Visual Haze Removal by a Unified Generative Adversarial Network

Yanwei Pang, *Senior Member, IEEE*, Jin Xie, and Xuelong Li, *Fellow, IEEE*

***Abstract*—Existence of haze significantly degrades visual quality and hence negatively affects the performance of visual surveillance, video analysis, and human–machine interaction. To remove haze from a visual signal, in this paper, we propose a generative adversarial network for visual haze removal called HRGAN. HRGAN consists of a generator network and a discriminator network. A unified network jointly estimating transmission maps, atmospheric light, and haze-free images (called UNTA) is proposed as the generator network of HRGAN. Instead of being optimized by minimizing the pixel-wise loss, HRGAN is optimized by minimizing a novel loss function consisting of pixel-wise loss, perceptual loss, and adversarial loss produced by a discriminator network. Classical model-based image dehazing algorithms consist of three separate stages: 1) estimating transmission map; 2) estimating atmospheric light; and 3) restoring haze-free image by using an atmospheric scattering model to process the transmission map and atmospheric light. Such a separate scheme is not guaranteed to achieve optimal results. On the contrary, UNTA performs transmission map estimation and atmospheric light estimation simultaneously to obtain joint optimal solutions. The experimental results on both synthetic and real-world image databases demonstrate that HRGAN outperforms the state-of-the-art algorithms in terms of both effectiveness and efficiency.**

***Index Terms*—Dehazing, visual quality improvement, generative adversarial network, convolutional neural network.**

I. Introduction

SEVERE weather conditions (e.g., fog, haze, and smoke) would significantly compromise the quality of the images acquired by the cameras. The performance of a lot of computer vision algorithms (e.g., tracking [1], object detection [2], and classification) would be adversely affected by the low-quality images [3]–[6]. So it has a great significance to study how to restore hazy images.

A large number of image dehazing methods have been brought forward [7]–[14]. According to whether or not to utilize physical models, these methods can be divided into two categories. One is model-based method, and the other is model-free method (e.g., IMDM [3]). The model-based method is a mathematical inversion process of restoring the haze-free image with the unknown factors (i.e. the transmission map and the atmospheric light). Because the physical-based analytical models can describe the composition of hazy images, the model-based dehazing methods can achieve state-of-art performance.

Although many model-based image dehazing methods have been proposed, most of these methods estimate the transmission map and the atmospheric light separately. Obviously, the separate manner cannot guarantee that final solutions are joint optimal solutions.

Recently, several Convolutional Neural Networks(CNN)-based image dehazing methods have been brought out [15], [16]. In these methods, CNN [17] is used to estimate the transmission map first, then traditional method is applied to estimate the atmospheric light, finally the transmission map and the atmospheric light are used to restore haze-free images via atmospheric scattering model [18]. Although these methods have made significant progresses, in fact, the transmission map and the atmospheric light are still estimated separately. Therefore, the aforementioned problem is not solved in these CNN-based methods.

In order to overcome aforementioned drawback, we propose a unified network which jointly estimates transmission map, atmospheric light, and the haze-free image called UNTA. That is, UNTA can obtain joint optimal solutions.

The optimization of traditional CNN-based image dehazing algorithms is to minimize the mean squared error (MSE) between the restored haze-free image and ground-truth images. The pixel-wise image difference can be decreased by decreasing the MSE. However, the less pixel-wise image difference cannot present better perceptual dehazed result. Instead of MSE, in this paper, we utilize a more effective loss which consists of pixel-wise loss (e.g., MSE), perceptual loss, and adversarial loss. The perceptual loss is the difference between the high-level features of restored haze-free image and ground-truth image. By minimizing perceptual loss, perceptual relevant differences of dehazed results can be decreased. In this paper, we propose a novel Generative Adversarial Network (GAN)-based framework for image haze removal (called HRGAN). Sample results of the proposed HRGAN are shown in Fig. 1. Similar to previous GAN, our network consists of two networks: a generator network and a discriminator network. The adversarial loss is produced by discriminator network. The adversarial loss pushes restored haze-free image to the realistic haze-free image.

Manuscript received July 16, 2018; revised October 4, 2018; accepted October 31, 2018. Date of publication November 9, 2018; date of current version October 29, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61632081, Grant 61871470, and Grant 61503274 and in part by Nokia. This paper was recommended by Associate Editor H. Li. *(Corresponding author: Xuelong Li.)*

Y. Pang and J. Xie are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: pyw@tju.edu.cn; jinxie@tju.edu.cn).

X. Li is with the School of Computer Science and Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: li@nwpu.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2018.2880223

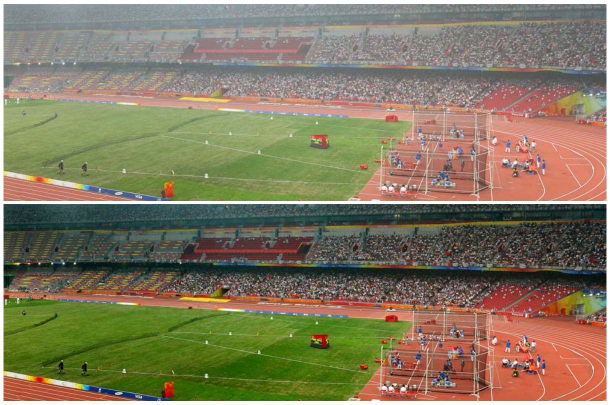


Fig. 1. Sample results of HRGAN. Top: the input hazy image. Bottom: the dehazed image.

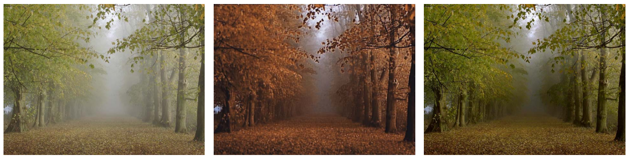


Fig. 2. Visual comparison between direct regression model and our proposed HRGAN. Left: input hazy image. Middle: dehazed result of direct regression model. Right: dehazed result of HRGAN.

The generator network of previous GAN-based image processing method [19] directly generates resulting images from input images. However, the direct regression model is not suitable for image dehazing. In this paper, we choose UNTA as generator network. As shown in Figure 2, this direct regression model may lead to serious color distortion. By contrast, HRGAN based on UNTA can generate visually appealing haze-free images. The main reason is that UNTA is based on the atmospheric scatting model. As described above, atmospheric scatting model can reveal physical characteristics of hazy images.

Li *et al.* [20] proposed a CNN-based framework which could directly generate haze-free image (referred to as AOD-Net). In their method, the transmission map and atmospheric light are unified into one variate, and CNN is used to solve this variate. Although their method has made great progress, the MSE is the only one loss in their method. As previously mentioned, the network trained by pixel-wise loss could lack high-frequency details of resulting haze-free images. Compared with our method, pixel-wise loss, perceptual loss, and adversarial loss are utilized to produce superior visual haze-free image. In addition, the running time of HRGAN is less than half of AOD-Net.

The novelty, contribution, and characteristic of the proposed method are as follows.

(1) We propose HRGAN which is a GAN-based image haze removal network. Compared with previous CNN-based method, HRGAN is optimized by an effective loss consisting of pixel-wise loss, perceptual loss calculated on feature maps of the VGG16 network [21], and adversarial loss produced by discriminator network.

(2) UNTA which can simultaneously estimate transmission map and atmospheric light is proposed as the generator network of HRGAN. Compared with previous model-based image dehazing methods, the UNTA has the capacity to obtain joint optimal solutions.

