hao, and L. Ma, "Saliency detection by deep network with boundary refinement and global context," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2018, pp. 1–6.

[52] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[53] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7151–7160.

[54] F. Yu, V. Koltun, and T. A. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 636–644.

[55] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.

[56] Z. Xu, X. Yang, X. Li, X. Sun, and P. Harbin, "Strong baseline for single image dehazing with deep features and instance normalization," in *Proc. Brit. Mach. Vis. Conf.*, 2018, p. 5.

[57] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. ECCV*, 2018, pp. 294–310.
[58] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[59] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[60] N. Ibtehaz and M. S. Rahman, "MultiResuNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, Jan. 2020.
[61] C. Chen, M. N. Do, and J. Wang, "Robust image and video dehazing with visual artifact suppression via gradient residual minimization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 576–591.
[62] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "AOD-Net: All-in-one dehazing network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4780–4788.
[63] D. Yang and J. Sun, "Proximal Dehaze-Net: A prior learning-based deep network for single image dehazing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 702–717.
[64] B. Li *et al.*, "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, pp. 492–505, 2019.
[65] Y. Zhang, L. Ding, and G. Sharma, "Hazerd: An outdoor scene dataset and benchmark for single image dehazing," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3205–3209.
[66] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-HAZE: A dehazing benchmark with real hazy and haze-free outdoor images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 754–762.
[67] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.
[68] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
[69] W. Ren *et al.*, "Gated fusion network for single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3253–3261.
[70] A. Singh, A. Bhave, and D. Prasad, "Single image dehazing for a variety of haze scenarios using back projected pyramid network," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 166–181.
[71] J. Pan *et al.*, "Physics-based generative adversarial models for image restoration and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2449–2462, Jul. 2021.
[72] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-net: Feature fusion attention network for single image dehazing," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, pp. 11908–11915.
[73] J.-P. Tarel and N. Hautière, "Fast visibility restoration from a single color or gray level image," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 2201–2208.
[74] K. Tang, J. Yang, and J. Wang, "Investigating haze-relevant features in a learning framework for image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 2995–3000.

**Shengdong Zhang** received the B.S. degree in computer science and technology from Shanxi University, Taiyuan, Shanxi, China, in 2009, the M.S. degree in mechanics from Peking University, Beijing, China, in 2012, and the Ph.D. degree from Wuhan University, Wuhan, China, in 2020.

His research interests include deep learning, computer vision, and image processing.

**Wenqi Ren** (Member, IEEE) received the Ph.D. degree from Tianjin University, Tianjin, China, in 2017.

He is an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. From 2015 to 2016, he was supported by the China Scholarship Council and working with Prof. M.-H. Yang as a joint-training Ph.D. student with the Electrical Engineering and Computer Science Department, University of California at Merced, Merced, CA, USA. His research interests include image processing and related high-level vision problems.

Dr. Ren received the Tencent Rhino Bird Elite Graduate Program Scholarship in 2017 and MSRA Star Track Program in 2018.

**Xin Tan** (Graduate Student Member, IEEE) received the B.Eng. degree in automation from Chongqing University, Chongqing, China in 2017. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China.

He has been a joint-Ph.D. student with the Department of Computer Science, City University of Hong Kong, Hong Kong, since 2019. His research interests lie in computer vision and deep learning, in particular, semantic segmentation and saliency detection.

**Zhi-Jie Wang** (Member, IEEE) received the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2015.

He is currently an Associate Professor with the College of Computer Science, Chongqing University (CQU), Chongqing, China. Before joining CQU, he worked with Sun Yat-sen University, Guangzhou, China, and The Hong Kong Polytechnic University, Hong Kong. He has published a set of research papers in venues, such as IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IJCAI, AAAI, ECAI, and ICME. His current research interests include data processing and analysis for spatial/temporal data and image/video data.

Dr. Wang is a member of CCF and ACM.

**Yong Liu** received the Ph.D. degree in computer science from Tianjin University, Tianjin, China, in 2016.

He is currently an Associate Professor with the Beijing Key Laboratory of Big Data Management and Analysis Methods, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China. His current research interests include large-scale machine learning, large-scale model selection, and auto machine learning.

**Jingang Zhang** received the B.S. degree and the M.S. degree in communication engineering from Xidian University, Xi'an, China, in 2005 and 2008, respectively, and the Ph.D. degree in communication and information systems from the University of Chinese Academy of Sciences, Beijing, China, in 2017.

He is an Associate Professor with the University of Chinese Academy of Sciences, where he is the Executive Director of the Intelligent Imaging Center. He has been engaged in the research of computational optical imaging technology for more than ten years, presided over more than ten national and ministerial-level scientific research projects, such as the National Natural Science Foundation of China and the Joint Foundation Program of the Chinese Academy of Sciences for equipment prefeasibility study, more than eight authorized invention patents and four software copyrights, and published over 20 academic papers and coauthored an academic book *Vector-Based Mathematics and Geometry*. His research interests include image denoising, deblurring, and dehazing, image/video analysis and enhancement, and related high-level vision problems.

**Xiaoqin Zhang** received the B.Sc. degree in electronic information science and technology from Central South University, Changsha, China, in 2005, and the Ph.D. degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010.

He is currently a Professor with Wenzhou University, Wenzhou, China. He has published more than 100 papers in international and national journals, and international conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *International Journal of Computer Vision*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, ICCV, CVPR, NIPS, IJCAI, AAAI, and ACM MM. His research interests are in pattern recognition, computer vision, and machine learning.

**Xiaochun Cao** (Senior Member, IEEE) received the B.E. and M.E. degrees in computer science from Beihang University, Beijing, China, in 1999 and 2002, respectively, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA, in 2006.

He has been a Professor with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, since 2012. After graduation, he spent about three years with ObjectVideo Inc., Reston, VA, USA, as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He has authored and coauthored more than 120 journal and conference papers.

Prof. Cao dissertation was nominated for the University of Central Florida's University-Level Outstanding Dissertation Award. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition in 2004 and 2010. He is on the Editorial Board of the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is a Fellow of IET.