Multi-level Fusion and Attention-guided CNN for Image Dehazing

Xiaoqin Zhang, Tao Wang, Wenhan Luo, Pengcheng Huang

***Abstract*—In this paper, we tackle the problem of single image dehazing with a convolutional neural network. Within this network, we develop a multi-level fusion module to utilize both low-level and high-level features. The low-level features help to recover finer details, and the high-level features discover abstract semantics. They are complementary in the restoring of clear images. Moreover, a Residual Mixed-convolution Attention Module (RMAM) with an attention block is proposed to guide the network to focus on important features in the learning process. In this RMAM, group convolution, depth-wise convolution, and point-wise convolution are mixed, and thus it is much faster than its counterparts. With these two modules, we thus have an end-to-end network without explicitly estimating the atmospheric light intensity and the transmission map in the classical atmosphere scattering model. Both qualitative and quantitative experimental studies are carried out on public datasets including RESIDE, DCPDN-TestA, and the real-world dataset. The extensive results demonstrate both the effectiveness and efficiency of the proposed solution to single image dehazing.**

***Index Terms*—Image dehazing; CNN; Multi-level fusion; Attention module**

I. Introduction

Haze has always been an annoying factor in the real world. For example, it results in the degradation of the quality of images captured by digital cameras. In some critical scenarios, haze can cause the failure of modern computer vision systems, such as autonomous driving systems and surveillance systems. Therefore, image dehazing plays an important role in many higher-level vision tasks [1–3].

The classical methods of physical models try to approximate the haze effect with the help of intermediate variables. For example, the classical atmospheric scattering model includes the atmospheric light intensity and the transmission map. By estimating these intermediate variables, the latent clear images can ultimately be derived. These methods achieve success to some extent, although the problem is not sufficiently solved. Physical model-based methods have several issues. First, as has been previously mentioned, physical models are approximations to the real-world haze process, so the justification of the employed models is questionable. Second, due to the separate steps in the physical model, the restoration quality of the final results heavily relies on the estimation of the intermediate variables, making the entire method difficult to tune.

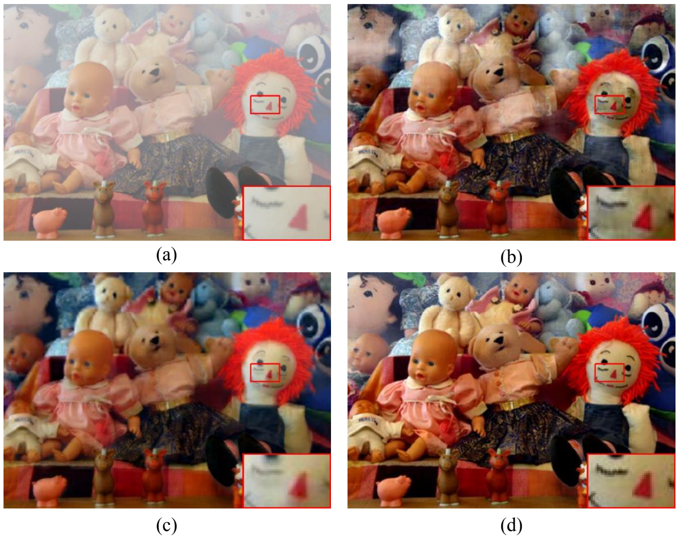

(a)
(b)
(c)
(d)

Figure 1. Comparison between different schemes given the same input hazy image (a). Images (b), (c), and (d) correspond to results from schemes using only low-level features (Stage-1 in the proposed network), only high-level features (Stage-4 in the proposed network), and both low-level and high-level features. For convenient comparison, a patch is cropped and its magnified version is placed in the bottom-right corner of each individual image. The face regions in the image (b) turns dark when only the low-level features are used. The face in the image (c) is blurry and is not as sharp as those in images (b) and (d), when only the high-level features are used. With the fused low-level and high-level features, the proposed method produces a clearer image with faithful color and rich details in image (d).

Due to the success of deep learning in various tasks, end-to-end solutions to image dehazing have been proposed. For instance, Ren *et al.* [4] propose a Gated Fusion Network (GFN) that combines several images into a single one to retain the most important features and uses a multi-scale fusion pipeline for image dehazing. Liu *et al.* [5] propose GridDehazeNet, a network of three modules that does not rely on the atmosphere scattering model and has demonstrated state-of-the-art performance on public datasets. While these types of end-to-end solutions are indeed successful, these methods have not fully solved the problem. The low-level features in a deep convolutional neural network are responsible for representing the details, while high-level features correspond to abstract and global semantics due to their relatively larger receptive field. However, for the specific image dehazing problem, the employed deep networks fail to cover both aspects mentioned above. Some also ignore the difference between the regions degraded by haze and the unaffected regions. Thus, their algorithms treat the spatial regions uniformly during the restoration process, which is contrary to human intuition. Moreover, the

X. Zhang, T. Wang and P. Huang are with the College of Computer Science and Artificial Intelligence, Wenzhou University, 325035, China (e-mail: {zhangxiaoqinnan, taowangzj, FShhppcc}@gmail.com).

W. Luo is with the Tencent AI Lab, Shenzhen, 518000, China (e-mail: whluo.china@gmail.com).

classical convolutional operation utilized in these networks is not efficient, making them both time-consuming and resource-consuming.

In light of this, we in this paper develop a deep network with two modules specifically for the issues described above. The first module is called the Multi-level Fusion Module (MFM), which fuses feature representations from different layers of different abstraction levels. With a set of this type of modules, both local finer details corresponding to low-level features and global semantics corresponding to high-level features can be learned, which are critical for removing haze in an image. Figure 1 depicts four images which are respectively an input hazy image, the dehazed result obtained by using only low-level features, the dehazed result obtained by using solely high-level features, and the dehazed result of our method fusing both low-level and high-level feature representations. Apparently, the result corresponding to the usage of only the low-level features is dark in some regions (see the magnified section of the toy face), which can be attributed to the lack of semantics. Using high-level features results in the removal of most of the haze, but fails to demonstrate fine details. In contrast, the proposed method restores the image with both semantics and finer details. Note that the work which is mostly similar to ours is GFN proposed by Ren *et al.* [4]. GFN is built on different image fusions with multi-scale, but this fusion strategy ignores the complementarity among features in different levels. We build the network on the principle of multi-level feature fusion to both utilize features in low and high levels, and achieve better performance of image dehazing.

The second module is a Residual Mixed-convolution Attention Module (RMAM). In this module, there are mixed types of convolutional operations, including group convolution, depth-wise convolution, and point-wise convolution. These mixed convolutional operations result in the superior efficiency of the proposed network as compared with its counterparts. We also implement an attention block based on point-wise and depth-wise convolutional operations. The attention mechanism in this block allows the network to focus more on the important features and ignore the redundant features.

With these two modules, we derive an end-to-end network, which directly recovers a clear image from a given hazy image efficiently, without relying on any physical model. To justify the proposed dehazing network, extensive experimental studies on three public datasets, including RESIDE [6], DCPDN-TestA [7] and the real-world dataset [8], are carried out. Specifically, ablation studies in terms of the MFM and RMAM are conducted to verify them individually. Both qualitative and quantitative comparisons with existing state-of-the-art methods are also conducted to demonstrate the superiority of the proposed method, and its efficiency is also verified.

In summary, the contributions of this work are three-fold:

1) We propose a Multi-level Fusion Module (MFM), which is able to adaptively employ different levels of features and use the complementation among them to effectively recover clear images from hazy images.

2) We develop an efficient Residual Mixed-convolution Attention Module (RMAM) with an attention block. The mixed convolutional operations make our network efficient, and the attention block drives our network to focus on more important features.