(3) HRGAN cannot only produce superior visual haze-free images but also be implemented very efficiently.

The rest of the paper is organized as follows. The related work are described in Section II. The proposed method is presented in III. Experimental results are presented in Section IV. Finally, Section V concludes the paper.

II. Related Work

In this section, we briefly review the literature for existing model-based image dehazing methods and Generative Adversarial Networks (GAN).

A. Single Image Dehazing

The model-based image dehazing methods are based on the atmospheric scattering model [18], [22], [23] which assumes that a hazy image I is composed of direct attenuations I_D and airlight I_A, respectively. Specifically, the atmospheric scattering model can be written as

$$\begin{aligned} I(x, y) &= I_D(x, y) + I_A(x, y) \\ &= J(x, y)t(x, y) + A(1 - t(x, y)) \end{aligned} \tag{1}$$

where $I(x, y)$ is the observed hazy image, $J(x, y)$ is the corresponding haze-free image, A represents the atmospheric light, $t(x, y)$ is the transmission map, and (x, y) is the position of the image.

When the atmosphere is homogeneous, the transmission map $t(x, y)$ can be described as

$$t(x, y) = e^{-\beta d(x,y)} \tag{2}$$

where $d(x, y)$ indicates the distance between the camera and the scene, and β represents the scattering coefficient of the atmosphere.

The model-based image dehazing method can be divided into handcrafted-feature-based method and CNN-based method.

The handcrafted-feature-based image dehazing methods are based on handcrafted-features. Generally speaking, these methods estimate the transmission map by hand-crafted features followed by estimating atmospheric light, finally restore haze-free image by using transmission map and atmospheric light via atmospheric scattering model. The main difference of these methods is the way to estimate transmission maps. For instance, He *et al.* [24] proposed a valid method based on dark channel prior (DCP) to estimate transmission maps. Meng *et al.* [25] presented a regularization method to estimate transmission maps by exploring the inherent boundary constraint. Tang *et al.* [26] proposed a learning-based method which uses the random forest [27] to learn the correlation between the transmission maps and four types of handcrafted features (i.e. multi-scale dark channel [24], multi-scale local max contrast [28], hue disparity [29], and multi-scale local

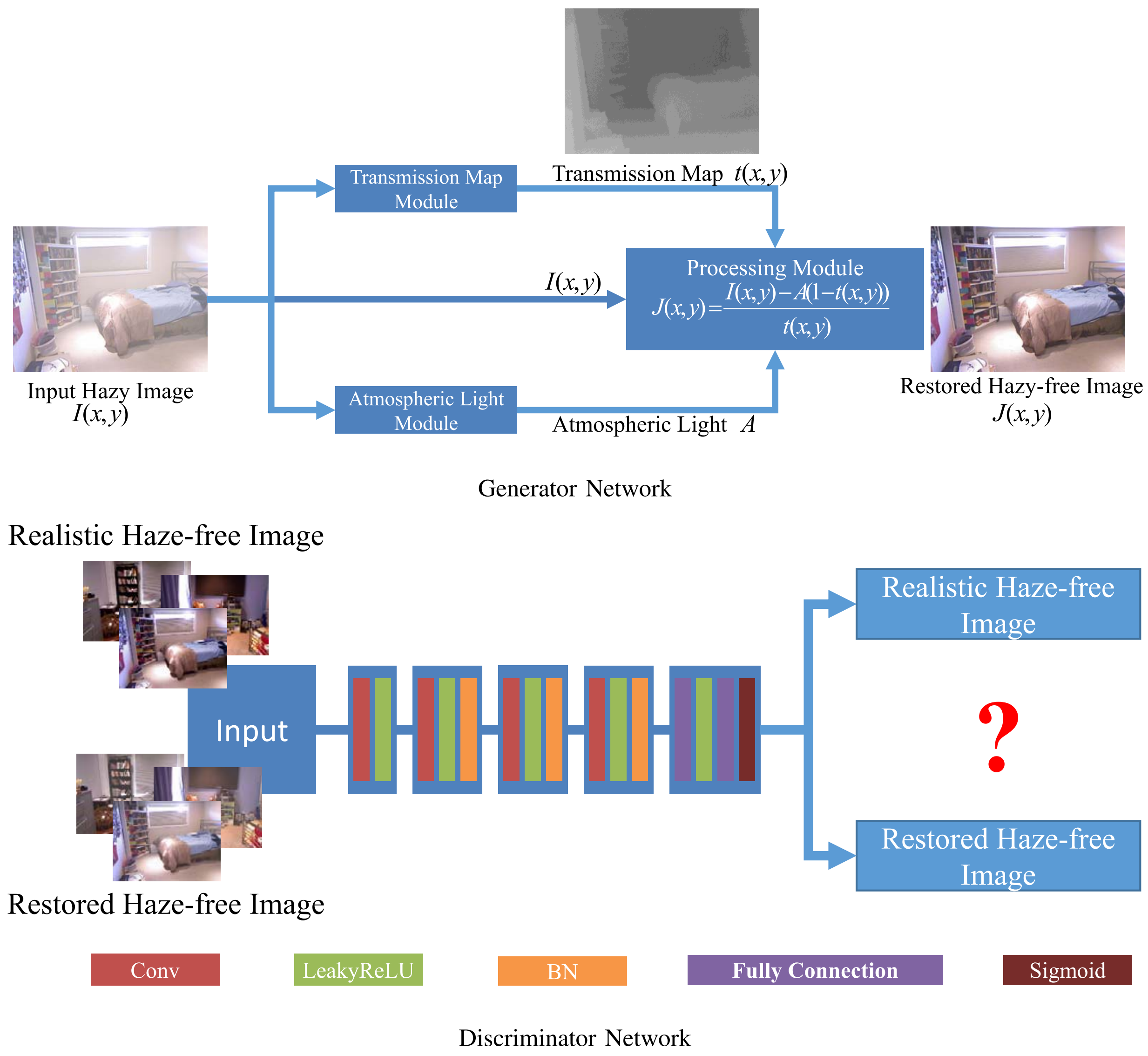


Fig. 3. The architecture of HRGAN. Top: Generator network. Bottom: Discriminator network.

max saturation). Zhu *et al.* [30] proposed to create a linear model based on a color attenuation prior for the depth map of the hazy image. In addition, Berman *et al.* [31] presented a non-local method based on the haze-line prior to estimate transmission maps.

Because of the great capacity of extracting features, CNN-based methods have received a lot of attention. The CNN-based methods can be divided into two categories. In the first category [15], [16], CNN is used to learn the mapping between hazy images and their corresponding transmission maps. Subsequently, the transmission maps and the atmospheric light estimated by traditional method are used to recover haze-free image via atmospheric scatting model. In another category [20], taking a hazy image as input, CNN could output a hazy-free image directly.

B. Generative Adversarial Networks

Goodfellow *et al.* [32] proposed Generative Adversarial Network (GAN). A typical GAN consists of two parts: a generator network and a discriminator network. The purpose of the generator network is to generate images which are used to make a fool of discriminator network, and the goal of the discriminator network is to distinguish the generating haze-free images from realistic haze-free images. Conditional Generative Adversarial Network (CGAN) is proposed by Mirza and Osindero [33]. The additional conditional information is added into traditional GAN, which makes the generator generate more effective results. GAN recently becomes one of the focus in the computer vision, and is applied in numerous tasks such as image super-resolution [19], image-to-image translation [34], text-to-image translation [35], and image inpainting [36].

