Received November 11, 2019, accepted November 27, 2019, date of publication December 2, 2019,
date of current version December 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2957057

AAGAN: Enhanced Single Image Dehazing With Attention-to-Attention Generative Adversarial Network

WENHUI WANG[1], ANNA WANG[1], QING AI[2], CHEN LIU[1], AND JINGLU LIU[1]
[1]College of Information Science and Engineering, Northeastern University, Shenyang 110819, China
[2]School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China

Corresponding author: Qing Ai (lyaiqing@126.com)

This work was supported in part by the Talent Cultivation Project of University of Science and Technology Liaoning of China under Grant 2018RC05, and in part by the National Natural Science Foundation of China (NSFC) under Grant 61973066.

ABSTRACT Due to the atmospheric scattering and absorption, hazy weather often occurs in our everyday life, thus reducing the visibility of scenes. Single image dehazing is considered as an ill-posed and challenging problem in computer vision. To restore visibility in inclement weather, we propose an attention-to-attention generative adversarial network (AAGAN) whose motivation is the human visual perceptual mechanism. More specifically, a dense channel attention model is embedded into the encoder. Moreover, its output is projected forward to the corresponding multiscale spatial attention model in the decoder. Both attention models form an attention-to-attention mechanism to implement attention projection, thus capturing global feature dependencies of the whole network. Besides, we analyze the dehazing mechanism based on the atmospheric scattering model, and then utilize an improved RaLSGAN to recover more realistic texture information and enhance visual contrast for different hazy scenes. Finally, in order to improve the visual performance of image restoration, we remove all the instance normalization layers to avoid unnecessary artifacts, and then introduce spectral normalization for all the convolution layers to stabilize the entire training process. Qualitative assessments and analyses demonstrate that our proposed approach can achieve remarkable dehazing performance on both synthetic and real-world scenes against previous state-of-the-art methods.

INDEX TERMS Image dehazing, GAN, attention projection, image restoration.

I. INTRODUCTION

In our daily life, inclement weather that people often encounter mainly comprises snow, rain, haze and mist. Haze and mist, as the most frequent weather, are caused by suspended particles and water drops. They can lead to some refracted atmospheric light to be absorbed and scattered, which makes the visibility of scenes obscure and dim. Since image dehazing can recover visibility, texture information and brightness of hazy scenes, it plays an important role in image restoration of computer vision. In particular, haze removal can be used as image preprocessing of many advanced vision tasks to improve their visual performance, such as object detection [1], face recognition [2], person re-identification [3] and semantic segmentation [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Hengyong Yu.

From the perspective of technology development, image dehazing can be generally classified into two categories: multi-image dehazing method and single-image dehazing one.

As to multi-image dehazing approaches, earlier researchers apply two or more images and even some additional auxiliary tools to remove haze due to lack of sufficient theories and efficient computing equipment. For one thing, the polarization-based dehazing methods [5]–[7] utilize two images and a polarizer device to remove haze for the same scene. For another, without any equipment, the multi-image methods [8]–[10] with simple constraints apply two or more photos to eliminate weather effects, and Deep Photo approach [11] employs either multiple images or 3D models registered to restore the scene. Even though these approaches achieve better performance on dehazing task, they are too complicated to implement on the hardware devices.

As to single-image dehazing approaches, people have further presented extensive methods to overcome deficiencies of multi-image dehazing ones, *i.e.* advanced physical models, powerful constraints and effective priors/hypotheses. Based on the locally uncorrelated surface shading, the refined image formation model in [12] estimates the transmission to remove haze. Scene depth modeling [13] evaluates the depth of scene points to recover image contrast. However, due to the conditional constraints of these physical models, they are difficult to implement in the real world. Tan [14] exploits Markov Random Fields (MRF) to enhance visual visibility with two basic constraints, but its results appear color oversaturation to some extent. To further overcome these drawbacks, the dark channel prior (DCP) in [15] separately estimates the transmission and the atmospheric light to remove haze in the regions without the sky. Although DCP has remarkable dehazing performance, it is difficult to adapt to sky areas or objects similar to the airlight. Except that, it is also time-consuming due to the employment of soft matting method [16]. Based on DCP, Meng *et al*. [17] proposes a regularization method based on the boundary constraint to accurately estimate the transmission for this issue. Besides, Liu *et al*. [18] segment single hazy image into sky and non-sky areas, and then apply multiscale opening dark channel model to recover the scene. Aside from this prior, researchers propose some more effective priors to remove haze, such as semi-inverse approach [19], color attenuation prior [20] and non-local prior [21].

In recent years, with the development of theoretical technology, classical machine methods and deep learning ones have been introduced to the image dehazing task. Firstly, in [22]–[24], a variety of approaches based on Markov random field (MRF) utilize depth information of the scene to achieve high-quality visual effects. Secondly, a unified variational model [25] is proposed to apply the total variation regularization, thus jointly estimating transmission map and recovering scene radiance. Thirdly, after systematically investigating numerous and various haze-relevant features, Tang *et al*. [26] employ the random forest (RF) framework to remove haze. Finally, in [27], Chen *et al*. propose a method based on the radial basis function (RBF) to recover scene radiance.

Although the above state-of-the-art methods make remarkable progress on the image dehazing task, some constraints and hypotheses that they depend on are very difficult to come true in the real world to some extent. Besides, some haze-relevant features and parameters are artificially defined, which leads to the absence of generalization capability and self-adaptive capability. As a result, deep learning method, as a self-learning and data-driven algorithm, has been widely introduced to remove haze. DehazeNet [28], MSCNN [29] and DPATN [30] estimate the medium transmission accurately with an end-to-end deep CNN; however, they still have shortages in the estimation of the atmospheric light to some extent. To solve this issue, AOD-Net [31] unifies the above two estimated parameters into one with a reformulated atmospheric scattering model. Besides, Zhang and Patel [32] apply two networks to predict the transmission map and the atmospheric light one, respectively. Even though this method removes haze efficiently, it tends to overestimate the atmospheric light to distort the color of the blue sky. To this end, AIPNet [33] offers an atmospheric illumination prior (AIP) to strengthen visual contrast with multiscale CNN. Furthermore, GFN [34] combines a hazy image with three derived images to restore a clean image via a gated fusion network. Similarly, in [35], Li *et al*. modify a conditional generative adversarial network (cGAN) to generate clean images using the VGG features and a gradient prior. Based on the classical atmospheric scattering model, the above image-to-image frameworks have better performance on haze removal; however, there are still some drawbacks for single hazy image restoration. In particular, the key to image dehazing is how to identify hazy areas precisely while ignoring other irrelevant ones, thus recovering scene radiance effectively.

As illustrated in Fig. 1, we present the system overview of our method and the remarkable dehazing effects. To remove haze effectively while recovering texture information and visual contrast, motivated by the human visual perceptual mechanism, we design an attention-to-attention generative adversarial network for dehazing task. In this paper, the main contributions of our AAGAN architecture are summarized as follows:

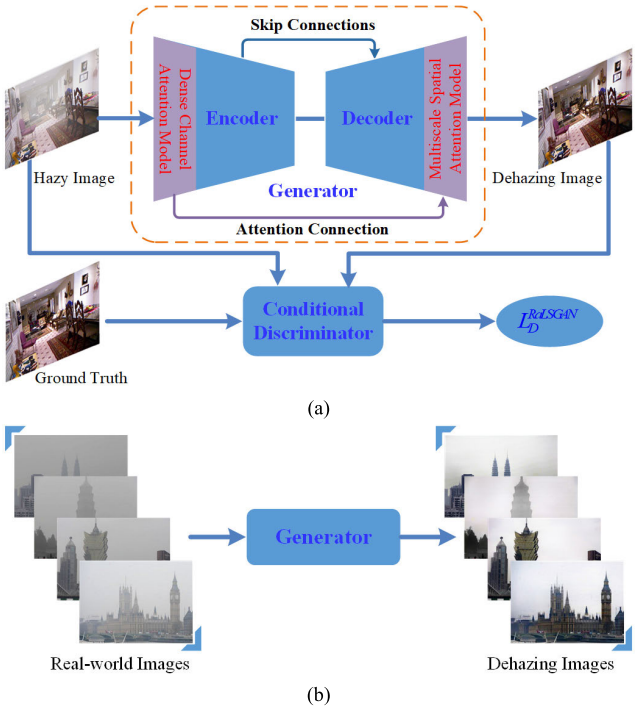

FIGURE 1. The system overview of the proposed AAGAN. (a) AAGAN model. (b) The real-world system function diagram.

