AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your article. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

Carefully check the page proofs (and coordinate with all authors); additional changes or updates WILL NOT be accepted after the article is published online/print in its final form. Please check author names and affiliations, funding, as well as the overall article for any errors prior to sending in your author proof corrections. Your article has been peer reviewed, accepted as final, and sent in to IEEE. No text changes have been made to the main part of the article as dictated by the editorial level of service for your publication.

AQ:1 = Please confirm or add details for any funding or financial support for the research of this article.

AQ:2 = Please confirm the location for Wenzhou University.

AQ:3 = Please provide the publisher location for Ref. [23].

AQ:4 = Please provide the organization name, organization location, and report no. for Ref. [30].

Hierarchical Feature Fusion With Mixed Convolution Attention for Single Image Dehazing

Xiaoqin Zhang, Jinxin Wang, Tao Wang, and Runhua Jiang

***Abstract*— Single image dehazing, which aims at restoring a haze-free image from its correspondingly unconstrained hazy scene, is a fundamental yet challenging task and has gained immense popularity recently. However, the images recovered by some existing haze-removal methods often contain haze, artifacts, and color distortions, which severely degrade the visual quality and have negative impacts on subsequent computer vision tasks. To this end, we propose a network combining multi-scale hierarchical feature fusion and mixed convolution attention to progressively and adaptively enhance the dehazing performance. The haze levels and image structure information are accurately estimated by fusing multi-scale hierarchical features, thus the model restores images with less remaining haze. The proposed mixed convolution attention mechanism is capable of reducing feature redundancy, learning compact and effective internal representations and highlighting task-relevant features, thus, it can further help the model estimate images with sharper textural details and more vivid colors. Furthermore, a deep semantic loss is also proposed to highlight essential semantic information in deep features. The experimental results show that the proposed method outperforms state-of-the-art haze removal algorithms.**

***Index Terms*— Image dehazing, hierarchical feature fusion, mixed convolution attention mechanism, deep learning.**

I. Introduction

IMAGES captured in hazy conditions often contain unclear content and degraded structural details. These low-visibility images are a hindrance to multiple subsequent high-level tasks [1]–[5], including autonomous driving, video surveillance, and visual object tracking. As a fundamental yet challenging technique, single image dehazing, which aims at restoring haze-removed images with less remaining haze, sharper structure details and vivid colors from low-visibility scenarios, will be beneficial to the application of these high-level tasks. Thus, single image dehazing has become an increasingly popular research topic recently.

Manuscript received September 13, 2020; revised January 6, 2021 and March 1, 2021; accepted March 12, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61922064 and Grant U2033210, in part by the Zhejiang Provincial Natural Science Foundation under Grant LR17F030001, and in part by the Project of Science and Technology Plans of Wenzhou City under Grant C20170008 and Grant ZG2017016. This article was recommended by Associate Editor J. Hou. *(Corresponding author: Xiaoqin Zhang.)*

The authors are with the College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou 325035, China (e-mail: jxwang@stu.wzu.edu.cn; zhangxiaoqinnan@gmail.com; taowangzj@gmail.com; ddghjikle1@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3067062.

Digital Object Identifier 10.1109/TCSVT.2021.3067062

The generation of haze in images can be described using the classical atmospheric scattering model [6], [7]:

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

where x represents the pixel position, I is the observed hazy image, J denotes the hazy-free radiance; A refers to the global atmospheric light and t denotes the medium transmission map. The physical model gives essential insights about image dehazing, however, without the knowledge of A and t, image dehazing based on this model becomes an under-determined estimation problem.

To estimate the unknowns A and t to solve the under-determined estimation problem, most dehazing methods use either physical grounded priors or data-driven ways. Specifically, He *et al.* [10] developed dark channel prior to obtain the transmission map. Zhu *et al.* [11] investigated an attenuation before recovering depth information for transmission map estimation. Wang *et al.* [12] proposed their method based on the physical model and the brightness components of the image. Although these prior-based methods have demonstrated their superiority, their assumptions do not accurately reflect the inherent properties of images. Thus, their performance tends to be limited. Recently, deep-learning-based methods have been employed to estimate the atmospheric light and transmission map. In DehazeNet [13], the transmission map is determined in an end-to-end manner. In DCPDN [14], the atmospheric light and the transmission map are estimated simultaneously.

With the strong representation power of convolutional neural networks, These aforementioned methods achieve a better dehazing performance. However, an inaccurate estimation of the transmission map or the atmospheric light would significantly interfere with the restoration of images. Unlike these methods, model-free methods [8], [9], [15]–[17] that directly restore haze-removed images from their hazy counterparts have demonstrated remarkable dehazing performance.

To accurately estimate the haze levels and generate images with sharp structural details, we propose hierarchical feature fusion schema for image dehazing. The hierarchical feature fusion in this work is investigated in a supervised manner with hazy-free ground truth images as learning targets. Using a composition of simple functions, the network first learns low-level features using simple functions of the input, these features are then aggregated level-by-level to generate increasingly rich feature representations. Based on these extracted intermediate feature representations, the target can be

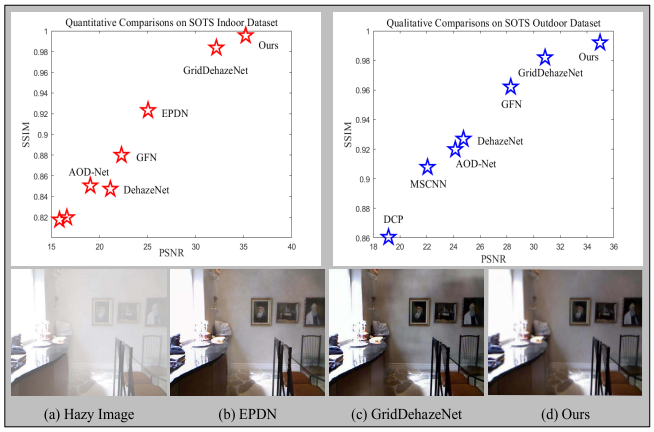


Fig. 1. Quantitative and qualitative results. The quantitative comparisons are shown in the top two images. From the qualitative results, EPDN [8] tends to underestimate the hazy level, thus the restored image contains remaining haze. In the present of artifacts and color distortions, GridDehazeNet [9] cannot generate visually pleasant haze-removed images. The proposed method achieves the best visual result.

accurately estimated with properly feature fusion mechanism and image reconstruction module.

When it comes to the feature sensitive image dehazing task, investigating appropriate strategies to transform the redundant parts in features into richer ones is a promising way to improve the dehazing performance. Boosting the representation power of hierarchical feature fusion, a mixed convolution attention module is designed to transform these ineffective and redundant features into valuable ones as well as adaptively highlight task-relevant features. When the attention module is appropriately utilized, we find that the dehazing model can generate haze-removed images with sharper structural details, less halo artifacts and color distortion.

The proposed model-free method has three stages. We first extract rich features at different depths of the network with hierarchical learning. Subsequently, these features are fused with several mixed convolution attention modules. Given the aggregated features and input hazy images, we accurately recover the haze-removed results. In addition, aiming at learning more semantic information, we propose a deep semantic loss to measure the semantic difference in deep features extracted from the dehazed results and their ground truth images. Extensive experimental results have shown that the proposed approach outperforms the state-of-the-art methods. Fig. 1 gives some quantitative and qualitative comparison results.

The main contributions of this work are summarized as follows:

- We propose a hierarchical feature fusion framework with attention mechanism to adaptively enhance the performance of the scale-sensitive image dehazing task.
- A novel spatial attention mechanism termed as mixed convolution attention module is proposed, which has the following advantages: reducing feature redundancy, learning compact and effective internal representations, and adaptively highlighting task-relevant features. Therefore, the dehazing model can generate images with less remaining haze, sharper textural details and vivid colors.
- Extensive experimental results on benchmark datasets and detailed analysis have demonstrated the effectiveness and superiority of the proposed dehazing method.

The rest of this paper is organised as follows. We first introduce the related work in Section II. Then we present the proposed dehazing method with detailed descriptions for hierarchical feature fusion, mixed convolution attention, and loss functions in Section III. Section IV describes experimental setup including datasets, evaluation metrics, implementation details, ablation study, model analysis, and extensive experimental results for performance evaluation and comparison. Finally, conclusions are given in Section V.

II. Related Work

A. Single Image Dehazing

Significant advances in single image dehazing have been witnessed in recent years. These methods can be roughly classified into two categories.

1) Prior-Based Image Dehazing Methods: Prior-based methods estimate the transmission maps and atmospheric light intensity based on the statistics of clear images. He *et al.* [10] propose DCP to estimate the transmission map of hazy images. Their assumption is that at least one color channel should have a very low intensity within a haze-free image that does not contain sky or bright regions. Zhu *et al.* [11] propose an efficient color attenuation prior. Using a linear model to build a bridge between the hazy image and its depth information, the method can estimate the transmission and restore the scene radiance to remove the haze. Berman *et al.* [18] propose their method based on a non-local prior. They assume that the colors of a haze-free image can be approximated by less compact and typical colors clustered in RGB space. Zhang *et al.* [19] introduced their dehazing method based on local physical features. Although these prior-based methods have provided meaningful insights for image dehazing, they may fail in cases where the priors or assumptions are invalid.

2) Learning-Based Image Dehazing Methods: With advancements in deep neural networks and the availability of large-scale datasets, data-driven approaches have received significant attention recently. Cai *et al.* [13] introduce DehazeNet to estimate the medium transmission map, which is then used to restore haze-free images based on the atmospheric scattering model. A multi-scale deep neural network is employed by Ren *et al.* [20] to estimate the scene transmission maps. They first predict the transmission maps using a coarse-scale network, then utilize a fine-scale network to refine the estimated results. Note that, Ancuti and Ancuti [21] are the first to design a fusion-based approach and demonstrate its effectiveness and potential for dehazing task. Li *et al.* [22] employ AOD-Net to directly restore dehazed images instead of estimating the atmospheric light and the transmission matrix. Zhang and Patel [14] propose DCPDN to simultaneously learn transmission map and atmospheric light in an end-to-end manner.

Without relying on the physical scattering model, Mei *et al.* [23] propose a progressive feature fusion network.

They utilize an U-like encoder-decoder network to learn the transformation from hazy images to their ground-truth images. Qu *et al.* [8] directly generate the haze-removed images with EPDN after embracing image dehazing as an image-to-image translation problem. Liu *et al.* [9] propose an end-to-end trainable network termed GridDehazeNet for haze removal. Zhang *et al.* [24] propose an efficient multi-scale single image dehazing approach using perceptual pyramid network, in order to directly learn the non-linear mapping between hazy images and their haze-free counterparts. The dehazing network is optimized with the L2 loss and perceptual loss. The differences between our work and [24] are analyzed as follows: First, Zhang *et al.* employ an encoder-decoder architecture based on residual and dense blocks to extract image features for image reconstruction, while we utilize the hierarchical features extracted from different depths of the network which will relatively provide more contextual and semantic information. Through fusing hierarchical features, the haze levels can be accurately estimated and the image structures will be clearly recovered. Second, Zhang *et al.* employ the multi-scale pyramid pooling module to handle the extracted features. In our work, the multi-scale protocols are conducted in two levels: (1). multi-scale hierarchical feature fusion; (2) mixed convolution attention for fine-grained multi-scale feature adjustment. After the two steps, more richer and task-relevant representations can be obtained. Third, Zhang *et al.* employ the perceptual loss to optimize their network. We not only employ the perceptual loss but also employ a newly proposed deep semantic loss, in which the Laplace operator is used to highlight semantic information in deep features.

B. Attention Mechanism

Inspired by the important role of attention in human perception, the attention mechanism has become a popular component in deep neural networks. Significant improvements in various tasks have been achieved in recent years. For example, Chen *et al.* [25] incorporate both spatial and channel-wise attention for image captioning. The proposed SCA-CNN model outperforms many other visual attention-based image captioning algorithms. Based solely on attention mechanism, Vaswani *et al.* [26] propose a transformer using stacked self-attention to draw global dependencies between the input and output for neural sequence transduction models. Woo *et al.* [27] propose a convolution block attention module with channel and spatial sub-modules to adaptively refine features. An attention on attention module is proposed by Huang *et al.* [28] to determine the relevance between attention results and queries. Aiming to exchange and aggregate information in a more flexible manner in dehazing task, Liu *et al.* [9] integrate GridDehazeNet with a channel-wise attention mechanism. Different from the attention introduced by GridDehazeNet, we propose a novel spatial attention with mixed convolution operations to highlight task-relevant features. With the proposed attention, which can help generate haze-removed images with vivid colors and sharper structure details, the performance is far superior to GridDehazeNet.

III. Methodology

Intuitively, restoring images to their haze-free state with excellent visibility requires rich feature representations. The hierarchical feature extraction, which is presented in Section III-A, is proposed to progressively extract multi-scale features at different depths of the deep networks. After capturing these rich features, multiple mixed convolution attention modules are employed as plug-and-play tools to boost the representation power of the network by focusing on important features and suppressing unnecessary ones. The mixed convolution attention module is detailed in Section III-C. Subsequently, we investigate an efficient way to aggregate these features, the processed features are then provided as inputs to the final image reconstruction module. The entire network architecture is presented in this section. Fig. 2 illustrates a typical architecture variant of the proposed dehazing network.

A. Feature Extraction

Generally, using features extracted from an isolated layer is inadequate. Features from earlier layers mainly focus on low-level information such as edges and shapes, which are essential for locating the positions of different objects and for restoring the structural details in images. While features from latter layers have meaningful and richer semantic information that is beneficial for better preserving the textural and color details. Through aggregating shallow and high-level features, the dehazing method can generate visual pleasant results with less remaining haze and sharper structures.

In this work, to aggregate hierarchical features captured from different layers, we apply several feature extraction blocks with multiple scales to attain this objective. Given a hazy image $I \in \mathbb{R}^{C\times H\times W}$ as input, the extracted hierarchical features can be expressed as:

$$F_s = f_s(F_{s-1}), \quad s = 1, \ldots, N, \tag{2}$$

where F_s denotes the features extracted by the s-th feature extraction block f_s, and $F_1 = f_1(I)$. N is the total number of scales utilized in this network.

The architecture of feature extraction block is shown at the bottom of Fig. 2. We first obtaining the downsampling features with a convolution layer; then compute the output features with residual mechanism:

$$F_{mid} = \downarrow F_{s-1} + \varphi(\delta_1(\varphi(\downarrow F_{s-1}))), \tag{3}$$

where $\downarrow$ indicates the downsampling operation; φ means the common convolution layer; δ_1 denotes the ReLU function; and F_{mid} is the middle output features. Similarly, the s-th features F_s can be computed with another residual module without the downsampling operation. The details are also shown in Fig. 2.

B. Feature Fusion

After extracting the hierarchical features F_s as expressed in Eq. (2), we refine them with mixed convolution attention modules:

$$\widehat{F}_s = \mathcal{ATT}(F_s), \quad s = 1, 2, \cdots, N, \tag{4}$$

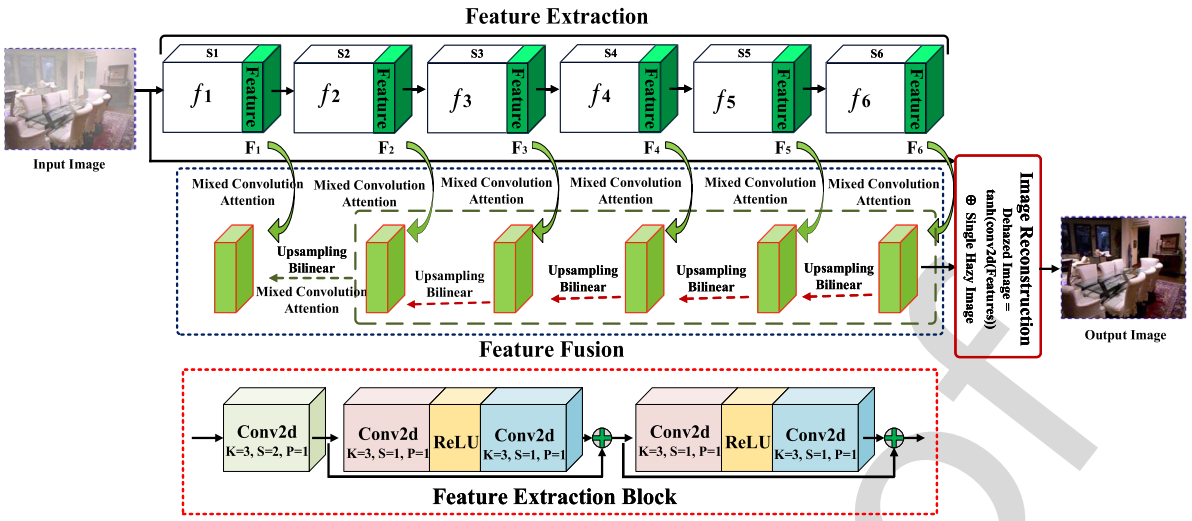


Fig. 2. A typical network variant of the proposed dehazing method. Here, the hierarchical features are extracted by six feature extraction blocks, then these extracted features are refined by six mixed convolution attention modules. These attention modules transform input features into other feature representations with richer task-relevant texture information. Finally, given the fused features and the hazy image as input, an image reconstruction module is utilized to accurately restore the haze-removed image. The green and black arrows indicate the processing of mixed convolution attention modules and the data flow in the network.

where $\mathcal{ATT}(\cdot)$ refers to the mixed convolution attention, $\widehat{F}_s$ is the refined result.