3) We implement an end-to-end network with the above two modules and the derived network achieves the best performance on the public datasets, compared with the state-of-the-art methods, both qualitatively and quantitatively.

The remainder of this paper is organized as follows. Section II presents a review of the existing related literature. The proposed method is introduced in detail in Section III. Experimental results including an ablation study and a comparison with existing methods are reported in Section IV. Section V draws the conclusions of this paper.

II. Related Work

A. Image Dehazing

Early works on image dehazing depend on external information, such as cues acquired from other sources [9], existing georeference models [10, 11], and using multiple images of the same scene taken under different weather conditions [12–14].

There is no external information available for single image dehazing. Therefore, this task is more challenging, and a wide variety of methods have been proposed to address it. A conventional solution is to first estimates the transmission map and the atmospheric light using certain assumptions or priors, then to restore the clear image via the atmosphere scattering model [8, 15, 16]. In the work by Fattal [8], a refined image formation model for surface shading and scene transmission is proposed to solve the image dehazing problem. Tan *et al.* [15] observe that haze-free images tend to have higher contrast than their hazy counterparts, and develop a local contrast maximization method for dehazing based on this observation. The dehazing method introduced in the work by He *et al.* [16] has shown remarkable effectiveness for haze removal. It is realized by employing the Dark Channel Prior (DCP) under the assertion that pixels in non-haze patches have low intensity in at least one color channel. Tang *et al.* [17] exploit a variety of haze-related priors using a random forest regressor. The color attenuation prior is developed in the literature [18] to recover depth information for the restoration of clear images. Although the aforementioned prior-based dehazing methods have achieved great success, their performance is influenced by the accuracy of the assumptions or priors. At the same time, these assumptions or priors may not reflect the inherent properties of natural images.

Recently, the rapid progress of deep learning technologies [19, 20] and the availability of large-scale synthetic datasets [17] have inspired various learning-based dehazing methods. In general, learning-based dehazing methods do not depend on priors and can be approximately divided into two classes: those which adopt deep learning technologies following the conventional solution mentioned previously, and those which directly recover clear images from their hazy counterparts using a convolutional neural network (CNN). For instance, Cai *et al.* [21] propose DehazeNet, which consists of a three-layer CNN, to directly estimate the transmission map from a given hazy image, and the coarse-to-fine manner is introduced into

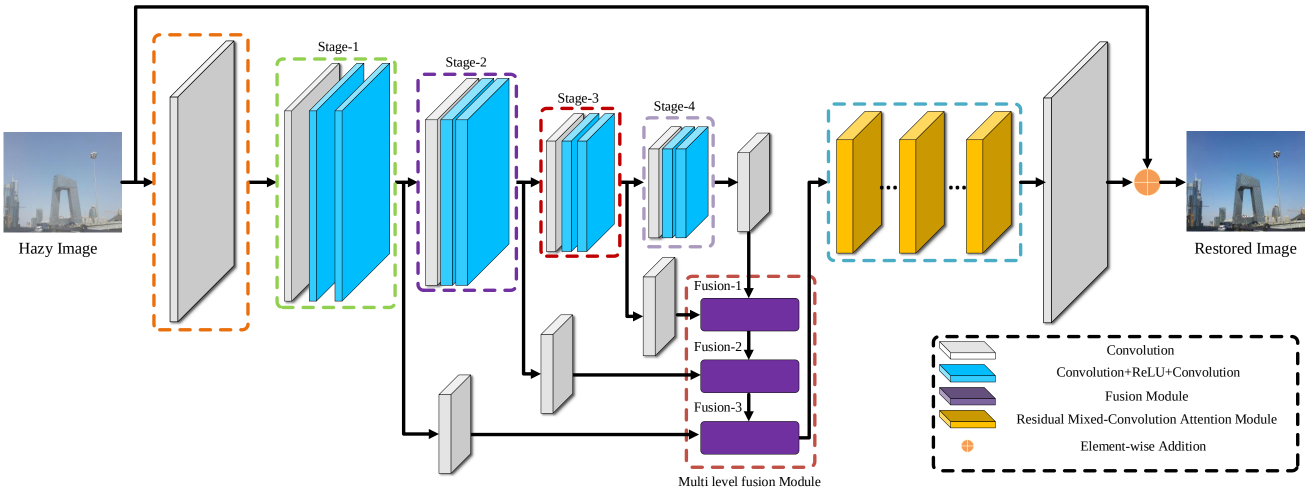


Figure 2. An overview of the proposed network for image dehazing. There are multiple stages in the network. Three fusion modules (represented by purple blocks) fuse feature representations from different levels. A residual mixed-convolution attention module (marked in orange) with an integrated attention block captures the residual, which is added to the input hazy image to produce the dehazed image.

the network for the estimation of the transmission map for haze removal. Specifically, a coarse-scale network is first adopted to estimate the transmission map, then a fine-scale network is designed to refine the transmission. Li *et al.* [22] suggest an all-in-one dehazing network (AOD-Net) via the reformulation of the scattering model, which computes a new variable containing both the atmospheric light and the transmission map. Ren *et al.* [23] propose a multi-scale convolutional neural networks (MSCNN) for image dehazing that learns the mapping between hazy images and their transmission maps. In contrast to these methods, some methods directly recover a clear image given a hazy image. For example, Qu *et al.* [24] transform the image dehazing problem into an image-to-image translation problem, and develop the Enhanced Pix2pix dehazing Network (EPDN) to directly restore a haze-free image. For example, an encoder-decoder network (GFN) and a novel fusion-based strategy are presented in the literature [4] to recover clear images from hazy images.

B. CNN with Attention Mechanism

The attention mechanism has an excellent ability to guide the model to adaptively process the most important components. It is widely employed in many computer vision tasks, including object tracking [25], image classification [26], and image restoration [27–29]. For example, a channel-attention is introduced into the network by Zhang *et al.* [30] for image super-resolution. A progressive attention based framework is proposed by Sun *et al.* [31]. This framework uses both channel-wise and spatial attention mechanisms. Liu *et al.* [5] design an end-to-end deep neural network for image dehazing, and apply an attention-based multi-scale estimation technique to guide efficient information exchange across different scales in the network. Li *et al.* [32] propose a level-aware progressive network (LAP-Net) for image dehazing. This network is supervised by different haze conditions, thus the network can dehaze progressively and is more adaptive to different haze conditions. This level-aware progressive method is essentially an attention mechanism. Zhang *et al.* [33] propose Pyramid Channel-based Feature Attention Network for image dehazing, which recovers the clear image with channel attention mechanism.

