Guided-Pix2Pix: End-to-End Inference and Refinement Network for Image Dehazing

Libin Jiao, Changmiao Hu, Lianzhi Huo, and Ping Tang

***Abstract*—Haze removal is still an essential prerequisite for image processing and computer vision tasks, and joint inference and refinement of transmission maps remain challenging in the physical scattering model-based haze removal methods. In this article, we propose an end-to-end learnable dehazing network, which is referred to as Guided-Pix2Pix, to jointly estimate and refine the transmission map and further dehaze images by the physical scattering equation. Instead of a two-stage model of predicting and postprocessing the transmission, Guided-Pix2Pix concatenates the trainable Pix2Pix backbone and differentiable guided filter as an embedded layer, which enables generating refined transmission maps in one feed-forward step, and then it substitutes these potential refinements into the physical scattering equation to restore dehazed images. To verify that our Guided-Pix2Pix can be embedded in both training and inference, we demonstrate that the guided filter layer is differentiable and capable of propagating both features forward and gradients backward. Furthermore, explicit derivatives with respect to the input of the guided filter are given, and the relationship between our derivation and that in the guided filter is also explored. Experiments show that our network is effective and robust in image dehazing, can alleviate the halo artifacts along edges, and has great generalization capability.**

***Index Terms*—Differentiable guided filter, end-to-end refinement of transmission map, image dehazing.**

I. Introduction

OUTDOOR images are usually contaminated by the turbid medium in the atmosphere [1], [2]. Optically, the light from the atmosphere or the surface of an object can be absorbed or scattered by the floating particles, which leads to the degradation of visibility. Consequently, the contrast and color fidelity are decayed within the degraded images. On the other hand, haze-free visibility is essentially required by automatic systems, including surveillance, intelligent vehicles, and outdoor object recognition [3], which makes dehazing an inevitable preprocessing. Therefore, haze removal is necessary for image processing and computer vision applications [1], [2].

Simply, the haze contamination can mathematically be described in (1), which is referred to as the physical scattering equation

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

in which J denotes the scene radiance, $t \in [0, 1]$ the medium transmission map, A the global atmospheric light, z the pixel coordinate, and I the observed intensity. The transmission map t describes the light portion that is not scattered and finally reaches the camera. t can be described by (2) if the atmospheric light A is homogenous

$$t(z) = e^{-\beta d(z)} \tag{2}$$

in which β is the scattering coefficient of the atmosphere and d is the function of the scene depth.

Apparently, haze removal for a single image is a challenging issue because estimating the two factors in the physical scattering equation is notoriously ill-posed. Atmospheric light can be simply estimated from the hazy image; for example, it can be given by the pixel with the highest color intensity [3]. But transmission maps are highly relevant to the precise or rough unknown depth information [1]. Therefore, many common methods focus on additional information or prior knowledge. For example, the dark channel prior [1], [2] assumed that there is at least one channel that is very dark in an outdoor image, and Fattal's method [4] assumed that the transmission and surface shading can be locally decomposed. The prior-based methods can be physically sound but may fail once the assumption is violated.

On the other hand, learning-based methods have made a striking achievement because of the utilization of large-scale labeled datasets and advanced informative feature extraction [5]–[7]. Many CNN-based structures [8]–[11] capture hierarchical representative features and impressively restore images; they have shown promising generalization capability. Either directly translating hazy-to-clear images or jointly estimating atmospheric light and transmission maps, learning-based methods can usually perform acceptably compared to hand-crafted prior-based methods if they have been well-trained on the high-quality hazy/clear datasets.

In addition, spatial refinement is necessary for dehazing methods when the transmission map is coarsely predicted. Transmission maps should match the spatial structure of hazy images, i.e., they should share similar contours and edges. Spatial refinement as postprocessing has been commonly used and is popularized by edge-preserving filters [12] and MRF-based methods [3], [13]. End-to-end refinement has been preliminarily explored in DCPDN [5] by minimizing gradient discrepancy. Therefore,

Manuscript received November 2, 2020; revised January 2, 2021 and February 3, 2021; accepted February 18, 2021. Date of publication February 23, 2021; date of current version March 22, 2021. This work was supported in part by the China Postdoctoral Science Foundation under Grant 2019M660852, in part by the Special Research Assistant Foundation of CAS, and in part by the National Natural Science Foundation of China under Grant 41971396. *(Corresponding author: Ping Tang.)*

The authors are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China (e-mail: jiaolb@aircas.ac.cn; hucm@aircas.ac.cn; huolz@aircas.ac.cn; tangping@aircas.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3061460

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

jointly estimating and refining transmission maps can be further explored.