Without loss of generality, we employ a network variant with 6 scales as an example, which is illustrated in Fig. 2. Note that, the variant can achieve the best performance on indoor scenarios. The details of ablation studies and model analysis are presented in Section IV-B. To fuse features $F_i, i = 1, \cdots, N$, we progressively upsample them with bilinear interpolation operations, so that the feature concatenation can be achieved in a clean manner given aligned features. The feature fusion process can be expressed as:

$$\widehat{F}_{out} = [\widehat{F}_1, \uparrow \mathcal{ATT}([\uparrow \widehat{F}_2, \cdots, \uparrow \widehat{F}_N])], \tag{5}$$

where $\uparrow$ is the upsampling operation, $[\cdot]$ denotes the feature concatenation operation, and $\widehat{F}_{out}$ represents the output fused features of an input image. In Eq. (5), we first apply another attention module to adjust the extracted features, then $\widehat{F}_1$ is concatenated to obtain the $\widehat{F}_{out}$. Finally, the haze-removed image $\widehat{J}$ can be correctly estimated using the image reconstruction module:

$$\widehat{J} = \delta_2(\varphi(\widehat{F}_{out})) \oplus I, \tag{6}$$

where δ_2 is the tanh function, $\oplus$ indicates the element-wise add operation, which refers to the network-level skip-connection of residual learning, and $\widehat{J}$ is the haze-removed image.

C. Mixed Convolution Attention

Extracting features and aggregating them based on hierarchical learning are essential for image dehazing; however, the kernel size used in the different convolution layers is often overlooked in attention mechanism for image dehazing. When designing their networks, people often simply choose 3×3 or 5×5 kernels for convolution layers. Although depthwise convolution [29] separately employ kernels to each individual channel of the input features, it ignores the effect of kernel sizes utilized in single layer. Besides, the redundancy in features has not been thoroughly investigated to boost the representation power. We find the similar or redundant features can be transformed into valuable ones and can help estimate haze-removed images with sharp textural details and vivid colors.

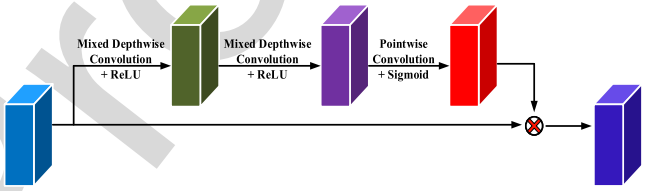


Fig. 3. The architecture of Mixed Convolution Attention Module (MCAM). The module is composed of mixed depthwise convolution, pointwise convolution, ReLU and Sigmoid operations. This module transforms any input features into other representations of the same size but with more compact, effective and task-relevant information. It significantly improves the dehazing performance.

In this work, the proposed attention module for adaptive feature refinement consists of mixed depthwise convolution and pointwise convolution operations, Fig. 3 depicts the detailed process. The attention module has three components, and the operations are specially designed to processing input information in a lightweight and sensitive manner. In Eq. (4) and Eq. (5), we use $\mathcal{ATT}(\cdot)$ to denote the mixed convolution attention module, which is expressed as follows:

$$\widehat{F}_s = \mathcal{ATT}(F_s) = F_s \odot W_{F_s}, \tag{7}$$

where $\widehat{F}_s$ represents the refined feature maps. W_{F_s} is the weight for each element in F_s, and $\odot$ denotes the element-wise multiplication operation. From the equation, we can see the mixed convolution attention module is a spatial attention mechanism, which spatially highlights task-relevant information rather than selecting key channels.

Fig. 3 shows that the W_{F_s} is obtained by two mixed depthwise convolutions, one pointwise convolution and activation

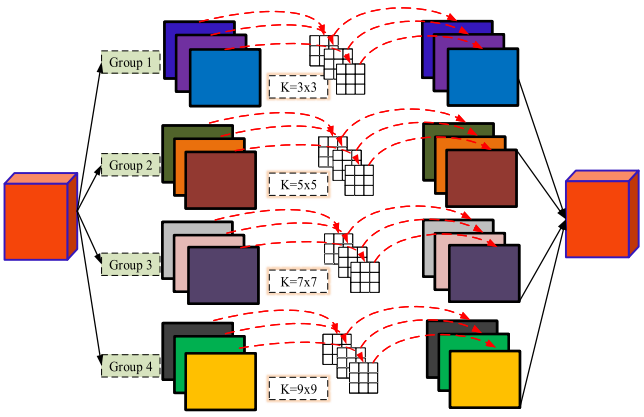


Fig. 4. The architecture of Mixed Depthwise Convolution Layer (MDCL). Without loss of generality, we first split the input feature maps into 4 groups, then perform depthwise convolution with different kernel sizes on each group of features. Finally, the output features are obtained by concatenating the 4 groups of processed features.

functions. We detail the operations as follows:

$$W_{F_s} = \delta_3(\kappa_p(\delta_1(\kappa_m(\delta_1(\kappa(F_s)))))). \tag{8}$$

In this equation, κ_m and κ_p refer to the mixed depthwise convolution and pointwise convolution. δ_1 and δ_3 refer to the ReLU and Sigmoid functions. We should note that the dimensions of both the channels and spatial axes are unaltered when information flows within the module. Therefore, the attention module can be integrated into any end-to-end trainable network architectures.

1) Mixed Depthwise Convolution: Different from conventional convolution operations, one convolution kernel of depthwise convolution is responsible for one channel, and one channel is convolved by only one convolution kernel. The proposed mixed depthwise convolution is composed of multiple depthwise convolution operations with different kernel sizes. The details of the operations for 4 groups of features are illustrated in Fig. 4. More specifically, the input feature maps F^{in} are partitioned into G groups. For each group of feature maps, we employ a depthwise convolution with different kernel sizes to process them. The kernel size for each group is formulated as: $k \times k$, $k = 2 \times (g-1) + 3$, where $g = 1 \cdots G$, and k is the kernel size. The output features F^{out} can be represented as:

$$F^{out} = [dw(F^{in}_{g=i}; k_{g=i})], \quad i = 1, \cdots, G, \tag{9}$$

where $dw(\cdot)$ is the depthwise convolution operation, $F^{in}_{g=i}$ refers to the i-th group of input feature maps and similarly $k_{g=i}$ is the kernel size of group i, and $k_{g=i} = 2 \times (i-1) + 3$. $[\cdot]$ is also the feature concatenation operation for the G groups of feature maps. Besides, the output features F^{out} can also be processed by an appropriate activation function.

2) Pointwise Convolution: To effectively utilize the information of the input feature maps in the same spatial location, the pointwise convolution operation with the kernel size of 1×1 is applied. The pointwise convolution can be seen as a special case of the mixed depthwise convolution when G is set as the number of channels of the input feature maps and all convolution operations have the same kernel size 1×1. Given the input feature maps F^{in}, the pointwise convolution operation can be formulated as:

$$F^{out} = [dw(F^{in}_{g=i}; k = 1)], \quad i = 1, \cdots, C, \tag{10}$$

where C is the total number of channels in the feature maps F^{in} and the $[\cdot]$ indicates the feature concatenation operation as well.

D. Loss Functions

To better optimize the proposed dehazing network, four loss functions are employed to measure the quantitative difference between the haze-removed images and their ground truth images. The four losses are mean square error (MSE), smooth L_1 loss, perceptual loss, and deep semantic loss. The proposed deep semantic loss is utilized to emphasize necessary semantic information in deep features. We first introduce the four losses, respectively, then use them to obtain the joint loss function.

1) Mean Square Error: The MSE is utilized to accurately capture most of the low frequencies in the images. This information is crucial for recovering high-quality images. The loss function is defined as:

$$L_{mse} = \frac{1}{CHW} \left\| \widehat{J} - J \right\|^2, \tag{11}$$

where $\widehat{J}$ represents the restored image and J refers to the corresponding ground truth image; and C, H, W refer to the number of channels, height and width, respectively.

2) Smooth L_1 Loss: L_1 norm can enforce the correctness at low frequencies as well. Moreover, $SmoothL_1$ loss is less sensitive to outliers and can alleviate the gradient explosion problem. The loss function is expressed as:

$$L_{smo} = \frac{1}{CHW} \psi(\widehat{J} - J),$$
$$where\ \psi(\varepsilon) = \begin{cases} 0.5\varepsilon^2, & \text{if } |\varepsilon| < 1, \\ |\varepsilon| - 0.5, & \text{otherwise.} \end{cases} \tag{12}$$

3) Perceptual Loss: Aiming at enforcing the network to recover images with low-to-high level semantic fidelity and high standard visual quality, we measure the difference in features between $\widehat{J}$ and J using perceptual loss:

$$L_{per} = \sum_{l \in \{4,9,16\}} \frac{1}{C_l H_l W_l} \left\| \phi_l(\widehat{J}) - \phi_l(J) \right\|^2, \tag{13}$$

where ϕ_l is the l-th feature extractor corresponding to the VGG16 network, the C_l, H_l, and W_l denote the number of channels, height and width of the feature maps, respectively, extracted from the l-th layer of the VGG16.

4) Deep Semantic Loss: Inspired by the perceptual loss which measures the difference between feature maps, Laplace operator [30] is employed to highlight the semantic details in extracted features with levels from low to high. Moreover, the Laplace operator is formulated in the deep semantic loss in an end-to-end manner. The loss function can be expressed as follows:

$$L_{sem} = \sum_{l \in \{4,9,16\}} \frac{l}{C_l H_l W_l} \left\| \delta_2(\xi(\phi_l(\widehat{J}))) - \delta_2(\xi(\phi_l(J))) \right\|_1, \tag{14}$$

where $\xi(\cdot)$ denotes the Laplace operator used to extract detailed semantic information for both the estimated image $\widehat{J}$ and the ground truth J. δ_2 is the tanh function. The L_1 norm is used to measure the difference in deep semantic information between $\widehat{J}$ and J.

5) Total Loss: During training, all the models are optimized by minimizing the following loss function L_{tot}:

$$L_{tot} = L_{mse} + L_{smo} + \lambda_1 \cdot L_{sem} + \lambda_2 \cdot L_{per}. \quad (15)$$

The λ_1 and λ_2 are used to control the interaction of these loss components.

IV. Experiments

In this section, we first introduce the dehazing datasets, evaluation metrics along with some implementation details about training and evaluating the proposed dehazing method. Then, detailed ablation studies and analysis are conducted to make these proposed components and our statements more convincing. Finally, we report the evaluation results of extensive experiments conducted on both synthetic and real-world benchmarks. We compare and analyze the proposed approach in terms of quantitative accuracy and visual quality with several state-of-the-art methods.

A. Experimental Settings

1) Datasets: Generally, it is prohibitively expensive to collect a large number of real-world hazy images and their haze-free counterparts. Therefore, we train and evaluate the proposed method on a synthetic dataset, namely RESIDE [31]. The RESIDE dataset includes synthetic hazy images in both indoor and outdoor scenarios. The indoor training set (**ITS**) contains a total of 13990 hazy indoor images, generated from 1399 haze-free images with $\beta \in [0.6, 0.8]$ and $A \in [0.7, 1.0]$ based on the atmospheric scattering model; the depth maps are obtained from the NYN Depth V2 [32] and Middlebury Stereo datasets. The outdoor training set (**OTS**) contains a total of 296695 hazy outdoor images, generated from 8477 haze-free images. For both the indoor and outdoor scenarios, we train the proposed model on the ITS and OTS datasets, respectively. For model evaluation, we adopt the synthetic objective testing set (**SOTS**), which contains 500 pairs of indoor images and 500 pairs of outdoor ones.

The challenge [33] contains two real-world image dehazing datasets: **I-HAZE** [34] and **O-HAZE** [35]. The I-HAZE dataset contains 35 pairs of hazy images and the corresponding haze-free images of various indoor scenes. While the O-HAZE dataset includes 45 pairs of hazy images and the corresponding ground truth images of various outdoor scenes. To keep the fairness of comparison, we finetune the proposed model using the training set and evaluate it on the corresponding testing part, following the setup of the NTIRE 2018 Image Dehazing Challenge.

2) Evaluation Metrics: In this work, the peak signal to noise ratio (**PSNR** [36]) and the structural similarity index (**SSIM** [37]) are utilized to evaluate the quality of the restored images. Furthermore, we compare the subjective visual effect of the restored images with other algorithms on the SOTS, I-HAZE, and O-HAZE datasets.

3) Implementation Details: In implementation, we utilize the Adam [38] algorithm with hyper-parameters β_1 and β_2 set as 0.5 and 0.999, respectively. The initial learning rate is set as 0.0001 and decayed by $\gamma = 0.1$ for every 50 epochs. The models are trained for about 150 epochs on indoor images and about 10 epochs on outdoor images. The trade-off hyper-parameters both λ_1 and λ_2 in the total loss function Eq. (15) are set as 0.01 and 0.04. We conduct all the training, testing, and model analysis on the same platform with an Intel Xeon Silver 4114 CPU, 32 GB RAM and a single NVIDIA RTX 2080 Ti GPU.

B. Ablation Study and Model Analysis

In this section, aiming at analyzing the performance of the proposed method, we conduct extensive ablation studies to investigate the effectiveness of scales used in hierarchical feature fusion and architecture settings for mixed convolution attention module. All variants of the proposed method mentioned below are trained and evaluated on the same experimental setting for fair comparison. The loss Eq. (15) is employed to optimize these variants. Table I lists the evaluation results on both indoor and outdoor datasets. For convenience, the evaluation results of GridDehazeNet are also listed here to compare the dehazing performance.

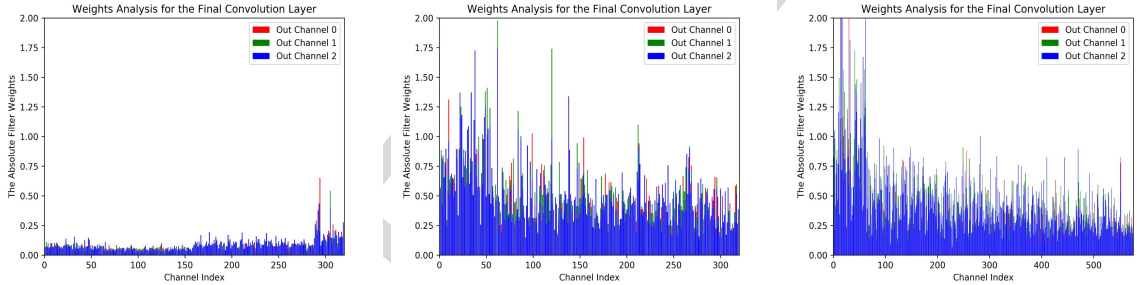
1) Model Introduction: In Table I, S4, S5 and S6 indicate the number of scales utilized for hierarchical feature fusion, while G2, G4, G6 and G8 denote the number groups employed in the mixed convolution attention module. For each group of features, the corresponding mixed depthwise convolution has its kernel size, the details are shown in Section III-C.1. The Baseline only contains 4 feature extraction blocks and an image reconstruction block. The Baseline has the simplest network architecture without hierarchical feature fusion or any mixed convolution attention modules. The Baseline+S4 means four scales of features are fused in this model without employing the attention mechanism. Besides, S4G2 indicates that the model not only has feature fusion for features extracted by four feature extraction blocks but also is equipped with multiple mixed convolution attention modules. G2 means that the input features in attention modules are split into 2 groups. The other model settings in this table have similar definitions for both the indoor and outdoor scenarios.

2) Hierarchical Feature Fusion Analysis: To validate the effectiveness of the hierarchical feature fusion mechanism, experiments and analysis are conducted on both indoor and outdoor datasets. Compared with Baseline, Baseline+S4 achieves a performance gain of approximately 1dB in terms of the PSNR. However, the variant cannot obtain the best performance only equipped with multi-scale feature fusion. On the indoor dataset, the effect of the number of scales in the hierarchical feature fusion is investigated through variants S4G2, S5G2 and S6G2. We can see that the dehazing performance increases rapidly as we increase the scales for feature extraction and fusion. On the outdoor dataset, all the variants have obtained better performance than GridDehazeNet. The S4G2 is able to achieve the best dehazing result.