III. Proposed Method

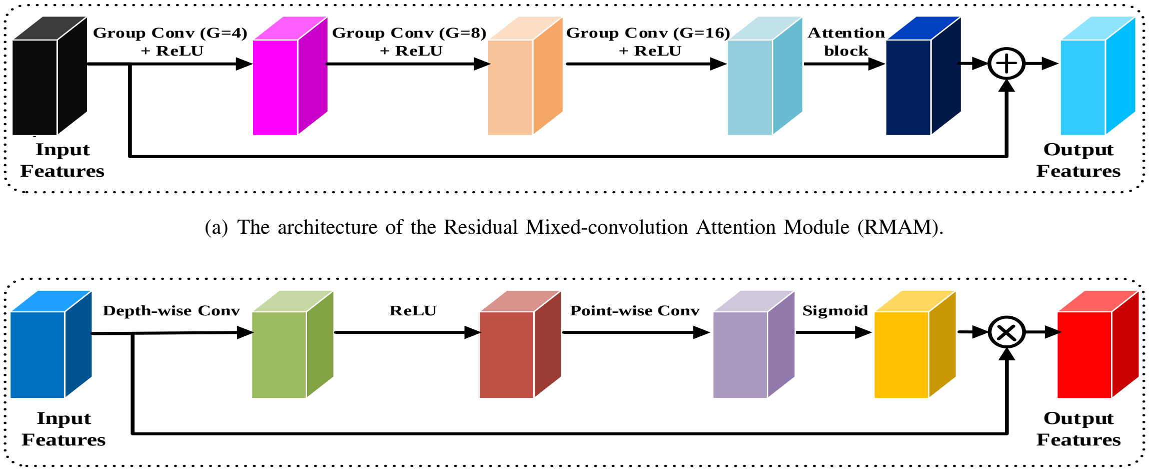
A. Network Architecture

Figure 2 illustrates an overview of the proposed network for image dehazing. As it is shown, a hazy image is first processed with a convolutional layer, after which there are four stages in the network. In each stage, features are extracted by a feature extraction module of convolutional layers with the addition of ReLU and convolutional layers, which are marked as blue in the figure. Three fusion modules, marked in purple, are adopted to fuse features from the different levels, ranging from Stage-1 to Stage-4. The orange color indicates the proposed Residual Mixed-convolution Attention Module (RMAM), within which there is an attention block. The RMAM is utilized to learn the residual regarding the given hazy image. A clear image is then restored with the derived residual. The following sections detail the individual components.

B. Feature Extraction Module

As mentioned previously, there are four stages in the network. In each stage, there is a feature extraction module that is implemented as a sequence of convolutional layers, ReLU, and convolutional layers.

Specifically, the feature extraction begins from a 3×3 convolutional layer that transforms the given hazy image into 16 feature maps. These feature maps are then processed by the following four stages to acquire features on different levels. In detail, each stage consists of four layers. The first layer is a 3×3 convolution with the stride of 2. It is used to decrease the resolution of the feature maps to 1/2 and double the width (the number of channels). The second and third layers include a 3×3 convolution, a ReLU activation function, and a 3×3



(a) The architecture of the Residual Mixed-convolution Attention Module (RMAM).

(b) The architecture of the attention block.

Figure 3. (a) illustrates the architecture of the Residual Mixed-convolution Attention Module (RMAM), where "G=4" indicates that the group number in the group convolution is 4. (b) shows the details of the attention block integrated into RMAM.

convolution, respectively. The fourth layer is 1×1 convolution, which reduces the width of the features produced by the third layer to 64 as the output for each stage.

C. Multi-level Fusion Module

Low-level features (*i.e.* Stage-1 and Stage-2 in the proposed network) typically represent local cues, such as edges and patterns. With the increase of the receptive field, higher-level features are capable of capturing semantics in a global range. Due to the abstraction ability of different levels, CNNs have achieved success in different tasks. The concept of the fusion of different levels of features has been applied in applications of object detection [2], object tracking [34], image restoration [35] and *etc*. However, for the specific image dehazing task, the advantage of multi-level feature fusion is ignored by the existing methods. It is not difficult to observe that, if we use only low-level features from the image, semantics in a more global range are not well restored though the details are primarily maintained, as shown in Figure 1 (b). On the contrary, if we use only the high-level features, finer details will be missing. As it is shown in Figure 1(c), the recovered image is not as sharp as the images in Figures 1 (b) and (d). For example, the eyebrow of the toy looks blurry as compared to its counterparts.

The advance of feature fusion inspires us to propose a multi-level fusion module for feature learning of image dehazing. As it is shown in Figure 2, from top to bottom, there are three feature fusion modules. The first module fuses features from the high level (Stage-4) and low level (Stage-3). The derived feature is taken as a high-level feature and then fused with the low-level feature from Stage-2 by the second feature fusion module. Again, the resulting feature is treated as a high-level feature and fused with the low-level feature from Stage-1, by the third feature fusion module. For each feature fusion module, given high-level and low-level features, the fusion is implemented as the element-wise multiplication between these two kinds of features. The fused feature will be forwarded to layers of convolution, BatchNorm, and ReLU before being processed by the next fusion module.

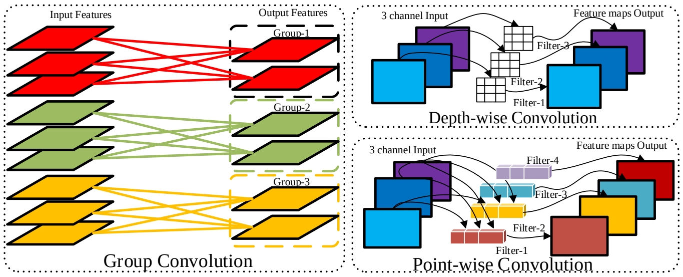


Figure 4. Illustration of the mixed types of convolutions used in the proposed residual mixed-convolution attention module, including group convolution, depth-wise convolution and point-wise convolution.

Specifically, let us denote the high-level and low-feature as f_h and f_l respectively for each fusion module. First, two 3×3 convolutional layers are applied to f_h and f_l individually. Then these features are fused by the multiplication operation, and the fused features capture the properties of both f_h and f_l. Finally, the fused features are processed by a 3×3 convolution to obtain the final output. BatchNorm and ReLU are added after every convolutional layer. The whole process is defined as,

$$f_{output} = S(F_h(f_h) * F_l(f_l)), \tag{1}$$

where $S(\cdot)$, $F_h(\cdot)$, and $F_l(\cdot)$ all represent structures composed of convolution, BatchNorm and ReLU. f_{output} is the final output feature. The fusion module presents a completely symmetric structure, in which f_l contributes details to f_h, and f_h embeds its semantics to f_l.

D. Residual Mixed-convolution Attention Module

The set of multi-level feature fusion modules outputs features representing both local details and global semantics in the image. With these features, we further propose a residual mixed-convolution attention module for learning a residual. The learned residual is element-wisely added to the given hazy image to restore a clear image.

The top portion of Figure 3 illustrates the structure of the proposed residual mixed-convolution attention module. There are three sequential group convolutional layers, followed by an attention block. The given features are processed by them and added to the residual to obtain the output features. Figure 4 presents an illustration of group convolution. In general, the input features are divided into groups, and the convolution operation is separately applied to the individual groups. Due to the division of groups, the FLOPs is greatly reduced. Thus, the efficiency is improved. The group numbers of the sequential group convolutional layers are respectively 4, 8, and 16. This configuration is determined by our experimental study, which will be detailed later in the experiment section.

Following the sequence of group convolutional layers, we develop an attention block. The structure of the attention block is shown in the bottom portion of Figure 3. The attention block exploits the following two steps to accomplish the attention mechanism. The first step utilizes depth-wise convolution followed by ReLU and point-wise convolution followed by Sigmoid to obtain the weights of features. The second step uses the obtained weights to multiply the original input features. Specifically, the core idea behind depth-wise and point-wise convolutions is to split the filtering and combination steps into two steps. Depth-wise convolution works on every channel of an input (using a single filter for each channel), and point-wise convolution is a regular 1×1 convolution mapping the channels processed by the depth-wise convolution into a new channel space. These convolutions increase the efficiency of the model. For example, Jing *et al.* [36] design a model based on these convolutions for faster inference in image generation. Figure 4 shows depth-wise convolution and point-wise convolution. By doing the following steps, we can have a weight map with the same dimension as the input features. The weight map is applied to the input features with element-wisely multiplication to output the final features.

The weight map derived from the attention module guides the network to abandon the redundant features and focus on the more important features. The adopted depth-wise and point-wise convolutional operations increase the efficiency of this module. In addition to group convolution, there are mixed convolutional layers in this module, after which the module is named for.