(1) We propose an attention-to-attention generative network (AAGAN) and extend attention mechanism to the dehazing task, thus directly mapping hazy images to haze-free ones.

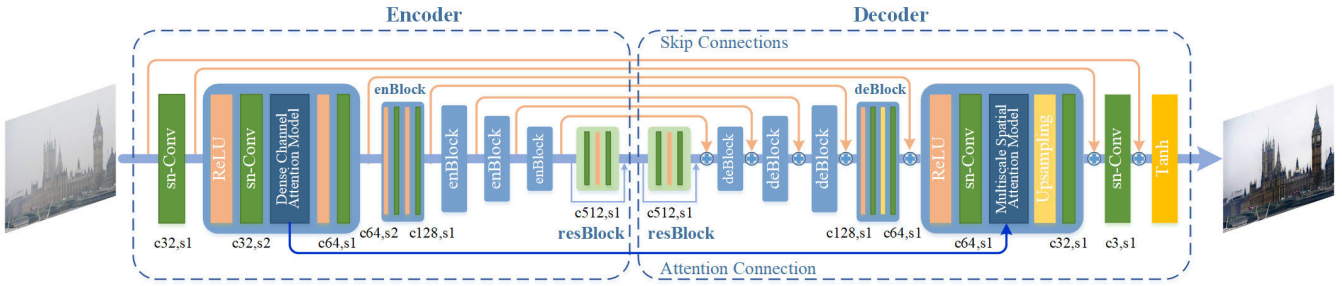

FIGURE 2. The encoder-decoder network architecture of AAGAN with the corresponding channel number of feature maps (c) and stride (s) signified for each convolutional layer. Skip connections implement elementwise sum among feature maps, and an attention connection connects the dense channel attention model and the multiscale spatial attention model. To avoid unpleasant artifacts, we removes all the instance normalization layers in AAGAN, and employs spectral normalization (sn) for all the convolutional layers.

(2) To focus accurately on hazy areas in the channel dimension and the spatial one, we design a dense channel attention model in the encoder and a multiscale spatial attention one in the decoder, respectively.

(3) Inspired by interactive projection connections in the human visual system, we present attention projection to capture long-range dependencies of the whole network, thus efficiently restoring deficient texture information.

(4) To generate high-quality images, we remove all the instance normalization layers and utilize spectral normalization for all the convolution layers. Meanwhile, an improved RaLSGAN is introduced to restore realistic scenes according to our proposed dehazing mechanism.

In this paper, the remaining sections are organized as follows. In Section II, the knowledge of atmospheric scattering model and attention mechanism is reviewed. In Section III, details of the proposed AAGAN, attention models and spectral normalization for RaLSGAN are presented, respectively. Subsequently, qualitative and quantitative evaluations and comparisons are analyzed in Section IV. Finally, our conclusion and discussion are demonstrated in Section V.

II. RELATED WORKS

In this section, some important relevant knowledge and literature are reviewed as follows.

A. ATMOSPHERIC SCATTERING MODEL

To effectively and feasibly describe the formation process of the atmospheric scattering model, the model can be further presented in [15], [37] as follows:

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

where $\boldsymbol{I}$, $\boldsymbol{J}$ and t are the hazy image, the corresponding scene radiance and the medium transmission, respectively; $\boldsymbol{A}$ is the global atmospheric light; x is the pixel location. More specifically, $t(x) = e^{-\beta d(x)}$ is the transmission map on the assumption that the atmosphere is homogenous, where d is the scene depth and β is the scattering coefficient of the atmosphere. Consequently, $\boldsymbol{J}(x)t(x)$ is expressed as direct attenuation [14] that the scene radiance is decayed by the medium; $\boldsymbol{A}(1-t(x))$ is defined as the airlight [14] that the scattered light causes the color shift for the scene.

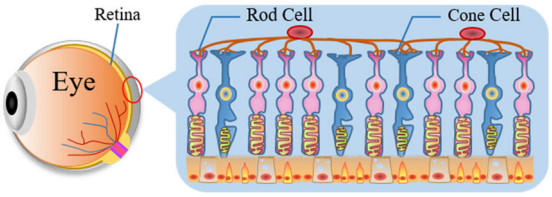

FIGURE 3. Physiological structure of the eye and the retina.

B. ATTENTION MECHANISM

Attention mechanism has such excellent ability to capture global dependencies in some semantic contexts that it is widely employed in diverse tasks [38]–[40]. At first, in order to gain long-range dependencies of the input sequence, the attention model [38] plays a significant role in machine translation. Secondly, influenced by this, attention model is gradually introduced into computer vision, such as image super-resolution [40], image classification [41], image captioning [42] and image segmentation [43]. Thirdly, in [44], self-attention model is regarded as a non-local operation to establish spatial-temporal dependencies among video frames. Finally, Zhang *et al.* [39] introduce self-attention mechanism to image generation with GAN framework to achieve a verisimilar visual effect. However, self-attention models in [39] and [44] cost excessive GPU memory. In particular, it is very challenging for GPU memory to produce the attention map with shape $HW \times HW$, where H and W denote height and width of feature maps, respectively.

As we know, when the human visual system processes visual information from the outside world, it does not treat all the information equally but shows some specificities. As shown in Fig. 3, we take retinal cells as an example that photoreceptor cells consist of rod cells and cone ones. Although they are located at the same cell layer, they have completely different functions. They have their own strengths and weaknesses in perceiving the outside world. However, they can compensate each other through neural connections,

which can realize an efficient and reasonable resource distribution for retinal imaging. Inspired by this, we extend the attention mechanism to image dehazing, and elaborately design two attention models similar to the layout of rod cells and cone ones, *i.e*. dense channel attention model and multiscale spatial attention one. To be specific, they are embedded at the same level of the encoder and the decoder respectively, which can deal with different feature information. Furthermore, for example, in [41] and [42], the channel attention and the spatial one only pay more attention to low-level feature information of feature maps. Compared with these previous works, our attention models not only process contextual feature information in low-level layers, but also fuse attention-aware information in high-level ones with an attention connection. Finally, both dense channel attention model and multiscale spatial attention one can capture long-range dependencies of the entire network with attention projection. Extensive experiments and analyses demonstrate the effectiveness of our proposed method.

III. PROPOSED METHODS

In this section, the encoder-decoder network architecture of AAGAN is presented, and then two kinds of attention models are demonstrated in detail, *i.e*. dense channel attention model (DCAM) and multi-scale spatial attention model (MSAM). Furthermore, two attention models implement attention projection with an attention connection to capture the long-range dependencies of the entire network. At last, we represent how to stabilize and optimize the training strategies of AAGAN to restore realistic scenes.

A. NETWORK ARCHITECTURE

For single image dehazing task, significant details in haze regions are paid more attention. In other words, our method mainly focuses on valuable haze-feature information while ignoring irrelevant information in hazy areas. Moreover, the previous methods [28]–[35] mostly employ full convolutional network that only draws contextual information within a local neighborhood, which is inefficient for capturing the long-range dependencies in hazy areas. To this end, as shown in Fig. 2, the framework of AAGAN is proposed to put conscious attention to image restoration in hazy areas with the attention mechanism. Besides, AAGAN can adaptively draw comprehensive contextual information, thus recovering hazy areas with global haze-feature information.

AAGAN, as a residual encoder-decoder network, comprises three kinds of function blocks, such as encoding block (enBlock), residual block (resBlock) and decoding block (deBlock). More importantly, motivated by the excellent performance of EDSR [45] and ESRGAN [46], AAGAN removes all the instance normalization (IN) [47] layers to avoid undesirable artifacts, thus enhancing visual performance. However, if IN layers are directly removed, extensive experiments demonstrate that the whole network easily appears exploding gradients due to its very deep network structure. The main reason could be that, compared with ESRGAN [46], the size of feature maps in AAGAN always vary, thus causing the constant changes of feature distribution. To resolve this intractable issue, we introduce spectral normalization [48] for all the convolutional layers in both the generator and the discriminator to stabilize the entire training network.

For AAGAN, the major working process of haze removal is described as follows. To begin with, five encoding blocks are followed by the first convolutional layer as an initial feature extractor, which constantly encodes high-frequency information of the input hazy image. Meanwhile, low-frequency information produced by each encoding block is delivered to the corresponding decoding block through a skip connection. Subsequently, the encoding relevant-haze features are further refined by two residual blocks. After that, these feature information is continually decoded to reconstruct scene radiance via five decoding blocks in the decoder. Finally, after the output of the last convolutional layer and the original hazy image are fused by elementwise sum, a clear scene will be recovered through a final Tanh activation layer.