TABLE I

THE EFFECTIVENESS OF UTILIZING DIFFERENT SCALES FOR HIERARCHICAL FEATURE FUSION AND DIVERSE ARCHITECTURE SETTINGS FOR MIXED CONVOLUTION ATTENTION MECHANISM. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE COLORS, RESPECTIVELY. ↑ MEANS THAT THE BETTER ALGORITHM SHOULD ACHIEVE A HIGHER SCORE FOR THIS METRIC. "√" INDICATES WE SELECT THE NETWORK SETTING

Models	S4	S5	S6	G2	G4	G6	G8	PSNR (↑)	SSIM (↑)	Param. Num
Models trained and evaluated on indoor images										
GridDehazeNet								32.16	0.9836	-
Baseline								27.98	0.9683	1.296 M
Baseline+S4	√							28.98	0.9771	1.301 M
S4G2	√			√				31.59	0.9922	1.436 M
S4G4	√				√			32.49	0.9933	1.465 M
S4G6	√					√		32.23	0.9919	1.505 M
S4G8	√						√	32.58	0.9921	1.562 M
S4G2	√			√				31.59	0.9922	1.436 M
S5G2		√		√				33.64	0.9940	2.293 M
S6G2			√	√				35.17	0.9954	3.182 M
S6G4			√		√			35.22	0.9954	3.236 M
Models trained and evaluated on outdoor images										
GridDehazeNet								30.86	0.9819	-
Baseline								28.82	0.9616	1.296 M
Baseline+S4	√							30.93	0.9836	1.301 M
S4G2	√			√				34.98	0.9920	1.436 M
S4G4	√				√			33.64	0.9905	1.465 M
S4G6	√					√		33.29	0.9897	1.505 M
S4G8	√						√	32.96	0.9889	1.562 M
S5G2		√		√				33.44	0.9894	2.293 M
S5G4		√			√			32.95	0.9888	2.334 M
S5G6		√				√		32.85	0.9872	2.391 M
S5G8		√					√	32.48	0.9871	2.472 M



(a) Baseline+S4 (b) S4G4 (c) S6G4

Fig. 5. The absolute weight analysis of the final convolution layer for different model variants. The final convolution layers in Baseline+S4 and S4G4 have 320 input feature channels and 3 output channels. The final convolution layer in S6G4 has 576 input channels and the number of output channels is 3. All convolution filters have the same kernel size 3×3. The absolute weight for each filter is obtained by calculating the sum of the absolute values of the 9 weights in the filter. In this figure, these convolution filters are split into 3 groups according to the number of output channels.

3) Mixed Convolution Attention Analysis: Based upon the evaluation results of S4G2, S4G4, S4G6, and S4G8 on the indoor dataset, we find the mixed convolution attention modules can significantly improve the dehazing performance. For example, S4G8 achieves PSNR and SSIM scores of 32.58 dB and 0.9921, which obtains a performance improvement of approximately 3.60dB\0.015 over the Baseline+S4. In addition, the number of groups in the mixed depthwise convolution also affects the performance on indoor images. When the feature maps are divided into 8 groups, the dehazing model achieves the best evaluation results between the four variants. Due to the balance between model complexity and dehazing performance, S6G4 is selected as the best dehazing model on indoor images. However, on outdoor dataset, the evaluation performance slightly decreases as we increase the number of groups in the attention module. This may be caused by overfitting on the train images when models have too strong presentation power. Thus, on outdoor images, we select the S4G2 as our final model.

TABLE II

THE EFFECTIVENESS OF DEEP SEMANTIC LOSS. THE "-" INDICATES WE DO NOT USE THE DEEP SEMANTIC LOSS WHEN TRAINING THE MODELS. ALL MODELS HERE ARE TRAINED FOR 100 EPOCHS. WE SELECT THEIR BEST EVALUATION RESULTS AND REPORT THEM IN THIS TABLE

Models	PSNR (↑)	SSIM (↑)
S6G2-	34.6758	0.9948
S6G2	35.0675	0.9953
S6G4-	34.1381	0.9947
S6G4	34.5392	0.9918

4) Model Parameter Analysis: Additionally, the mixed convolution attention module is lightweight as well as capable of significantly improving the dehazing performance. From the number of parameters in S4G2, S4G4, S4G6, and S4G8,

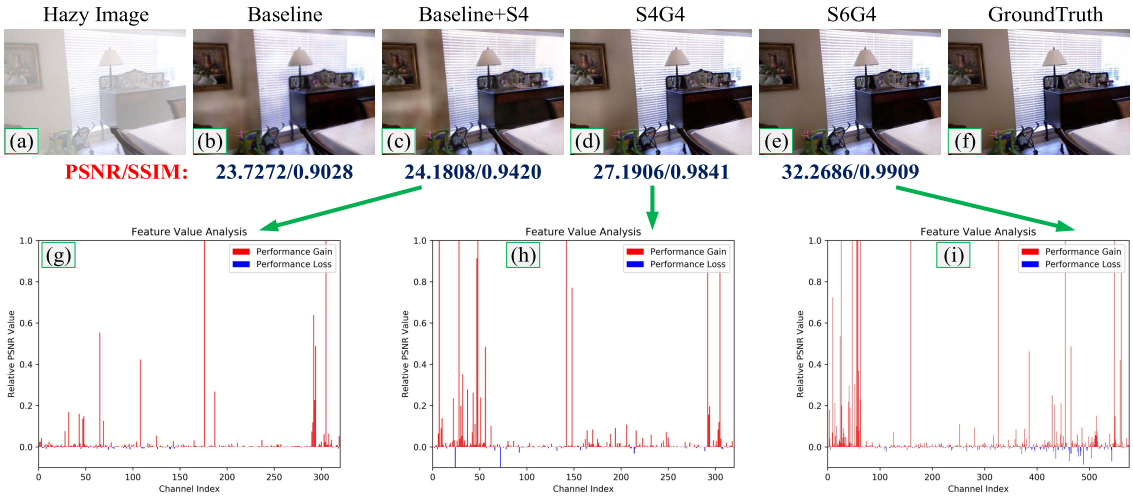


Fig. 6. Visual Comparisons and Feature Value Analysis. The first row shows the visual comparison results between different schemes given the same input hazy image. For each variant, we list the PSNR and SSIM values under its restored image. Images (g), (h), and (i) give the value of each channel of the input feature maps for Baseline+S4, S4G4, S6G4, respectively. The positive value of each channel indicates that it has positive impact on the dehazing performance; while a negative value of each channel means that the feature channel decreases the dehazing performance.

TABLE III

QUANTITATIVE EVALUATION ON THE INDOOR SCENARIOS FROM THE SOTS DATASET. THE RED AND BLUE TEXTS INDICATE THE BEST AND SECOND-BEST PERFORMANCES. ↑ MEANS THAT THE BETTER ALGORITHM SHOULD ACHIEVE A HIGHER SCORE FOR THIS METRIC

Methods	DCP	DehazeNet	MSCNN	AOD-Net	DCPDN	GFN	EPDN	GridDehazeNet	Ours
PSNR (↑)	16.62	21.14	19.84	19.06	15.85	22.30	25.06	32.16	35.21
SSIM (↑)	0.8197	0.8472	0.8327	0.8504	0.8175	0.8800	0.9232	0.9836	0.9954

TABLE IV

QUANTITATIVE EVALUATION ON THE OUTDOOR SCENARIOS FROM THE SOTS DATASET. THE RED AND BLUE TEXTS INDICATE THE BEST AND SECOND-BEST PERFORMANCES. ↑ MEANS THAT THE BETTER ALGORITHM SHOULD ACHIEVE A HIGHER SCORE FOR THIS METRIC

Methods	DCP	DehazeNet	MSCNN	AOD-Net	GFN	GridDehazeNet	Ours
PSNR (↑)	19.14	24.75	22.06	24.14	28.29	30.86	34.98
SSIM (↑)	0.8605	0.9269	0.9078	0.9198	0.9621	0.9819	0.9920

TABLE V

QUANTITATIVE EVALUATION ON THE I-HAZE AND O-HAZE DATASETS IN TERMS OF THE PSNR AND SSIM. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE COLORS, RESPECTIVELY. ↑ MEANS THE BETTER ALGORITHM SHOULD ACHIEVE A HIGHER SCORE FOR THIS METRIC

Datasets	Methods	DCP	AOD-Net	MSCNN	GFN	PFFNet	GridDehazeNet	DuRN	Ours
I-HAZE	PSNR (↑)	14.43	13.98	15.22	15.84	16.01	17.22	21.23	21.40
	SSIM (↑)	0.752	0.732	0.755	0.751	0.740	0.732	0.842	0.887
O-HAZE	PSNR (↑)	16.78	15.03	17.56	18.16	23.33	20.91	22.00	23.64
	SSIM (↑)	0.653	0.539	0.650	0.671	0.869	0.726	0.820	0.886

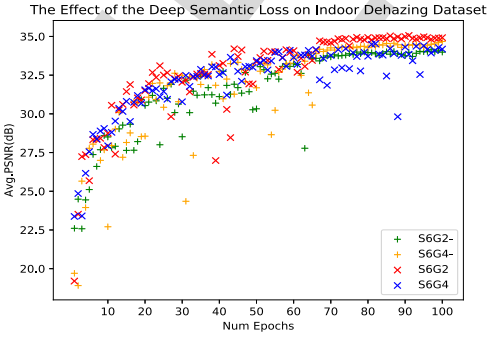


Fig. 7. Average evaluation results during training on indoor images. Models can achieve higher performance with the deep semantic loss.

we can see that the model complexity does not change too much as we increase the number of groups in the mixed convolution attention modules. In addition, by comparing Baseline+S4 and S4G2, a performance gain of approximately 2.61 dB PSNR value has been achieved with only 0.135 million additional learnable parameters.

5) Feature Redundancy Analysis: To make our experiments more convincing and verify the rationality of our statements, we investigate the feature redundancy in networks by analysing the learned filter weights of the final convolution layer in Baseline+S4, S4G4 and S6G4. To analyze the impact of each filter, we compute its absolute weights and the analysis results are shown in Fig. 5. The absolute weights of each filter is obtained by computing the sum of the absolute values of its 9 weights. From the plot, several observations can be made: (1). Fig. 5 (a) indicates that feature usage cannot be effectively performed by Baseline+S4 and the model indeed has many redundant features (with abundant low

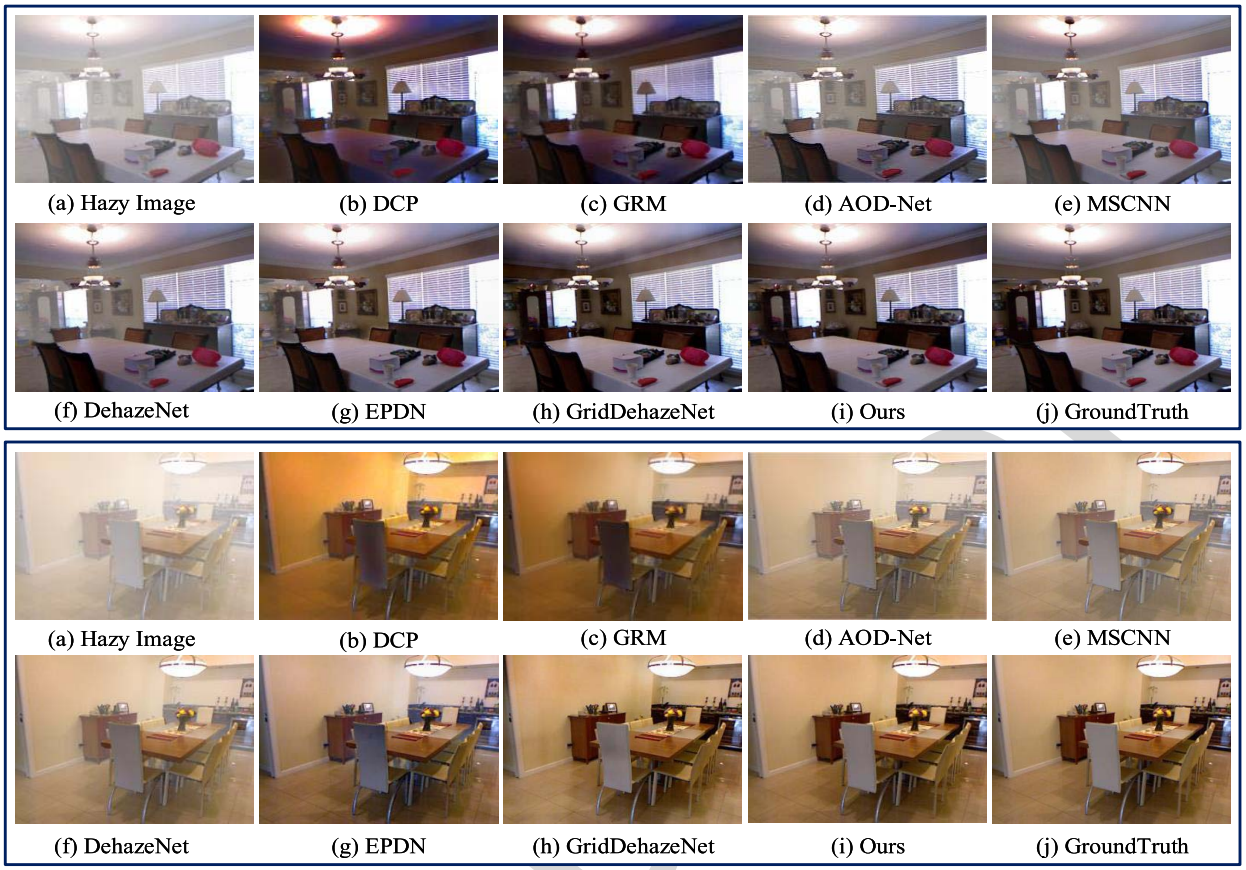


Fig. 8. Examples of haze-removed images on indoor scenarios. The proposed method generates more visually pleasant results than other state-of-the-art algorithms. **Best viewed in color and zoom in for better visibility.**

absolute weights). (2). With our proposed attention mechanism, compact internal feature representations are effectively learned and the feature redundancy is well reduced (the absolute weights are improved). (3). The learned features have more diversity (the absolute weights are diversified).

6) Visual Comparisons and Feature Value Analysis: The visual comparisons between Baseline, Baseline+S4, S4G2, and S6G4 are illustrated in the first row of Fig. 6. Based on the images Fig. 6 (c) and Fig. 6 (d) of Baseline+S4 and S4G4, respectively, we find that the proposed mixed convolution attention mechanism effectively alleviates the artifacts and color distortion problems in images, and the recovered structure details are sharper as well. Further, the S6G4 generates more visually pleasant haze-removed result, the artifacts and color distortions are completely resolved as we increase the scales for the hierarchical feature fusion and select appropriate number of groups for the proposed attention modules.

To investigate the relationship between each channel of the input feature maps with the final performance, we compute the value of each channel. Note that the features analysed here are the input features for image reconstruction module. We first compute the PSNR value of recovered images with all input feature channels, and the PSNR value is set as the baseline for evaluating the effectiveness of each channel. Then, we compute a new PSNR value with all channels as input except a specific channel, the value of the specific channel is defined as the difference between the baseline PSNR value

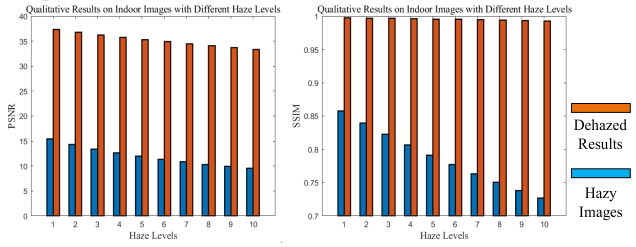


Fig. 9. Evaluation results on indoor images with different haze levels.

and the new computed PSNR value. Channels with positive values indicate that they have positive impacts on the dehazing performance. While channels with negative values mean that these channels will decrease the dehazing performance. From Fig. 6 (g), (h), and (i), we can know that feature redundancy has negative impacts on the dehazing performance to some extent. Models with the proposed mechanism can reduce feature redundancy and transform some ineffective features into valuable ones, therefore the performance is improved. Furthermore, the improvement of dehazing performance is the comprehensive result of multiple channels.