An end-to-end learning-based dehazing model is expected as it facilitates fitting the haze distribution and the spatial structure of the hazy image. To this end, we propose an end-to-end CNN-based network in this article to achieve learning-based image dehazing, in which transmission maps can be jointly estimated and refined, and further dehazed images can be accordingly restored by the physical scattering equation. The dehazing network is derived from the deep guided filter [14], but the perceptual and adversarial losses are employed to train the network, which is our distinct contribution. In practice, we customize the Guided-Pix2Pix network that concatenates the trainable Pix2Pix backbone and differentiable guided filter as an embedded layer. In particular, coarse transmission maps can be generated by the trainable Pix2Pix backbone, and to achieve the end-to-end refinement, a differentiable guided filter is embedded into the network as a layer, which has been proved to be capable of propagating features forward and gradients backward. The atmospheric light is predicted by the pixel with the highest color intensity, and the dehazed image can be finally restored by the physical scattering equation, given the transmission map and the atmospheric light. In the training phase, the L1 loss is used as the perception loss, and the Markovian discriminator is employed in adversarial training to improve the visual quality of the dehazed images. Furthermore, explicit derivatives with respect to the input of the guided filter are given, and the relationship between our derivation and that in the guided filter is also explored. Our method can improve the visibility of hazy images, alleviate the halo artifacts along edges, and is robustly generalized to other hazy images, in terms of our thorough experiments. Therefore, our main contributions in this article are listed as follows.

1) An end-to-end transmission map estimation and haze removal network is proposed. Our network can jointly generate transmission maps and dehazed images instead of directly yielding clear images, which is able to alleviate halo artifacts in terms of visual assessment.
2) End-to-end transmission map refinement is embedded into the network. A differentiable guided filter is employed as a layer to refine the transmission map estimation yielded by the backbone network, which significantly improves the visual dehazing performance and enhances the generalization of the network. The perceptual and adversarial losses are employed to promote the net to generate realistic images, which is the distinct contribution against the vanilla guided filter [12] and the deep guided filter [14].
3) Explicit derivatives with respect to the input of the guided filter layer are given in this article. Furthermore, the relationship between our derivation and that in [12] is explored in appendix, which is supplementary for [12].

The rest of the article is organized as follows. Section II reviews related work regarding prior-based and learning-based dehazing methods. The proposed Guided-Pix2Pix is described in Section III. Section IV presents the experiments on several benchmark datasets, including RESIDE-SOTS indoor/outdoor, RICE, and I/O-HAZE datasets, and visual and quantitative evaluations are thoroughly performed, including comparisons against state-of-the-art methods, the effect of the guided filter layer, the hyperparameter sensitivity test with respect to radius r and ϵ, and the generalization capability of the method. Section V concludes this article.

II. Related Work

We review the haze removal techniques in this section, which can be categorized as image prior-based and learning-based methods. Additionally, we also review some typical methods of refining the transmission.

A. Image Prior-Based Methods

Dehazing methods attempt to remove haze by maximizing the color contrast and restraining the oversaturation of the color intensity in the hazy image. Image prior-based methods tend to decomposite the physical scattering equation and remove the haze under the predefined assumption. Typical dehazing methods focus on adjusting the color intensity or estimating the transmission. Fattal [4] estimated the scene albedo and further inferred the medium transmission by decomposing the transmission and surface shading. The dark channel prior (DCP) method [1], [2] assumed that at least one channel had very low color intensity at some pixels and inferred the transmission accordingly, and the color ellipsoid prior method [15] formally explained the color intensity distribution of DCP. Tan *et al.* [3] maximized the local contrast to remove haze and build an MRF to smooth the estimation. Kratz and Nishino [16] inferred the scene albedo and the depth field with a factorial MRF. Color-line [17] was proposed to dehaze in terms of the 1-D color distribution within a small patch. Berman *et al.* [18] characterized haze-free images by a nonlocal patch prior. Prior-based methods can usually achieve promising results but they can work under their strict assumptions, which leads to unstable performance.