III. Proposed Method

A. Network Architecture

The architecture of the proposed HRGAN is illustrated in Figure 3. The network consists of two networks: a generator network and a discriminator network.

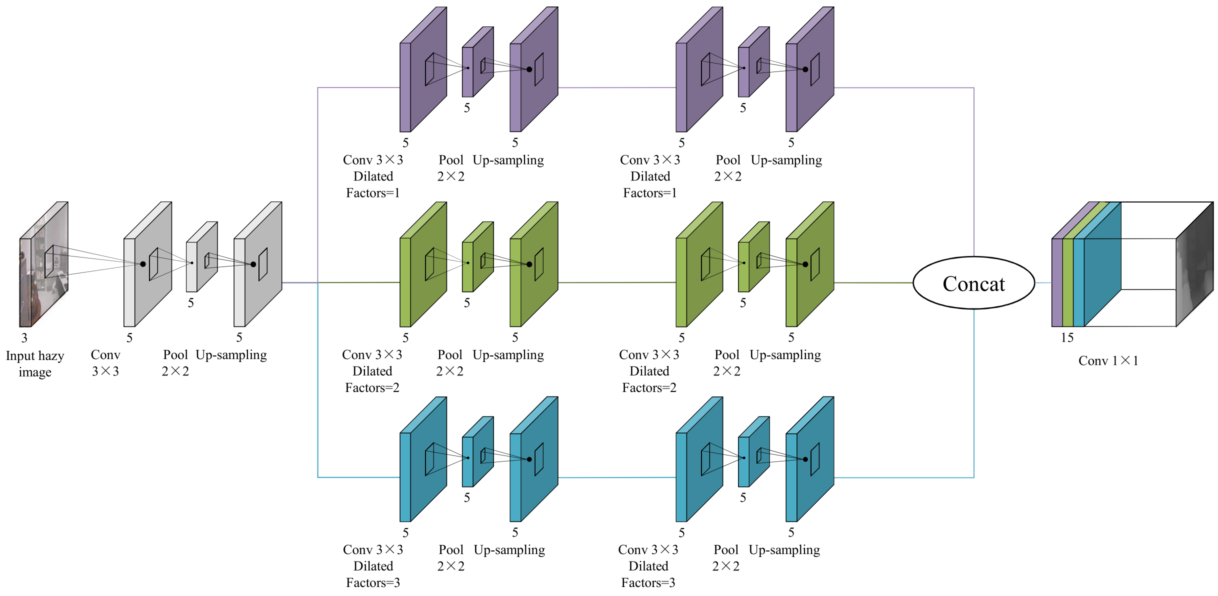


Fig. 4. The architecture of transmission maps module.

1) Generator Network Architecture: The generator network aims to generate hazy-free image. The generator network takes the hazy image as input, and produces hazy-free image. As shown in the top part of Figure 3, the generator network consists of three components: transmission map module, atmospheric light module, and processing module.

a) Transmission map module: The task of the transmission map module is to estimate transmission maps. The architecture of transmission map module is illustrated in Figure 4. The dilated convolution [37] achieves great success in semantic segmentation [38]. Inspired by the success, three parallel dilated convolutional layers with different dilated factors are used to extract multi-scale features. It is known that different dilation factor can extract different scales of features. With dilated factors being 1, 2, and 3, the 3 parallel branches of transmission map architecture can extract features of small-scale, middle-scale, and large-scale, respectively. The features extracted for each dilated factor are processed in separate branches and fused to generate the final result. As the same as the method proposed by Ren *et al.* [16], pooling layers and up-sampling layers are used after each convolutional layers. The down-sampling factor of the pooling layer is 2. The up-sampling factor of the up-sampling layers is 2. In the last convolutional layers, the 1×1 convolutional filter is utilized to perform a linear transformation of the multi-scale feature maps produced by multi-branch dilated convolution. Compared with traditional convolutional filter, the network parameters of dilated convolutional filter is much fewer. For example, the receptive field of traditional 7×7 convolutional filter is 7×7. In contrast, 3×3 dilated convolutional filter with dilated factors 3 has the same receptive field. The parameter number of each traditional convolution filter is 49. By comparison, the parameter number of each dilated convolutional filter is only 9.

b) Atmospheric light module: Atmospheric light module aims to estimate atmospheric light A in Eq.(1). As shown in Fig. 5, the atmospheric light module consists of one convolutional layer, one sigmoid activation layer, and one pooling layer. W and H are the dimensions of the input image. The size of convolutional filters is $h \times w$, the convolution stride is fixed to 1 pixel, and the padding is 0 pixel. Max-pooling is performed over a $(W - (w - 1)) \times (H - (h - 1))$ window.

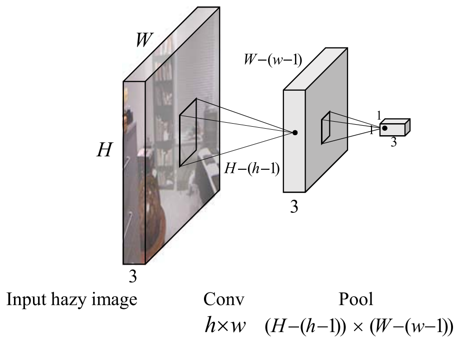


Fig. 5. The architecture of atmospheric light module.

c) Processing module: From Eq.(1), the haze-free image $J(x, y)$ can be formulated as

$$J(x, y) = \frac{I(x, y) - A(1 - t(x, y))}{t(x, y)}. \tag{3}$$

The transmission map module and the atmospheric light module produce the transmission map $t(x, y)$ and the atmospheric light A, respectively. The purpose of processing module is to combine the transmission map $t(x, y)$, the atmospheric light A, and the hazy image $I(x, y)$ to restore haze-free image $J(x, y)$ from Eq. (3).

2) Discriminator Network Architecture: The discriminator network is utilized to distinguish generated haze-free images from realistic images. On the contrary, the generator network is utilized to fool the discriminator network. Following the structure proposed in [32], in this paper, the generator network and discriminator network are alternately updated by solving the min-max problem:

$$\min_{G}\max_{D}\ \mathbb{E}_{J_{real}\sim p_{train}(J_{real})}[\log D(J_{real})] + \mathbb{E}_{I\sim p_G(I)}[\log(1-D(G(I))] \quad (4)$$

where I represents input hazy image, J_{real} is realistic hazy-free image, $G(\bullet)$ represents generator network, and $D(\bullet)$ is discriminator network.

The architecture of discriminator network is shown in the bottom part of Figure 3. It consists of five convolutional layers with 3×3 convolutional filters, LeakyRELU activation layers, batch normalization layers [39], two fully connection layer and sigmoid activation layer. The number of output channels of the five convolutional layers is 64, 64, 128, 256, and 256, respectively. The five convolutional layers are followed by two fully connection layers: the first has 512 channels, the second performs 2-way classification and thus contains 2 channels (one for realistic haze-free image, the other for generated haze-free image).

B. Loss Function

Pixel-wise Euclidean loss, adversarial loss, and perceptual loss are utilized to form the loss function. The loss function L is formulated as

$$L = L_E + \lambda_A L_A + \lambda_P L_P \quad (5)$$

where L_E is pixel-wise euclidean loss, L_A is adversarial loss which is from the discriminator network, L_P represents perceptual loss, and λ_A and λ_P are respectively the weights of adversarial loss and perceptual loss.

1) Euclidean Loss: The pixel-wise euclidean loss is composed of two components, one is the euclidean distance between the generated haze-free images and its corresponding ground-truth images, the other is the euclidean distance between the estimated transmission maps and its corresponding ground-truth transmission maps.