7) Effectiveness of Deep Semantic Loss: To validate the effectiveness of the deep semantic loss, several check experiments are conducted. In table II, S6G2- means that the model is optimized without the deep semantic loss, while S6G2 employs all four loss functions. S6G4- and S6G4 have

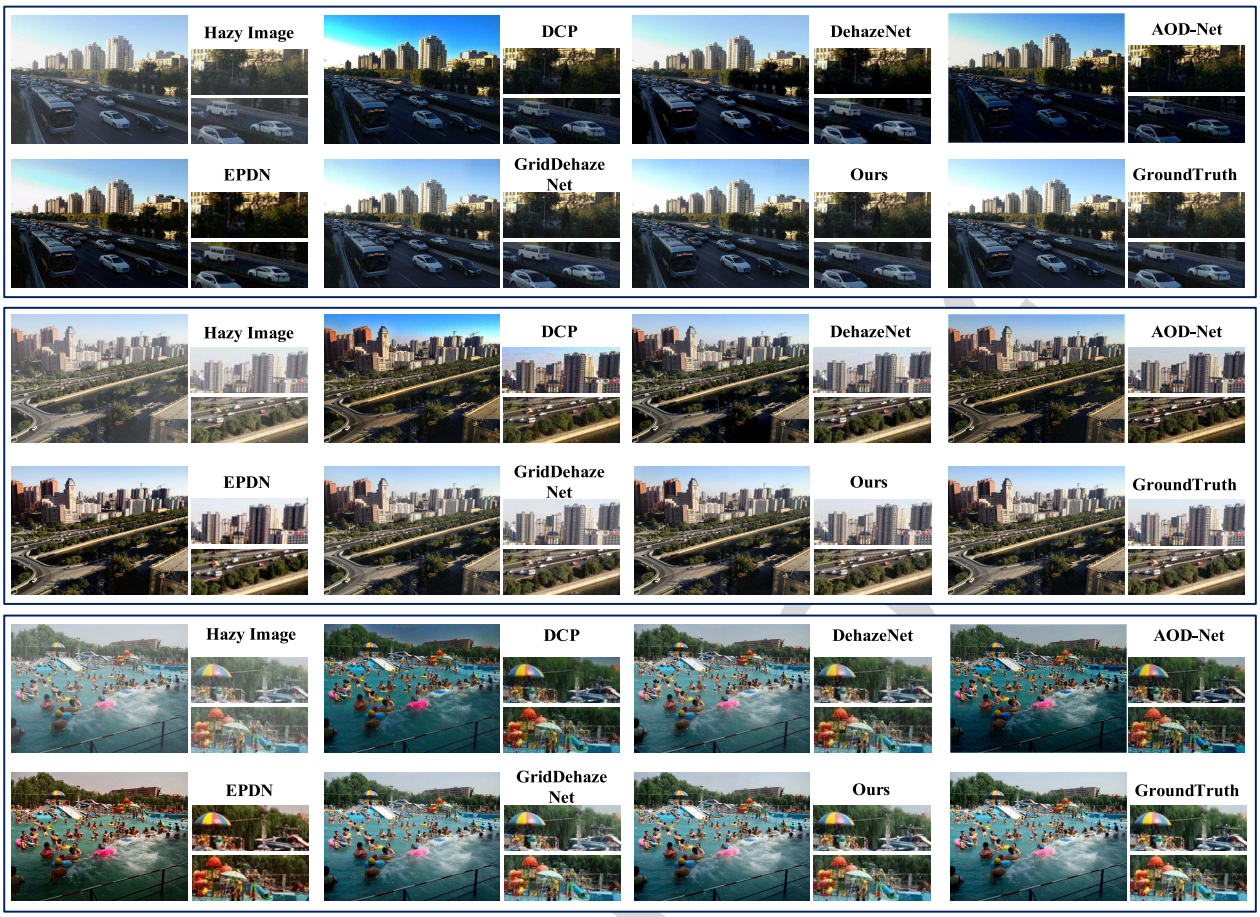


Fig. 10. Examples of haze-removed images on outdoor scenarios. For convenient comparison, we crop two patches from each image and magnify them to show the details. **Best viewed in color and zoom in for better visibility.**

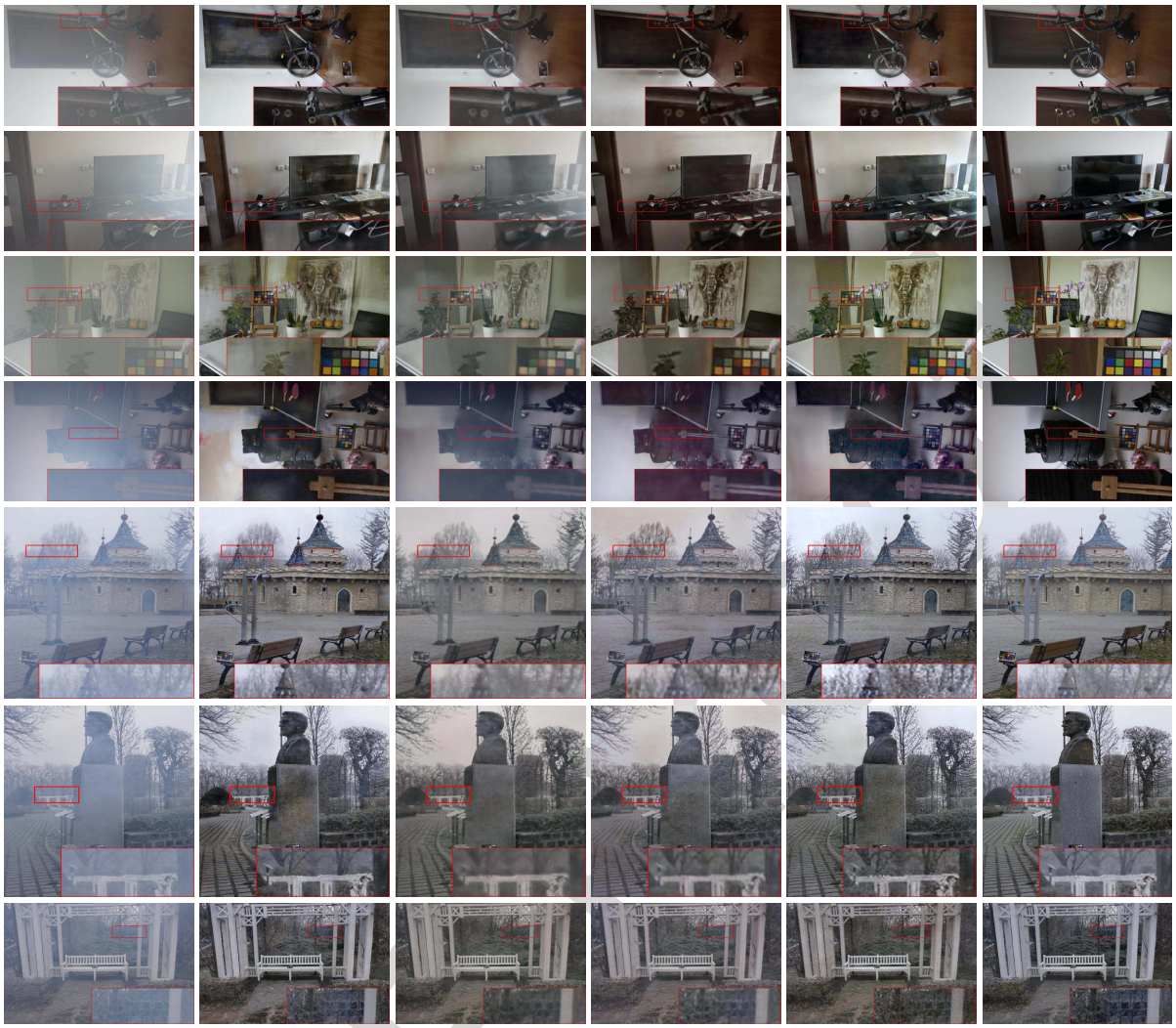
similar meanings as well. Note that, S6G2 and S6G4 are the most powerful variants for indoor scenarios, thus the experiment settings here are rigorous enough to investigate the effect of the deep semantic loss. All the models here are trained for 100 epochs, and we select the best evaluation results from them. From the evaluation results, we can see that the proposed deep semantic loss is able to help model optimization and improve the dehazing performance. The average evaluation results during training on indoor images are also illustrated in Fig. 7. The convergence speed is also accelerated with the loss function.

C. Performance Evaluation on Synthetic Images

To validate the effectiveness of the proposed dehazing method, extensive experimental comparisons between the proposed approach and several state-of-the-art methods are conducted on synthetic indoor and outdoor images. These compared methods include hand-crafted prior method (DCP [10]) and learning-based approaches (DehazeNet [13], MSCNN [20], AOD-Net [22], DCPDN [14], GFN [39], EPDN [8], and GridDehazeNet [9]).

Table III lists the evaluation results on indoor images. As expected, the dehazing method DCP does not perform well, which means that their proposed prior-based strategy does not effectively fit the problem. By estimating the medium transmission map to restore the haze-free images using an end-to-end architecture, DehazeNet achieves a performance improvement of 4.52 dB in terms of the PSNR metric over DCP. Similarly, MSCNN is also based on the atmospheric scattering model. MSCNN first estimates the transmission maps using a coarse-to-fine network, and then restores the haze-removed images with a fine-scale network. The evaluation results also demonstrate its effectiveness. DCPDN learns the transmission map and atmospheric light in an end-to-end manner, but it does not achieve a remarkable performance on the evaluation dataset. Without relying on the atmospheric scattering model, AOD-Net, EPDN and GridDehazeNet achieve better dehazing performance. This indicates that the estimation of the transmission map or the atmospheric light from a single hazy input is not a trivial task; while model-free methods that directly learn the map between a hazy input and its corresponding clean result perform better.

As listed in Table III, EPDN gives the PSNR and SSIM scores of 25.06 dB and 0.9232. GridDehazeNet achieves the performance of 32.16 dB on PSNR and 0.9836 on SSIM. Our proposed dehazing algorithm achieves a performance improvement of 3.05 dB in terms of the PSNR over GridDehazeNet, and the SSIM value is 0.9954. The evaluation results significantly outperforms all the compared dehazing methods, thus demonstrating the effectiveness of the proposed



(a) Hazy Image (b) PFFNet (c) GridDehazeNet (d) DuRN (e) Ours (f) GroundTruth

Fig. 11. Comparison of qualitative results between state-of-the-art methods and the proposed approach on the I-HAZE and O-HAZE datasets. The indoor results are shown in the four upper rows, and the outdoor results are shown in the bottom three rows. For convenient comparison, a patch is cropped and its magnified version is placed in the bottom-right corner of each individual image. **Best viewed in color and zoom in for better visibility.**

hierarchical feature fusion and mixed convolution attention mechanism. Furthermore, the evaluation results on synthetic outdoor dataset also demonstrate the effectiveness of the proposed algorithm. As listed in table IV, our approach achieves the best dehazing performace. To the best of our knowledge, we are the first to report a dehazing performance with approximately 34.98 dB of PSNR and 0.9920 of SSIM on the SOTS outdoor dataset.

Fig. 8 and Fig. 10 illustrate the comparisons between our method and other methods in terms of the visual quality on the synthetic images. As shown in Fig. 8, the images recovered by DCP have severe color distortion and look fairly hazy. Although the GRM outputs severe color-distorted images, it tends to estimate the haze level more accurately, thus the restored images have less haze. AOD-Net, MSCNN, and DehazeNet effectively alleviate the color distortion problem, while they tend to underestimate the haze level of the input images. Detailed information such as textures and edges in their recovered images is unsatisfactory as well. When it comes to EPDN, it can generate more visually pleasant results than these aforementioned methods; however, the restored images still have some remaining haze when the input images contain severe haze. Besides, it cannot effectively recover the structures and details due to loss of high-level semantics information. GridDehazeNet effectively overcomes the color distortion problem and generates dehazed images closer to the haze-free images. However, it tends to cause some artifacts in the background parts when removing thick haze. Compared with these aforementioned methods, our haze-removed images are free of major artifacts and color distortions and have the best visual effect.

The evaluation results about the effect of haze levels on the model performance are shown in Fig. 9. As the haze level increases, the PSNR values slightly decrease for both the haze-removed results and input hazy images. However, the proposed method still can perform well on images with the highest haze level. The SSIM values of the dehazed results are high and largely constant, whereas the SSIM scores of

(a) Hazy Image (b) Our Result (c) GroundTruth

Fig. 12. Visualization of an imperfect example on an outdoor scenario with heavily haze. The proposed method will generate some distortions and the recovered image as shown is polluted by chaotic colors.

the input hazy images decrease dramatically. The results indicate that the proposed method can generate visually pleasant images with sharper structures and meaningful textural details.

Fig. 10 illustrates the qualitative results on the SOTS outdoor dataset. The output of DCP severely suffers from color distortions and the description of the sky looks unreal. DehazeNet and AOD-Net slightly alleviate the color distortion problem; however, their results are typically darker than the ground truth images. Although EPDN does not output low-brightness images, the color fidelity of some restored results and detailed information such as textures, edges and the blue sky are unsatisfactory. GridDehazeNet and our method are superior in image details and color fidelity. Overall, the proposed method can generate more visually pleasant results than other methods.

D. Performance Evaluation on Real-World Images

To make the model evaluation more convincing, we conduct comparison experiments between the proposed approach and several state-of-the-art methods on real-world images. These methods include DCP [10], AOD-Net [22], MSCNN [20], GFN [39], PFFNet [23], GridDehazeNet [9], and DuRN [40]. For a fair comparison, we obtain the evaluation results of PFFNet, GridDehazeNet, and DuRN using their released codes and models on the same train and test datasets. The other algorithms are evaluated with their provided pre-trained models. Table V lists the quantitative results on the I-HAZE and O-HAZE datasets. The proposed method reports better PSNR and SSIM results than other algorithms. Compared with DuRN and our method, GridDehazeNet has a slightly poor performance on real hazy images, which indicates its overfitting on the large-scale training set. Our method achieves similar results with DuRN on the I-HAZE dataset and obtains a performance improvement of 1.62 dB PSNR value over DuRN on the O-HAZE dataset. The proposed method gives a much higher SSIM score than the other methods, indicating it effectiveness of preserving meaningful texture and structure information in the restored images.

Fig. 11 illustrates the qualitative comparison on the I-HAZE and O-HAZE testing sets. The restored images from I-HAZE are shown in the top four rows, and the bottom three rows show the results from O-HAZE. We highlight some regions in the images to show detailed information. In Fig. 11, PFFNet is far from achieving realistic performance due to the unsatisfactory image details and color fidelity on indoor scenarios, and a large number of image contents are polluted by chaotic colors. Furthermore, it can achieve significant visual results on outdoor real-world images. For GridDehazeNet, a significant amount of haze remains unremoved in both the indoor and outdoor images. It also suffers from the color distortion problem. DuRN can estimate the haze level accurately; however, it cannot effectively recover clear structures and details from images containing severe haze. Our proposed method is found to generalize well on realistic hazy images in terms of visual quality, and can better preserve details and color information in haze-removed images.

E. Limitation

The proposed method is not robust enough for few super-hard scenarios. Fig. 12 illustrates an example: when the haze is extremely thick and the image boundary is unclear in the input hazy scenario, the method will generate images with some distortions and chaotic colors. The limitation might be alleviated by fully optimizing the model parameters.

V. Conclusion

In this study, we developed a hierarchical feature fusion dehazing method with a novel proposed mixed convolution attention mechanism to progressively and adaptively improve the dehazing performance. The haze levels could be accurately estimated by fusing multi-scale hierarchical features; thus, the proposed method could generate images almost without remaining haze. The structure information in recovered images was also clearly preserved as we increase the scales in feature fusion. Moreover, the mixed convolution attention module was designed to reduce feature redundancy, transform some ineffective features into valuable ones and highlight task-relevant features for reconstructing visually pleasant haze-removed images. The deep semantic loss can facilitate the optimization of the learnable parameters as well. Extensive experimental results demonstrated that the proposed approach achieved superior dehazing performance on both synthetic and real-world images compared with several state-of-the-art algorithms.

References

[1] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "DeepDriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2722–2730.

[2] X. Zhang, D. Wang, Z. Zhou, and Y. Ma, "Robust low-rank tensor recovery with rectification and alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 238–255, Jan. 2021.

[3] Z. Jia, H. Wang, R. E. Caballero, Z. Xiong, J. Zhao, and A. Finn, "A two-step approach to see-through bad weather for surveillance video quality enhancement," *Mach. Vis. Appl.*, vol. 23, no. 6, pp. 1059–1082, Nov. 2012.

[4] L. Ren, J. Lu, Z. Wang, Q. Tian, and J. Zhou, "Collaborative deep reinforcement learning for multi-object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 508–602.

[5] X. Zhang, W. Hu, N. Xie, H. Bao, and S. Maybank, "A robust tracking system for low frame rate video," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 279–304, Dec. 2015.

[6] E. J. McCartney, *Optics of the Atmosphere: Scattering by Molecules and Particles*, vol. 421. New York, NY, USA: Wiley, 1976.

[7] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," *Int. J. Comput. Vis.*, vol. 48, no. 3, pp. 233–254, 2002.

[8] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced Pix2pix dehazing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8160–8168.

[9] X. Liu, Y. Ma, Z. Shi, and J. Chen, "GridDehazeNet: Attention-based multi-scale network for image dehazing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7314–7323.

[10] 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.

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

[12] J. Wang, K. Lu, J. Xue, N. He, and L. Shao, "Single image dehazing based on the physical model and MSRCR algorithm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2190–2199, Sep. 2018.

[13] 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, no. 11, pp. 5187–5198, Nov. 2016.

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

[15] J.-L. Yin, Y.-C. Huang, B.-H. Chen, and S.-Z. Ye, "Color transferred convolutional neural networks for image dehazing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3957–3967, Nov. 2020.

[16] D. Zhao, L. Xu, L. Ma, J. Li, and Y. Yan, "Pyramid global context network for image dehazing," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Nov. 9, 2020, doi: 10.1109/TCSVT.2020.3036992.

[17] 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.*, early access, Dec. 22, 2020, doi: 10.1109/TCSVT.2020.3046625.

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

[19] Y. Zhang, P. Wang, Q. Fan, F. Bao, X. Yao, and C. Zhang, "Single image numerical iterative dehazing method based on local physical features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3544–3557, Oct. 2020.

[20] 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.*, Oct. 2016, pp. 154–169.

[21] C. O. Ancuti and C. Ancuti, "Single image dehazing by multi-scale fusion," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3271–3282, Aug. 2013.

[22] 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. (ICCV)*, Oct. 2017, pp. 4770–4778.

[23] K. Mei, A. Jiang, J. Li, and M. Wang, "Progressive feature fusion network for realistic image dehazing," in *Proc. Asian Conf. Comput. Vis.* Springer, 2018, pp. 203–215.

[24] H. Zhang, V. Sindagi, and V. M. Patel, "Multi-scale single image dehazing using perceptual pyramid deep network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 902–911.

[25] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.

[26] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[27] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[28] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4634–4643.

[29] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[30] N. S. Trudinger, "Elliptic partial differential equations of second order," Tech. Rep., 1983.

[31] B. Li *et al.*, "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2019.