E. Loss Function

To train the proposed network, we adopt a composited loss function. This loss function is composed of two terms. The first term is the traditional Mean Square Error (MSE) loss, which measures the deviation between the restored image and the corresponding ground truth image. Formally, it is written as,

$$\mathcal{L}_{mse} = \frac{1}{WH}\Sigma_{i,j,c}(I^{re}_{i,j,c} - I^{gt}_{i,j,c})^2 \, , \tag{2}$$

where W and H represent the width and height of a concerned image, I^{re} and I^{gt} are the recovered image and the ground truth corresponding to the concerned given image, i and j index the pixel location in the image, and c ranges from 1 to 3 to cover the RGB channels.

The second term is a perceptual measurement that quantizes the difference between I^{re} and I^{gt} in terms of the perceptual property. Similar to an existing work [5], we also use the seminal VGG16 network [37] pre-trained on the large-scale ImageNet to extract the features from I^{re} and I^{gt}. To be specific, the features from the stages of *Conv1-2, Conv2-2,* and *Conv3-3* are employed to calculate the difference. *Convn-m* means the m layer of the n-*th* stage in VGG16. This perceptual loss term is formulated as,

$$\mathcal{L}_{per} = \frac{1}{C_k W_k H_k}\Sigma_k ||\Phi_k(I^{re}) - \Phi_k(I^{gt})|| \, , \tag{3}$$

where $\{\Phi_k(\cdot), k = 1, 2, 3\}$ are feature extractors corresponding to the three stages of VGG16 mentioned above, C_k, W_k and H_k correspond to the number of channels, width and height of the feature maps from $\Phi_k(\cdot)$.

The final loss is a composition of the above terms, and is defined as,

$$\mathcal{L} = \mathcal{L}_{mse} + \alpha * \mathcal{L}_{per} \, , \tag{4}$$

where α is a weight parameter to balance the two terms.

IV. Experiments

A. Datasets

To verify the proposed method for image dehazing, we carry out experimental studies on three datasets, including RESIDE [6], DCPDN-TestA [7], and the real-world dataset [8]. Note that there are five subsets in the RESIDE dataset. Following the settings of existing studies [5], we use the subsets of Indoor Training Set (ITS), Outdoor Training Set (OTS) for training, and use the subsets of Synthetic Objective Testing Set (SOTS) and Real World Task-driven Testing Set (RTTS) for testing. The DCPDN-TestA dataset contains 400 image pairs with denser haze than those in SOTS. To verify the generalization ability of the proposed model, we synthesize 200 (hazy/clean) images from both the Middlebury stereo database (40) [38] and the Sun3D dataset (160) [39]. We name it as TestB. Moreover, we perform comparisons using the real-world hazy image dataset in [8] to demonstrate the performance of the proposed method for real-world scenery.

B. Implementation Details

To train the proposed network, the Adam optimizer [40] is adopted with a mini-batch size of 1, where β_1 and β_2 are fixed with the values of 0.5 and 0.999, respectively. The learning rate is set as 0.0001, and the hyper-parameter α in the loss function is set as 0.01. The network is trained on the ITS subset for 100 epochs, and on the OTS subset for 10 epochs.

The network is trained and tested using the PyTorch 1.0.0 [41] framework and Python 3.5. Specifically, all experiments are conducted on a server with Ubuntu. The hardware configuration is as follows: an Intel CPU with 2.60GHz, 128GB of RAM, and an Nvidia GeForce RTX 2080Ti GPU. The

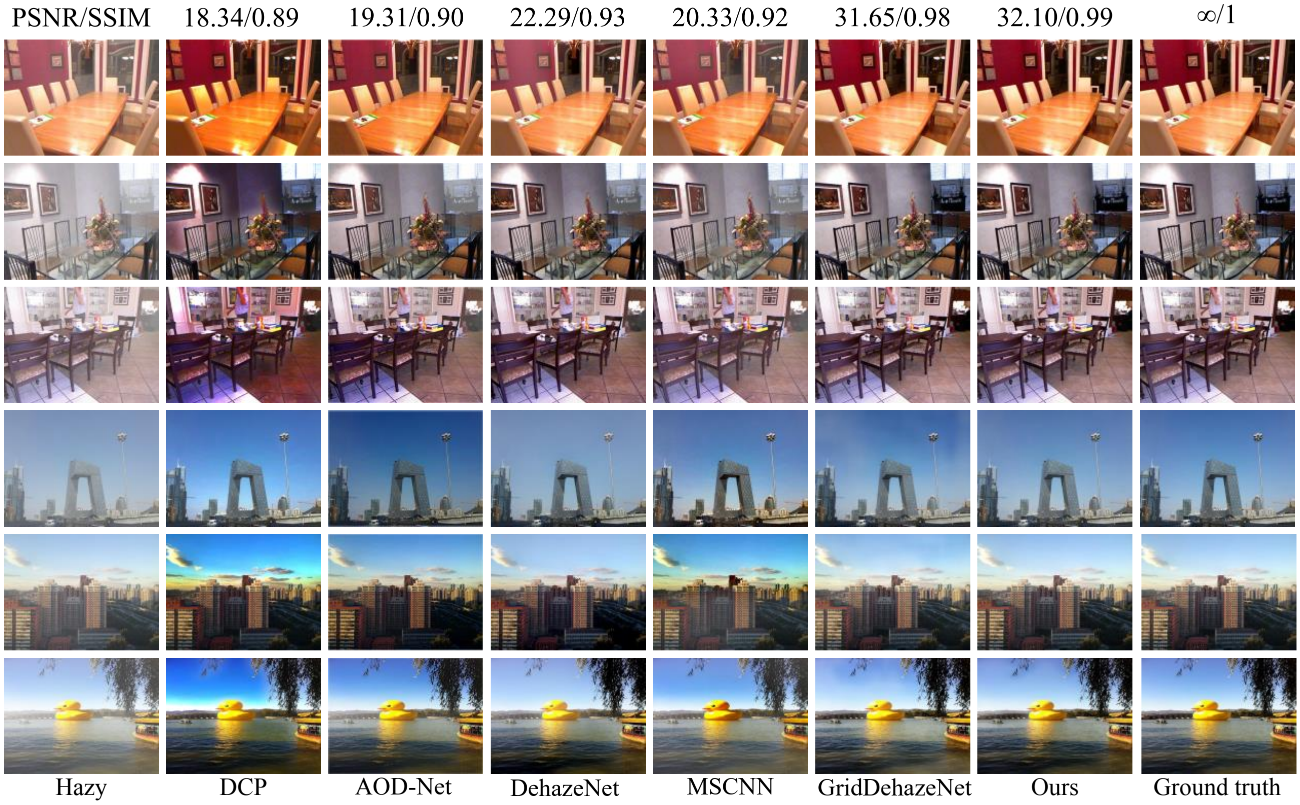


Figure 5. Qualitative comparison results between our method and the state-of-the-art approaches on SOTS. Each row shows one example.

Nvidia CUDA 9.0 toolkit and cuDNN 7.5 are adopted. For fair comparisons, the quantitative results of PSNR and SSIM in this paper are calculated on RGB channels.

C. Ablation Studies

To verify the effectiveness of the different modules in the proposed network, and to investigate how the configurations of these modules influence the final dehazing results, we carry out the following ablation studies.

1) Multi-level fusion module analysis: We firstly test how the different fusion settings affect the dehazing performance. To make the experiment study feasible, we define a basic model at first. This basic model is a naive model without any fusion modules. Namely, *the features from Stage-4 in Figure 2 are forwarded into the residual mixed-convolution attention module directly*. After defining this basic model, we further have the following four variants.