In Fig. 2, the architecture of each function block is presented in detail as follows. Note that, in the proposed AAGAN, all the convolutional layers employ kernels of the same size 3×3 except those of attention models. Firstly, the encoding block incorporates two groups of Rectified Linear Unit (ReLU) [49] and convolutional layer. The first group down-samples the feature maps with stride 2, and then the second doubles the number of output channels, thus efficiently saving GPU memory. Secondly, the decoding block also consists of two parts. The former part utilizes a convolutional layer to reduce the number of input channels, and then the latter employs an upsampling layer followed by a convolutional one to double the size of output feature maps. Thirdly, two convolutional layers are combined with a ReLU layer to form the residual block [50] with a residual connection. Finally, DCAM and MSAM can be separately embedded into any corresponding function block in the encoder and the decoder, because they do not change the size of feature maps. Considering the overall performance of AAGAN, DCAM and MSAM in Section III-B are embedded into the first encoding block and the final decoding block, respectively. More importantly, two attention models employ attention projection to capture global contextual information of the entire network.

B. ATTENTION MODELS

1) DENSE CHANNEL ATTENTION MODEL

To capture long-range dependencies in the channel dimension, we introduce dense CNNs unit with dense connectivity [51], which constantly iterates extensive channel-based feature information. As illustrated in Fig. 4(a), DCAM establishes the interdependencies among channel-based feature maps and pays more attention to feature information in hazy areas. We feed the initial feature map $X_d \in \mathbb{R}^{C\times H\times W}$ into DCAM, and then obtain a new feature map $X_0 \in \mathbb{R}^{C\times H\times W}$ through a convolutional layer followed by ReLU layer.

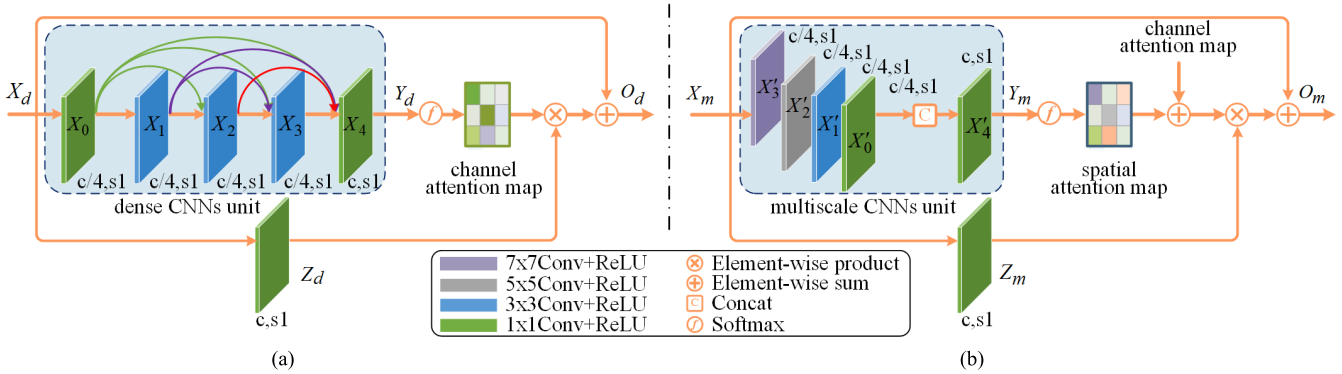

FIGURE 4. Attention models of the proposed AAGAN with corresponding channel number of feature maps (c) and stride (s) signified for each convolutional layer. (a) Dense channel attention model. (b)Multiscale spatial attention model.

Subsequently, the final feature map $Y_d \in \mathbb{R}^{C\times H\times W}$ is generated by 4 convolutional operations, which can be formulated as

$$\begin{cases} X_0 = F^{(1)}(X_d) \\ X_i = F^{(3)}([X_0, X_1, \cdots, X_{i-1}]) \qquad i \in [1, K] \\ Y_d = F^{(1)}(X_K), \end{cases} \tag{2}$$

where X_i represents the i^{th} haze-feature maps generated from all the preceding layers. K is the total number of feature maps in dense CNNs unit, *i.e.* $K = 4$. $[X_0, X_1, \cdots, X_{i-1}]$ denotes the concatenation of feature maps generated by each convolution layer. $F^{(k)}(\cdot)$ can be viewed as a composite function of operations, *i.e.* convolution and ReLU, and k indicates the filter size ($k \times k$). To reduce the computational burden, the number of the first four output feature maps is reduced to 1/4 (indicated as the growth rate in [51]). Finally, the number of the output feature maps is restored to the initial number of the input X_d by the final 1×1 convolution.

Afterwards, a softmax layer is exploited to calculate the channel attention map $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_i, \cdots, \alpha_C) \in \mathbb{R}^{C\times H\times W}$

$$\alpha_i = \frac{exp(Y_{d_j})}{\sum_{j=1}^{H\times W} exp(Y_{d_j})} \tag{3}$$

where α_i denotes the channel attention weights at the j^{th} position in each feature map.

Meanwhile, similar to the generation of X_0, we obtain another feature map $Z_d \in \mathbb{R}^{C\times H\times W}$ on the bypass. At last, we implement an elementwise multiplication between α and Z_d, and then their output result is multiplied by a scale parameter ε and add the original feature map $X_d \in \mathbb{R}^{C\times H\times W}$. Finally, the final output $O_d \in \mathbb{R}^{C\times H\times W}$ is expressed as

$$O_d = \varepsilon(\alpha \cdot Z_d) + X_d \tag{4}$$

where ε is initialized as 0, and then it is adaptively adjusted with training times. It can be seen that the output O_d is regarded as a weighted sum of extensive channel-based features and original input ones. As a result, massive feature mappings with dense connectivity can take global haze-feature information in the channel dimension, thus recognizing hazy areas precisely.

2) MULTISCALE SPATIAL ATTENTION MODELS

To further obtain diverse contextual information in the spatial dimension, we extend the multiscale CNNs unit [52] to multiscale spatial attention. More importantly, multiscale CNNs unit adapts to stimulate multiscale receptive fields of the human visual system. Consequently, in Fig. 4(b), MSAM can extract adaptively a large variety of spatial feature information and establish spatial feature correlation in hazy regions. Note that we only discuss the working mechanism of MSAM without any attention connection, while MSAM with the attention connection is presented in Section III-B.3 in detail. For MSAM, the first step is that the input feature map $X_m \in \mathbb{R}^{C\times H\times W}$ is fed into four parallel convolutional layers with diverse filter sizes, such as 1×1, 3×3, 5×5, 7×7, respectively. Secondly, their output channels are reduced to 1/4, and then their output feature maps are concatenated together. These various spatial feature information is fused further by a 1×1 convolution, which can be presented as follows

$$\begin{cases} X'_l = F^{(k)}(X_m), k \in \{1, 3, 5, 7\} \\ X'_L = F^{(1)}([X'_0, X'_1, \cdots, X'_l, \cdots, X'_{L-1}]) \qquad l \in [0, L-1] \\ Y_m = F^{(1)}(X'_L), \end{cases} \tag{5}$$

where, similar to DCAM, l indicates l^{th} single-scale feature maps, L denotes the total number of feature maps in multiscale CNNs unit, *i.e.* $L = 4$.

Thirdly, we employ a softmax layer to achieve the spatial attention map $\beta = (\beta_1, \beta_2, \cdots, \beta_i, \cdots, \beta_C) \in \mathbb{R}^{C\times H\times W}$

$$\beta_i = \frac{exp(Y_{m_j})}{\sum_{j=1}^{H\times W} exp(Y_{m_j})} \tag{6}$$

where β_i denotes the spatial attention weights at the j^{th} position in each feature map.

Ultimately, a new feature map $Z_m \in \mathbb{R}^{C\times H\times W}$ is generated by a 1×1 convolution on another branch, which is multiplied

by β in an elementwise way. Subsequently, we multiply their result by a learnable parameter ζ whose initial value is 0, and then execute an elementwise sum operation with $X_m \in \mathbb{R}^{C\times H\times W}$ to acquire the final output

$$O_m = \zeta(\beta \cdot Z_m) + X_m \tag{7}$$

As analyzed above, a large amount of different spatial feature information can capture long-range contextual information in the multiscale spatial dimension.