The pixel-wise euclidean loss is calculated as:

$$\begin{aligned} L_E &= L_J + \lambda_t L_t \\ &= \frac{1}{CWH}\sum_{c=1}^{C}\sum_{x=1}^{W}\sum_{y=1}^{H}(G(I)_{c,x,y} - J_{c,x,y})^2 \\ &\quad + \lambda_t \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}(G_t(I)_{x,y} - t_{x,y})^2 \end{aligned} \quad (6)$$

where I is the input hazy image, L_J represents the loss of the haze-free image, L_t represents the loss of the transmission map, and λ_t is the weights of L_t. C, W, and H are the dimensions of the input image. c, x, and y are the location of the input image. And the function $G(\bullet)$ and $G_t(\bullet)$ is to generate the haze-free image and the transmission map, respectively.

2) Adversarial Loss: The task of adversarial loss is to make haze-free images produced by generator network much closer to realistic haze-free images. When training generator network, the min-max problem (4) is reduced to minimize $\log(1-D(G(I)))$. At the beginning of the training stage, $\log(1-D(G(I)))$ could saturate [32]. Because $\log(D(G(I)))$ can provide stronger gradients during training stage, we maximize $\log(D(G(I)))$ to train generator network instead of training generator network to minimize $\log(1-D(G(I)))$. The adversarial loss L_A would be minimized during training stage. For N training images, L_A can be defined as:

$$L_A = \sum_{n=1}^{N} -\log D(G(I_i)) \quad (7)$$

where $D(G(I_i))$ is the probability that the dehazed image $G(I_i)$ is a realistic haze-free image.

3) Perceptual Loss: Perceptual loss based on high-level features extracted from pertained network is wildly used in image super-resolution [40]. In addition, perceptual losses measure image visual similarities more effectively than pixel-wise loss. Inspired by this, in this paper, we define a perceptual loss to increase perceptual similarities between restored haze-free images and realistic images. The perceptual loss can be written as:

$$L_P = \frac{1}{C_f W_f H_f}\sum_{c=1}^{C_f}\sum_{w=1}^{W_f}\sum_{y=1}^{H_f}(\phi(J)_{c,x,y} - \phi(G(I))_{c,x,y})^2 \quad (8)$$

where C_f, W_f and H_f are the dimensions of the respective feature maps within the VGG-16 network [21] and the effect of ϕ is to obtain the feature maps from the VGG-16 networks.

C. Optimization

We optimize the transmission map $t(x,y)$ and atmospheric light A using Stochastic Gradient Descent (SGD) with momentum. The gradients of loss L with respect to transmission map $t(x,y)$ and atmospheric light A are computed respectively as:

$$\begin{aligned} \frac{\partial L}{\partial t(x,y)} &= \frac{\partial L}{\partial J(x,y)}\frac{\partial J(x,y)}{\partial t(x,y)} \\ &= \frac{\partial L}{\partial J(x,y)}\frac{-I(x,y)+A}{t^2(x,y)} \\ \frac{\partial L}{\partial t(x,y)} &= \frac{\partial L}{\partial J(x,y)}\frac{\partial J(x,y)}{\partial A} \\ &= \frac{\partial L}{\partial J(x,y)}\frac{1-A}{t(x,y)} \end{aligned} \quad (9)$$

where $\frac{\partial L}{\partial J(x,y)}$ is calculated in the loss layer. The gradients $\frac{\partial L}{\partial t(x,y)}$ and $\frac{\partial L}{\partial A}$ are passed down to the transmission map module and atmospheric light module respectively to update the network parameters with standard back-propagation.

IV. Exprimental Results

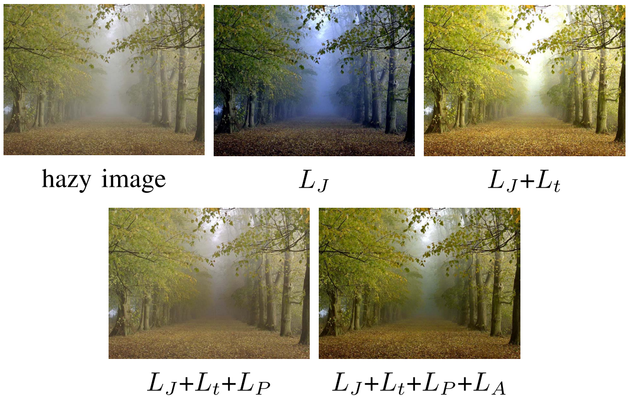
A. Datasets

We synthesize hazy image using haze-free image and its corresponding depth map from the NYU2 Depth dataset [41]. For each image, the depth $d(x,y)$ and scattering coefficient β are used to calculate transmission map $t(x,y)$ using Eq. (2). Next, a haze-free image, the atmospheric light A, and the transmission map $t(x,y)$ are used to

TABLE I

AVERAGE PSNR, SSIM, AND MSE OF HRGAN WITH DIFFERENT LOSS FUNCTION ON INDOOR TEST SYNTHETIC HAZY DATASETS. √ MEANS THAT THE CORRESPONDING LOSS TERM IS USED

L_E			L_P	L_A	PSNR(dB)	SSIM	MSE(10^{-2})
L_J	L_t	L_{al}					
√					21.0843	0.9333	0.9101
√	√				21.2883	0.9373	0.8674
√	√		√		21.8538	0.9457	0.7715
√	√		√	√	22.4068	0.9495	0.6959
√	√	√	√	√	22.3998	0.9489	0.6966



hazy image L_J L_J+L_t

L_J+L_t+L_P L_J+L_t+L_P+L_A

Fig. 6. The dehazed results of HRGAN with different loss function.

synthesize hazy image via the atmospheric scatting model (i.e. Eq. (1)). The atmospheric light A is assumed to be uniform globally. We set the atmospheric light $A = [a, a, a]$, where $a \in [0.7, 1.0]$, and select the scattering coefficient $\beta \in \{0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6\}$. One thousand haze-free images are randomly chosen from the NYU2 Depth dataset. For each hazy-free image, we create ten training images by using randomly sampled scatting coefficient β and atmospheric light A to synthesize hazy images. Finally, we have 10000 training images.

We create an **indoor** test synthetic dataset containing 300 images which is generated by using images and its corresponding depth maps from the Middlebury stereo dataset [42]–[44]. In addition, 500 outdoor synthetic hazy images from SOTS dataset [45] are used as **outdoor** test synthetic dataset. All these test images are **not** used in the training stage.

B. Experiment Settings

We train the networks on an NVIDIA TITANX GPU. The proposed method is implemented using the MatConvNet toolbox [46]. All the training images are resized to 320 × 240. We set the parameters of batch-size, weight decay, and momentum to 10, 0.001, and 0.9, respectively. The initial learning rates of transmission map module and atmospheric light module are 10^{-6} and 10^{-3}, respectively. And the learning rates of both modules decrease by factor of 10 after every 20 epochs. Training stage stops at 80 epochs. The parameters are initialized as follows: $\lambda_t = 1$, $\lambda_A = 10^2$, and $\lambda_P = 5 \times 10^{-4}$. The kernel size $h \times w$ of convolutional layer in atmospheric light module is set to be 15×15. As the same as traditional GAN [32], the generator network and discriminator network are alternately updated.

To quantitatively assess image dehazing methods, three metrics are used to evaluate the performance on synthetic images: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [47], and Mean Squared Errors (MSE). Because there are no ground-truth images for real-world images, the performance on real-world images is evaluated visually and subjectivity.

