Hindawi
Scientific Programming
Volume 2022, Article ID 2056662, 10 pages
https://doi.org/10.1155/2022/2056662

Research Article

A Multistage with Multiattention Network for Single Image Dehazing

Bin Hu, Mingcen Gu, and Yuehua Li

School of Information Science and Technology, Nantong University, Nantong, Jiangsu, China

Correspondence should be addressed to Bin Hu; hubin@ntu.edu.cn

Received 19 December 2021; Revised 8 March 2022; Accepted 22 April 2022; Published 9 May 2022

Academic Editor: Jianping Gou

For single image dehazing, an end-to-end multistage with multiattention network is proposed in this paper. The network contains two different stages, in which the first stage uses an encoder-decoder subnet to obtain contextual features, and the second stage adopts a single-scale pipeline to provide spatial image details. At each stage, ground-truth supervision is provided, and an attention mechanism is used between the two stages, so the features learned from the previous stage will be refined before passing to the next stage. A basic multiattention unit that combines channel attention, spatial attention, and pixel attention is designed to earn more weight from important features, and a positional normalization that normalizes exclusively across channels is used in the multiattention unit to learn more weight from important features. Experimental results in several benchmarks indicate that the proposed network outperforms the state-of-the-art methods both quantitatively and qualitatively.

1. Introduction

Image dehazing is a challenging task in the field of image restoration. Since there are infinite feasible solutions, it is a highly ill-posed problem. The atmosphere scattering model [1, 2] proposes a simple and effective formula to solve the problem.

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

where $I(x)$ is the hazy image, $J(x)$ stands for clear image, the global atmospheric light A represents the intensity of the scattered light of the scene, and the transmission map $t(x)$ describes the attenuation in intensity.

Let the clear image $J(x)$ be the output, and the formula (1) can be re-written as follows:

$$J(x) = \frac{1}{t(x)} I(x) - A\frac{1}{t(x)} + A. \tag{2}$$

It can be observed from formula (2) that the goal of single image dehazing is to restore the clear image $J(x)$ from the hazy image $I(x)$ by estimating A and $t(x)$. Since only $I(x)$ is known, it is difficult to restore the clear image $J(x)$.

Recent decades have witnessed significant progress in image dehazing, and lots of techniques have been proposed. Early works were mostly based on priors such as the atmosphere scattering model, and these methods often try to design hand-crafted features to learn $t(x)$ in formula (2). However, the methods are easily sensitive to image variations such as changes in viewpoints, illumination, and scenes [3]. With the success of deep convolution neural network (CNN) in the community of image processing and computer vision, lots of image dehazing methods based on CNN have been proposed [4–6], which can directly regress the intermediate transmission map or the final haze-free image. Compared to early methods based on hand-designed features, CNN-based methods achieve superior performance with robustness.

The net design is a primary reason of the superior performance achieved by CNN-based methods. Lots of network modules are introduced for image dehazing including residual dense connections [7, 8], attention mechanisms [8, 9], encoder-decoder [10, 11], and generative models [12, 13]. Nevertheless, most of them are single-stage models. On the other hand, multistage models have been shown to be more effective than single-stage models in different vision tasks such as segmentation and pose-estimation. Recently, few efforts have adopted multistage networks to solve image deblurring and image deraining [14–16]. We analyze those methods and find there are several bottlenecks that prevent the performance. First, the existing multistage networks use same architecture in different stage, either an encoder-decoder architecture or a single-scale pipeline. The encoder-decoder [11] architecture provides broad contextual information but lacks image spatial details, and the single-scale pipeline is effective in preserving spatially accurate but unreliable in extracting semantical information. We combine the two architectures in a multistage network for image dehazing. As far as we know, this is the first attempt to solve this problem. Second, we do not naively pass the output of the previous stage to the next stage [15]. A ground-truth supervision is provided in the first stage to refine the feature map before moving to the next stage. Third, most attention modules (MAU) are single and limited, such as channel attention, which can extract the interdependencies among channels but lacks spatial information. We first combine different attention mechanisms to address the limits. The proposed multiattention combines channel attention, spatial attention, and pixel attention to extract more important information.

In summary, the main contributions of the work are as follows:

(1) We employ a multistage network, which combines two different architectures. The proposed multistage network is capable of extracting broad contextual and spatially detailed information.

(2) At each stage, a ground-truth supervision is provided and an attention mechanism is used among adjacent stages. By the supervision of ground-truth image, the features learned from the previous stage will be refined before moving to the next stage.

(3) A multiattention unit (MAU) is proposed that combines channel attention in channel-wise, spatial attention in spatial-wise, and pixel attention in pixel-wise to earn more weight from important features.

(4) The positional normalization [17] (PONO), which is position-dependent and reveals structural information at this particular layer of the deep net, is adopted to improve the training performance.

2. Related Work

Most dehazing approaches follow the similar three-step methodology based on the atmospheric scattering model: (1) estimating the transmission map $t(x)$ by the hazy image samples; (2) estimating the global atmospheric light A using empirical methods; and (3) computing the clear image $J(x)$ according to formula (3). Most of the work focuses on the first step. There are two ways to estimate $t(x)$: physically grounded priors and fully data-driven approaches.

Early methods based on physically grounded priors often require multiple images from the same scene under different conditions [2, 18–20]. However, these methods do not work when there is only one image for a scene. The dark channel prior (DCP) [21] is the most successful prior-based method and is followed by many successors. Gibson et al. [22] adopted a standard median filter to improve the DCP computing speed. An effective contextual regularization based on boundary constraints is proposed in [23] to restore the hazy image. Based on depth estimation, a color attenuation prior [24] is proposed for haze removal. Berman et al. [25] assume that an image contains only several hundreds of distinct colors and proposed a nonlocal method. However, the prior is computationally expensive and unreliable.

With the success of deep learning in diverse computer vision tasks, the data-driven dehazing approaches have become popular. To avoid estimating the parameters inaccurately and designing hand-crafted features, algorithms use convolutional neural networks (CNNs) to directly learn $t(x)$ from data.