- Basic model + Fusion-1: the basic model with the Fusion-1 module. As Fusion-1 corresponds to the fusion of the highest level of features, this will reveal the influence of high-level features.
- Basic model + Fusion-3: the basic model with the Fusion-3 module. This variant will demonstrate the effect of low-level features.
- Basic model + Fusion-1 + Fusion-2: compared with the first variant, this model will reveal the influence of the additional Fusion-2 module.
- Full model: all the Fusion modules in different levels are used in the full model. This can be considered as the basic model + Fusion-1 + Fusion-2 + Fusion-3, as Figure 2 shows.

Table I
THE RESULTS OF DIFFERENT FUSION VARIANTS ON THE INDOOR SUBSET OF SOTS IN TERMS OF PSNR AND SSIM. THE BEST PERFORMANCE IS MARKED IN BOLD.

Fusion variant	Indoor (PSNR/SSIM)
Basic model	27.6085/0.9688
Basic model + Fusion-1	31.9678/0.9891
Basic model + Fusion-3	29.8106/0.9824
Basic model + Fusion-1 + Fusion-2	32.9488/0.9880
Full model	**33.4737/0.9941**
Full model_w/o perceptual loss	33.1248/0.9935

The experiment is carried out on the Indoor subset of SOTS. Both PSNR and SSIM metrics are used to quantify the results. Table I reports the comparison results. We have the following findings. 1) The basic model, though without dedicated fusion modules, achieves a reasonable performance. 2) Fusing high-level features greatly improves the performance, if we compare the first variant with the basic model. This verifies the effectiveness of the global semantics from the high-level features. 3) Solely fusing low-level features also results in improvements in terms of the PSNR and SSIM, suggesting the benefits from the low-level features. However, the quantity of improvement from high-level features is more significant than that from the low-level features, if we compare the first and the second variants. This might indicate that the high-level features are more important than the low-level features. 4) With the additional Fusion-2 module, the performance is

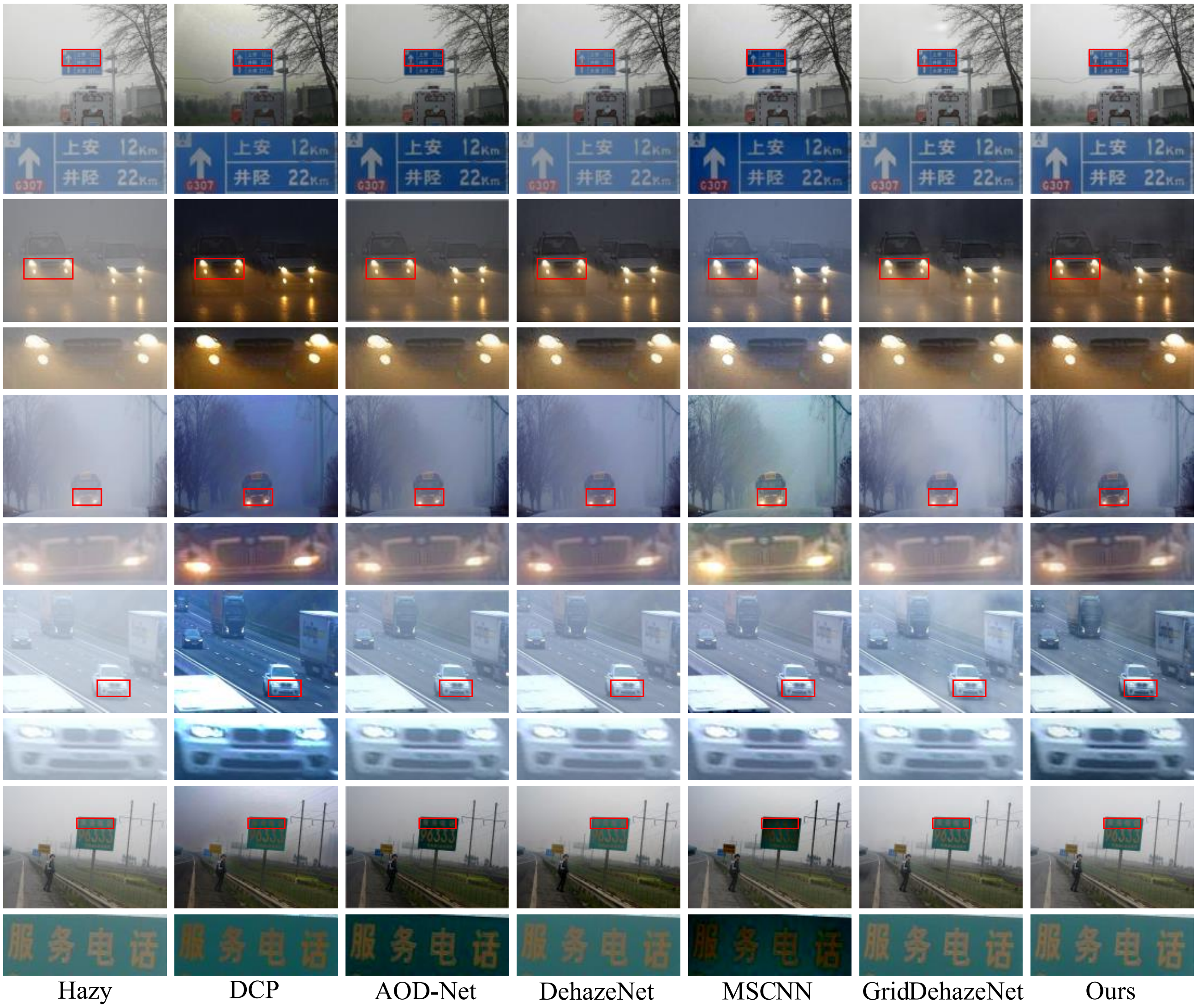


Figure 6. Exemplar images from RTTS for qualitative comparison. The ground truth is not available in this scenery. From top to bottom, every two rows presents one example.

further improved in terms of PSNR, which is also a sign of improvement resulted from fused features. 5) The full model achieves the best performance regarding both PSNR and SSIM. This demonstrates that the multi-level fusion module indeed benefits the image dehazing performance. In addition, to verify the effectiveness of the perceptual loss, we design a variant model without the perceptual loss (*i.e.* Full model_w/o perceptual loss). As shown in Table I, the improvement gains in PSNR and SSIM are 0.3489 and 0.0006 with the help of perceptual loss.

2) Residual mixed-convolution attention module analysis: The residual mixed-convolution attention module (RMAM) plays an important role in the proposed network. We thus conduct analysis to investigate how its structure can affect the performance. As presented in Figure 4, there are three group convolutional layers and an attention block in the entire RMAM. Thus, we incrementally increase the number of the group convolutional layers, and vary the number of groups in the group layers. We also remove the attention block to test the performance.

Table II
THE RESULTS OF DIFFERENT CONFIGURATIONS OF RMAM ON THE INDOOR SUBSET OF SOTS IN TERMS OF PSNR AND SSIM. THE BEST PERFORMANCE IS MARKED IN BOLD.

RMAM configuration	Indoor (PSNR/SSIM)
0	29.2687/0.9804
4	31.6488/0.9888
4-8	32.7163/0.9892
4-8-16	**33.4737/0.9941**
1-1-1	30.4713/0.9899
w/o_attention	30.0364/0.9845

We test our method on the Indoor subset of SOTS and report the results in Table II. Again, PSNR and SSIM are employed for quantitative comparison. Specifically, we report the performances of the following variants. In Table II, “0” indicates the removal of all the group convolutional layers in the RMAM, *i.e.*, only the attention block remains. “4” indicates the addition of one group convolutional layer, and

Table III
THE RESULTS OF DIFFERENT NUMBERS OF RMAM ON THE INDOOR SUBSET OF SOTS IN TERMS OF PSNR AND SSIM. THE BEST PERFORMANCE IS MARKED IN BOLD.