C. Ablation Study

Table I shows the average PSNR, SSIM, and MSE of HRGAN with the different loss function on indoor test synthetic datasets. √ means that the corresponding loss term is used. L_J is always used. $L_{al} = \frac{1}{C}\sum_{c=1}^{C}(G_{al}(I)_c - A_c)^2$ is added into the Euclidean Loss $L_E = L_J + \lambda_t L_t + \lambda_{al} L_{al}$, where L_{al} represents the loss of atmospheric light A, λ_{al} is the weights of L_{al}, and the function $G_{al}(\bullet)$ is to generate the atmospheric light. There are the following observations from Table I: (1) The transmission map euclidean loss L_t is beneficial to the dehazed results. (2) By utilizing the perceptual loss L_P and the adversarial loss L_A, the PSNR and SSIM becomes higher and the MSE becomes lower. Thus, we know that both the perceptual loss L_P and the adversarial loss L_A can improve dehazed results. (3) The atmospheric light loss L_{al} cannot improve the dehazed results. Because, from Eq. 1, we know the atmospheric light can be solved by the input hazy image, the output dehazed images, and the transmission map. When we supervise the L_J and L_t, L_{al} is supervised implicitly.

The dehazed results with different loss function are shown in Figure 6. It can be observed that the results without L_t have significant color distortions. The reason is that there is a strong correlation between transmission map module and atmospheric light module. An inaccurate transmission map estimation would lead to an inaccurate atmospheric light estimation. An inaccurate atmospheric light estimation tends to change the color of the dehazed result. By observing the last three images in Figure 6, we can find that the effect of L_P and L_A is to make the dehazed result visually appealing.

Table II shows the average PSNR and SSIM of HRGAN with different number of parallel branches and different dilated factors in the transmission maps module. From Table II, we can easily find that compared with the module with five parallel branches, three parallel branches has similar SSIM and PSNR. However, the module with five parallel branches has more number of parameters and longer running time. Compared with the module with one parallel branch, the dehazed performance of the module with three parallel branches is much better. Thus, the module with three branches makes a good balance of dehazed performance and speed. From Table II, we can find that with the increase of dilated factors, the dehazed results would become worse. The main reason is that too large dilated factor would lead to block artifacts.

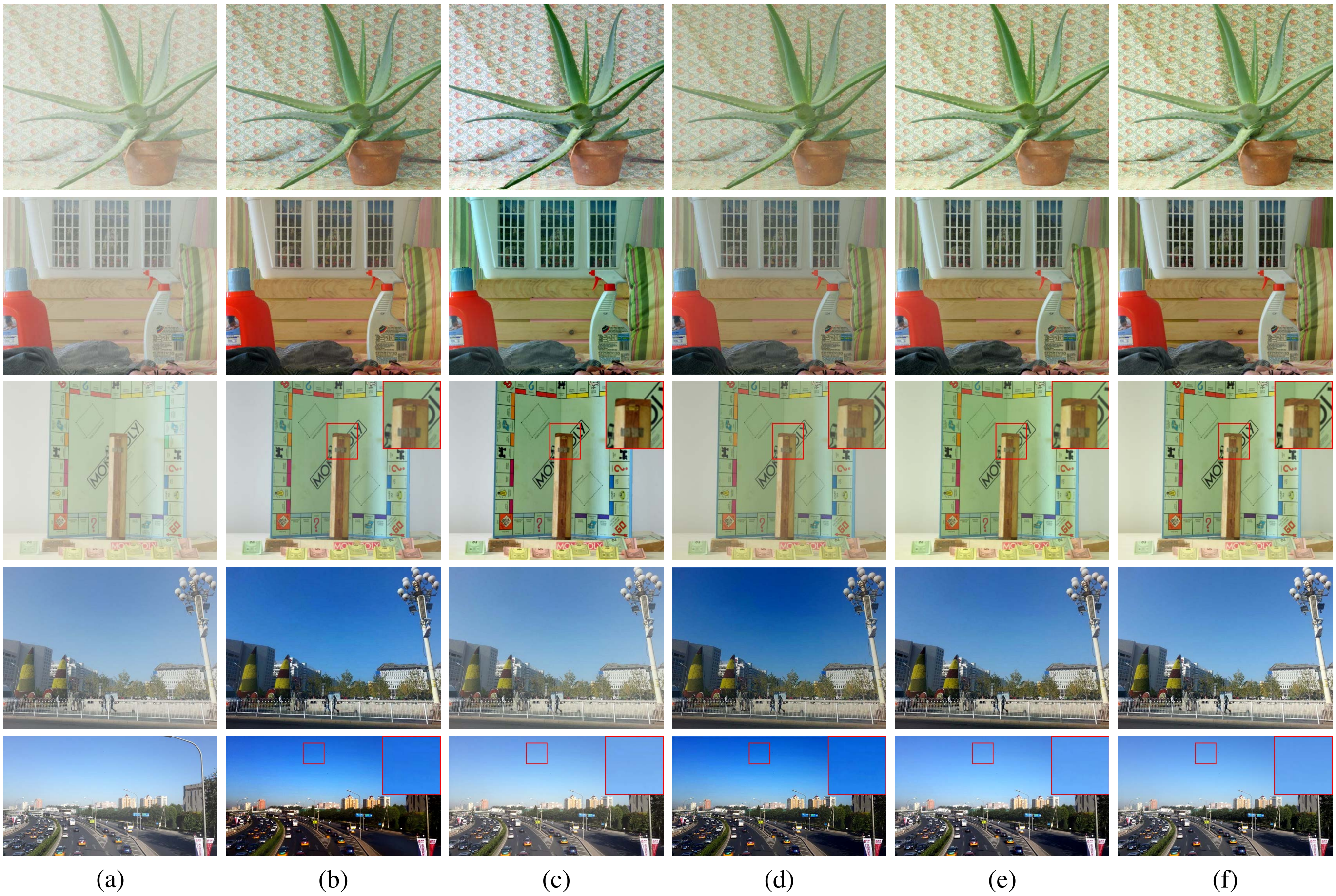


Fig. 7. Comparison of different methods on test synthetic hazy images: **First:** Aloe. **Second:** Laundry. **Third:** Monopoly, **Fourth:** Buildings, **Fifth:** Road. The first three hazy images are from indoor test synthetic hazy datasets, the last two hazy images are from outdoor test synthetic hazy datasets. (a) Synthetic hazy images. (b) CAP [30]. (c) MSCNN [16]. (d) AOD-Net [20]. (e) HRGAN. (f) Ground-truth images. (Red rectangles in the top-right corner is the zoom-in views.)