Single-Stage Networks: currently, most single image dehazing methods are based on single-stage networks. The AOD-Net [26] is the first end-to-end network to generate clean images directly. It is a lightweight CNN, but still performs much better than prior-based methods. The EPDN [27] adopts a generative adversarial network to solve the image dehazing without relying on the physical scattering model. Zhao et al. proposed a weakly supervised refinement framework called RefineDNet [28], which can outperform the weakly supervised methods but is weak than the supervised networks. The DehazeFlow [29] proposes a conditional normalizing flow based framework for single image dehazing.

MultiStage Networks: the existing multistage networks usually use the identical architecture in different stages, such as the Grid DehazeNet [8] and the gated fusion network [30]. The information generated by the previous stage always naively flows to the next stage to refine the restored image [3]. However, a common practice is to use the same subnetwork for each stage may yield a suboptimal result, and the naive connection between adjacent stages is also a bottleneck, as shown in our experiments.

Attention: attention mechanisms are widely used in both high-level computer vision tasks, including image classification [31] and object detection [32], and low-level computer vision tasks such as image dehazing [8, 9], deraining [16], and deblurring [14, 15]. The main idea is to capture long-range interdependencies in channel-wise, spatial-wise, or pixel-wise.

3. Proposed Method

We mainly discuss the detail of the proposed network MSNet in this section. The MSNet is a multistage with multiattention network, and it is a trainable end-to-end network that does not rely on the atmosphere model. The MSNet consists of two stages, as shown in Figure 1, of which the first stage is based on the encoder-decoder network which learns the contextual information, and the second stage is a single-scale pipeline to provide the spatial image details. Inspired by [33], a supervised attention block (SAB) is used between the two stages. By the supervision of clear

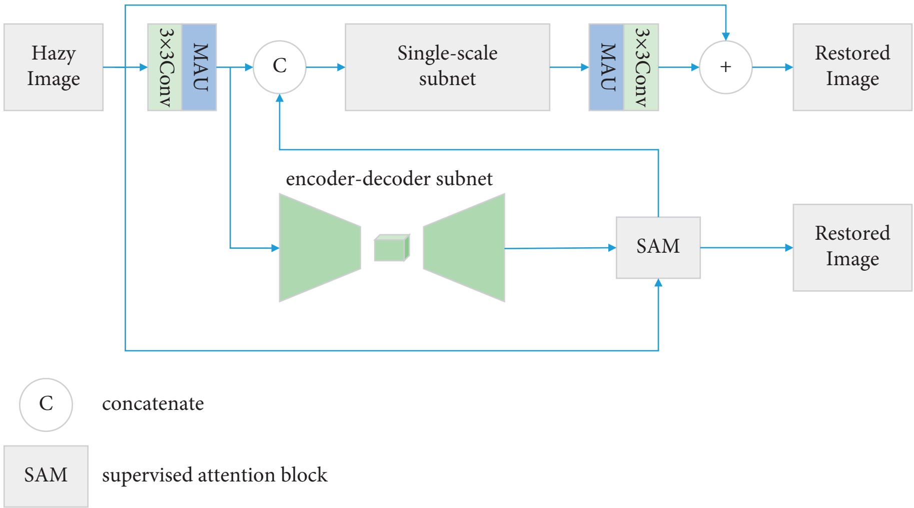

FIGURE 1: The architecture of the multistage network. The first stage uses an encoder-decoder net to extract contextualized information, and the second stage employs a multiattention single-scale network to output spatial features. A supervised attention block is used between the two stages to refine the features of the first stage.

images, the feature maps in the first stage are refined by SAB before flowing to the second stage.

3.1. Multiattention Unit. In our framework, a multiattention unit (MAU) is proposed as the basic unit. The architecture of basic MAU is depicted in Figure 2, and it consists of two convolution layers, a local residual learning and a multiattention block. The convolution layers are activated by ReLU, and the second convolution layer adopts positional normalization (PONO) with moment shortcuts (MS) [17] to normalize the activations. A global residual learning connects the input feature and the output feature. With local residual learning and global residual learning, the low-frequency regions from the input features can be learned through the skip connection.

The multiattention block combines channel attention, pixel attention, and spatial attention, so it can provide additional ability in dealing with nonlocal and local information, and the representational ability of CNNs is expanded. The architecture of MAU is depicted in Figure 3.

3.1.1. Channel Attention. Usually, a network uses a number of convolutional layers to capture the neighboring spatial dependencies within local receptive fields. However, the global spatial patterns also need to be considered under the complicated nonuniform condition. When the neighborhoods of the image contain strong hazy component, the contextual information from clear regions may be required. Recently, a channel attention module [31] has been proposed to capture richer nonlocal features by modeling the interdependencies among channels. Thus, we propose the channel attention module to extract nonlocal context features, and the different weighted information from the different channel feature maps will be learned by the channel attention module.

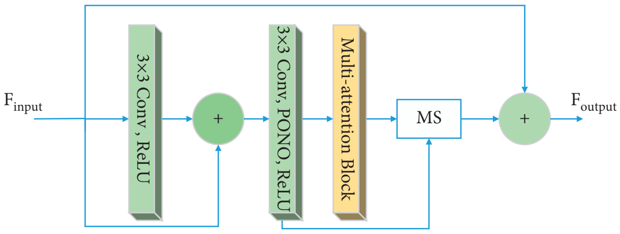

FIGURE 2: Multiattention unit (MAU) uses two convolution layers and a multiattention block.

Firstly, a global average pooling is used to capture the channel-wise global spatial features:

$$g_c = H_p(F_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i, j), \tag{3}$$

where H_p means the global average pooling function and $X_c(i, j)$ is the value of cth channel of input X_c at position (i, j). And the dimension of the feature map changes from $C \times H \times W$ to $C \times 1 \times 1$, C denotes the channels, and $H \times W$ is the size of the feature map.