3) ATTENTION PROJECTION

To capture long-range dependencies of the entire network, we establish the interactive relationship of both attention models with an attention connection according to interactive projection of the human visual system. To be specific, prior attention pays more attention to high-frequency information, while posterior attention mainly emphasizes low-frequency one. Posterior attention is combined with prior attention to generate refined novel attention, which accurately restores significant texture information and efficiently removes residual haze. For example, as illustrated in Fig. 5, we utilize Canny operator to extract the effective edges of prior attention map I_{pri}, posterior attention map I_{pos} and novel attention map I_{nov}. Note that I_{pri} is considered as a binary region of interest (ROI) mask of I_{pos} and I_{nov}. Apparently, compared with the effective edges of I_{pos}, those of I_{nov} remarkably increase in the same ROI region. This demonstrates that attention projection is conducive to improve the missing edge information of posterior attention. More importantly, Fig 5. can precisely visualize the interaction effect of three kinds of attention maps. Additionally, attention projection of real-world hazy images is shown in Fig. 6. Finally, more discussions and analyses of relevant experiments are presented in Section IV-D.

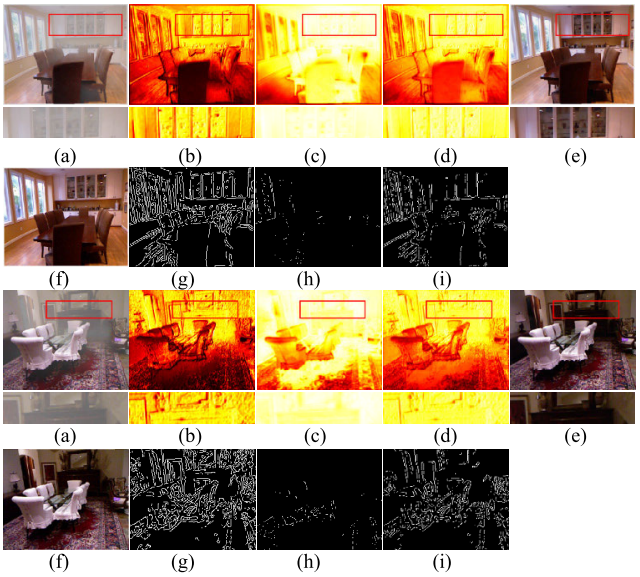

FIGURE 5. Attention projection of synthetic images. (a) Hazy images. (b) Prior attention map. (c) Posterior attention map. (d) Novel attention map. (e) Our dehazed results. (f) Ground truth. (g) Effective edges of prior attention map. (h) Effective edges of posterior attention map. (i) Effective edges of novel attention map. (Best observed in color and amplification factor. The higher the brightness is, the larger the weight value is.)

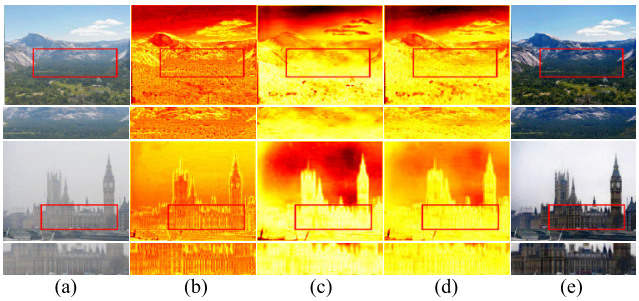

FIGURE 6. Attention projection of real-world hazy images. (a) Hazy images. (b) Prior attention maps. (c) Posterior attention maps. (d) Novel attention maps. (e) Our results. (Best observed in color and amplification factor. The higher the brightness is, the larger the weight value is.)

Generally, extensive channel attention information produced by DCAM in the encoder is projected forward to MSAM in the decoder to enhance spatial attention information, which effectively compensates the missing high-frequency details (such as effective edges and texture information) of MSAM during the decoding process. In Fig. 4(b), this fusing process of attention information can be formulated as

$$\widetilde{\beta} = \tau \cdot \alpha + (1-\tau)\beta \tag{8}$$

where $\tau \in [0, 1]$ is the learnable interpolation parameter; α and β can be regarded as prior attention and posterior attention respectively, which produces novel attention $\tilde{\beta}$ jointly.

Substituting (8) into (7), finally we can acquire

$$O_m = \zeta(\widetilde{\beta} \cdot Z_m) + X_m \tag{9}$$

It demonstrates that DCAM and MSAM are located at the symmetrically same level in the encoder and the decoder, respectively. Moreover, they show their own specificities in similar hazy areas while compensating the missing high-frequency information of MSAM, which is consistent with Section II-B.

C. SPECTRAL NORMALIZATION FOR RELATIVISTIC AVERAGE LEAST SQUARES GENERATIVE ADVERSARIAL NETWORK

Based on the atmospheric scattering model (1), it demonstrates that both hazy image $\boldsymbol{I}$ and clean image $\boldsymbol{J}$ maintain a spatial dependency in Fig. 7, which can be represented as

$$\Delta Q(x) = I(x) - J(x) \tag{10}$$

where $\Delta Q(x)$ can be regarded as the hazy concentration which estimates the dehazing effect. Consequently, we utilize the discriminator of the conditional generative adversarial network [53] to maintain their spatial relationship better. Furthermore, with the guidance of the discriminator estimating the dehazing effect $\Delta\tilde{Q}(x)$, the generator will continuously compel $\tilde{J}(x)$ to approach to $J(x)$ until they cannot be distinguished. To achieve this idea, we introduce the relativistic

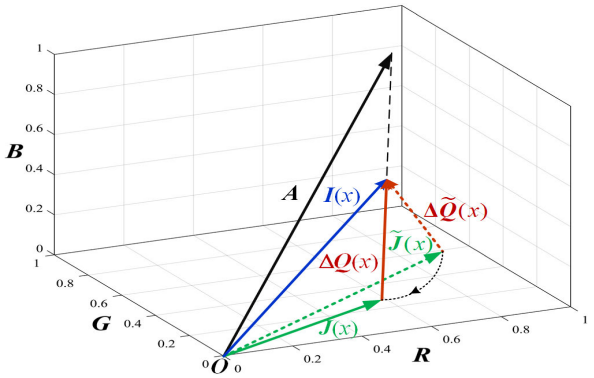

FIGURE 7. Space vector diagram of the atmospheric scattering model.

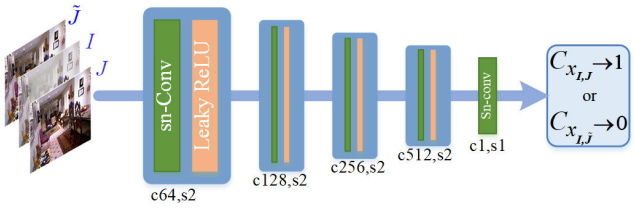

FIGURE 8. The network architecture of the conditional discriminator similar to the generator. I, J and "J" indicates a hazy image, a ground-truth image and a dehazed image, respectively.

average least squares discriminator (RaLSD) [54] to optimize the entire training process better. Meanwhile, similar to the generator in AAGAN, we also remove all the instance normalization [47] layers in the discriminator to extract better feature information, and then introduce spectral normalization [48] to stabilize the training process of AAGAN.

Compared with the standard discriminator in Dehaze-cGAN [35], RaLSD [54] estimates the probability that true dehazing effect Q(x) is relatively better than dehazing one $\tilde{Q}(x)$. In Fig. 8, the discriminator loss can be expressed as

$$\begin{aligned} L_D^{RaLSGAN} &= E_{x_{I,J}}[(C_{x_{I,J}} - E_{x_{I,\tilde{J}}}(C_{x_{I,\tilde{J}}}) - 1)^2] \\ &\quad + E_{x_{I,\tilde{J}}}[(C_{x_{I,\tilde{J}}} - E_{x_{I,J}}(C_{x_{I,J}}) + 1)^2], \\ C_{x_{I,J}} &= C(x_I, x_J), C_{x_{I,\tilde{J}}} = C(x_I, x_{\tilde{J}}) \end{aligned} \tag{11}$$

where $C(\cdot)$ denotes the output of discriminator without the final sigmoid layer; x_I, x_J and $x_{\tilde{J}}$ are the input hazy image, the original clean image and the output dehazed image, respectively, where $x_{\tilde{J}} = G(x_I)$. Similarly, the adversarial loss of the corresponding generator can be defined as

$$\begin{aligned} L_G^{RaLSGAN} &= E_{x_{I,\tilde{J}}}[(C_{x_{I,\tilde{J}}} - E_{x_{I,J}}(C_{x_{I,J}}) - 1)^2] \\ &\quad + E_{x_{I,J}}[(C_{x_{I,J}} - E_{x_{I,\tilde{J}}}(C_{x_{I,\tilde{J}}}) + 1)^2] \end{aligned} \tag{12}$$

As above, it is clear that $L_G^{RaLSGAN}$ incorporates x_I, x_J and $x_{\tilde{J}}$ at the same time. Consequently, the generator in AAGAN achieves better gradient optimization from these data in the adversarial training process, which promotes $x_{\tilde{J}}$ in accordance with x_J.