[32] P. K. Nathan Silberman, D. Hoiem, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.

[33] C. Ancuti *et al.*, "NTIRE 2018 challenge on image dehazing: Methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 891–901.

[34] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "I-HAZE: A dehazing benchmark with real hazy and haze-free indoor images," 2018, *arXiv:1804.05091*. [Online]. Available: http://arxiv.org/abs/1804.05091

[35] 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/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 754–762.

[36] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electron. Lett.*, vol. 44, no. 13, pp. 800–801, Jun. 2008.

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[39] W. Ren *et al.*, "Gated fusion network for single image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3253–3261.

[40] X. Liu, M. Suganuma, Z. Sun, and T. Okatani, "Dual residual networks leveraging the potential of paired operations for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7007–7016.

Xiaoqin Zhang received the B.Sc. degree in electronic information science and technology from Central South University, 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, China, in 2010. He is currently a Professor with Wenzhou University, 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, *IJCV*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, IEEE TRANSACTIONS ON COMPUTERS, ICCV, CVPR, NIPS, IJCAI, AAAI, and among others. His research interests are in pattern recognition, computer vision, and machine learning.

Jinxin Wang received the B.Sc. degree in information and computing science with Wenzhou University. He is currently pursuing the degree with the College of Computer Science and Artificial Intelligence, Wenzhou University, China. His research interests include image restoration, reinforcement learning, and statistical learning theory.

Tao Wang received the B.Sc. degree in information and computing science from Hainan Normal University, China, in 2018. He is currently pursuing the degree with the College of Computer Science and Artificial Intelligence, Wenzhou University, China. 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.

Runhua Jiang received the B.Sc. degree with the Department of Information Science, Tianjin University of Finance and Economy, China. He is currently pursuing the degree in computer software and theory with the College of Computer Science and Artificial Intelligence, Wenzhou University, China. His research interests include several computer vision tasks, such as image/video restoration, crowd counting, visual understanding, and video question answering.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your article. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

Carefully check the page proofs (and coordinate with all authors); additional changes or updates WILL NOT be accepted after the article is published online/print in its final form. Please check author names and affiliations, funding, as well as the overall article for any errors prior to sending in your author proof corrections. Your article has been peer reviewed, accepted as final, and sent in to IEEE. No text changes have been made to the main part of the article as dictated by the editorial level of service for your publication.

AQ:1 = Please confirm or add details for any funding or financial support for the research of this article.

AQ:2 = Please confirm the location for Wenzhou University.

AQ:3 = Please provide the publisher location for Ref. [23].

AQ:4 = Please provide the organization name, organization location, and report no. for Ref. [30].

Hierarchical Feature Fusion With Mixed Convolution Attention for Single Image Dehazing

Xiaoqin Zhang, Jinxin Wang, Tao Wang, and Runhua Jiang

***Abstract*— Single image dehazing, which aims at restoring a haze-free image from its correspondingly unconstrained hazy scene, is a fundamental yet challenging task and has gained immense popularity recently. However, the images recovered by some existing haze-removal methods often contain haze, artifacts, and color distortions, which severely degrade the visual quality and have negative impacts on subsequent computer vision tasks. To this end, we propose a network combining multi-scale hierarchical feature fusion and mixed convolution attention to progressively and adaptively enhance the dehazing performance. The haze levels and image structure information are accurately estimated by fusing multi-scale hierarchical features, thus the model restores images with less remaining haze. The proposed mixed convolution attention mechanism is capable of reducing feature redundancy, learning compact and effective internal representations and highlighting task-relevant features, thus, it can further help the model estimate images with sharper textural details and more vivid colors. Furthermore, a deep semantic loss is also proposed to highlight essential semantic information in deep features. The experimental results show that the proposed method outperforms state-of-the-art haze removal algorithms.**

***Index Terms*— Image dehazing, hierarchical feature fusion, mixed convolution attention mechanism, deep learning.**

I. Introduction

IMAGES captured in hazy conditions often contain unclear content and degraded structural details. These low-visibility images are a hindrance to multiple subsequent high-level tasks [1]–[5], including autonomous driving, video surveillance, and visual object tracking. As a fundamental yet challenging technique, single image dehazing, which aims at restoring haze-removed images with less remaining haze, sharper structure details and vivid colors from low-visibility scenarios, will be beneficial to the application of these high-level tasks. Thus, single image dehazing has become an increasingly popular research topic recently.

Manuscript received September 13, 2020; revised January 6, 2021 and March 1, 2021; accepted March 12, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61922064 and Grant U2033210, in part by the Zhejiang Provincial Natural Science Foundation under Grant LR17F030001, and in part by the Project of Science and Technology Plans of Wenzhou City under Grant C20170008 and Grant ZG2017016. This article was recommended by Associate Editor J. Hou. *(Corresponding author: Xiaoqin Zhang.)*

The authors are with the College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou 325035, China (e-mail: jxwang@stu.wzu.edu.cn; zhangxiaoqinnan@gmail.com; taowangzj@gmail.com; ddghjikle1@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3067062.

Digital Object Identifier 10.1109/TCSVT.2021.3067062

The generation of haze in images can be described using the classical atmospheric scattering model [6], [7]:

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

where x represents the pixel position, I is the observed hazy image, J denotes the hazy-free radiance; A refers to the global atmospheric light and t denotes the medium transmission map. The physical model gives essential insights about image dehazing, however, without the knowledge of A and t, image dehazing based on this model becomes an under-determined estimation problem.

To estimate the unknowns A and t to solve the under-determined estimation problem, most dehazing methods use either physical grounded priors or data-driven ways. Specifically, He *et al.* [10] developed dark channel prior to obtain the transmission map. Zhu *et al.* [11] investigated an attenuation before recovering depth information for transmission map estimation. Wang *et al.* [12] proposed their method based on the physical model and the brightness components of the image. Although these prior-based methods have demonstrated their superiority, their assumptions do not accurately reflect the inherent properties of images. Thus, their performance tends to be limited. Recently, deep-learning-based methods have been employed to estimate the atmospheric light and transmission map. In DehazeNet [13], the transmission map is determined in an end-to-end manner. In DCPDN [14], the atmospheric light and the transmission map are estimated simultaneously.

With the strong representation power of convolutional neural networks, These aforementioned methods achieve a better dehazing performance. However, an inaccurate estimation of the transmission map or the atmospheric light would significantly interfere with the restoration of images. Unlike these methods, model-free methods [8], [9], [15]–[17] that directly restore haze-removed images from their hazy counterparts have demonstrated remarkable dehazing performance.

To accurately estimate the haze levels and generate images with sharp structural details, we propose hierarchical feature fusion schema for image dehazing. The hierarchical feature fusion in this work is investigated in a supervised manner with hazy-free ground truth images as learning targets. Using a composition of simple functions, the network first learns low-level features using simple functions of the input, these features are then aggregated level-by-level to generate increasingly rich feature representations. Based on these extracted intermediate feature representations, the target can be

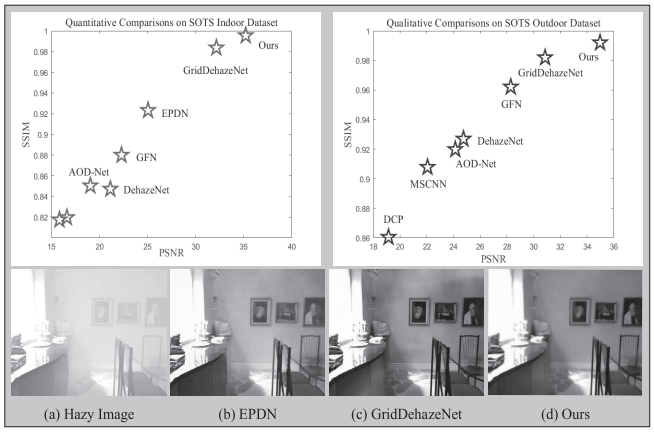


Fig. 1. Quantitative and qualitative results. The quantitative comparisons are shown in the top two images. From the qualitative results, EPDN [8] tends to underestimate the hazy level, thus the restored image contains remaining haze. In the present of artifacts and color distortions, GridDehazeNet [9] cannot generate visually pleasant haze-removed images. The proposed method achieves the best visual result.

accurately estimated with properly feature fusion mechanism and image reconstruction module.

When it comes to the feature sensitive image dehazing task, investigating appropriate strategies to transform the redundant parts in features into richer ones is a promising way to improve the dehazing performance. Boosting the representation power of hierarchical feature fusion, a mixed convolution attention module is designed to transform these ineffective and redundant features into valuable ones as well as adaptively highlight task-relevant features. When the attention module is appropriately utilized, we find that the dehazing model can generate haze-removed images with sharper structural details, less halo artifacts and color distortion.

The proposed model-free method has three stages. We first extract rich features at different depths of the network with hierarchical learning. Subsequently, these features are fused with several mixed convolution attention modules. Given the aggregated features and input hazy images, we accurately recover the haze-removed results. In addition, aiming at learning more semantic information, we propose a deep semantic loss to measure the semantic difference in deep features extracted from the dehazed results and their ground truth images. Extensive experimental results have shown that the proposed approach outperforms the state-of-the-art methods. Fig. 1 gives some quantitative and qualitative comparison results.

The main contributions of this work are summarized as follows:

- We propose a hierarchical feature fusion framework with attention mechanism to adaptively enhance the performance of the scale-sensitive image dehazing task.
- A novel spatial attention mechanism termed as mixed convolution attention module is proposed, which has the following advantages: reducing feature redundancy, learning compact and effective internal representations, and adaptively highlighting task-relevant features. Therefore, the dehazing model can generate images with less remaining haze, sharper textural details and vivid colors.
- Extensive experimental results on benchmark datasets and detailed analysis have demonstrated the effectiveness and superiority of the proposed dehazing method.

The rest of this paper is organised as follows. We first introduce the related work in Section II. Then we present the proposed dehazing method with detailed descriptions for hierarchical feature fusion, mixed convolution attention, and loss functions in Section III. Section IV describes experimental setup including datasets, evaluation metrics, implementation details, ablation study, model analysis, and extensive experimental results for performance evaluation and comparison. Finally, conclusions are given in Section V.

II. Related Work

A. Single Image Dehazing

Significant advances in single image dehazing have been witnessed in recent years. These methods can be roughly classified into two categories.

1) Prior-Based Image Dehazing Methods: Prior-based methods estimate the transmission maps and atmospheric light intensity based on the statistics of clear images. He *et al.* [10] propose DCP to estimate the transmission map of hazy images. Their assumption is that at least one color channel should have a very low intensity within a haze-free image that does not contain sky or bright regions. Zhu *et al.* [11] propose an efficient color attenuation prior. Using a linear model to build a bridge between the hazy image and its depth information, the method can estimate the transmission and restore the scene radiance to remove the haze. Berman *et al.* [18] propose their method based on a non-local prior. They assume that the colors of a haze-free image can be approximated by less compact and typical colors clustered in RGB space. Zhang *et al.* [19] introduced their dehazing method based on local physical features. Although these prior-based methods have provided meaningful insights for image dehazing, they may fail in cases where the priors or assumptions are invalid.

2) Learning-Based Image Dehazing Methods: With advancements in deep neural networks and the availability of large-scale datasets, data-driven approaches have received significant attention recently. Cai *et al.* [13] introduce DehazeNet to estimate the medium transmission map, which is then used to restore haze-free images based on the atmospheric scattering model. A multi-scale deep neural network is employed by Ren *et al.* [20] to estimate the scene transmission maps. They first predict the transmission maps using a coarse-scale network, then utilize a fine-scale network to refine the estimated results. Note that, Ancuti and Ancuti [21] are the first to design a fusion-based approach and demonstrate its effectiveness and potential for dehazing task. Li *et al.* [22] employ AOD-Net to directly restore dehazed images instead of estimating the atmospheric light and the transmission matrix. Zhang and Patel [14] propose DCPDN to simultaneously learn transmission map and atmospheric light in an end-to-end manner.

Without relying on the physical scattering model, Mei *et al.* [23] propose a progressive feature fusion network.

They utilize an U-like encoder-decoder network to learn the transformation from hazy images to their ground-truth images. Qu *et al.* [8] directly generate the haze-removed images with EPDN after embracing image dehazing as an image-to-image translation problem. Liu *et al.* [9] propose an end-to-end trainable network termed GridDehazeNet for haze removal. Zhang *et al.* [24] propose an efficient multi-scale single image dehazing approach using perceptual pyramid network, in order to directly learn the non-linear mapping between hazy images and their haze-free counterparts. The dehazing network is optimized with the L2 loss and perceptual loss. The differences between our work and [24] are analyzed as follows: First, Zhang *et al.* employ an encoder-decoder architecture based on residual and dense blocks to extract image features for image reconstruction, while we utilize the hierarchical features extracted from different depths of the network which will relatively provide more contextual and semantic information. Through fusing hierarchical features, the haze levels can be accurately estimated and the image structures will be clearly recovered. Second, Zhang *et al.* employ the multi-scale pyramid pooling module to handle the extracted features. In our work, the multi-scale protocols are conducted in two levels: (1). multi-scale hierarchical feature fusion; (2) mixed convolution attention for fine-grained multi-scale feature adjustment. After the two steps, more richer and task-relevant representations can be obtained. Third, Zhang *et al.* employ the perceptual loss to optimize their network. We not only employ the perceptual loss but also employ a newly proposed deep semantic loss, in which the Laplace operator is used to highlight semantic information in deep features.

B. Attention Mechanism

Inspired by the important role of attention in human perception, the attention mechanism has become a popular component in deep neural networks. Significant improvements in various tasks have been achieved in recent years. For example, Chen *et al.* [25] incorporate both spatial and channel-wise attention for image captioning. The proposed SCA-CNN model outperforms many other visual attention-based image captioning algorithms. Based solely on attention mechanism, Vaswani *et al.* [26] propose a transformer using stacked self-attention to draw global dependencies between the input and output for neural sequence transduction models. Woo *et al.* [27] propose a convolution block attention module with channel and spatial sub-modules to adaptively refine features. An attention on attention module is proposed by Huang *et al.* [28] to determine the relevance between attention results and queries. Aiming to exchange and aggregate information in a more flexible manner in dehazing task, Liu *et al.* [9] integrate GridDehazeNet with a channel-wise attention mechanism. Different from the attention introduced by GridDehazeNet, we propose a novel spatial attention with mixed convolution operations to highlight task-relevant features. With the proposed attention, which can help generate haze-removed images with vivid colors and sharper structure details, the performance is far superior to GridDehazeNet.

III. Methodology

Intuitively, restoring images to their haze-free state with excellent visibility requires rich feature representations. The hierarchical feature extraction, which is presented in Section III-A, is proposed to progressively extract multi-scale features at different depths of the deep networks. After capturing these rich features, multiple mixed convolution attention modules are employed as plug-and-play tools to boost the representation power of the network by focusing on important features and suppressing unnecessary ones. The mixed convolution attention module is detailed in Section III-C. Subsequently, we investigate an efficient way to aggregate these features, the processed features are then provided as inputs to the final image reconstruction module. The entire network architecture is presented in this section. Fig. 2 illustrates a typical architecture variant of the proposed dehazing network.

A. Feature Extraction

Generally, using features extracted from an isolated layer is inadequate. Features from earlier layers mainly focus on low-level information such as edges and shapes, which are essential for locating the positions of different objects and for restoring the structural details in images. While features from latter layers have meaningful and richer semantic information that is beneficial for better preserving the textural and color details. Through aggregating shallow and high-level features, the dehazing method can generate visual pleasant results with less remaining haze and sharper structures.

In this work, to aggregate hierarchical features captured from different layers, we apply several feature extraction blocks with multiple scales to attain this objective. Given a hazy image $I \in \mathbb{R}^{C\times H\times W}$ as input, the extracted hierarchical features can be expressed as:

$$F_s = f_s(F_{s-1}), \quad s = 1, \ldots, N, \tag{2}$$

where F_s denotes the features extracted by the s-th feature extraction block f_s, and $F_1 = f_1(I)$. N is the total number of scales utilized in this network.

The architecture of feature extraction block is shown at the bottom of Fig. 2. We first obtaining the downsampling features with a convolution layer; then compute the output features with residual mechanism:

$$F_{mid} = \downarrow F_{s-1} + \varphi(\delta_1(\varphi(\downarrow F_{s-1}))), \tag{3}$$

where $\downarrow$ indicates the downsampling operation; φ means the common convolution layer; δ_1 denotes the ReLU function; and F_{mid} is the middle output features. Similarly, the s-th features F_s can be computed with another residual module without the downsampling operation. The details are also shown in Fig. 2.