Then, two convolution layers are applied to get the weights from different channels, and the first convolution layer uses PONO to normalize the activation.

$$C_f = \sigma(\mathrm{Conv}(\delta(\mathrm{Conv}(g_c)))), \tag{4}$$

where σ stands for the sigmoid function that is used to activate the first convolution layer and δ is the ReLU function used to activate the second convolution layer.

Finally, the weight of the channel F_c^* is computed by element-wise multiplying the input F_{input} and C_f.

$$F_c^* = F_{\text{input}} \otimes C_f. \tag{5}$$

3.1.2. Pixel Attention. The variant hazy pixels may distribute in the whole image, so we adopt a pixel attention module to get the variant features from the image in pixel-wise. The

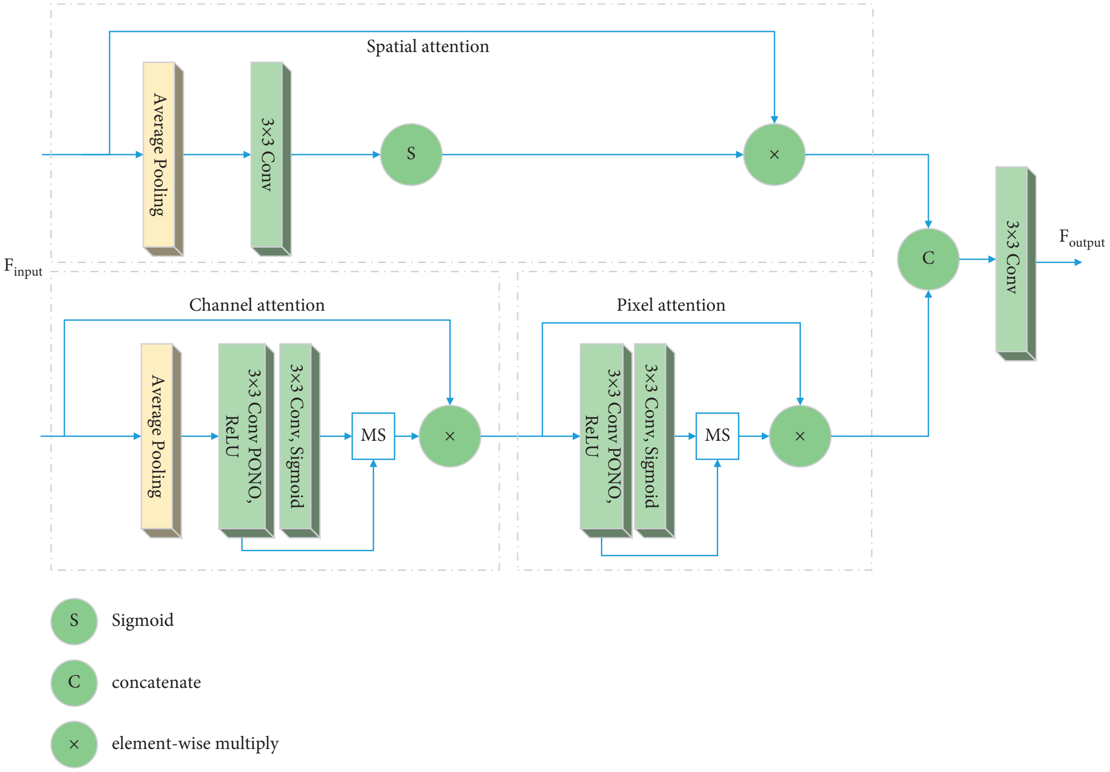

Figure 3: Multiattention block contains different attention mechanism to expand the representational ability of the network.

module is applied to learn weights in an adaptive way from pixels, and the network can learn more informative features from thick-hazed pixels and high-frequency image regions.

The architecture of the pixel attention module is depicted in Figure 3, it consists of two convolution layers and a sigmoid activation function, and the first convolution layer uses PONO to normalize the activations.

$$C_p = \sigma\left(\mathrm{Conv}\left(\delta\left(\mathrm{Conv}\left(F_c^*\right)\right)\right)\right). \tag{6}$$

Then, we element-wise multiply F_c^* and C_p as the output of the channel-pixel attention map:

$$F_{CP} = F_c^* \otimes C_p. \tag{7}$$

3.1.3. Spatial Attention. Spatial attention is designed to exploit the spatial attention map from the input convolutional features F_{input}. The spatial attention module first applies global average pooling on F_{input} along the channel dimensions and outputs a feature map $f \in \mathbb{R}^{H\times W}$. The feature f is then passed through a convolution layer and sigmoid activation to get the spatial attention feature $f_{SA} \in \mathbb{R}^{H\times W}$.

Finally, the spatial attention map f_{SA} and channel-pixel attention map F_{CP} are concatenated, and then the concatenated feature map is passed through a convolution layer to obtain the multiattention map.

3.2. Encoder-Decoder Subnetwork. The encoder-decoder subnetwork is based on the standard U-Net [34] as shown in Figure 4, each scale of the subnet uses an UBlock, which contains several MAUs to extract feature maps, and two down-sampling layers are adopted to reduce the size of the input map to reduce the computation. The skip connections are also processed by an UBlock and then concatenated with the decoder layer. The skipped connections enhance the detailed information of the image. The down-sampling and up-sampling are implemented by a convolution layer.

3.3. Single-Scale Subnet. The single-scale subnet in the second stage consists of several multiattention groups (MAGs), each of which contains several MAUs and a shortcut, and the module is depicted in Figure 5. With the dense attention modules, the net can generate high-resolution and enriched detailed features from the input.

3.4. Supervised Attention Block. Inspired by [30], a supervised attention block (SAB) is used between the two stages, and the architecture of SAB is shown as Figure 6. The SAB uses a ground-true image to supervise the feature maps at the encoder-decoder stage. With the supervision of the ground-truth, the encoder-decoder stage will provide more informative features to the next stage.