TABLE II

AVERAGE PSNR AND SSIM OF HRGAN WITH DIFFERENT NUMBER OF PARALLEL BRANCHES AND DIFFERENT DILATED FACTORS IN THE TRANSMISSION MAPS MODULE ON INDOOR TEST SYNTHETIC HAZY DATASETS. THE NUMBER IN THE **BRACKET** IS THE DILATED FACTORS. FOR EXAMPLE, (1, 2, 3) MEANS THE MODULE CONSISTS OF THREE PARALLEL BRANCHES, THE DILATED FACTORS OF PARALLEL BRANCHES IS 1, 2, AND 3

Dilated Factors	PSNR(dB)	SSIM
(1)	21.01	0.9255
(1,2,3)	22.41	0.9495
(1,3,5)	22.33	0.9451
(1,5,9)	21.65	0.9355
(1,2,3,4,5)	22.39	0.9491

TABLE III

AVERAGE PSNR AND SSIM OF THE PROPOSED HRGAN WITH DIFFERENT WEIGHTS OF LOSS FUNCTION ON INDOOR TEST SYNTHETIC HAZY DATASETS

λ_P	λ_A	PSNR(dB)	SSIM
5×10^{-4}	10^2	22.41	0.9495
1	10^2	20.55	0.9213
5×10^{-4}	1	21.02	0.9287

Table III shows the average PSNR and SSIM of HRGAN with different loss weights on indoor test synthetic hazy datasets. From Table III, we can easily find that with increase of λ_P, the value of PSNR and SSIM would be decrease. The reasons is that instead of reducing the difference of the pixel-level, the perceptual loss is used to reduce the difference of high-frequency information. As we know, the effect of discriminator network is to make the generated dehazed images more similar to ground-truth. Therefore, with the increase λ_A, the value of PSNR and SSIM is increase.

D. Quantitative Results on Synthetic Images

Table IV and Table V compares our proposed HRGAN with DCP [24], BCCR [25], CAP [30], NLD [31], MSCNN [16], AOD-Net [20] in terms of PSNR and SSIM on indoor test synthetic hazy datasets and outdoor test synthetic hazy datasets, respectively.

It can be observed from Table IV and Table V that our proposed HRGAN rank first in terms of PSNR and SSIM on both two datasets. As described in Section III, HRGAN is optimized by an effective loss consisting of pixel-wise loss, perceptual loss, and adversarial loss. By minimizing pixel-wise loss, HRGAN can get high PSNR performance. The perceptual loss and adversarial loss can make HRGAN get great SSIM.

Figure 7 shows the dehazed results produced by different methods on test synthetic hazy datasets. Figure 7(a) presents

TABLE IV

AVERAGE PSNR AND SSIM ON **INDOOR** TEST SYNTHETIC HAZY DATASETS

Metrics	DCP [24]	BCCR [25]	CAP [30]	NLD [31]	MSCNN [16]	AOD-Net [20]	HRGAN
PSNR(dB)	18.03	16.14	21.33	18.10	19.46	20.47	**22.41**
SSIM	0.7595	0.7225	0.8859	0.7642	0.8037	0.9363	**0.9495**

TABLE V

AVERAGE PSNR AND SSIM ON **OUTDOOR** TEST SYNTHETIC HAZY DATASETS

Metrics	DCP [24]	BCCR [25]	CAP [30]	NLD [31]	MSCNN [16]	AOD-Net [20]	HRGAN
PSNR(dB)	18.54	17.71	23.95	19.52	21.73	24.08	**25.84**
SSIM	0.7100	0.7236	0.8692	0.7328	0.8313	0.8726	**0.9214**

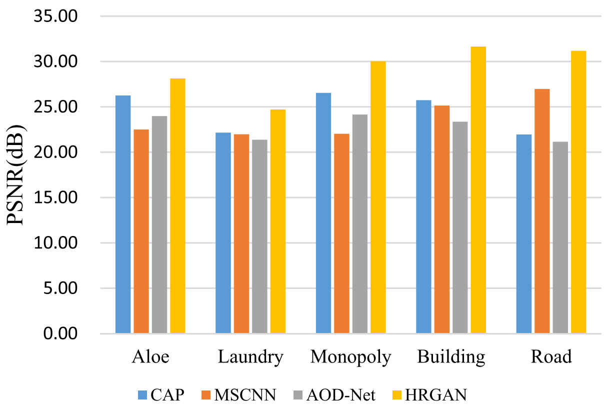


Fig. 8. PSNR of the dehazed images shown in Figure 7.

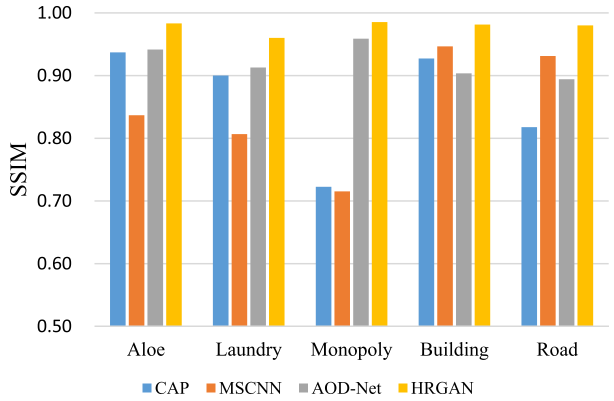


Fig. 9. SSIM of the dehazed images shown in Figure 7.

the hazy images which are from the test synthetic datasets. Figure 7(b)-7(e) shows the results of CAP [30], MSCNN [16], AOD-Net [20], and our proposed HRGAN, respectively. Figure 7(f) gives the ground-truth images.

By observing the dehazed results in Figure 7(b), we can find that the dehazed results generated by CAP have some color distortions (*e.g.*, the fourth and fifth line in Figure 7(b)). We note that the dehazed results of MSCNN have some remaining haze by observing the dehazed results in Figure 7(c). We can find that the dehazed results of AOD-Net have some remaining haze (*e.g.*, the first and second line in Figure 7(e)), and some color distortions (*e.g.*, the sky in the fourth and fifth images). In contrast, the dehazed results of our proposed HRGAN in Figure 7(f) is more visually appealing and closer to ground-truth haze-free images.

Figure 8 and Figure 9 show the PSNR and SSIM of the dehazed results produced by different algorithms on the five images in the Figure 7. It can be easily found that HRGAN achieves the greatest PSNR and SSIM for all the five images.

In summary, our proposed HRGAN achieves the best performance subjectively and objectively against the state-of-art dehazing methods on both indoor and outdoor synthetic hazy images.

E. Qualitative Results on Real-World Images

Figure 10 demonstrates the dehazed hazy-free images and transmission maps restored by HRGAN. Because most of the image haze removal algorithms can obtain nice visual performance on general real-world images, it is difficult to rank them. To demonstrate the superiority of our method, we evaluate our proposed algorithm against the state-of-art image haze removal algorithms (CAP [30], DehazeNet [15], MSCNN [16], AOD-Net [20]) using five highly challenging real-world image shown in Figure 11.

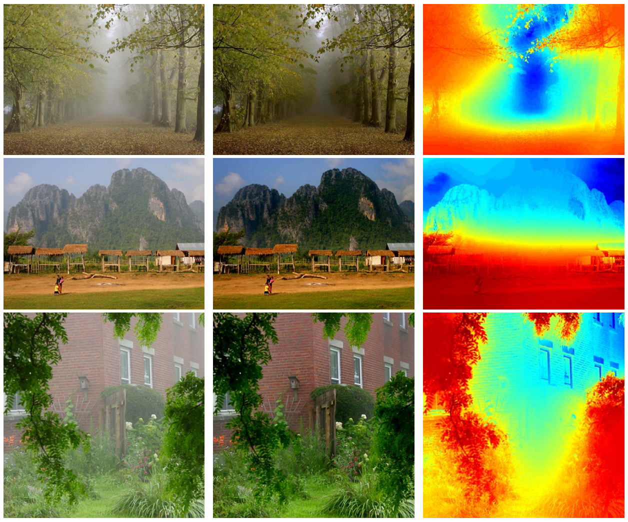


Fig. 10. The dehazed results of HRGAN. Left: input hazy images. Middle: the restored haze-free images. Right: the restored transmission maps. (Best viewed in color).