Number of RMAM	Indoor (PSNR/SSIM)
0	29.2222/0.9798
3	31.9902/0.9869
6	**33.4737/0.9941**
9	31.2202/0.9918

Table IV
THE RESULTS OF DIFFERENT ESTIMATION STRATEGIES ON BOTH THE INDOOR AND OUTDOOR SUBSETS OF SOTS IN TERMS OF PSNR AND SSIM. THE DIRECT ESTIMATION STRATEGY CONSISTENTLY OUTPERFORMS THE INDIRECT ESTIMATION STRATEGY.

Strategy	Indoor		Outdoor	
	PSNR	SSIM	PSNR	SSIM
Indirect	30.5477	0.9843	30.4507	0.9794
Direct	**33.4737**	**0.9941**	**32.0743**	**0.9841**

the group number is set as 4. “4-8” means that there are two group convolutional layers, and the group numbers of them are 4 and 8 individually. “4-8-16” indicates that there are three group convolutional layers with group numbers of 4, 8, and 16, respectively. This setting is the default configuration in the full model. “1-1-1” is the configuration of three group convolutional layers with respective group numbers of 1, 1, and 1. In this case, the group convolution degrades into ordinary convolution. Finally, “w/o_attention” indicates the setting of removing the attention block. In this scenario, we keep the “4-8-16” configuration of the group convolutional layers.

The results suggest the following. 1) removing all the group convolutional layers decreases the overall performance to a large extent, which demonstrates the importance of all the group convolutional layers. 2) When incrementally adding group convolutional layers, the performance also improves accordingly, as indicated by the results of “4” and “4-8”. 3) The best performance was achieved with the presence of all three group convolutional layers. We suspect this is an optimal configuration of the RMAM. 4) By changing the group number in the sequential group convolutional layers from “4-8-16” to “1-1-1”, the group convolution degenerates into the ordinary convolution. The decreases in both PSNR and SSIM indicates that the group convolution is more useful than the normal convolution. 5) The attention block is also important, as the PSNR and SSIM values drop when the attention block is removed.

3) Number of RMAM: We further conduct an analysis of the performance in terms of the number of the RMAM. The analysis is conducted on the Indoor subset of SOTS and the results are presented in Table III. We firstly include no RMAM, and then include more RMAM, with numbers ranging from 3 to 6, and finally to 9. Intuitively, with the addition of RMAM, the network captures a better residual, consequently resulting in better dehazing performance. However, it was also found that the performance did not consistently improve with the increase of the number of RMAM. Specifically, when the number of RMAM is 6, the performance is the best. With more RMAM, *e.g.*, 9, the results begin to decrease. This might have been caused by overfitting as a result of the deeper structure of 9 RMAM.

Table V
THE COMPARISON RESULTS BETWEEN THE PROPOSED METHOD AND THE EXISTING STATE-OF-THE-ART METHODS FOR IMAGE DEHAZING ON BOTH THE INDOOR AND OUTDOOR SUBSETS OF SOTS. THE PROPOSED METHODS ACHIEVES THE BEST PERFORMANCE IN TERMS OF BOTH PSNR AND SSIM.

Method	Indoor		Outdoor	
	PSNR	SSIM	PSNR	SSIM
DCP	16.61	0.8546	19.14	0.8605
DehazeNet	19.82	0.8209	24.75	0.9269
MSCNN	19.84	0.8327	22.06	0.9078
AOD-Net	20.51	0.8162	24.14	0.9198
GFN	24.91	0.9186	28.29	0.9621
GridDehazeNet	32.16	0.9836	30.86	0.9819
Ours	**33.47**	**0.9941**	**32.07**	**0.9841**

4) Different estimation strategies: As mentioned previously, the proposed network is an end-to-end solution to image dehazing, which is able to directly estimate the latent clean image without relying on any physical model. The advantages of end-to-end solutions have been demonstrated in existing studies. Nevertheless, we still compare the direct estimation strategy (*i.e.*, the end-to-end estimation network) with the indirect estimation strategy. To ensure a fair comparison, the indirect estimation is also accomplished by the proposed network. Specifically, two branches of convolution with the same structure were added to the network to respectively estimate the atmospheric light and the transmission map. Then, with the estimated atmosphere light and transmission map, the clear image was recovered by the atmospheric scattering model.

This comparison is conducted on both the Indoor and the Outdoor subsets of SOTS, and the results are reported in Table IV. In both scenarios, the direct estimation strategy outperforms the indirect one. The difference coincides with the advantages of the direct estimation mentioned above.

D. Comparisons with State-of-the-art Dehazing Methods

With the above ablation studies, we have already demonstrated the effectiveness of the different components in our method. To further verify it, we carry out quantitative and qualitative comparisons with the existing state-of-the-art methods on both synthetic and real-world data.

1) Synthetic data: We firstly compare our method with the existing methods on SOTS, which is composed of an Indoor subset and an Outdoor subset. The compared methods include DCP [16], DehazeNet [21], MSCNN [23], AOD-Net [22], GFN [4], and GridDehazeNet [5]. All of the compared methods are driven by data except DCP, which is a prior-based method. Among the data-driven methods, DehazeNet and MSCNN use the indirect estimation strategy. They estimate the final images by intermediately estimating the atmosphere light and the transmission map. AOD-Net, GFN, and GridDehazeNet directly estimate the dehazed image given a hazy input image.

Table V exhibits the results in terms of PSNR and SSIM. DCP is a reasonable baseline that is worse than the data-driven methods. The methods of direct estimation consistently

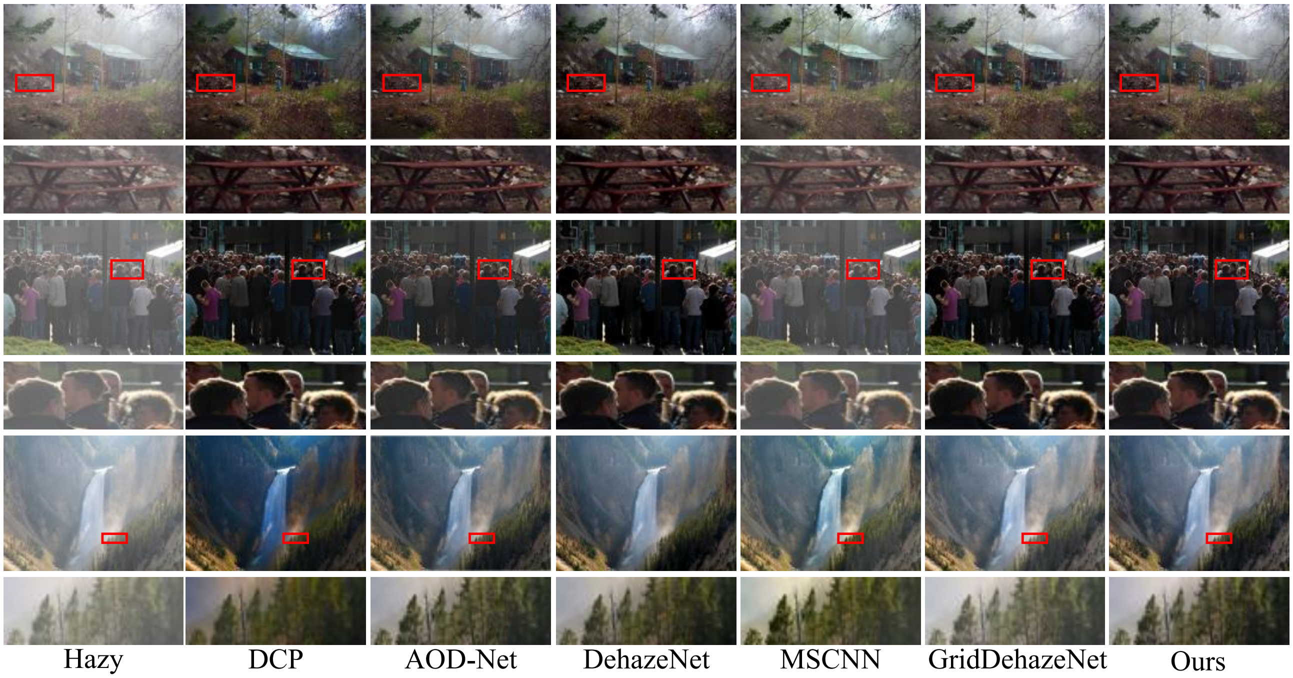


Figure 7. Qualitative results on the real-world dataset. From top to bottom, every two rows represents one example. Note that there is no ground truth for images from the real-world dataset.