D. LOSS FUNCTION

To achieve better dehazing performance, the generator in AAGAN combines adversarial loss with content loss and perceptual loss. Specifically, first, the content loss, which assesses effectively differences between dehazed images and clean ones with Manhattan distance, can be expressed as

$$L_C = E_{x_I} \left\| G(x_I) - x_J \right\|_1 \tag{13}$$

Besides, perceptual loss can restore high-frequency feature information better, thus generating more realistic scenes. Similar to SRGAN [55], it can be indicated as

$$L_P = E_{x_I} \left\| V(G(x_I)) - V(x_J) \right\|_2^2 \tag{14}$$

where $V(\cdot)$ denotes the output feature maps before the final maxpooling layer in the pre-trained VGG-19 [56] network.

In the end, the overall loss of our proposed generator is formulated as

$$L_G = \lambda_C L_C + \lambda_P L_P + \lambda_R L_G^{RaLSGAN} \tag{15}$$

where λ_C, λ_P and λ_R are the weight coefficients of each loss term, respectively. Equation (15) is minimized continuously to optimize the generator G.

IV. EXPERIMENTS

To demonstrate the effectiveness of AAGAN, we make quantitative and qualitative comparisons with several state-of-the-art dehazing algorithms on synthetic datasets, real-world images and other scene images. The details are as follows.

A. EXPERIMENT DETAILS

In AAGAN, the input and output sizes of the generator are $256 \times 256 \times 3$, while the input size of the discriminator is $256 \times 256 \times 6$ and its output dimension is $32 \times 32 \times 1$. In the entire training process, $\lambda_C = 100$, $\lambda_P = 100$ and $\lambda_R = 0.1$ are set in our extensive experiments. Besides, the proposed AAGAN with the learning rate of 2×10^{-4} and batch size of 1 is trained by ADAM optimization policy. Finally, our model is trained for 318500 iterations in Tensorflow running on Windows 10, which takes about 93 hours on the Dell Precision Tower 7910 workstation with an NVIDIA 1080Ti GPU, a 2.2GHz Intel Xeon E5-2650 CPU and a 32G RAM.

B. TRAINING DATASET

Considering that it is very difficult to gain extensive pairs of haze-free image and hazy one in the real world, we are motivated by previous methods [29], [32], [34], [35], and synthesize the training dataset with NYU2 dataset [57] to train AAGAN. For this dataset, similar to [34], we utilize 1300 pairs of clean images and depth ones. Besides, we employ the guided image filtering method [58] to remove artifacts for all depth maps, which are added 1% Perlin noise to enhance the robustness of AAGAN. Subsequently, we crop randomly 9100 pairs of haze-free patches and depth ones whose size are 256×256 from the training dataset. Based on

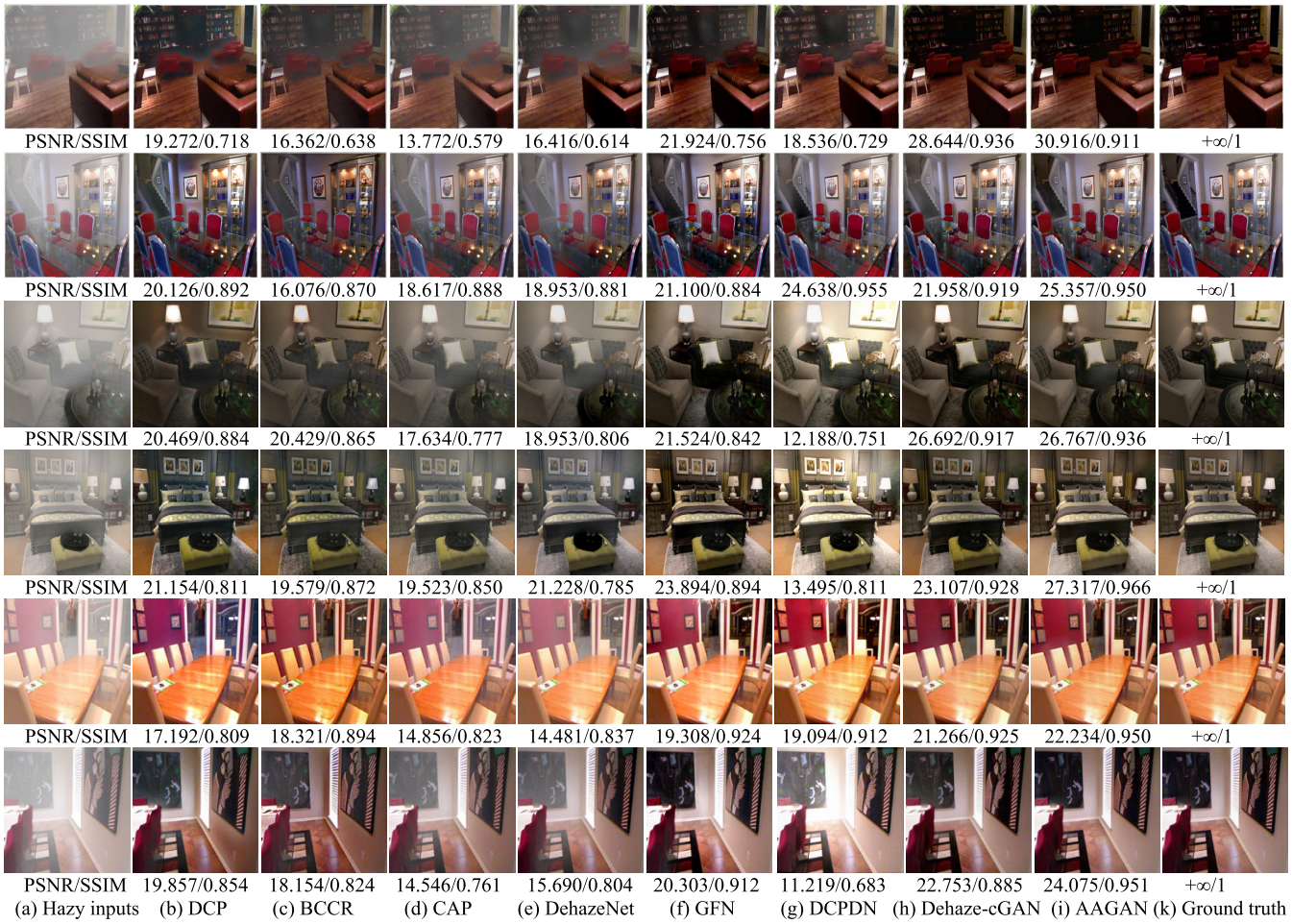

FIGURE 9. Dehazing comparisons on the NYU2, SUN3D and RESIDE synthetic datasets. The first two rows of comparison results are on the NYU2 data-set, the middle two rows are on the SUN3D dataset and the final two rows are on the SOTS dataset of RESIDE.

the physical model (1), we synthesize the corresponding hazy patches using random scattering coefficient $\beta \in [0.6, 1.8]$ and atmospheric light $A = [\eta, \eta, \eta]$ with $\eta \in [0.5, 1]$. Finally, we obtain 9100 pairs of haze-free patches and corresponding hazy ones to train our model.

C. QUANTITATIVE EVALUATION

1) EFFECTIVENESS ON SYNTHETIC DATASET

To evaluate the effectiveness of our proposed approach, we synthesize two datasets with the above method: NYU2 synthetic dataset and SUN3D synthetic one. The former dataset includes 149 images except the 1300 training ones from NYU2 dataset [57], while the latter has 150 ones from SUN3D dataset [59]. More importantly, we select two full-reference important criteria to evaluate these two datasets, *i.e.* Peak Signal to Noise Ratio (PSNR) and Structure Similarity (SSIM). Besides, RESIDE [60] is introduced to further demonstrate the dehazing effectiveness of our method with two no-reference metrics, such as spatial-spectral entropy-based quality (SSEQ) [61] and blind image integrity notator using DCT statistics (BLIINDS-II) [62], except for two full-reference ones. To be specific, we randomly sample 100 hazy images from the SOTS dataset of RESIDE [60], and then make extensive comparisons with the evaluation method provided by Li [60].