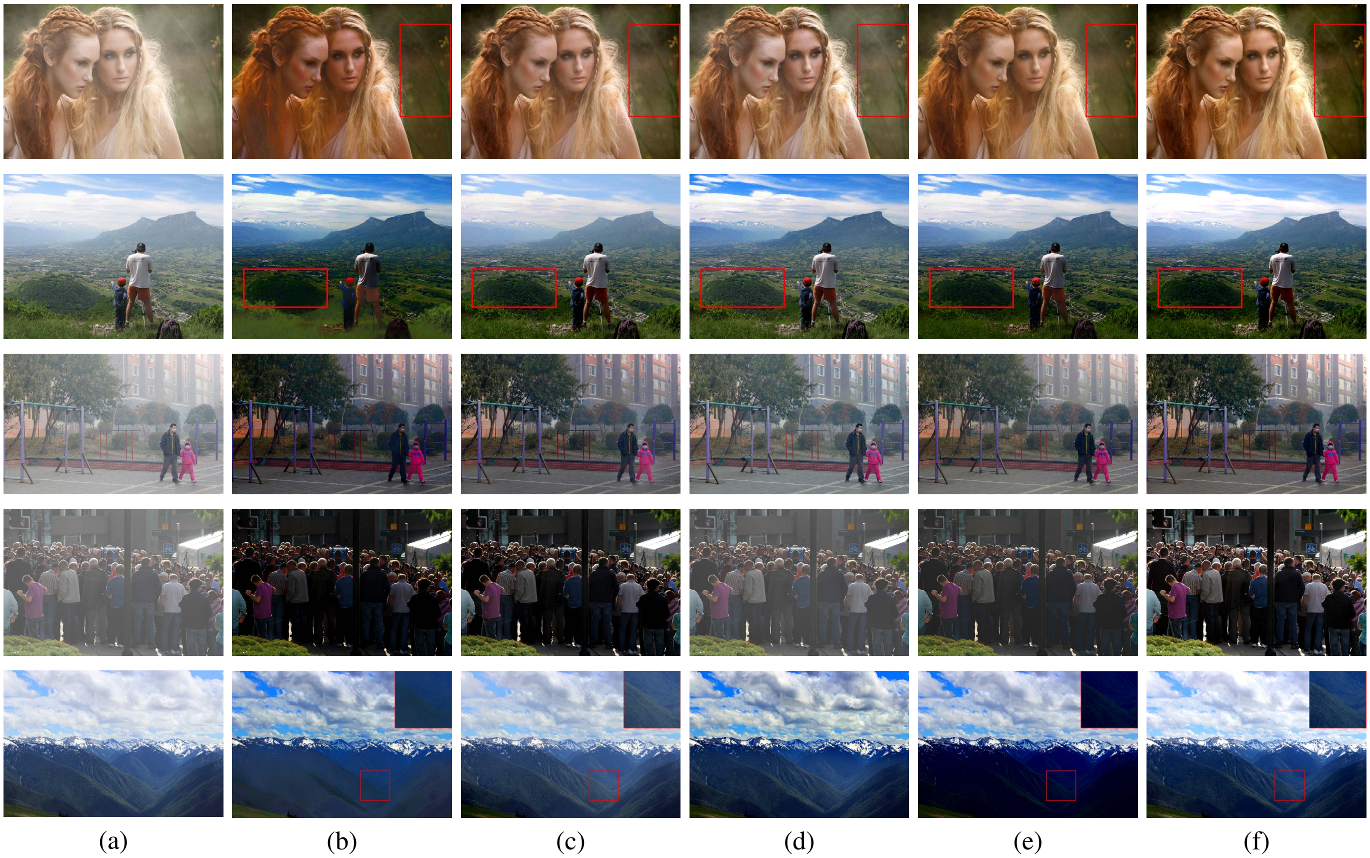


Fig. 11. Comparison of different methods on real-world images. (a) The hazy images. (b) CAP [30]. (c) DehazeNet [15]. (d) MSCNN [16]. (e) AOD-Net [20]. (f) HRGAN. (Red rectangles are used to highlight the improvements obtained by HRGAN. Red rectangles in the top-right corner is the zoom-in views).

The blind image quality assessment (BIQA) models can be used to evaluate the performance of dehazed results of real-world images [48]–[53]. However, the current image quality models are mainly designed for degraded images, so the evaluation performance of dehazed results is unsatisfactory. Therefore, the BIQA models are not used to evaluate the dehazed results in our paper.

As shown in figure 11(b), CAP may blur image textures (e.g., as shown in the fifth line of Figure 11(b), the details of the mountain are lost). And shown in the fourth line of Figure 11(b), the dehazed result is much darker than it should be. DehazeNet produces undesirable results in regions with heavy hazes (e.g., as shown in the second and fifth line of Figure 11(c), there are remaining haze in the region of distant mountains). As show in the second, third, and fourth line of MSCNN, the dehazed results of MSCNN have some remaining haze. In addition, as shown in the fifth line of Figure 11(d), the colors of the sky region are over saturated. The dehazed results of AOD-Net [20] sometimes may result in color distortion (e.g., as illustrated in the fifth line of Figure 11(e), the mountain region is much darker than it should be). In addition, there are some remaining haze in the third and fourth line of Figure 11(e)). In contrast, the dehazed results of HRGAN (shown in Figure 11(f)) achieve higher visual quality and less color distortions.

CAP [30] is based on the handcrafted features. Because handcrafted features are weak to perform image dehazing, the dehazed results are not satisfactory. Compared with handcrafted features, the features learned by CNN-based method [15], [16] include more various kinds of information. However, the effective features are only used to estimate transmission maps instead of producing haze-free image. For this reason, as stated in Section I, the dehazed results of these two CNN-based methods are not optimal. Because the AOD-Net optimize the network by minimizing only pixel-wise loss, the dehazed results cannot achieve high visual quality.

F. Running Time

Efficiency is important for a computer vision system [54], [55]. The running time comparison on CPUs with DCP [24] (accelerated by the guided image filtering [56]), BCCR [25], CAP [30], DehazeNet [15], MSCNN [16] and our proposed HRGAN is shown in Table VI. All the methods are implemented in MATLAB, on the same machine (Intel(R) Core(TM) i7-4790 CPU @3.60GHz, and 16 GB memory). It can be seen from Table VI that our proposed HRGAN is much faster than other methods. In addition, AOD-Net [20] is implemented in PyCaffe. With four different image resolution 640×480, 800×600, 1024×768, and 1600×1200, AOD-Net costs 1.108, 1.72, 3.252, and 6.298 seconds, respectively. It can be observed in Table VI that our proposed HRGAN

TABLE VI

COMPARISON OF AVERAGE RUNNING TIME ON CPUs (IN SECONDS)

Image Resolution	DCP [24]	BCCR [25]	CAP [30]	DehazeNet [15]	MSCNN [16]	HRGAN
640×480	2.89	1.93	0.85	1.21	1.36	**0.39**
800×600	4.50	2.76	1.31	1.89	1.92	**0.59**
1024×768	7.59	4.481	2.15	3.41	3.33	**1.00**
1600×1200	18.34	10.43	5.23	9.21	6.69	**2.50**

TABLE VII

COMPARISON OF AVERAGE RUNNING TIME ON GPUs (IN SECONDS)

Image Resolution	MSCNN [16]	AOD-Net [20]	HRGAN
640×480	0.932	0.079	**0.077**
800×600	1.332	0.114	**0.092**
1024×768	1.985	0.154	**0.114**
1600×1200	3.663	0.298	**0.204**

costs less than half of the running time of AOD-Net. The running time comparison on GPUs with MSCNN [16], AOD-Net [20], and our proposed HRGAN is shown in Table VII. All the results are tested on the NIVIDIA TITANX. It can be found that our method is faster than other CNN-based methods, especially the size of input image is large. The number of parameter of DehazeNet, MS-CNN, AOD-Net and our proposed HRGAN is 8.2K, 8.0K, 1.7K, and 3.5K, respectively. It can be found that compared with DehazeNet and MS-CNN, the model size of our proposed HRGAN is smaller. In addition, from Table VI and VII, we can know our proposed HRGAN is faster than DehazeNet and MS-CNN. Although the model size of our proposed HRGAN is bigger than AOD-Net, the running time (shown in VI and VII) of our proposed HRGAN is faster than AOD-Net on both CPUs and GPUs. The high efficiency of HRGAN mainly benefits from the fact that the atmospheric light module based on light-weight CNN significantly simplifies the estimation of atmospheric light.