Table VI
THE COMPARISON RESULTS BETWEEN THE PROPOSED METHOD AND THE EXISTING STATE-OF-THE-ART METHODS FOR IMAGE DEHAZING ON THE DCPDN-TESTA DATASET. THE PROPOSED METHOD ACHIEVES THE BEST PERFORMANCE OF BOTH PSNR AND SSIM.

Method	DCPDN-TestA	
	PSNR	SSIM
DCP	13.91	0.8642
DehazeNet	20.89	0.8860
MSCNN	18.90	0.8314
AOD-Net	21.48	0.8481
DCPDN	29.27	0.9560
GridDehazeNet	24.53	0.9209
Ours	**30.31**	**0.9847**

perform better than the indirect ones. Among the methods of direct estimation, the GridDehazeNet is much better than others. However, ours beats it with the absolute advantage of 1.31 and 1.21 in terms of PSNR in Indoor and Outdoor subsets, respectively.

To make the comparison intuitive, we show exemplar images from SOTS are presented in Figure 5. The first column exhibits the input hazy images, and the subsequent columns show results from different methods and the ground truth images. The top three rows represent images from the indoor subset, and the bottom three ones correspond to the outdoor subset. In general, the results of DCP usually suffer from the color distortion problem (*e.g.* the skin of the man turns red in the third row, and the sky is too blue in the fifth and sixth rows). In addition, the dehazed images of DCP are usually darker than the ground truth images. Although the other comparison methods do not suffer from color distortion, they fail to remove haze in some regions of the images. For example, in the second row, the surface of the chair should be brown and black. However, most methods produce a white color instead. This is probably because these methods fail in the restoration of details. On the contrary, the proposed method does not suffer from these problems and produces dehazed images that look more like the ground truth images.

Moreover, we also test our method on a more challenging synthetic dataset, DCPDN-TestA. In general, the synthesized haze artifacts in DCPDN-TestA are more severe than the case in SOTS, which poses more difficulties to image dehazing methods. Both PSNR and SSIM are used for quantitative comparison. As presented in Table VI, the values of PSNR and SSIM are not as high as those of SOTS in Table V. This also confirms that the DCPDN-TestA is more difficult than the SOTS. Despite this, the proposed method consistently outperforms others and achieves the best performance. The proposed method achieves the gain with 1.04 in terms of PSNR and 0.0287 regarding SSIM compared with the second place method DCPDN [7], which is a densely connected encoder-decoder structure for estimating the transmission map.

To demonstrate the generalization ability of the proposed model, we also test the proposed model on the TestB dataset. The comparison results are shown in Table VII. In which the digital values are the averages of the results on TestB in terms of PSNR and SSIM. The proposed method achieves the best performance in terms of PSNR and SSIM. DCPDN ranks the second regarding PSNR. GridDehazeNet obtains the second in SSIM. The distance between the best and the second best is 1.2 and 0.0172 in terms of PSNR and SSIM.

2) Real-world data: By the above comparison, the advance of our method over the existing state-of-the-art methods has been demonstrated. However, there are obvious differences between synthetic hazy images and real-world hazy images.

Table VII
THE COMPARISON RESULTS BETWEEN THE PROPOSED METHOD AND THE EXISTING STATE-OF-THE-ART METHODS FOR IMAGE DEHAZING ON THE TESTB DATASET. THE PROPOSED METHOD ACHIEVES THE BEST PERFORMANCE OF BOTH PSNR AND SSIM.

Method	TestB	
	PSNR	SSIM
DCP	12.56	0.8064
DehazeNet	19.90	0.8378
MSCNN	17.84	0.8135
AOD-Net	20.24	0.8260
DCPDN	24.61	0.8948
GridDehazeNet	23.68	0.9208
Ours	**25.81**	**0.9380**

Thus, the advance in terms of real-world images still needs to be further verified. To this end, we accomplish experiments on the real-world images from RTTS and the real-world dataset. Note that, as the ground truth images corresponding to the hazy images are not available in these two datasets, we are only able to conduct a qualitative comparison, as Figure 6 and Figure 7 show. In both figures, we crop a patch and zoom in to provide better visibility.

In Figure 6, DCP and MSCNN suffer from color distortion and generate undesired artifacts. For instance, the sky of the third example of DCP becomes blue, and that of MSCNN turns green. It can also be observed that, DehazeNet, MSCNN, and AOD-Net fail to remove the haze completely and suffer from blur (see the surface of the car in the fourth example). The GridDehazeNet cannot restore well when facing a dense haze scene (*e.g.*, the background, and the car in the fourth example). The proposed method removes haze effectively in heavily hazy scenes and achieves the best visual quality.

As shown in Figure 7, the results of DCP and AOD-Net exhibit severe color distortions, thus, they acquire the low quality of outputs (*e.g.*, see the tree and the sky in the third example). For DehazeNet and MSCNN, the haze is still unremoved and the output also suffers from the color distortion problem. Although GridDehazeNet works well, its outputs lost details and thus it restores unclear images. The proposed method is found to be capable of dehazing well whether the haze is dense or light, and can faithfully restore the textural details.

E. Running-time Analysis

The mixed convolution plays an important role in the RMAM. Furthermore, it also leads to the improvement of efficiency due to the property of the different types of convolution as compared with the ordinary convolution. With the configuration of hardware described in Section IV-B, we test the speed of our method on SOTS and compare it with the speeds of existing state-of-the-art methods. The results are presented in Figure 8. To make the comparison fair, we extensively run all the compared methods on the same machines and count the running time.

The proposed method dehazes an image with only 0.01s, which is nearly 100 FPS in accomplishing the dehazing task and much faster than the counterparts. The second fast is AOD-Net, the runtime of which is 0.03s. The third fast is GridDehazeNet. The fourth place is DCPDN, the runtime of which is 0.26s. All the other methods are much slower. We emphasize that the fastest running time of our method results from the dedicated design of the mixed convolutions in the RMAM.

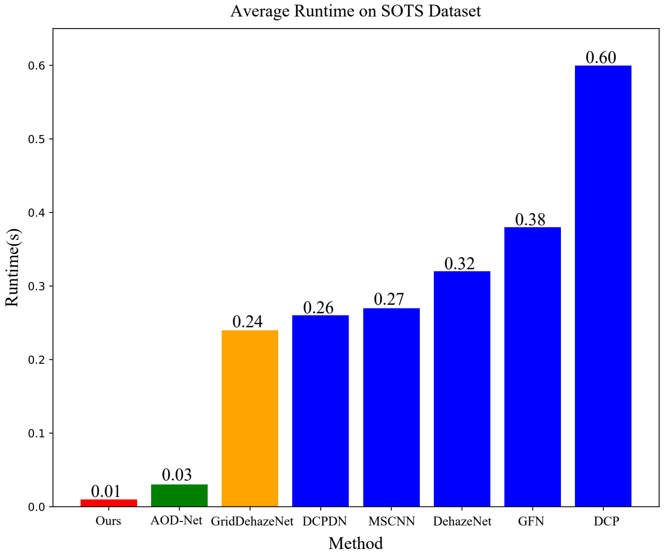


Figure 8. The comparison between the runtime of the different methods on SOTS. Note that the runtime is counted with the same hardware configuration for fairness.