Based on these datasets we compare our algorithm with several state-of-the-art methods, such as DCP [15], BCCR [17], CAP [20], DehazeNet [28], GFN [34], Dehaze-cGAN [35] and DCPDN [32]. As illustrated in Fig. 9, we show the dehazing effectiveness of the above methods. For one thing, DCP [15], BCCR [17] and CAP [20] are based on priors to remove haze. DCP [15] and BCCR [17] could cause color distortions and obscurity in some hazy areas to some extent, and CAP [20] cannot effectively remove dense haze using a linear model. For another, the rest algorithms take advantage of CNNs learning methods. DehazeNet [28] has better progress on estimating the transmission, but some residual haze remains. Based on three kinds of prior information, GFN [33] can produce some artifacts in some hazy regions to some extent. DCPDN [31] tends to overestimate the atmospheric light, thus resulting in local overexposure for some hazy regions. Dehaze-cGAN [34] is closer to the ground truth for haze removal except for slight artifacts. Compared with the above seven methods, our proposed method achieves

TABLE 1. Average results of PSNR, SSIM on the NYU2 synthetic dataset.

Metric	DCP [15]	BCCR [17]	CAP [20]	DehazeNet [28]	GFN [34]	DCPDN [32]	Dehaze-cGAN [35]	AAGAN
SSIM	0.854	0.793	0.848	0.858	0.885	0.954	0.955	**0.968**
PSNR	16.994	13.583	18.279	18.324	20.357	26.641	25.376	**27.291**

TABLE 2. Average results of PSNR, SSIM on the SUN3D synthetic dataset.

Metric	DCP [15]	BCCR [17]	CAP [20]	DehazeNet [28]	GFN [34]	DCPDN [32]	Dehaze-cGAN [35]	AAGAN
SSIM	0.809	0.808	0.846	0.860	0.841	0.831	0.922	**0.937**
PSNR	18.653	16.914	19.784	20.979	20.474	16.227	23.256	**23.758**

TABLE 3. Average results of PSNR, SSIM, SSEQ and BLIINDS-II on the SOTS dataset of RESIDE.

Metric	DCP [15]	BCCR [17]	CAP [20]	DehazeNet [28]	GFN [34]	DCPDN [32]	Dehaze-cGAN [35]	AAGAN
SSIM	20.867	17.879	18.461	19.743	20.385	15.946	22.053	**23.687**
PSNR	0.892	0.856	0.859	0.885	0.907	0.835	0.911	**0.945**
SSEQ	74.197	77.015	75.801	75.226	**78.212**	74.756	71.179	75.465
BLIINDS-II	85.695	87.665	85.975	85.840	85.785	84.635	80.475	**88.455**

FIGURE 10. Qualitative hazy images and corresponding dehazed ones.

the best performance on haze removal. The main reason is that our method focuses on hazy areas to remove haze accurately with attention models, and restores realistic texture information and visual contrast using an improved RaLSGAN [54].

In the end, the massive comparison results in Table 1, Table 2 and Table 3 present that AAGAN gets the highest PSNR, SSIM and BLIINDS-II against other methods, and achieves the moderate SSEQ. Overall, AAGAN has the powerful ability to remove haze and achieves the best visual performance against other state-of-the-art methods.

2) EFFECTIVENESS ON REAL-WORLD DATASET

To further demonstrate the effectiveness of our method, we select several challenging hazy images to achieve comprehensive comparisons with the above state-of-the-art dehazing methods. As illustrated in Fig. 11, all the dehazing approaches have achieved good performance on the dehazing effect to some extent. As we know, it is very difficult to remove haze for sky regions and white objects similar to haze. For example, sky regions are oversaturated and distorted by DCP [15] that is prone to overestimate the transmission. Similarly, DCPDN [32] tends to overestimate the local atmospheric light to cause over-enhancement for the blue sky in the final three rows. BCCR [17] causes color distortions for sky areas, thus reducing visual contrast. Secondly, as shown in Fig. 11(b, c, f), gray rocks have some color distortions in the second row. Except for these sensitive areas, the ability to remove haze is also a significant evaluation index. According to a closer examination for light haze, even though the first seven methods could remove most of the haze in Fig. 11, we discover that there is still some residual haze in some red boxes. Overall, our method can capture long-range dependencies to recognize hazy areas precisely, thus efficiently removing residual haze and improving visual contrast.

At present, image restoration for dense hazy scenes is still very intractable due to the low quality of dense hazy images.

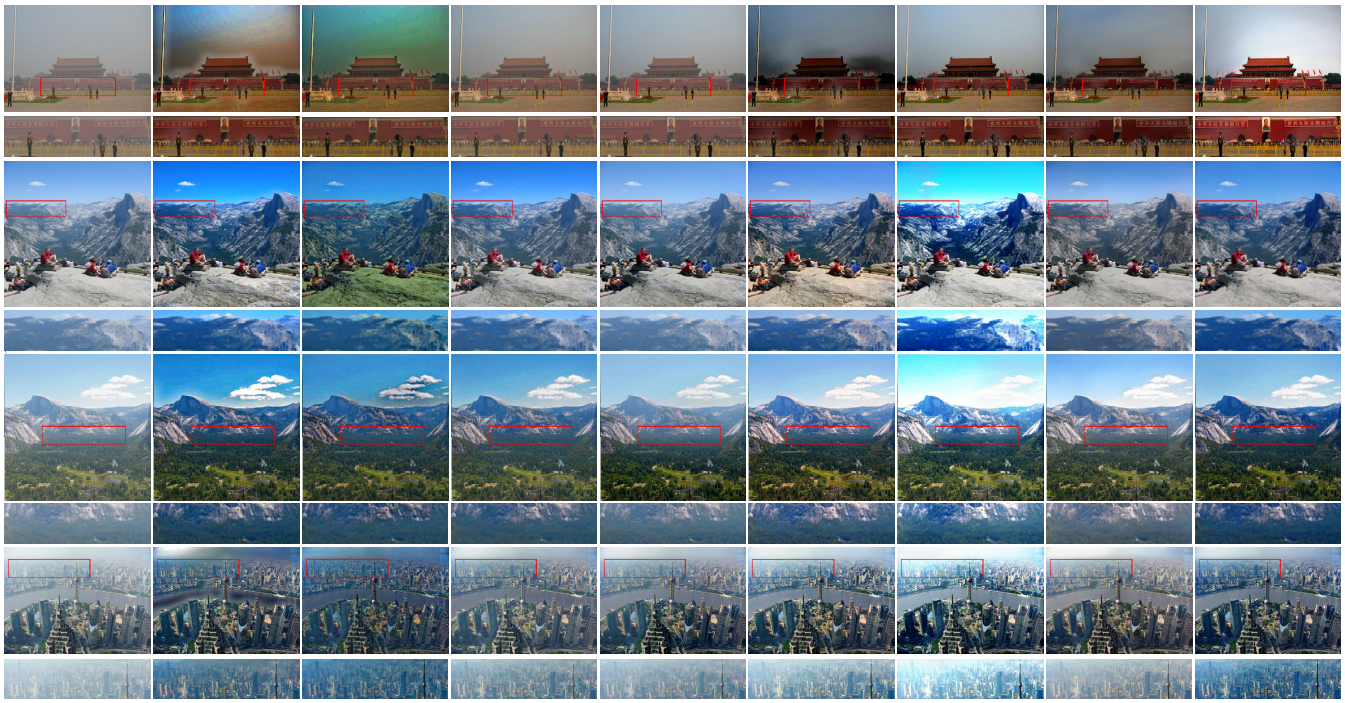

FIGURE 11. Comprehensive comparisons of numerous dehazing methods on real-world hazy images.

TABLE 4. Average results of SSEQ and BLIINDS-II for real-world dehazed images in Fig. 11.

Metric	DCP [15]	BCCR [17]	CAP [20]	DehazeNet [28]	GFN [34]	DCPDN [32]	Dehaze-cGAN [35]	AAGAN
SSEQ	79.339	77.170	78.727	79.920	**83.308**	78.809	79.809	77.303
BLIINDS-II	80.500	81.250	78.250	78.375	79.750	79.875	81.000	**84.125**

FIGURE 12. Visual enhancement for halation scenes.