B. Feature Fusion

After extracting the hierarchical features F_s as expressed in Eq. (2), we refine them with mixed convolution attention modules:

$$\widehat{F}_s = \mathcal{ATT}(F_s), \quad s = 1, 2, \cdots, N, \tag{4}$$

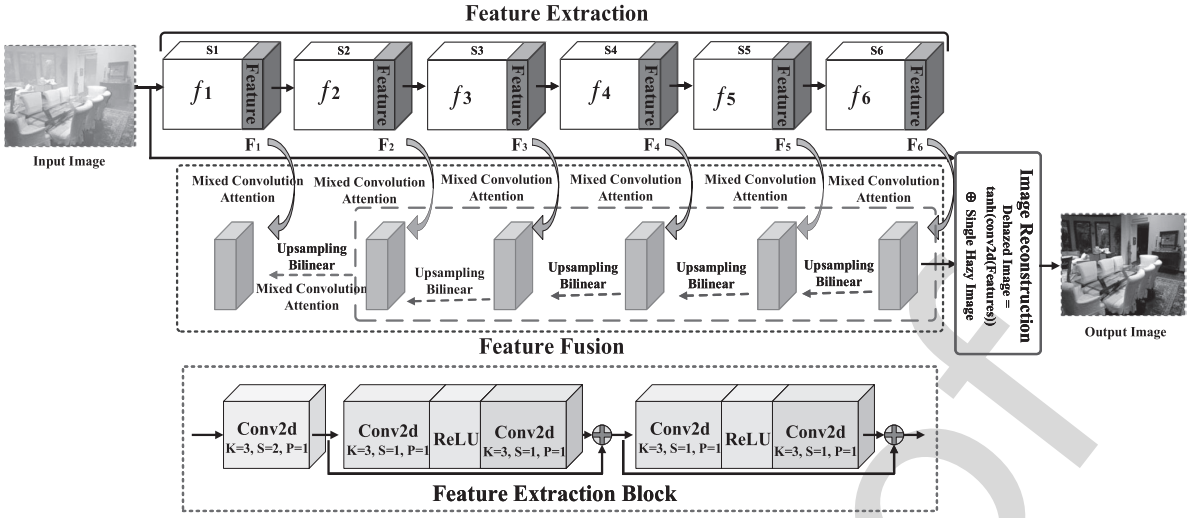


Fig. 2. A typical network variant of the proposed dehazing method. Here, the hierarchical features are extracted by six feature extraction blocks, then these extracted features are refined by six mixed convolution attention modules. These attention modules transform input features into other feature representations with richer task-relevant texture information. Finally, given the fused features and the hazy image as input, an image reconstruction module is utilized to accurately restore the haze-removed image. The green and black arrows indicate the processing of mixed convolution attention modules and the data flow in the network.

where $\mathcal{ATT}(\cdot)$ refers to the mixed convolution attention, $\widehat{F}_s$ is the refined result.

Without loss of generality, we employ a network variant with 6 scales as an example, which is illustrated in Fig. 2. Note that, the variant can achieve the best performance on indoor scenarios. The details of ablation studies and model analysis are presented in Section IV-B. To fuse features $F_i, i = 1, \cdots, N$, we progressively upsample them with bilinear interpolation operations, so that the feature concatenation can be achieved in a clean manner given aligned features. The feature fusion process can be expressed as:

$$\widehat{F}_{out} = [\widehat{F}_1, \uparrow \mathcal{ATT}([\uparrow \widehat{F}_2, \cdots, \uparrow \widehat{F}_N])], \tag{5}$$

where $\uparrow$ is the upsampling operation, $[\cdot]$ denotes the feature concatenation operation, and $\widehat{F}_{out}$ represents the output fused features of an input image. In Eq. (5), we first apply another attention module to adjust the extracted features, then $\widehat{F}_1$ is concatenated to obtain the $\widehat{F}_{out}$. Finally, the haze-removed image $\widehat{J}$ can be correctly estimated using the image reconstruction module:

$$\widehat{J} = \delta_2(\varphi(\widehat{F}_{out})) \oplus I, \tag{6}$$

where δ_2 is the tanh function, $\oplus$ indicates the element-wise add operation, which refers to the network-level skip-connection of residual learning, and $\widehat{J}$ is the haze-removed image.

C. Mixed Convolution Attention

Extracting features and aggregating them based on hierarchical learning are essential for image dehazing; however, the kernel size used in the different convolution layers is often overlooked in attention mechanism for image dehazing. When designing their networks, people often simply choose 3×3 or 5×5 kernels for convolution layers. Although depthwise convolution [29] separately employ kernels to each individual channel of the input features, it ignores the effect of kernel sizes utilized in single layer. Besides, the redundancy in features has not been thoroughly investigated to boost the representation power. We find the similar or redundant features can be transformed into valuable ones and can help estimate haze-removed images with sharp textural details and vivid colors.

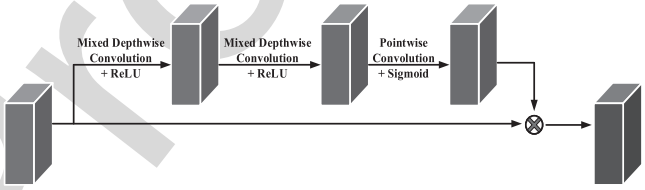


Fig. 3. The architecture of Mixed Convolution Attention Module (MCAM). The module is composed of mixed depthwise convolution, pointwise convolution, ReLU and Sigmoid operations. This module transforms any input features into other representations of the same size but with more compact, effective and task-relevant information. It significantly improves the dehazing performance.

In this work, the proposed attention module for adaptive feature refinement consists of mixed depthwise convolution and pointwise convolution operations, Fig. 3 depicts the detailed process. The attention module has three components, and the operations are specially designed to processing input information in a lightweight and sensitive manner. In Eq. (4) and Eq. (5), we use $\mathcal{ATT}(\cdot)$ to denote the mixed convolution attention module, which is expressed as follows:

$$\widehat{F}_s = \mathcal{ATT}(F_s) = F_s \odot W_{F_s}, \tag{7}$$

where $\widehat{F}_s$ represents the refined feature maps. W_{F_s} is the weight for each element in F_s, and $\odot$ denotes the element-wise multiplication operation. From the equation, we can see the mixed convolution attention module is a spatial attention mechanism, which spatially highlights task-relevant information rather than selecting key channels.

Fig. 3 shows that the W_{F_s} is obtained by two mixed depthwise convolutions, one pointwise convolution and activation

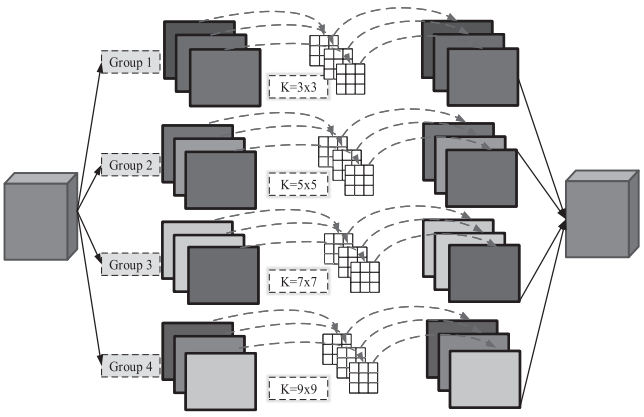


Fig. 4. The architecture of Mixed Depthwise Convolution Layer (MDCL). Without loss of generality, we first split the input feature maps into 4 groups, then perform depthwise convolution with different kernel sizes on each group of features. Finally, the output features are obtained by concatenating the 4 groups of processed features.

functions. We detail the operations as follows:

$$W_{F_s} = \delta_3(\kappa_p(\delta_1(\kappa_m(\delta_1(\kappa(F_s)))))). \tag{8}$$

In this equation, κ_m and κ_p refer to the mixed depthwise convolution and pointwise convolution. δ_1 and δ_3 refer to the ReLU and Sigmoid functions. We should note that the dimensions of both the channels and spatial axes are unaltered when information flows within the module. Therefore, the attention module can be integrated into any end-to-end trainable network architectures.

1) Mixed Depthwise Convolution: Different from conventional convolution operations, one convolution kernel of depthwise convolution is responsible for one channel, and one channel is convolved by only one convolution kernel. The proposed mixed depthwise convolution is composed of multiple depthwise convolution operations with different kernel sizes. The details of the operations for 4 groups of features are illustrated in Fig. 4. More specifically, the input feature maps F^{in} are partitioned into G groups. For each group of feature maps, we employ a depthwise convolution with different kernel sizes to process them. The kernel size for each group is formulated as: $k \times k,\ k = 2 \times (g-1) + 3$, where $g = 1 \cdots G$, and k is the kernel size. The output features F^{out} can be represented as:

$$F^{out} = [dw(F^{in}_{g=i}; k_{g=i})], \quad i = 1, \cdots, G, \tag{9}$$

where $dw(\cdot)$ is the depthwise convolution operation, $F^{in}_{g=i}$ refers to the i-th group of input feature maps and similarly $k_{g=i}$ is the kernel size of group i, and $k_{g=i} = 2 \times (i-1) + 3$. $[\cdot]$ is also the feature concatenation operation for the G groups of feature maps. Besides, the output features F^{out} can also be processed by an appropriate activation function.

2) Pointwise Convolution: To effectively utilize the information of the input feature maps in the same spatial location, the pointwise convolution operation with the kernel size of 1×1 is applied. The pointwise convolution can be seen as a special case of the mixed depthwise convolution when G is set as the number of channels of the input feature maps and all convolution operations have the same kernel size 1×1. Given the input feature maps F^{in}, the pointwise convolution operation can be formulated as:

$$F^{out} = [dw(F^{in}_{g=i}; k = 1)], \quad i = 1, \cdots, C, \tag{10}$$

where C is the total number of channels in the feature maps F^{in} and the $[\cdot]$ indicates the feature concatenation operation as well.

D. Loss Functions

To better optimize the proposed dehazing network, four loss functions are employed to measure the quantitative difference between the haze-removed images and their ground truth images. The four losses are mean square error (MSE), smooth L_1 loss, perceptual loss, and deep semantic loss. The proposed deep semantic loss is utilized to emphasize necessary semantic information in deep features. We first introduce the four losses, respectively, then use them to obtain the joint loss function.

1) Mean Square Error: The MSE is utilized to accurately capture most of the low frequencies in the images. This information is crucial for recovering high-quality images. The loss function is defined as:

$$L_{mse} = \frac{1}{CHW} \left\| \widehat{J} - J \right\|^2, \tag{11}$$

where $\widehat{J}$ represents the restored image and J refers to the corresponding ground truth image; and C, H, W refer to the number of channels, height and width, respectively.

2) Smooth L_1 Loss: L_1 norm can enforce the correctness at low frequencies as well. Moreover, $SmoothL_1$ loss is less sensitive to outliers and can alleviate the gradient explosion problem. The loss function is expressed as:

$$L_{smo} = \frac{1}{CHW} \psi(\widehat{J} - J),$$
$$where\ \psi(\varepsilon) = \begin{cases} 0.5\varepsilon^2, & \text{if } |\varepsilon| < 1, \\ |\varepsilon| - 0.5, & \text{otherwise.} \end{cases} \tag{12}$$

3) Perceptual Loss: Aiming at enforcing the network to recover images with low-to-high level semantic fidelity and high standard visual quality, we measure the difference in features between $\widehat{J}$ and J using perceptual loss:

$$L_{per} = \sum_{l \in \{4,9,16\}} \frac{1}{C_l H_l W_l} \left\| \phi_l(\widehat{J}) - \phi_l(J) \right\|^2, \tag{13}$$

where ϕ_l is the l-th feature extractor corresponding to the VGG16 network, the C_l, H_l, and W_l denote the number of channels, height and width of the feature maps, respectively, extracted from the l-th layer of the VGG16.

4) Deep Semantic Loss: Inspired by the perceptual loss which measures the difference between feature maps, Laplace operator [30] is employed to highlight the semantic details in extracted features with levels from low to high. Moreover, the Laplace operator is formulated in the deep semantic loss in an end-to-end manner. The loss function can be expressed as follows:

$$L_{sem} = \sum_{l \in \{4,9,16\}} \frac{l}{C_l H_l W_l} \left\| \delta_2(\xi(\phi_l(\widehat{J}))) - \delta_2(\xi(\phi_l(J))) \right\|_1, \tag{14}$$

where $\xi(\cdot)$ denotes the Laplace operator used to extract detailed semantic information for both the estimated image $\widehat{J}$ and the ground truth J. δ_2 is the tanh function. The L_1 norm is used to measure the difference in deep semantic information between $\widehat{J}$ and J.

5) Total Loss: During training, all the models are optimized by minimizing the following loss function L_{tot}:

$$L_{tot} = L_{mse} + L_{smo} + \lambda_1 \cdot L_{sem} + \lambda_2 \cdot L_{per}. \quad (15)$$

The λ_1 and λ_2 are used to control the interaction of these loss components.

IV. Experiments

In this section, we first introduce the dehazing datasets, evaluation metrics along with some implementation details about training and evaluating the proposed dehazing method. Then, detailed ablation studies and analysis are conducted to make these proposed components and our statements more convincing. Finally, we report the evaluation results of extensive experiments conducted on both synthetic and real-world benchmarks. We compare and analyze the proposed approach in terms of quantitative accuracy and visual quality with several state-of-the-art methods.

A. Experimental Settings

1) Datasets: Generally, it is prohibitively expensive to collect a large number of real-world hazy images and their haze-free counterparts. Therefore, we train and evaluate the proposed method on a synthetic dataset, namely RESIDE [31]. The RESIDE dataset includes synthetic hazy images in both indoor and outdoor scenarios. The indoor training set (**ITS**) contains a total of 13990 hazy indoor images, generated from 1399 haze-free images with $\beta \in [0.6, 0.8]$ and $A \in [0.7, 1.0]$ based on the atmospheric scattering model; the depth maps are obtained from the NYN Depth V2 [32] and Middlebury Stereo datasets. The outdoor training set (**OTS**) contains a total of 296695 hazy outdoor images, generated from 8477 haze-free images. For both the indoor and outdoor scenarios, we train the proposed model on the ITS and OTS datasets, respectively. For model evaluation, we adopt the synthetic objective testing set (**SOTS**), which contains 500 pairs of indoor images and 500 pairs of outdoor ones.

The challenge [33] contains two real-world image dehazing datasets: **I-HAZE** [34] and **O-HAZE** [35]. The I-HAZE dataset contains 35 pairs of hazy images and the corresponding haze-free images of various indoor scenes. While the O-HAZE dataset includes 45 pairs of hazy images and the corresponding ground truth images of various outdoor scenes. To keep the fairness of comparison, we finetune the proposed model using the training set and evaluate it on the corresponding testing part, following the setup of the NTIRE 2018 Image Dehazing Challenge.

2) Evaluation Metrics: In this work, the peak signal to noise ratio (**PSNR** [36]) and the structural similarity index (**SSIM** [37]) are utilized to evaluate the quality of the restored images. Furthermore, we compare the subjective visual effect of the restored images with other algorithms on the SOTS, I-HAZE, and O-HAZE datasets.

3) Implementation Details: In implementation, we utilize the Adam [38] algorithm with hyper-parameters β_1 and β_2 set as 0.5 and 0.999, respectively. The initial learning rate is set as 0.0001 and decayed by $\gamma = 0.1$ for every 50 epochs. The models are trained for about 150 epochs on indoor images and about 10 epochs on outdoor images. The trade-off hyper-parameters both λ_1 and λ_2 in the total loss function Eq. (15) are set as 0.01 and 0.04. We conduct all the training, testing, and model analysis on the same platform with an Intel Xeon Silver 4114 CPU, 32 GB RAM and a single NVIDIA RTX 2080 Ti GPU.

B. Ablation Study and Model Analysis

In this section, aiming at analyzing the performance of the proposed method, we conduct extensive ablation studies to investigate the effectiveness of scales used in hierarchical feature fusion and architecture settings for mixed convolution attention module. All variants of the proposed method mentioned below are trained and evaluated on the same experimental setting for fair comparison. The loss Eq. (15) is employed to optimize these variants. Table I lists the evaluation results on both indoor and outdoor datasets. For convenience, the evaluation results of GridDehazeNet are also listed here to compare the dehazing performance.

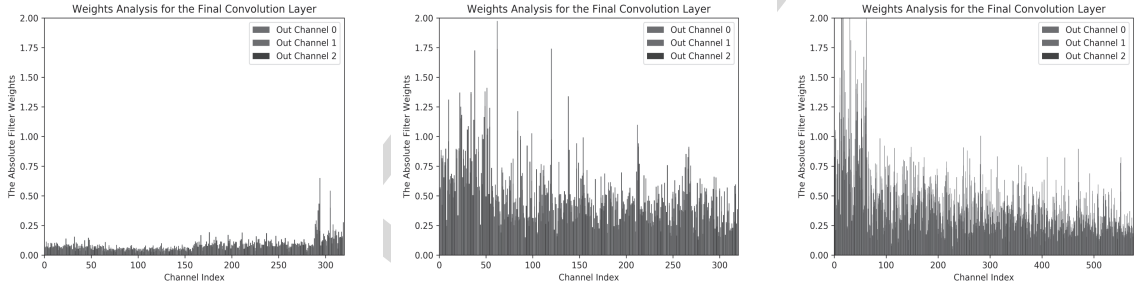
1) Model Introduction: In Table I, S4, S5 and S6 indicate the number of scales utilized for hierarchical feature fusion, while G2, G4, G6 and G8 denote the number groups employed in the mixed convolution attention module. For each group of features, the corresponding mixed depthwise convolution has its kernel size, the details are shown in Section III-C.1. The Baseline only contains 4 feature extraction blocks and an image reconstruction block. The Baseline has the simplest network architecture without hierarchical feature fusion or any mixed convolution attention modules. The Baseline+S4 means four scales of features are fused in this model without employing the attention mechanism. Besides, S4G2 indicates that the model not only has feature fusion for features extracted by four feature extraction blocks but also is equipped with multiple mixed convolution attention modules. G2 means that the input features in attention modules are split into 2 groups. The other model settings in this table have similar definitions for both the indoor and outdoor scenarios.