V. CONCLUSION

In this paper, we have proposed a GAN-based image haze removal network called HRGAN. HRGAN consists of two networks: a generator network and a discriminator network. The generator network of HRGAN is a unified network jointly estimating transmission map, atmospheric light, and haze-free image. Apart from pixel-wise loss, adversarial loss produced by the discriminator network and perceptual loss are utilized in optimization task. Experimental results demonstrate that HRGAN achieves remarkably high efficiency and outperforms state-of-art methods on both synthetic and real-world images.

REFERENCES

[1] Z. Li, J. Zhang, K. Zhang, and Z. Li, "Visual tracking with weighted adaptive local sparse appearance model via spatio-temporal context learning," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4478–4489, Sep. 2018.

[2] J. Cao, Y. Pang, and X. Li, "Learning multilayer channel features for pedestrian detection," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3210–3220, 2017.

[3] X. Lian, Y. Pang, and A. Yang, "Learning intensity and detail mapping parameters for dehazing," *Multimedia Tools Appl.*, vol. 77, no. 12, pp. 15695–15720, 2018.

[4] X. Fan, Y. Wang, X. Tang, R. Gao, and Z. Luo, "Two-layer Gaussian process regression with example selection for image dehazing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2505–2517, Dec. 2017.

[5] Y. Pang, L. Ye, X. Li, and J. Pan, "Incremental learning with saliency map for moving object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 640–651, Mar. 2018.

[6] Y. Niu, H. Zhang, W. Guo, and R. Ji, "Image quality assessment for color correction based on color contrast similarity and color value difference," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 4, pp. 849–862, Apr. 2018.

[7] M. Ju, C. Ding, D. Zhang, and Y. J. Guo, "BDPK: Bayesian dehazing using prior knowledge," *IEEE Trans. Circuits Syst. Video Technol.*, to be published. [Online]. Available: https://ieeexplore.ieee.org/document/8464077, doi: 10.1109/TCSVT.2018.2869594.

[8] C.-H. Son and X.-P. Zhang, "Near-infrared fusion via color regularization for haze and color distortion removals," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3111–3126, Nov. 2018.

[9] W. Ren *et al.*, "Gated fusion network for single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.

[10] J.-M. Guo, J.-Y. Syue, V. R. Radzicki, and H. Lee, "An efficient fusion-based defogging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4217–4228, Sep. 2017.

[11] T. M. Bui and W. Kim, "Single image dehazing using color ellipsoid prior," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 999–1009, Feb. 2018.

[12] J.-P. Tarel and N. Hautière, "Fast visibility restoration from a single color or gray level image," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2201–2208.

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

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

[15] 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.

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

[17] H. Sun and Y. Pang, "GlanceNets—Efficient convolutional neural networks with adaptive hard example mining," *Sci. China Inf. Sci.*, vol. 61, no. 10, p. 109101, 2018.

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

[19] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 105–114.

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

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[22] S. G. Narasimhan and S. K. Nayar, "Chromatic framework for vision in bad weather," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2000, pp. 598–605.

[23] S. K. Nayar and S. G. Narasimhan, "Vision in bad weather," in *Proc. IEEE Conf. Comput. Vis.*, vol. 2. Sep. 1999, pp. 820–827.

[24] 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.

[25] 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.*, Dec. 2013, pp. 617–624.

[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.*, Jun. 2014, pp. 2995–3002.

[27] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

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

[29] 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. Asian Conf. Comput. Vis.*, 2010, pp. 501–514.

[30] 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.

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

[32] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[33] M. Mirza and S. Osindero. (2014). "Conditional generative adversarial nets." [Online]. Available: https://arxiv.org/abs/1411.1784

[34] 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.*, Jul. 2017, pp. 5967–5976.

[35] H. Zhang *et al.*, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5908–5916.

[36] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2536–2544.

[37] F. Yu and V. Koltun. (2015). "Multi-scale context aggregation by dilated convolutions." [Online]. Available: https://arxiv.org/abs/1511.07122

[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (2016). "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs." [Online]. Available: https://arxiv.org/abs/1606.00915

[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[40] J. Justin, A. Alexandre, and F.-F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

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

[42] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2003, pp. 195–202.

[43] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[44] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[45] B. Li *et al.* (2017). "Benchmarking single image dehazing and beyond." [Online]. Available: https://arxiv.org/abs/1712.04143

[46] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 689–692.

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

[48] K. Ma, W. Liu, and Z. Wang, "Perceptual evaluation of single image dehazing algorithms," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 3600–3604.

[49] Q. Wu *et al.*, "Blind image quality assessment based on multichannel feature fusion and label transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 425–440, Mar. 2016.

[50] J. Yao, W. Lu, L. He, and X. Gao, "Rank learning for dehazed image quality assessment," in *Proc. CCCV*, 2017, pp. 295–308.

[51] Q. Wu *et al.*, "Blind image quality assessment based on rank-order regularized regression," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2490–2504, Nov. 2017.

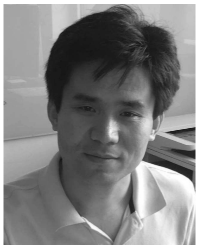
[52] Z. Chen, T. Jiang, and Y. Tian, "Quality assessment for comparing image enhancement algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3003–3010.

[53] Q. Wu, H. Li, K. N. Ngan, and K. Ma, "Blind image quality assessment using local consistency aware retriever and uncertainty aware evaluator," *IEEE Trans. Circuits Syst. Video Technologe*, vol. 28, no. 9, pp. 2078–2089, Sep. 2018.

[54] Y. Pang, J. Cao, and X. Li, "Cascade learning by optimally partitioning," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4148–4161, Dec. 2017.

[55] Y. Pang, J. Cao, and X. Li, "Learning sampling distributions for efficient object detection," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 117–129, Jan. 2017.

[56] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.



Yanwei Pang (M'07–SM'09) received the Ph.D. degree in electronic engineering from the University of Science and Technology of China in 2004. He is currently a Professor with Tianjin University, China. His current research interests include object detection, image recognition, image processing, deep learning, and their applications in self-driving cars, unmanned surface vessel, visual surveillance, human–machine interaction, and biometrics, in which areas he has published more than 100 scientific papers including more than 32 IEEE transactions papers.

Jin Xie received the B.S. degree in electronic engineering from Tianjin University, Tianjin, China, in 2016, where he is currently pursuing the Ph.D. degree under the supervision of Prof. Y. Pang. His research interests include machine learning and computer vision.

Xuelong Li (M'02–SM'07–F'12) is a Full Professor with the School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China.