SAB takes the output $F_{\mathrm{input}} \in \mathbb{R}^{H\times W\times C}$ from encoder-decoder as the input, where $H \times W$ is the dimension of the features and C denotes the channel's number. After processed by a 1 × 1 convolution, the F_{input} is added to the input hazy image to obtain the dehazed image $I_d \in \mathbb{R}^{H\times W\times 3}$, and a ground-truth image is provided here to predict the dehazed image. Then I_d is processed by a convolution layer with a sigmoid activation to generate the attention maps M. Then,

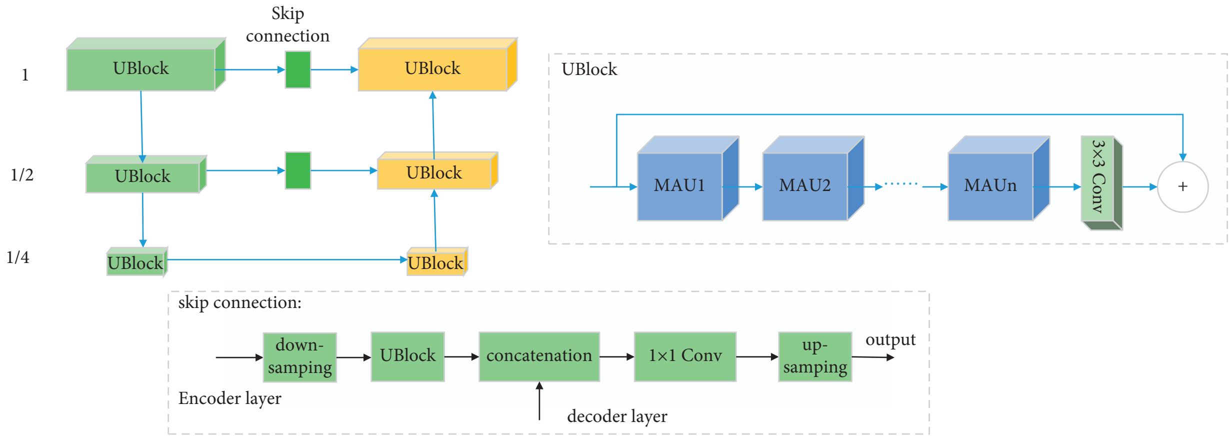

Figure 4: Encoder-decoder subnetwork is based on U-Net, and down-sampling layer is used to reduce the computation. The down-sampling and up-sampling are implemented by a 3 × 3 convolution.

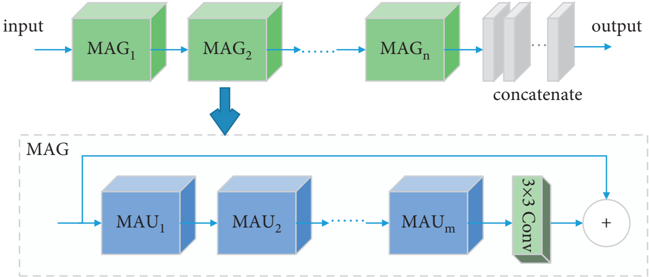

Figure 5: Single-scale subnet consists of several multiattention groups, and each group contains several multiattention units, and the features of each group are concatenated as the output.

we element-wise multiply M and transformed F_{input} that processed by a convolution layer. Finally, a shortcut is used to generate the output, which will pass to the next stage.

3.5. Positional Normalization and Moment Shortcut. Although normalizing inputs is considered to be one of the tricks for training the network, several normalization methods have been proposed to improve the performance, such as batch normalization. Different from the prior normalization scheme, the positional normalization (PONO) is position-dependent and reveals structural information at this particular layer of the deep net. It normalizes exclusively over the channels at all spatial positions, so it is translation, scaling, and rotation invariant. The PONO computes the mean μ and standard deviation σ in the layer:

$$\mu_{b,h,w} = \frac{1}{C}\sum_{c=1}^{C} X_{b,c,h,w},$$
$$\sigma_{b,h,w} = \sqrt{\frac{1}{C}\sum_{c=1}^{C}\left(X_{b,c,h,w} - \mu_{b,h,w}\right)^2 + \varepsilon}, \tag{8}$$

where ε is the small stability constant.

Moment Shortcut (MS) fast-forward the PONO information μ and σ as shown in Figure 7.

The two moments of the activations (μ, σ) are extracted from the early layer and are sent to the corresponding layer later as

$$\text{MS}(x) = \gamma F(x) + \beta, \tag{9}$$

where F denotes the intermediate layers, and γ and β are predicted from μ and σ via a shallow convolution layer.

3.6. Loss Function. The perceptual loss, mean squared error (MSE), GAN loss, and L_2 loss is widely used in many dehazing networks. The research in [35] points that the smooth L_1 loss provides better PSNR and SSIM metrics in many image restoration tasks. So we use the smooth L_1 loss to train the network:

$$L_s = \frac{1}{N}\sum_{x=1}^{N}\sum_{i=1}^{3} F_s\left(\hat{J}_i(x) - J_i(x)\right), \tag{10}$$

where

$$F_s(e) = \begin{cases} 0.5e^2, & \text{if } |e| < 1 \\ |e| - 0.5, & \text{otherwise.} \end{cases} \tag{11}$$

$\hat{J}_i(x)$ stands for the intensity of the ith color channel of pixel x in the dehazed image, and N is the total count of pixels of the image.