Consequently, in order to further validate the effectiveness of our method, we make some challenging comparisons for dense hazy scenes in Fig. 13. Apparently, it is very difficult for DCP [15], BCCR [17], CAP [20] and DehazeNet [28] methods to remove dense haze. Besides, GFN [33] and Dehaze-cGAN [34] approaches could handle with dense haze to some extent, but produce excessive artifacts for dense hazy images. DCPDN [31] can effectively remove most of the dense haze, but cannot further improve visual contrast of the scenes. In contrast, the proposed method is able to remove dense haze and enhance visual contrast. In the whole comparison process of light and dense hazy scenes, our method can have the best performance on the dehazing task and gain excellent visual quality against the previous methods. For fair and objective comparison, we introduce two no-reference metrics to evaluate the dehazing effective-ness for all the methods in Table 4 and Table 5. Although our method obtains the moderate SSEQ, it still achieves the best BLIINDS-II performance for real-world hazy images. In addition, for this issue, RESIDE [60] also addresses that objective image quality assessment (IQA) algorithms still are limited for the visual quality of dehazed images. Therefore, in Fig. 10, we present more dehazing results generated by AAGAN to demonstrate the effective dehazing performance of our method.

3) EFFECTIVENESS ON HALATION SCENE DATASET

Likewise, our proposed method can also be appropriate for halation scenes to verify its generalization capability effectively. As shown in Fig. 12, halation could be easily produced when extremely intense light (*e.g.* sunlight and lamplight) is scattered. Even though our model is trained with the hazy

TABLE 5. Average results of SSEQ and BLIINDS-II for real-world dehazed images in Fig. 13.

Metric	DCP [15]	BCCR [17]	CAP [20]	DehazeNet [28]	GFN [34]	DCPDN [32]	Dehaze-cGAN [35]	AAGAN
SSEQ	71.784	71.337	68.624	69.193	75.552	71.899	**76.571**	74.887
BLIINDS-II	79.357	81.714	78.500	78.500	82.357	85.286	85.714	**88.500**

(a) Hazy inputs (b) DCP (c) BCCR (d) CAP (e) DehazeNet (f) GFN (g) DCPDN (k) Dehaze-cGAN (i) AAGAN

FIGURE 13. Comprehensive comparisons of numerous dehazing methods for the inclement weather, where dense hazy images are from the Internet.

TABLE 6. Effectiveness of the proposed method on the NYU2 synthetic dataset.

Metric	AAGAN -wo-AC	AAGAN -wo-AM	AAGAN -IN	SGAN	AAGAN
SSIM	0.965	0.948	0.935	0.958	**0.968**
PSNR	26.700	24.402	24.699	25.512	**27.291**

TABLE 7. Effectiveness of the proposed method on the SUN3D synthetic dataset.

Metric	AAGAN -wo-AC	AAGAN-wo-AM	AAGAN -IN	SGAN	AAGAN
SSIM	0.931	0.918	0.914	0.929	**0.937**
PSNR	23.162	22.549	22.791	22.997	**23.758**

dataset, it can alleviate the effect of halation and enhance the visibility of scenes to a great extent. Therefore, it demonstrates that our method has better robustness for relevant atmospheric scattering issues.

D. ANALYSES AND DISCUSSIONS

For the proposed method, we further conduct quantitative and qualitative analyses and discussions for the effectiveness of each part within AAGAN. To this end, we evaluate the relevant variants of AAGAN to verify this issue with the same network parameters in Table 6, Table 7 and Fig. 14. Finally, we discuss the environmental limitations of AAGAN.

1) EFFECTIVENESS OF ATTENTION MECHANISM

To validate the effectiveness of the attention connection and the attention models, we have two variants: (a) AAGAN without any attention connection (AAGAN-wo-AC), and (b) AAGAN without any attention model (AAGAN-wo-AM).

By comparisons in Table 6 and Table 7, AAGAN-wo-AC is more than AAGAN-wo-AMs on the SSIM and PSNR

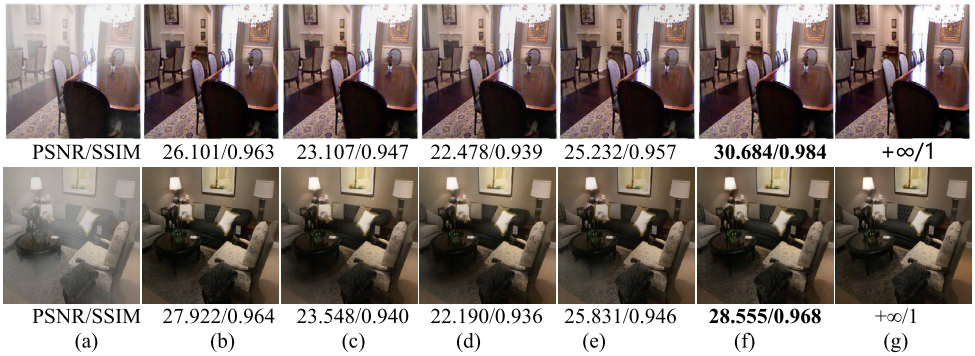

FIGURE 14. The effective comparisons of relevant variants of AAGAN. The first row of comparison results is on the NYU2 dataset and the last row is on the SUN3D one. (a) Hazy inputs. (b) AAGAN-wo-AC. (c) AAGAN-wo-AM. (d) AAGAN-IN. (e) SGAN. (f) AAGAN. (g) Ground truth.

metric. This demonstrates that the attention models enhance the dehazing ability of AAGAN. Subsequently, based on AAGAN-wo-AC, AAGAN utilizes an attention connection to capture long-range dependencies of the entire network. Compared with AAGAN-wo-AC and AAGAN-wo-AMs, AAGAN achieves the best performance on the SSIM and PSNR metric in Table 6 and Table 7.

2) EFFECTIVENESS OF SPECTRAL NORMALIZATION

Considering that instance normalization (IN) [47] can cause some undesired artifacts, we remove all the instance normalization layers and introduce spectral normalization (SN) [48] for all the convolution (Conv) layers, thus ensuring the training stability of the whole network. To validate the effectiveness of spectral normalization, we have the variant AAGAN-IN, which employs Conv layer and IN one instead of Conv layer with SN in AAGAN. Note that the attention models of AAGAN-IN only utilize Conv layer to avoid extensive artifacts.

We discover that AAGAN has better performance than AAGAN-IN on the SSIM and PSNR metric with a considerable margin in Table 6 and Table 7. Similarly, the same result also occurs in Fig. 14(d) and Fig. 14(f). The main reason is that instance normalization can easily make some impact on high-frequency feature information to degrade image quality. Consequently, AAGAN has the excellent ability to restore image relative to AAGAN-IN.

3) EFFECTIVENESS OF RELATIVISTIC AVERAGE LEAST SQUARES GENERATIVE ADVERSARIAL NETWORK

According to the dehazing analysis in Section III-C, we utilize an improved RaLSGAN [54] to optimize the training stability, recover more realistic texture information and enhance visual contrast. To demonstrate the dehazing ability of AAGAN, we compare AAGAN with the standard GAN (SGAN) on the SSIM and PSNR metric in Table 6 and Table 7. Obviously, AAGAN is superior to SGAN for haze removal. Furthermore, compared with SGAN, AAGAN can better strengthen visual contrast to restore high-quality scene in Fig. 15(a).

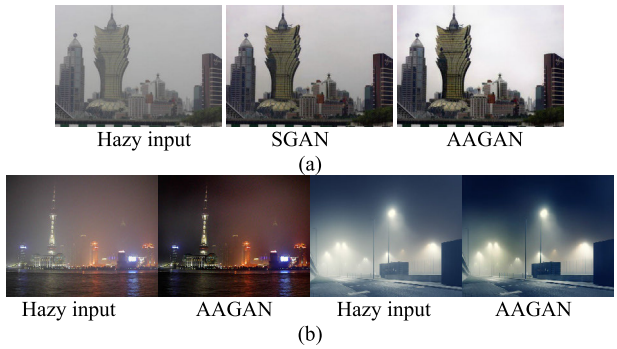

FIGURE 15. Dehazing comparisons during the day and the night, respectively.

4) LIMITATIONS

The classical atmospheric scattering model only depends on the sun as the main light source, so it is not suitable for nighttime images dehazing due to the influence of non-uniform, varicolored and non-homogenous lights produced by multiple light sources at night. Consequently, as illustrated in Fig. 15(b), the proposed AAGAN is trained with the synthetic dataset generated by the atmospheric scattering model, which does not adapt to nighttime images dehazing. In the future, we will study more adaptive physical model to implement the dehazing task for hazy scenes with different light sources.