2) Hierarchical Feature Fusion Analysis: To validate the effectiveness of the hierarchical feature fusion mechanism, experiments and analysis are conducted on both indoor and outdoor datasets. Compared with Baseline, Baseline+S4 achieves a performance gain of approximately 1dB in terms of the PSNR. However, the variant cannot obtain the best performance only equipped with multi-scale feature fusion. On the indoor dataset, the effect of the number of scales in the hierarchical feature fusion is investigated through variants S4G2, S5G2 and S6G2. We can see that the dehazing performance increases rapidly as we increase the scales for feature extraction and fusion. On the outdoor dataset, all the variants have obtained better performance than GridDehazeNet. The S4G2 is able to achieve the best dehazing result.

TABLE I

THE EFFECTIVENESS OF UTILIZING DIFFERENT SCALES FOR HIERARCHICAL FEATURE FUSION AND DIVERSE ARCHITECTURE SETTINGS FOR MIXED CONVOLUTION ATTENTION MECHANISM. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE COLORS, RESPECTIVELY. ↑ MEANS THAT THE BETTER ALGORITHM SHOULD ACHIEVE A HIGHER SCORE FOR THIS METRIC. "√" INDICATES WE SELECT THE NETWORK SETTING

Models	S4	S5	S6	G2	G4	G6	G8	PSNR (↑)	SSIM (↑)	Param. Num
Models trained and evaluated on indoor images										
GridDehazeNet								32.16	0.9836	-
Baseline								27.98	0.9683	1.296 M
Baseline+S4	√							28.98	0.9771	1.301 M
S4G2	√			√				31.59	0.9922	1.436 M
S4G4	√				√			32.49	0.9933	1.465 M
S4G6	√					√		32.23	0.9919	1.505 M
S4G8	√						√	32.58	0.9921	1.562 M
S4G2	√			√				31.59	0.9922	1.436 M
S5G2		√		√				33.64	0.9940	2.293 M
S6G2			√	√				35.17	0.9954	3.182 M
S6G4			√		√			35.22	0.9954	3.236 M
Models trained and evaluated on outdoor images										
GridDehazeNet								30.86	0.9819	-
Baseline								28.82	0.9616	1.296 M
Baseline+S4	√							30.93	0.9836	1.301 M
S4G2	√			√				34.98	0.9920	1.436 M
S4G4	√				√			33.64	0.9905	1.465 M
S4G6	√					√		33.29	0.9897	1.505 M
S4G8	√						√	32.96	0.9889	1.562 M
S5G2		√		√				33.44	0.9894	2.293 M
S5G4		√			√			32.95	0.9888	2.334 M
S5G6		√				√		32.85	0.9872	2.391 M
S5G8		√					√	32.48	0.9871	2.472 M



(a) Baseline+S4 (b) S4G4 (c) S6G4

Fig. 5. The absolute weight analysis of the final convolution layer for different model variants. The final convolution layers in Baseline+S4 and S4G4 have 320 input feature channels and 3 output channels. The final convolution layer in S6G4 has 576 input channels and the number of output channels is 3. All convolution filters have the same kernel size 3×3. The absolute weight for each filter is obtained by calculating the sum of the absolute values of the 9 weights in the filter. In this figure, these convolution filters are split into 3 groups according to the number of output channels.

3) Mixed Convolution Attention Analysis: Based upon the evaluation results of S4G2, S4G4, S4G6, and S4G8 on the indoor dataset, we find the mixed convolution attention modules can significantly improve the dehazing performance. For example, S4G8 achieves PSNR and SSIM scores of 32.58 dB and 0.9921, which obtains a performance improvement of approximately 3.60dB\0.015 over the Baseline+S4. In addition, the number of groups in the mixed depthwise convolution also affects the performance on indoor images. When the feature maps are divided into 8 groups, the dehazing model achieves the best evaluation results between the four variants. Due to the balance between model complexity and dehazing performance, S6G4 is selected as the best dehazing model on indoor images. However, on outdoor dataset, the evaluation performance slightly decreases as we increase the number of groups in the attention module. This may be caused by overfitting on the train images when models have too strong presentation power. Thus, on outdoor images, we select the S4G2 as our final model.

TABLE II

THE EFFECTIVENESS OF DEEP SEMANTIC LOSS. THE "-" INDICATES WE DO NOT USE THE DEEP SEMANTIC LOSS WHEN TRAINING THE MODELS. ALL MODELS HERE ARE TRAINED FOR 100 EPOCHS. WE SELECT THEIR BEST EVALUATION RESULTS AND REPORT THEM IN THIS TABLE

Models	PSNR (↑)	SSIM (↑)
S6G2-	34.6758	0.9948
S6G2	35.0675	0.9953
S6G4-	34.1381	0.9947
S6G4	34.5392	0.9918

4) Model Parameter Analysis: Additionally, the mixed convolution attention module is lightweight as well as capable of significantly improving the dehazing performance. From the number of parameters in S4G2, S4G4, S4G6, and S4G8,

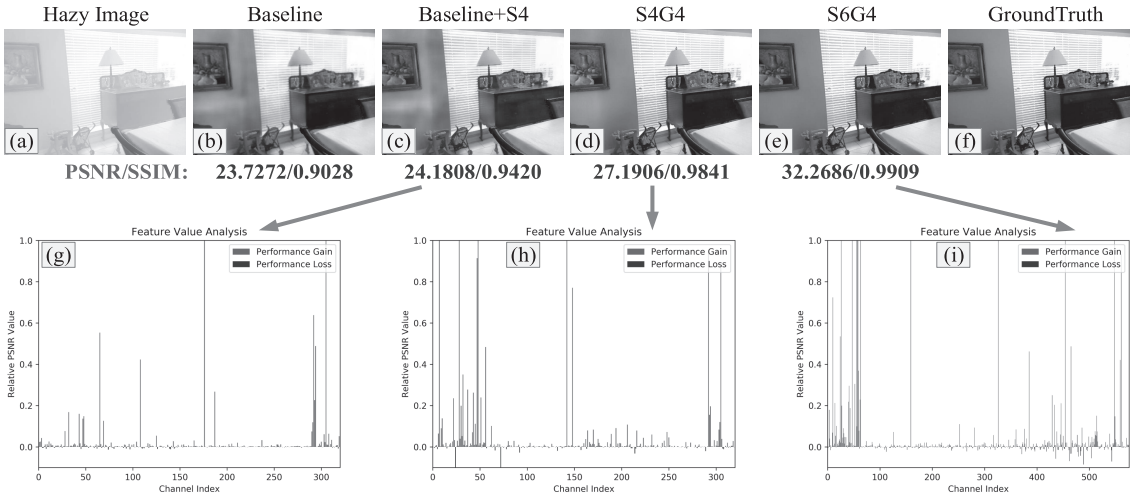


Fig. 6. Visual Comparisons and Feature Value Analysis. The first row shows the visual comparison results between different schemes given the same input hazy image. For each variant, we list the PSNR and SSIM values under its restored image. Images (g), (h), and (i) give the value of each channel of the input feature maps for Baseline+S4, S4G4, S6G4, respectively. The positive value of each channel indicates that it has positive impact on the dehazing performance; while a negative value of each channel means that the feature channel decreases the dehazing performance.

TABLE III

QUANTITATIVE EVALUATION ON THE INDOOR SCENARIOS FROM THE SOTS DATASET. THE RED AND BLUE TEXTS INDICATE THE BEST AND SECOND-BEST PERFORMANCES. ↑ MEANS THAT THE BETTER ALGORITHM SHOULD ACHIEVE A HIGHER SCORE FOR THIS METRIC

Methods	DCP	DehazeNet	MSCNN	AOD-Net	DCPDN	GFN	EPDN	GridDehazeNet	Ours
PSNR (↑)	16.62	21.14	19.84	19.06	15.85	22.30	25.06	32.16	35.21
SSIM (↑)	0.8197	0.8472	0.8327	0.8504	0.8175	0.8800	0.9232	0.9836	0.9954

TABLE IV

QUANTITATIVE EVALUATION ON THE OUTDOOR SCENARIOS FROM THE SOTS DATASET. THE RED AND BLUE TEXTS INDICATE THE BEST AND SECOND-BEST PERFORMANCES. ↑ MEANS THAT THE BETTER ALGORITHM SHOULD ACHIEVE A HIGHER SCORE FOR THIS METRIC

Methods	DCP	DehazeNet	MSCNN	AOD-Net	GFN	GridDehazeNet	Ours
PSNR (↑)	19.14	24.75	22.06	24.14	28.29	30.86	34.98
SSIM (↑)	0.8605	0.9269	0.9078	0.9198	0.9621	0.9819	0.9920

TABLE V

QUANTITATIVE EVALUATION ON THE I-HAZE AND O-HAZE DATASETS IN TERMS OF THE PSNR AND SSIM. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE COLORS, RESPECTIVELY. ↑ MEANS THE BETTER ALGORITHM SHOULD ACHIEVE A HIGHER SCORE FOR THIS METRIC

Datasets	Methods	DCP	AOD-Net	MSCNN	GFN	PFFNet	GridDehazeNet	DuRN	Ours
I-HAZE	PSNR (↑)	14.43	13.98	15.22	15.84	16.01	17.22	21.23	21.40
	SSIM (↑)	0.752	0.732	0.755	0.751	0.740	0.732	0.842	0.887
O-HAZE	PSNR (↑)	16.78	15.03	17.56	18.16	23.33	20.91	22.00	23.64
	SSIM (↑)	0.653	0.539	0.650	0.671	0.869	0.726	0.820	0.886

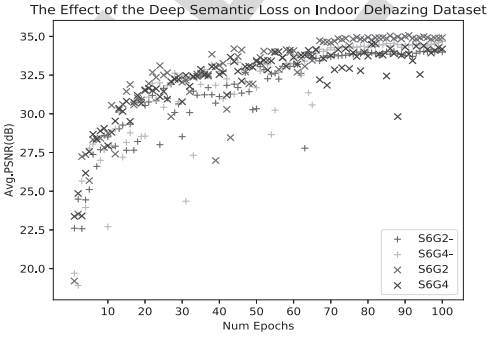


Fig. 7. Average evaluation results during training on indoor images. Models can achieve higher performance with the deep semantic loss.

we can see that the model complexity does not change too much as we increase the number of groups in the mixed convolution attention modules. In addition, by comparing Baseline+S4 and S4G2, a performance gain of approximately 2.61 dB PSNR value has been achieved with only 0.135 million additional learnable parameters.

5) Feature Redundancy Analysis: To make our experiments more convincing and verify the rationality of our statements, we investigate the feature redundancy in networks by analysing the learned filter weights of the final convolution layer in Baseline+S4, S4G4 and S6G4. To analyze the impact of each filter, we compute its absolute weights and the analysis results are shown in Fig. 5. The absolute weights of each filter is obtained by computing the sum of the absolute values of its 9 weights. From the plot, several observations can be made: (1). Fig. 5 (a) indicates that feature usage cannot be effectively performed by Baseline+S4 and the model indeed has many redundant features (with abundant low

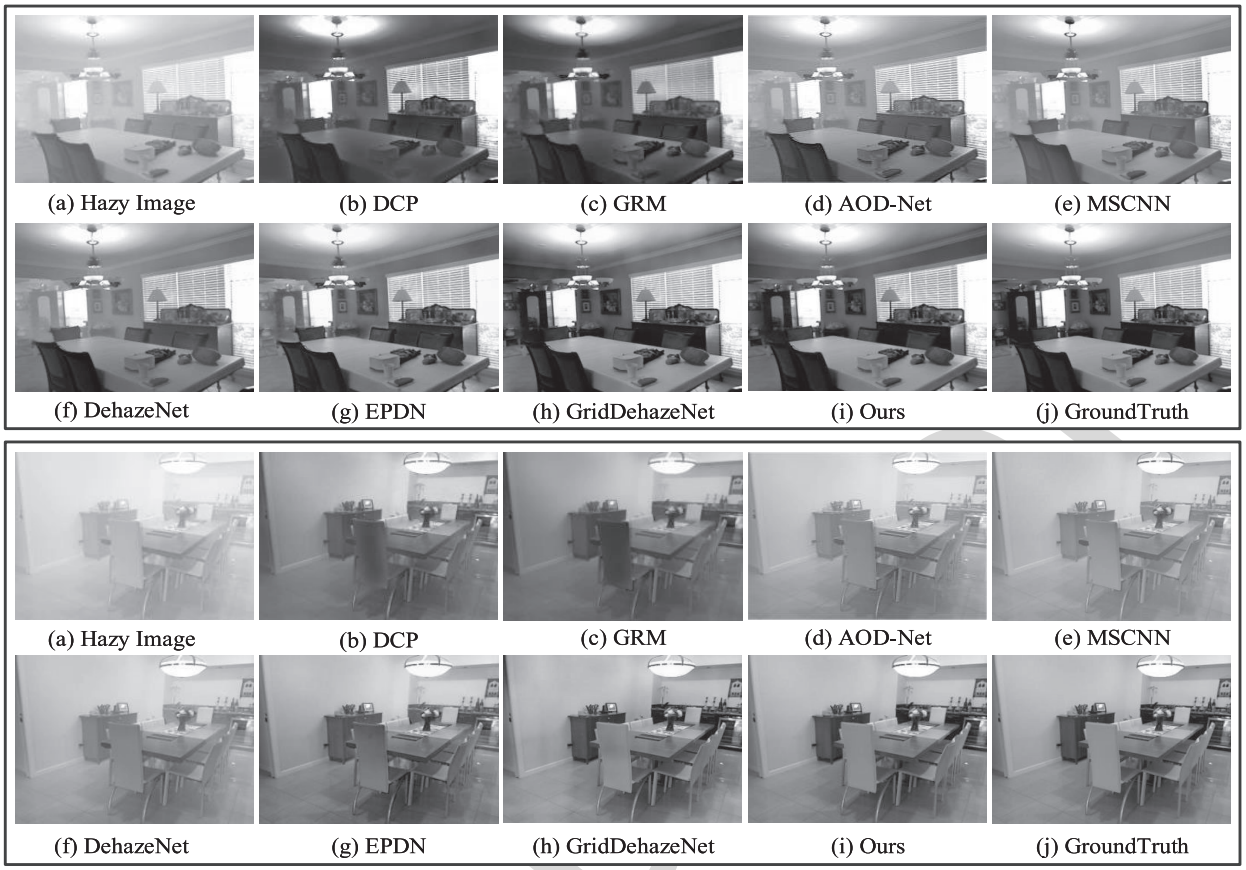


Fig. 8. Examples of haze-removed images on indoor scenarios. The proposed method generates more visually pleasant results than other state-of-the-art algorithms. **Best viewed in color and zoom in for better visibility.**

absolute weights). (2). With our proposed attention mechanism, compact internal feature representations are effectively learned and the feature redundancy is well reduced (the absolute weights are improved). (3). The learned features have more diversity (the absolute weights are diversified).

6) Visual Comparisons and Feature Value Analysis: The visual comparisons between Baseline, Baseline+S4, S4G2, and S6G4 are illustrated in the first row of Fig. 6. Based on the images Fig. 6 (c) and Fig. 6 (d) of Baseline+S4 and S4G4, respectively, we find that the proposed mixed convolution attention mechanism effectively alleviates the artifacts and color distortion problems in images, and the recovered structure details are sharper as well. Further, the S6G4 generates more visually pleasant haze-removed result, the artifacts and color distortions are completely resolved as we increase the scales for the hierarchical feature fusion and select appropriate number of groups for the proposed attention modules.

To investigate the relationship between each channel of the input feature maps with the final performance, we compute the value of each channel. Note that the features analysed here are the input features for image reconstruction module. We first compute the PSNR value of recovered images with all input feature channels, and the PSNR value is set as the baseline for evaluating the effectiveness of each channel. Then, we compute a new PSNR value with all channels as input except a specific channel, the value of the specific channel is defined as the difference between the baseline PSNR value

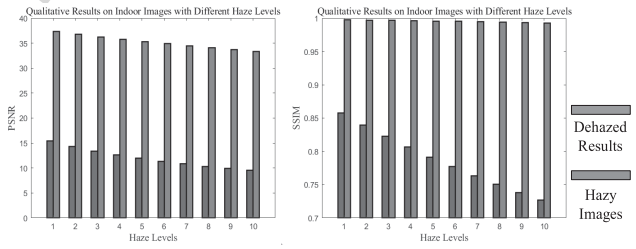


Fig. 9. Evaluation results on indoor images with different haze levels.

and the new computed PSNR value. Channels with positive values indicate that they have positive impacts on the dehazing performance. While channels with negative values mean that these channels will decrease the dehazing performance. From Fig. 6 (g), (h), and (i), we can know that feature redundancy has negative impacts on the dehazing performance to some extent. Models with the proposed mechanism can reduce feature redundancy and transform some ineffective features into valuable ones, therefore the performance is improved. Furthermore, the improvement of dehazing performance is the comprehensive result of multiple channels.