At each stage, there is a ground-truth to predict, so we add the losses from each stage to optimize the net:

$$L = L_{s1} + L_{s2}. \tag{12}$$

4. Experiment Results

4.1. Dataset. We evaluate the proposed network on three benchmarks including RESIDE [36], Dense-Haze [37], and real-world dataset [38]. RESIDE contains both indoor and outdoor synthetic hazy images, which are collected from depth datasets [39] and stereo datasets [40]. After data

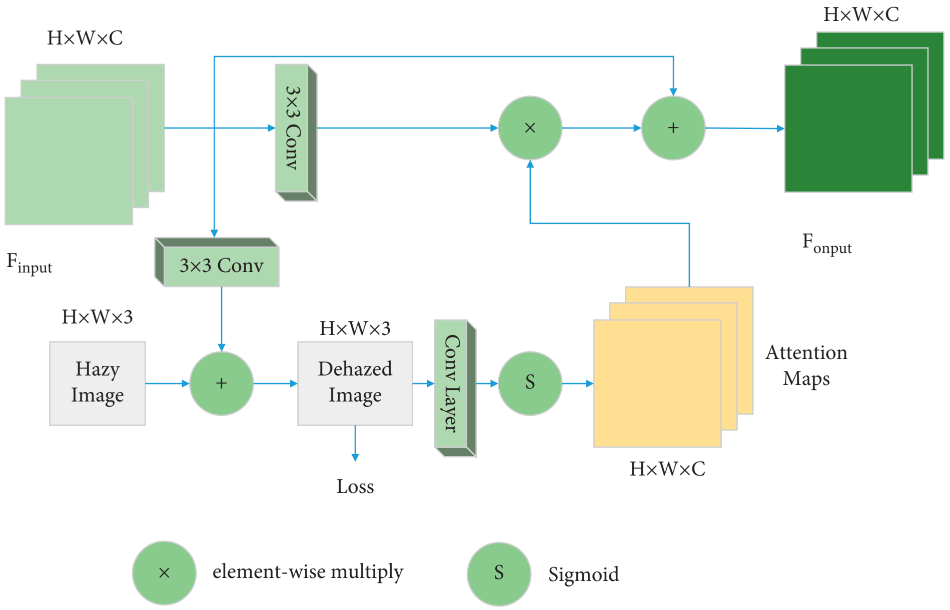

FIGURE 6: Supervised attention block.

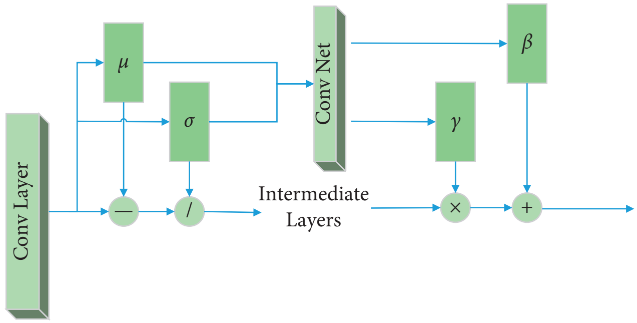

FIGURE 7: PONO-MS uses the extracted mean and standard deviation.

cleaning, the Indoor Training Set (ITS) contains 1399 clear images and 13,990 hazy images, and the hazy ones are generated by the clear images with global atmosphere light $A \in [0.7, 1.0]$ and scatter parameters $t \in [0.6, 1.8]$. The Outdoor Training Set (OTS) contains 2061 clear images and 72,135 hazy images generated by the clear images with $A \in [0.8, 1.0]$ and $t \in [0.04, 0.2]$. The Synthetic Objective Testing Set (SOTS) of RESIDE is used for testing, and the SOTS contains 500 indoor images and 500 outdoor images. The images of Dense-Haze and the real-world dataset [38] are collected from the real world.

4.2. Training Settings. We resize the size of training images to 240 × 240 with 3 channels, randomly rotate the images by 90,180,270°, and horizontal flip the images for data augmentation. We choose the Adam optimizer for accelerated training, where β_1 and β_2 take the default values of 0.9 and 0.999, respectively. In the encoder-decoder subnet, each UBlock contains 3 MAUs. In the single-scale subnet of the second stage contains 3 MAGs, each of which consists of 8 MAUs. The channel number of preprocessing convolution layer is set to 64, and the channel number of input and output in MAU are both 64. We adopt the cosine annealing strategy [41] to adjust the learning rate η_t from the initial value $\eta = 1 \times 10^{-4}$ to 0 by following the cosine function:

TABLE 1: Quantitative comparisons on SOTS and Dense-Haze.

Method	Indoor		Outdoor		Dense-haze	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DCP	16.62	0.8179	19.13	0.8148	10.06	0.3856
AOD-Net	19.06	0.8504	20.29	0.8765	13.14	0.4144
GCANet	30.23	0.9800	—	—	13.21	0.4253
GFN	22.30	0.8800	21.55	0.8444	13.96	0.4274
RefineDNet	24.23	0.9431	—	—	14.14	0.4386
DehazeNet	21.14	0.8472	22.46	0.8514	13.84	0.4252
GridDehazeNet	32.16	0.9836	30.86	0.9819	13.31	0.3681
FFA-Net	36.39	0.9886	33.57	0.9849	14.39	0.4524
Ours	37.01	0.9912	34.19	0.9897	15.06	0.4636

$$\eta_t = \frac{1}{2}\left(1 + \cos\left(\frac{t\pi}{T}\right)\right)\eta, \tag{13}$$

where T is the total number of training batches and t is the current training batch. The batch size is set to 4, and we evaluate the model every 5000 steps, the total steps are set to 1,000,000.

4.3. Results and Analysis. In this section, we compare MSNET with recent state-of-the-art image dehazing algorithms which are DCP, AOD-Net, DehazeNet, GCANet, RefineDNet, DehazeNet, GridDehazeNet, and FFA-Net quantitatively and qualitatively. Following those methods, we use peak signal to noise ratio (PSNR) and structure similarity (SSIM) for quantitative assessment of the dehazed outputs, and the outputs higher is better. And the