V. CONCLUSION

In this paper, we have proposed a stable and enhanced attention-to-attention generative adversarial network (AAGAN) to remove haze, which can also be extended to image segmentation, style transfer, image super-resolution, *etc*. Motivated by the attention mechanism of the human visual system, we elaborately design dense channel attention model (DCAM) and multiscale spatial attention model (MSAM) to pay close attention to hazy areas. They can

capture long-range dependencies from the input feature maps in the channel dimension and the spatial one, respectively. More importantly, we present the attention projection, which utilizes an attention connection to connect two attention models to acquire global feature dependencies of the whole network.

To further improve image quality and enhance the stability of the proposed network, we remove all the Instance normalization layers and employ spectral normalization for all the convolution layers. Next, we analyze the dehazing mechanism based on the atmospheric scattering model, and then leverage an improved RaLSGAN to recover more realistic texture information and visual contrast for different hazy scenes. Finally, the remarkable comparisons demonstrate that our proposed algorithm has the best performance on haze removal against the state-of-the-art methods. In the future, we will study the adaptive physical model to implement nighttime images dehazing, and continue to further improve image quality for dense hazy scenes under extreme weather conditions.

REFERENCES

[1] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 385–400.

[2] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *Int. J. Comput. Vis.*, vol. 127, no. 2, pp. 115–142, Feb. 2019.

[3] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Comput. Surv.*, vol. 46, no. 2, pp. 29:1–29:37, Nov. 2013.

[4] F. S. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, and J. M. Alvarez, "Effective use of synthetic data for urban scene semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 86–103.

[5] Y. Y. Schechner, S. G. Narasimhan, and S. K. Nayar, "Instant dehazing of images using polarization," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2001, pp. I-325–I-332.

[6] Y. Y. Schechner, S. G. Narasimhan, and S. K. Nayar, "Polarization-based vision through haze," *Appl. Opt.*, vol. 42, no. 3, pp. 511–525, Jan. 2003.

[7] L. Shen, Y. Zhao, Q. Peng, J. C.-W. Chan, and S. G. Kong, "An iterative image dehazing method with polarization," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1093–1107, May 2019.

[8] S. K. Nayar and S. G. Narasimhan, "Vision in bad weather," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep. 1999, pp. 820–827.

[9] S. G. Narasimhan and S. K. Nayar, "Chromatic framework for vision in bad weather," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2000, pp. 598–605.

[10] S. G. Narasimhan and S. K. Nayar, "Contrast restoration of weather degraded images," *IEEE Trans. Pattern Anal. Mach. Learn.*, vol. 25, no. 6, pp. 713–724, Jun. 2003.

[11] J. Kopf, J. Kopf, B. Neubert, B. Chen, M. Cohen, D. Cohen-Or, O. Deussen, M. Uyttendaele, and D. Lischinski, "Deep photo: Model-based photograph enhancement and viewing," *ACM Trans. Graph.*, vol. 27, no. 5, pp. 116-1–116-10, Dec. 2008.

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

[13] N. Hautiere, J.-P. Tarel, and D. Aubert, "Towards fog-free in-vehicle vision systems through contrast restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

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

[15] 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.

[16] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 228–242, Feb. 2008.

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

[18] Y. Liu, H. Li, and M. Wang, "Single image dehazing via large sky region segmentation and multiscale opening dark channel model," *IEEE Access*, vol. 5, pp. 8890–8903, 2017.

[19] C. O. Ancuti, C. Ancuti, C. Hermans, and P. Bekaert, "A fast semi-inverse approach to detect and remove the haze from a single image," in *Proc. 10th ACCV*, 2011, pp. 501–514.

[20] 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.

[21] 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.

[22] L. Kratz and K. Nishino, "Factorizing scene albedo and depth from a single foggy image," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Sep/Oct. 2009, pp. 1701–1708.

[23] K. Nishino, L. Kratz, and S. Lombardi, "Bayesian defogging," *Int. J. Comput. Vis.*, vol. 98, no. 3, pp. 263–278, Jul. 2012.

[24] Y. Wang and C. Fan, "Single image defogging by multiscale depth fusion," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4826–4837, Nov. 2014.

[25] Y. Liu, J. Shang, L. Pan, A. Wang, and M. Wang, "A unified variational model for single image dehazing," *IEEE Access*, vol. 7, pp. 15722–15736, 2019.

[26] K. Tang, J. Yang, and J. Wang, "Investigating haze-relevant features in a learning framework for image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2995–3002.

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

[28] 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.

[29] 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. (ECCV)*, Sep. 2016, pp. 154–169.

[30] R. Liu, X. Fan, M. Hou, Z. Jiang, Z. Luo, and L. Zhang, "Learning aggregated transmission propagation networks for haze removal and beyond," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 2973–2986, Oct. 2019.

[31] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "AOD-Net: All-in-one dehazing network," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4780–4788.

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

[33] A. Wang, W. Wang, J. Liu, and N. Gu, "Aipnet: Image-to-image single image dehazing with atmospheric illumination prior," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 381–393, Jan. 2019.

[34] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang, "Gated fusion network for single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3253–3261.

[35] R. Li, J. Pan, Z. Li, and J. Tang, "Single image dehazing via conditional generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8202–8211.

[36] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 184–199.

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

[38] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Sep. 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473

[39] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," May 2018, *arXiv:1805.08318*. [Online]. Available: https://arxiv.org/abs/1805.08318

[40] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2018, pp. 294–310.

[41] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6450–6458.

[42] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "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.
[43] Z. Wu, X. Han, Y.-L. Lin, M. G. Uzunbas, T. Goldstein, S. N. Lim, and L. S. Davis, "DCAN: Dual channel-wise alignment networks for unsupervised scene adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 535–552.
[44] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2018, pp. 7794–7803.
[45] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.
[46] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Sep. 2018, pp. 63–79.
[47] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," Jul. 2016, *arXiv:1607.08022*. [Online]. Available: https://arxiv.org/abs/1607.08022
[48] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. ICLR*, 2018, pp. 1–15.
[49] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. AISTATS*, 2011, pp. 315–323.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[51] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269.
[52] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[53] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.
[54] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," Jul. 2018, *arXiv:1807.00734*. [Online]. Available: https://arxiv.org/abs/1807.00734
[55] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.
[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.
[57] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 746–760.
[58] F. Kou, W. Chen, C. Wen, and Z. Li, "Gradient domain guided image filtering," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4528–4539, Nov. 2015.
[59] J. Xiao, A. Owens, and A. Torralba, "SUN3D: A database of big spaces reconstructed using SfM and object labels," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1625–1632.
[60] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2019.
[61] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Process., Image Commun.*, vol. 29, no. 8, pp. 856–863, 2014.
[62] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

WENHUI WANG received the B.S. and M.S. degrees from the North China University of Science and Technology, Tangshan, China, in 2010 and 2013, respectively. He is currently pursuing the Ph.D. degree in pattern recognition and intelligent system with the School of Information Science and Engineering, Northeastern University, China. His research interests are in computer vision, image processing, and machine learning.

ANNA WANG received the B.S., M.S., and Ph.D. degrees in measurement technology and instruments from Northeastern University, Shenyang, China, in 1982, 1988, and 2001, respectively. Since 1994, she has been a Full Professor and a Ph.D. Supervisor with the School of Information Science and Engineering, Northeastern University. Her main research interests include computer vision, machine learning, self-driving technology, and electric network control of distributed generation systems.

QING AI received the B.S. and M.S. degrees, in 2003 and 2007, respectively. He is currently pursuing the Ph.D. degree in pattern recognition and intelligent system with Northeastern University, China. He is currently an Associate Professor with the University of Science and Technology, Liaoning, China. His research interests include pattern recognition, support vector machines, optimization theory and applications, and fault diagnosis.

CHEN LIU received the B.S. and M.S. degrees, in 2015 and 2018, respectively. He is currently pursuing the Ph.D. degree in pattern recognition and intelligent system with the School of Information Science and Engineering, Northeastern University, China. His research interests mainly include computer vision and 3D construction.

JINGLU LIU received the B.S. and M.S. degrees, in 2013 and 2015, respectively. He is currently pursuing the Ph.D. degree in power electronics and drives with the School of Information Science and Engineering, Northeastern University, China. His current research interests include machine learning, machine vision, and self-driving.