V. CONCLUSION

In this paper, we have proposed a network with two effective modules. The first multi-level fusion module leverages features from multiple levels to benefit from the complementarity among them. The second residual mixed-convolution attention module drives our network to focus more on important features effectively. Our end-to-end network is also more efficient in dehazing compared with the existing state-of-the-art methods. Experimental studies on public datasets have verified its advantages over the state-of-the-art methods for image dehazing.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China [grant nos. 61922064, U2033210], in part by the Zhejiang Provincial Natural Science Foundation [grant nos. LR17F030001, LQ19F020005], in part by the Project of science and technology plans of Wenzhou City [grant nos. C20170008, ZG2017016].

REFERENCES

[1] X. Zhang, W. Hu, N. Xie, H. Bao, and S. Maybank, “A robust tracking system for low frame rate video,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 279–304, 2015.

[2] B. Xu and Z. Chen, “Multi-level fusion based 3d object detection from monocular images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2345–2353, 2018.

[3] X. Zhang, D. Wang, Z. Zhou, and Y. Ma, “Robust low-rank tensor recovery with rectification and alignment,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 238–255, 2021.

[4] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang, "Gated fusion network for single image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3253–3261, 2018.

[5] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7314–7323, 2019.

[6] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2018.

[7] Z. He and P. V. M, "Densely connected pyramid dehazing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3194–3203, 2018.

[8] R. Fattal, "Dehazing using color-lines," *ACM Transactions on Graphics*, vol. 34, no. 1, pp. 1–14, 2014.

[9] S. G. Narasimhan and S. K. Nayar, "Interactive (de) weathering of an image using physical models," in *Proceedings of the IEEE Workshop on Color and Photometric Methods in Computer Vision*, vol. 6, no. 6.4, p. 1, 2003.

[10] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 617–624, 2013.

[11] C. O. Ancuti, C. Ancuti, C. Hermans, and P. Bekaert, "A fast semi-inverse approach to detect and remove the haze from a single image," in *Proceedings of the Asian Conference on Computer Vision*, pp. 501–514, 2010.

[12] Y. Y. Schechner, S. G. Narasimhan, and S. K. Nayar, "Instant dehazing of images using polarization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 325–332, 2001.

[13] S. Shwartz, E. Namer, and Y. Y. Schechner, "Blind haze separation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1984–1991, 2006.

[14] S. G. Narasimhan and S. K. Nayar, "Chromatic framework for vision in bad weather," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 598–605, 2000.

[15] R. T. Tan, "Visibility in bad weather from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[16] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 2341–2353, 2010.

[17] K. Tang, J. Yang, and J. Wang, "Investigating haze-relevant features in a learning framework for image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2995–3000, 2014.

[18] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3522–3533, 2015.

[19] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu, "Stacked deconvolutional network for semantic segmentation," *IEEE Transactions on Image Processing*, 2019.

[20] Z. Huang, J. Fan, S. Cheng, S. Yi, X. Wang, and H. Li, "Hms-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion," *IEEE Transactions on Image Processing*, 2019.

[21] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, pp. 5187–5198, 2016.

[22] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4770–4778, 2017.

[23] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proceedings of European Conference on Computer Vision*, pp. 154–169, 2016.

[24] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8160–8168, 2019.

[25] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, and S. Maybank, "Learning attentions: residual attentional siamese network for high performance online visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4854–4863, 2018.

[26] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1487–1500, 2017.

[27] X. Zhang, R. Jiang, T. Wang, P. Huang, and L. Zhao, "Attention-based interpolation network for video deblurring," *Neurocomputing*, 2020.

[28] Y. Hu, J. Li, Y. Huang, and X. Gao, "Channel-wise and spatial feature modulation network for single image super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[29] H. Wu, Z. Zou, J. Gui, W.-J. Zeng, J. Ye, J. Zhang, H. Liu, and Z. Wei, "Multi-grained attention networks for single image super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[30] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of European Conference on Computer Vision*, pp. 286–301, 2018.

[31] F. Sun, W. Li, and Y. Guan, "Self-attention recurrent network for saliency detection," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 30 793–30 807, 2019.

[32] Z. Deng, L. Zhu, X. Hu, C.-W. Fu, X. Xu, Q. Zhang, J. Qin, and P.-A. Heng, "Deep multi-model fusion for single-image dehazing," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2453–2462, 2019.

[33] X. Zhang, T. Wang, J. Wang, G. Tang, and L. Zhao, "Pyramid channel-based feature attention network for image dehazing," *Computer Vision and Image Understanding*, vol. 2197-198, p. 103003, 2020.

[34] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3074–3082, 2015.

[35] J. Li, F. Fang, J. Li, K. Mei, and G. Zhang, "Mdcn: Multi-scale dense cross network for image super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[36] Y. Jing, X. Liu, Y. Ding, X. Wang, E. Ding, M. Song, and S. Wen, "Dynamic instance normalization for arbitrary style transfer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4369–4376, 2020.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

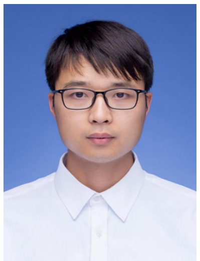
[38] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proceedings of German Conference on Pattern Recognition*, pp. 31–42. Springer, 2014.

[39] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 567–576, 2015.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

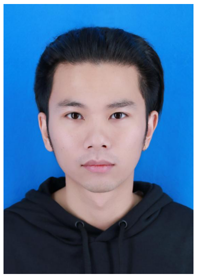
[41] A. Paszke, S. Gross, S. Chintala, and G. Chanan, "Pytorch," *Computer Software. Vers. 0.3*, vol. 1, 2017.

Xiaoqin Zhang received the B.Sc. degree in electronic information science and technology from Central South University, China, in 2005 and Ph.D. degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China, in 2010. He is currently a professor in Wenzhou University, China. His research interests are in pattern recognition, computer vision and machine learning. He has published more than 100 papers in international and national journals, and international conferences, including IEEE T-PAMI, IJCV, IEEE T-IP, IEEE T-IE, IEEE T-C, ICCV, CVPR, NIPS, IJCAI, AAAI, and among others.



Tao Wang is currently a graduate student at College of Computer Science and Artificial Intelligence, Wenzhou University, China. He received the B.Sc. degree in information and computing science from Hainan Normal University, China, in 2018. His research interests include several topics in computer vision and machine learning, such as object tracking/detection, image/video quality restoration, adversarial learning, image-to-image translation and reinforcement learning.

Wenhan Luo received the Ph.D. degree from Imperial College London, UK, 2016, M.E. degree from Institute of Automation, Chinese Academy of Sciences, China, 2012 and B.E. degree from Huazhong University of Science and Technology, China, 2009. His research interests include several topics in computer vision and machine learning, such as motion analysis (especially object tracking), image/video quality restoration, object detection and recognition, reinforcement learning.



Pengcheng Huang is currently a graduate student majoring in computer software and theory at College of Computer Science and Artificial Intelligence, Wenzhou University, China. He received his bachelor's degree in department of modern science and technology at China metrology university, China. His research interests including image and video processing, pattern recognition and machine learning.