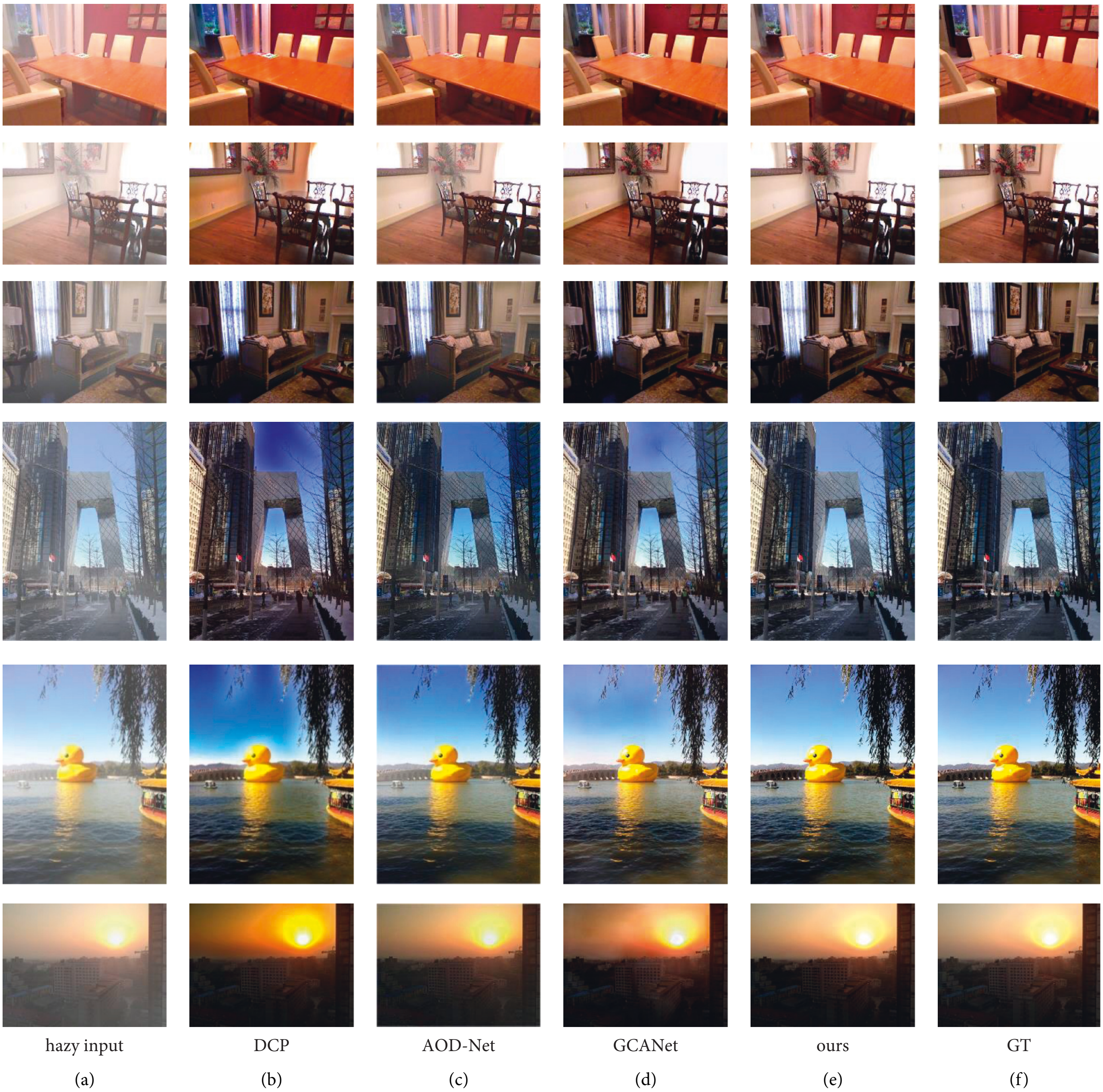

Figure 8: Qualitative comparisons on SOTS (a) hazy input; (b) DCP; (c) AOD-Net; (d) GCANet; (e) ours; (f) GT.

quantitative comparison results on SOTS and Dense-Haze are shown in Table 1. Among those methods, DCP is a prior-based, and it is often regarded as the baseline, and the others are deep learning based.

From Table 1, we can observe that the results of data-based method are better than the result of DCP, which is a prior-based method. Among the data-based methods, AOD-Net is simplest network, so the result value is low, but still much higher than DCP. The RefineDNet is a weakly supervised method, so its results are not good as the other supervised methods. Compared to FFA-Net, our results increased by about 1.8% on SOTS and about 4.6% on Dense-Haze because of the multistage and multiattention mechanisms used.

The qualitative comparisons of visual effect on SOTS are shown in Figure 8. We select three images from the outdoor dataset and the indoor dataset, respectively, and the upper three rows are indoor results; the left three rows are outdoor results. The first column is the hazy input, the last column is the ground-truth, and the middle columns are the dehazed results from DCP, AOD-Net, GCANet, and MSNet (ours), respectively. From the results, we can see that the DCP method suffers from severe color distortion extremely, especially in the blue sky and the halo of the sun in the last image, and it loses some details. AOD-Net cannot remove all the hazy regions from the hazy image because of its simple network architecture. In the second row of Figure 8, the fog near the tree is still there. And in the fifth row of figure 8,