7) Effectiveness of Deep Semantic Loss: To validate the effectiveness of the deep semantic loss, several check experiments are conducted. In table II, S6G2- means that the model is optimized without the deep semantic loss, while S6G2 employs all four loss functions. S6G4- and S6G4 have

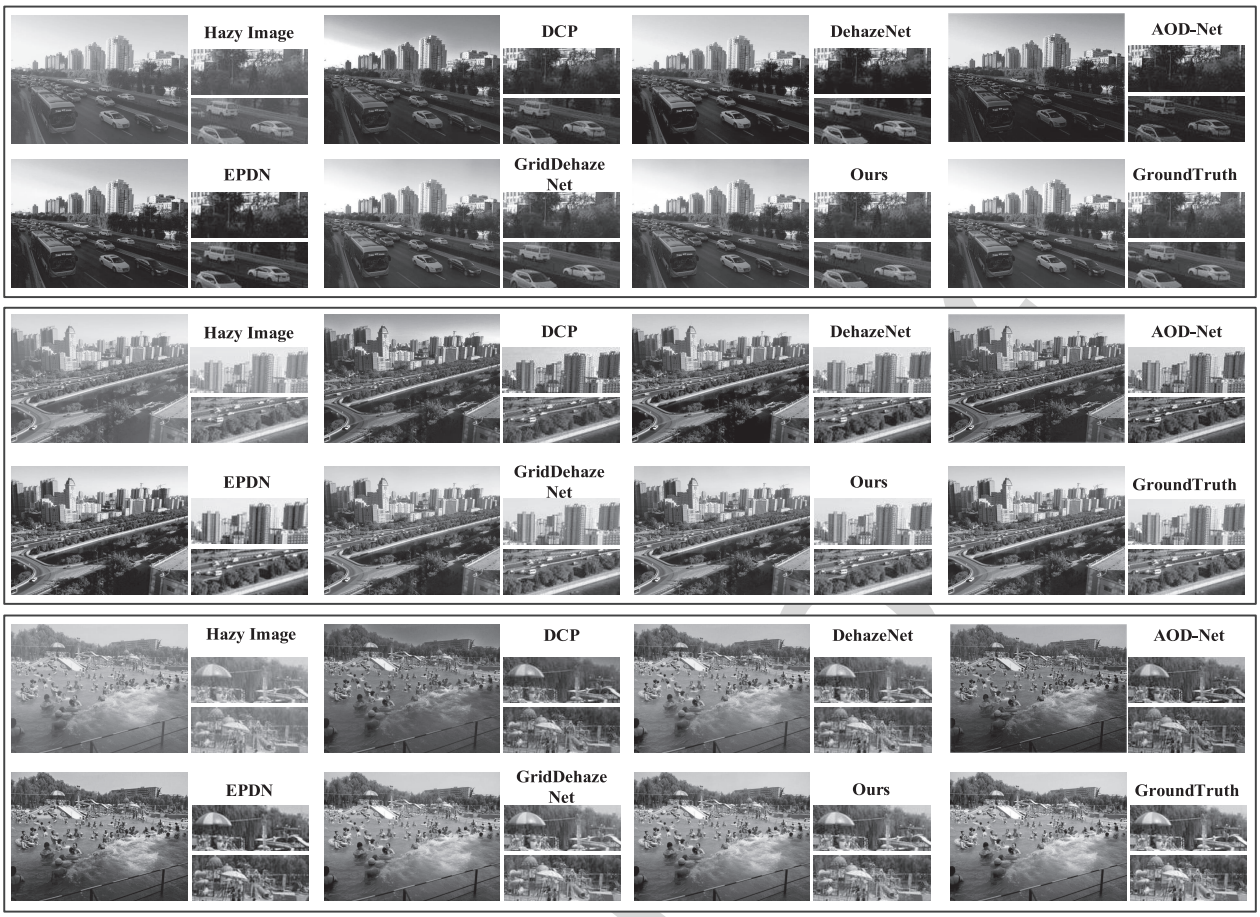


Fig. 10. Examples of haze-removed images on outdoor scenarios. For convenient comparison, we crop two patches from each image and magnify them to show the details. **Best viewed in color and zoom in for better visibility.**

similar meanings as well. Note that, S6G2 and S6G4 are the most powerful variants for indoor scenarios, thus the experiment settings here are rigorous enough to investigate the effect of the deep semantic loss. All the models here are trained for 100 epochs, and we select the best evaluation results from them. From the evaluation results, we can see that the proposed deep semantic loss is able to help model optimization and improve the dehazing performance. The average evaluation results during training on indoor images are also illustrated in Fig. 7. The convergence speed is also accelerated with the loss function.

C. Performance Evaluation on Synthetic Images

To validate the effectiveness of the proposed dehazing method, extensive experimental comparisons between the proposed approach and several state-of-the-art methods are conducted on synthetic indoor and outdoor images. These compared methods include hand-crafted prior method (DCP [10]) and learning-based approaches (DehazeNet [13], MSCNN [20], AOD-Net [22], DCPDN [14], GFN [39], EPDN [8], and GridDehazeNet [9]).

Table III lists the evaluation results on indoor images. As expected, the dehazing method DCP does not perform well, which means that their proposed prior-based strategy does not effectively fit the problem. By estimating the medium transmission map to restore the haze-free images using an end-to-end architecture, DehazeNet achieves a performance improvement of 4.52 dB in terms of the PSNR metric over DCP. Similarly, MSCNN is also based on the atmospheric scattering model. MSCNN first estimates the transmission maps using a coarse-to-fine network, and then restores the haze-removed images with a fine-scale network. The evaluation results also demonstrate its effectiveness. DCPDN learns the transmission map and atmospheric light in an end-to-end manner, but it does not achieve a remarkable performance on the evaluation dataset. Without relying on the atmospheric scattering model, AOD-Net, EPDN and GridDehazeNet achieve better dehazing performance. This indicates that the estimation of the transmission map or the atmospheric light from a single hazy input is not a trivial task; while model-free methods that directly learn the map between a hazy input and its corresponding clean result perform better.

As listed in Table III, EPDN gives the PSNR and SSIM scores of 25.06 dB and 0.9232. GridDehazeNet achieves the performance of 32.16 dB on PSNR and 0.9836 on SSIM. Our proposed dehazing algorithm achieves a performance improvement of 3.05 dB in terms of the PSNR over GridDehazeNet, and the SSIM value is 0.9954. The evaluation results significantly outperforms all the compared dehazing methods, thus demonstrating the effectiveness of the proposed

(a) Hazy Image (b) PFFNet (c) GridDehazeNet (d) DuRN (e) Ours (f) GroundTruth

Fig. 11. Comparison of qualitative results between state-of-the-art methods and the proposed approach on the I-HAZE and O-HAZE datasets. The indoor results are shown in the four upper rows, and the outdoor results are shown in the bottom three rows. For convenient comparison, a patch is cropped and its magnified version is placed in the bottom-right corner of each individual image. **Best viewed in color and zoom in for better visibility.**

hierarchical feature fusion and mixed convolution attention mechanism. Furthermore, the evaluation results on synthetic outdoor dataset also demonstrate the effectiveness of the proposed algorithm. As listed in table IV, our approach achieves the best dehazing performace. To the best of our knowledge, we are the first to report a dehazing performance with approximately 34.98 dB of PSNR and 0.9920 of SSIM on the SOTS outdoor dataset.

Fig. 8 and Fig. 10 illustrate the comparisons between our method and other methods in terms of the visual quality on the synthetic images. As shown in Fig. 8, the images recovered by DCP have severe color distortion and look fairly hazy. Although the GRM outputs severe color-distorted images, it tends to estimate the haze level more accurately, thus the restored images have less haze. AOD-Net, MSCNN, and DehazeNet effectively alleviate the color distortion problem, while they tend to underestimate the haze level of the input images. Detailed information such as textures and edges in their recovered images is unsatisfactory as well. When it comes to EPDN, it can generate more visually pleasant results than these aforementioned methods; however, the restored images still have some remaining haze when the input images contain severe haze. Besides, it cannot effectively recover the structures and details due to loss of high-level semantics information. GridDehazeNet effectively overcomes the color distortion problem and generates dehazed images closer to the haze-free images. However, it tends to cause some artifacts in the background parts when removing thick haze. Compared with these aforementioned methods, our haze-removed images are free of major artifacts and color distortions and have the best visual effect.

The evaluation results about the effect of haze levels on the model performance are shown in Fig. 9. As the haze level increases, the PSNR values slightly decrease for both the haze-removed results and input hazy images. However, the proposed method still can perform well on images with the highest haze level. The SSIM values of the dehazed results are high and largely constant, whereas the SSIM scores of

(a) Hazy Image (b) Our Result (c) GroundTruth

Fig. 12. Visualization of an imperfect example on an outdoor scenario with heavily haze. The proposed method will generate some distortions and the recovered image as shown is polluted by chaotic colors.

the input hazy images decrease dramatically. The results indicate that the proposed method can generate visually pleasant images with sharper structures and meaningful textural details.

Fig. 10 illustrates the qualitative results on the SOTS outdoor dataset. The output of DCP severely suffers from color distortions and the description of the sky looks unreal. DehazeNet and AOD-Net slightly alleviate the color distortion problem; however, their results are typically darker than the ground truth images. Although EPDN does not output low-brightness images, the color fidelity of some restored results and detailed information such as textures, edges and the blue sky are unsatisfactory. GridDehazeNet and our method are superior in image details and color fidelity. Overall, the proposed method can generate more visually pleasant results than other methods.

D. Performance Evaluation on Real-World Images

To make the model evaluation more convincing, we conduct comparison experiments between the proposed approach and several state-of-the-art methods on real-world images. These methods include DCP [10], AOD-Net [22], MSCNN [20], GFN [39], PFFNet [23], GridDehazeNet [9], and DuRN [40]. For a fair comparison, we obtain the evaluation results of PFFNet, GridDehazeNet, and DuRN using their released codes and models on the same train and test datasets. The other algorithms are evaluated with their provided pre-trained models. Table V lists the quantitative results on the I-HAZE and O-HAZE datasets. The proposed method reports better PSNR and SSIM results than other algorithms. Compared with DuRN and our method, GridDehazeNet has a slightly poor performance on real hazy images, which indicates its overfitting on the large-scale training set. Our method achieves similar results with DuRN on the I-HAZE dataset and obtains a performance improvement of 1.62 dB PSNR value over DuRN on the O-HAZE dataset. The proposed method gives a much higher SSIM score than the other methods, indicating it effectiveness of preserving meaningful texture and structure information in the restored images.

Fig. 11 illustrates the qualitative comparison on the I-HAZE and O-HAZE testing sets. The restored images from I-HAZE are shown in the top four rows, and the bottom three rows show the results from O-HAZE. We highlight some regions in the images to show detailed information. In Fig. 11, PFFNet is far from achieving realistic performance due to the unsatisfactory image details and color fidelity on indoor scenarios, and a large number of image contents are polluted by chaotic colors. Furthermore, it can achieve significant visual results on outdoor real-world images. For GridDehazeNet, a significant amount of haze remains unremoved in both the indoor and outdoor images. It also suffers from the color distortion problem. DuRN can estimate the haze level accurately; however, it cannot effectively recover clear structures and details from images containing severe haze. Our proposed method is found to generalize well on realistic hazy images in terms of visual quality, and can better preserve details and color information in haze-removed images.

E. Limitation

The proposed method is not robust enough for few super-hard scenarios. Fig. 12 illustrates an example: when the haze is extremely thick and the image boundary is unclear in the input hazy scenario, the method will generate images with some distortions and chaotic colors. The limitation might be alleviated by fully optimizing the model parameters.

V. Conclusion

In this study, we developed a hierarchical feature fusion dehazing method with a novel proposed mixed convolution attention mechanism to progressively and adaptively improve the dehazing performance. The haze levels could be accurately estimated by fusing multi-scale hierarchical features; thus, the proposed method could generate images almost without remaining haze. The structure information in recovered images was also clearly preserved as we increase the scales in feature fusion. Moreover, the mixed convolution attention module was designed to reduce feature redundancy, transform some ineffective features into valuable ones and highlight task-relevant features for reconstructing visually pleasant haze-removed images. The deep semantic loss can facilitate the optimization of the learnable parameters as well. Extensive experimental results demonstrated that the proposed approach achieved superior dehazing performance on both synthetic and real-world images compared with several state-of-the-art algorithms.

References

[1] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "DeepDriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2722–2730.

[2] X. Zhang, D. Wang, Z. Zhou, and Y. Ma, "Robust low-rank tensor recovery with rectification and alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 238–255, Jan. 2021.

[3] Z. Jia, H. Wang, R. E. Caballero, Z. Xiong, J. Zhao, and A. Finn, "A two-step approach to see-through bad weather for surveillance video quality enhancement," *Mach. Vis. Appl.*, vol. 23, no. 6, pp. 1059–1082, Nov. 2012.

[4] L. Ren, J. Lu, Z. Wang, Q. Tian, and J. Zhou, "Collaborative deep reinforcement learning for multi-object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 508–602.

[5] X. Zhang, W. Hu, N. Xie, H. Bao, and S. Maybank, "A robust tracking system for low frame rate video," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 279–304, Dec. 2015.

[6] E. J. McCartney, *Optics of the Atmosphere: Scattering by Molecules and Particles*, vol. 421. New York, NY, USA: Wiley, 1976.

[7] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," *Int. J. Comput. Vis.*, vol. 48, no. 3, pp. 233–254, 2002.

[8] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced Pix2pix dehazing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8160–8168.

[9] X. Liu, Y. Ma, Z. Shi, and J. Chen, "GridDehazeNet: Attention-based multi-scale network for image dehazing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7314–7323.

[10] 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.

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

[12] J. Wang, K. Lu, J. Xue, N. He, and L. Shao, "Single image dehazing based on the physical model and MSRCR algorithm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2190–2199, Sep. 2018.

[13] 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, no. 11, pp. 5187–5198, Nov. 2016.

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

[15] J.-L. Yin, Y.-C. Huang, B.-H. Chen, and S.-Z. Ye, "Color transferred convolutional neural networks for image dehazing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3957–3967, Nov. 2020.

[16] D. Zhao, L. Xu, L. Ma, J. Li, and Y. Yan, "Pyramid global context network for image dehazing," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Nov. 9, 2020, doi: 10.1109/TCSVT.2020.3036992.

[17] 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.*, early access, Dec. 22, 2020, doi: 10.1109/TCSVT.2020.3046625.

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

[19] Y. Zhang, P. Wang, Q. Fan, F. Bao, X. Yao, and C. Zhang, "Single image numerical iterative dehazing method based on local physical features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3544–3557, Oct. 2020.

[20] 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.*, Oct. 2016, pp. 154–169.

[21] C. O. Ancuti and C. Ancuti, "Single image dehazing by multi-scale fusion," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3271–3282, Aug. 2013.

[22] 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. (ICCV)*, Oct. 2017, pp. 4770–4778.

[23] K. Mei, A. Jiang, J. Li, and M. Wang, "Progressive feature fusion network for realistic image dehazing," in *Proc. Asian Conf. Comput. Vis.* Springer, 2018, pp. 203–215.

[24] H. Zhang, V. Sindagi, and V. M. Patel, "Multi-scale single image dehazing using perceptual pyramid deep network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 902–911.

[25] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.

[26] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[27] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[28] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4634–4643.

[29] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[30] N. S. Trudinger, "Elliptic partial differential equations of second order," Tech. Rep., 1983.

[31] B. Li *et al.*, "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2019.

[32] P. K. Nathan Silberman, D. Hoiem, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.

[33] C. Ancuti *et al.*, "NTIRE 2018 challenge on image dehazing: Methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 891–901.

[34] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "I-HAZE: A dehazing benchmark with real hazy and haze-free indoor images," 2018, *arXiv:1804.05091*. [Online]. Available: http://arxiv.org/abs/1804.05091

[35] 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/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 754–762.

[36] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electron. Lett.*, vol. 44, no. 13, pp. 800–801, Jun. 2008.

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[39] W. Ren *et al.*, "Gated fusion network for single image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3253–3261.

[40] X. Liu, M. Suganuma, Z. Sun, and T. Okatani, "Dual residual networks leveraging the potential of paired operations for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7007–7016.

Xiaoqin Zhang received the B.Sc. degree in electronic information science and technology from Central South University, 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, China, in 2010. He is currently a Professor with Wenzhou University, 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, *IJCV*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, IEEE TRANSACTIONS ON COMPUTERS, ICCV, CVPR, NIPS, IJCAI, AAAI, and among others. His research interests are in pattern recognition, computer vision, and machine learning.

Jinxin Wang received the B.Sc. degree in information and computing science with Wenzhou University. He is currently pursuing the degree with the College of Computer Science and Artificial Intelligence, Wenzhou University, China. His research interests include image restoration, reinforcement learning, and statistical learning theory.

Tao Wang received the B.Sc. degree in information and computing science from Hainan Normal University, China, in 2018. He is currently pursuing the degree with the College of Computer Science and Artificial Intelligence, Wenzhou University, China. 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.

Runhua Jiang received the B.Sc. degree with the Department of Information Science, Tianjin University of Finance and Economy, China. He is currently pursuing the degree in computer software and theory with the College of Computer Science and Artificial Intelligence, Wenzhou University, China. His research interests include several computer vision tasks, such as image/video restoration, crowd counting, visual understanding, and video question answering.