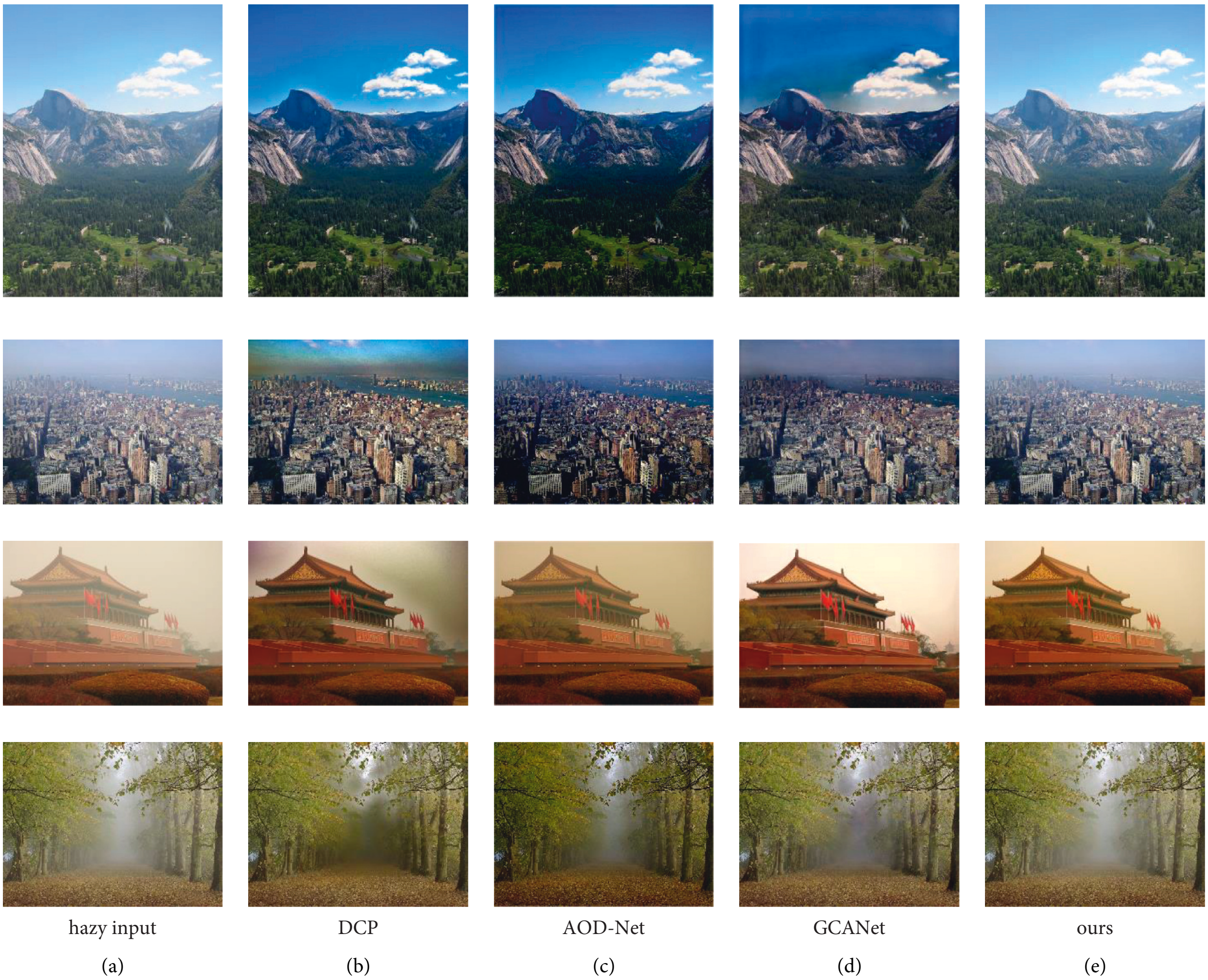

Figure 9: Qualitative comparisons on the real-world dataset (a) hazy input; (b) DCP; (c) AOD-Net; (d) GCANet; (e) ours.

Table 2: Ablation studies on SOTS.

ID	#Stages	Stage combination	PONO	SAB	PSNR
1	1	U-Net	—	—	31.82
	1	SC-Net	—	—	31.79
2	2	U-Net + U-Net	—	—	32.53
	2	SC-Net + SC-Net	—	—	32.57
	2	U-Net + SC-Net	—	—	32.69
3	2	U-Net + SC-Net	✗	✓	32.79
	2	U-Net + SC-Net	✓	✗	32.72
	2	U-Net + SC-Net	✓	✓	32.86

there still has fog near the bridge. And its brightness value is lower. GCANet also performs not well on the blue sky and sun halo in the last two rows of figure 8 especially. The images recovered from our network are almost entirely in line with real-scene information, especially, the restoration of blue sky and halo images is much better.

We further give the qualitative comparisons on the real-world dataset [38] in Figure 9, and the results are similar with those on the SOTS dataset. The DCP and GCANet still suffer from severe color distortions, such as the blue sky of rows 1 and 2, and AOD-Net cannot remove the haze completely, so the output images are of low brightness such as a result of row 3. Also, DCP cannot remove the haze completely, such as in the sky of row 3. Although none of the methods can completely remove the hazy regions such as the last row of Figure 9, other methods suffer from color distortions compare to ours. And the restoration of our method is more natural. Above all, our method is capable of performing in image details and color fidelity than other methods in general.

4.4. Ablation Studies. In this section, we present ablation experiments to discuss the different modules of our network. The factors below are mainly concerned: (1) the number of stages; (2) the choices of each stage; and (3) the SAB and PONO. Evaluation is performed on SOTS indoor dataset, and the images as training input are cropped to 48 × 48. The results are shown in Table 2. First, we compare the results of one-stage and two-stage without PONO and SAM. The results of the two-stage (ID 2) increased by an average of 2.3% than the results of the one-stage (ID 1). The results indicate the effectiveness of the two-stage network. Second, we prove the need of two different architectures in the two-stage network by the results of ID 2. While using two different architectures, the PSNR is increased. Finally, from the comparison of ID 3, "✗" in the table means the module is not used, and "✓" means the module is used. We can observe that while the PONO and SAM are used, the result is better.

5. Conclusion

In this work, we propose a multistage with multiattention network for image dehazing. The model consists of two different stages: one uses an encoder-decoder subnet to obtain contextual features, and the other adopts a single-scale pipeline to provide spatial image details. At each stage, a ground-truth supervision is provided, an attention mechanism is used between the two stages, and a multi-attention unit with positional normalization is proposed to stack the network. The results in several benchmarks show that the proposed network outperforms the state-of-the-art methods and have a great advantage over those methods in terms of image detail and color fidelity.

Data Availability

The data used in this paper are all from public data sets, including RESIDE, Dense-Haze, and real-world dataset. which can be found in each reference in the paper.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Jiangsu Industry Research Project BY2020552, and Nantong Science and Technology Program Project JC2020065.

References

[1] E. J. McCartney and F. F. Hall, "Optics of the atmosphere: scattering by molecules and particles," *Physics Today*, vol. 30, no. 5, pp. 76-77, 1977.

[2] S. Narasimhan and S. Nayar, "Chromatic framework for vision in bad weather," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2000)*, pp. 598–605, Hilton Head, SC, USA, June 2000.

[3] W. Ren, J. Pan, H. Zhang, X. Cao, and M.-H. Yang, "Single image dehazing via multiscale convolutional neural networks with holistic edges," *International Journal of Computer Vision*, vol. 128, no. 1, pp. 240–259, 2020.

[4] H. Ullah, K. Muhammad, M. Irfan et al., "Light-DehazeNet: a novel lightweight CNN architecture for single image dehazing," *IEEE Transactions on Image Processing*, vol. 30, pp. 8968–8982, 2021.

[5] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: an end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, 2016.

[6] G. Kim, S. Ha, and J. Kwon, "Adaptive patch based convolutional neural network for robust dehazing," in *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP 2018)*, pp. 2845–2849, Athens, Greece, October 2018.

[7] H. Dong, J. Pan, L. Xiang et al., "Multiscale boosted dehazing network with dense feature fusion," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2157–2167, Seattle, WA, USA, Jun. 2020.

[8] X. Liu, Y. Ma, Z. Shi, and J. Chen, "GridDehazeNet: attention-based multiscale network for image dehazing," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, pp. 7314–7323, Seoul, Korea (South), September 2019.

[9] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-net: feature fusion attention network for single image dehazing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 11908–11915, New York, NY, USA, February 2020.

[10] H. H. Yang and Y. Fu, "Wavelet u-net and the chromatic adaptation transform for single image dehazing," in *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2736–2740, IEEE, Taipei, Taiwan, September 2019.

[11] Q. Zhou, Y. Wang, Y. Fan et al., "AGLNet: towards real-time semantic segmentation of self-driving images via attention-guided lightweight network," *Applied Soft Computing*, vol. 96, Article ID 106682, 2020.

[12] Y. Dong, Y. Liu, H. Zhang, Y. Qiao, and Y. au, "FD-GAN: generative adversarial networks with fusion-discriminator for single image dehazing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10729–10736, New York, NY, USA, February 2020.

[13] Y. Z. Su, Z. G. Cui, C. He, A. H. Li, T. Wang, and K. Cheng, "Prior guided conditional generative adversarial network for single image dehazing," *Neurocomputing*, vol. 423, pp. 620–638, 2021.

[14] M. Suin, K. Purohit, and A. Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, pp. 3603–3612, Seattle, WA, USA, June 2020.

[15] H. Zhang, Y. Dai, H. Li, and P. Koniusz, "Deep stacked hierarchical multipatch network for image deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pp. 5971–5979, Long Beach, CA, USA, June 2019.

[16] D. Ren, W. Zuo, Q. Hu, P. Zhu, and D. Meng, "Progressive image deraining networks: a better and simpler baseline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pp. 3932–3941, Long Beach, CA, USA, June 2019.

[17] B. Li, F. Wu, K. Q. Weinberger, and S. J. Belongie, "Positional normalization," in *Proceedings of the Advances in Neural*

Information Processing Systems (NIPS 2018), pp. 1620–1632, Cambridge, MA, USA, December 2018.

[18] T. Treibitz and Y. Schechner, "Polarization: beneficial for visibility enhancement?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pp. 525–532, Miami, FL, USA, May 2009.

[19] S. G. Narasimhan and S. K. Nayar, "Contrast restoration of weather degraded images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 713–724, 2003.

[20] S. Shwartz, E. Namer, and Y. Schechner, "Blind haze separation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, pp. 1984–1991, New York, NY, USA, June 2006.

[21] K. Kaiming He, J. Xiaoou Tang, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2011.

[22] K. B. Gibson, D. T. Vo, and T. Q. Nguyen, "An investigation of dehazing effects on image and video coding," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 662–673, 2012.

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

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

[25] D. Berman, T. Treibitz, and A. Shai, "Non-local image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pp. 1674–1682, Las Vegas, NV, USA, June 2016.

[26] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: all-in-one dehazing network," in *Proceedings of the IEEE International Conference on Computer Vision (CVPR 2016)*, pp. 4770–4778, Las Vegas, NV, USA, June 2016.

[27] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in *Proceed-ings of the IEEE Conference on Computer Vision and Pat-tern Recognition*, pp. 8160–8168, Long Beach, CA, USA, June 2019.

[28] S. Zhao, L. Zhang, Y. Shen, and Y. Zhou, "RefineDNet: a weakly supervised refinement framework for single image dehazing," *IEEE Transactions on Image Processing*, vol. 30, pp. 3391–3404, 2021.

[29] H. Li, J. Li, D. Zhao, and L. Xu, "DehazeFlow: multiscale conditional flow network for single image dehazing," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2577–2585, Chengdu, China, May 2021.

[30] W. Ren, L. Ma, J. Zhang et al., "Gated fusion network for single image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pp. 3253–3261, Salt Lake City, UT, USA, June 2018.

[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[32] Q. Zhou, J. Wang, J. Liu, S. Li, W. Ou, and X. Jin, "RSANet: towards real-time object detection with residual semantic-guided attention feature pyramid network," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 77–87, 2021.

[33] S. W. Zamir, A. Arora, S. Khan et al., "Multistage progressive image restoration," in *Proceedings of the IEEE International Conference on Computer Vision (CVPR 2021)*, pp. 14821–14831, Nashville, TN, USA, June 2021.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U- Net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pp. 234–241, Munich, Germany, October 2015.

[35] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2017)*, pp. 136–144, Honolulu, HI, USA, July 2017.

[36] B. Li, W. Ren, D. Fu et al., "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2019.

[37] C. O. Ancuti, C. Ancuti, M. Sbert, and T. Radu, "Dense haze: a benchmark for image dehazing with dense-haze and haze-free images," in *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, September 2019.

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

[39] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proceedings of the European Conference on Computer Vision (ECCV 2012)*, pp. 746–760, Florence, Italy, October 2012.

[40] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, Madison, WI, USA, June 2003.

[41] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pp. 558–567, Long Beach, CA, USA, June 2019.