B. Learning-Based Methods

Learning-based methods build a mapping between hazy and haze-free images by brewing trainable models. They usually tend to optimize the model globally or locally on the labeled dataset. Typical learning-based methods can be categorized as hazy-to-clear methods and physical model-based methods in terms of directly generating clear images or estimating factors in the physical scattering equation (transmission maps and the atmospheric light). Hazy-to-clear methods attempt to disentangle dehazing from the physical scattering model, and they directly output clear images: end-to-end trainable CNN-based models employ a variety of state-of-the-art convolutional techniques to achieve desirable results, including DehazeNet [19], gated fusion network [20], and EPDN [7]. On the other hand, some learning-based methods aim to jointly estimate transmission maps and atmospheric light or only the former and remove haze by the physical scattering equation. Ren *et al.* [21] proposed a multiscale network image to yield transmission maps. Yang *et al.* [22] used physics-based disentanglement and adversarial training. DCPDN [5] predicted atmospheric light and transmission map together with two individual components, while AOD-Net [23] jointly estimated one factor merging atmospheric

light and transmission map. Proximal Dehaze-Net [24] unfolded the optimization of the dark channel and transmission priors to a network by proximal operators. LAP-Net [6] progressively dehazed images by fusing haze results at different stages. Naturally, learning-based methods depend definitely on the large-scale labeled dataset that affects positively the quality of dehazing.

Interestingly, generative adversarial nets (GAN) [25] have been commonly used in image processing, including image-to-image translation [10], [11], [26], single image super-resolution [27]–[31], and image inpainting [32]–[35]. Particularly, adversarial training based on the minimax game of GAN is able to improve the visual quality of counterfeited photo-realistic images, which has been an essential property used in image generation tasks. Some typical improvements on GAN focus on stabilizing its training, including LSGAN [36], Wasserstein GAN [37], [38], improved Wasserstein GAN [39], and SN-GAN [40]. Progressive GAN [41] trained GAN in a progressive fashion, which enabled high-resolution image generation with high visual and quantitative quality. BigGAN [42] was trained on a large-scale dataset and applied orthogonal regularization to the generator, which achieved the state of the art in class-conditional image synthesis. Accordingly, the Markovian discriminator [10] is employed in our method to improve the visual quality of dehazing images.

C. Transmission Map Refinement

Refining transmission is necessary for the physical model-based methods as the coarse transmission without refinement can cause a mismatch on the edges within the images. Methods of transmission refinement aim to convey the spatial structure of hazy images to the transmission. He and Siu [2] applied the soft matting algorithm [13] to the DCP method, and further, they tested the performance of the guided filter [12] in the refinement. A Markovian random field [43] was employed in [3]. Learning-based methods also involve similar strategies, including gradient discrepancy DCPDN [5] and the deep guided filter [14]. Our differentiable guided filter is derived from DGF [14] and embedded into the dehazing model to perform end-to-end dehazing.

III. Methodology

We introduce the overview of our physical scattering model-based dehazing network, its backbone net, feed forward and backward propagation of the differentiable guided filter layer, and adversarial training in this section.

A. Overview of the Dehaze Model

The haze contamination within an image can be mathematically formulated by

$$I(z) = J(z)t(z) + A(1 - t(z)) \tag{3}$$

in which J denotes the scene radiance, $t \in (0, 1]$ the medium transmission map, A the global atmospheric light, z the pixel coordinate, and I the observed intensity.

The ultimate goal of haze removal is to directly restore the haze-free image J from the given I or to restore J by (4) after estimating t and A

$$J(z) = \frac{I(z) - A(1 - t(z))}{t(z)} = \frac{I(z) - A}{t(z)} + A. \tag{4}$$

As shown in (4), physical model-based haze removal is highly ill-posed because the number of unknown variables (t, A, and J) is much more than the number of given variables (I). For brevity, the atmospheric light A can be roughly obtained from pixels that have the highest intensity in I, as is described in [3]. Specifically, the Y channel of the image is calculated from the RGB channels in I, described in

$$Y(z) = 0.299R(z) + 0.587G(z) + 0.114B(z). \tag{5}$$

The index of the pixel with the highest intensity is selected from the Y channel, described in

$$z^* = \arg\max_z Y(z). \tag{6}$$

The atmospheric light A is obtained at the pixel z^*

$$A = [R(z^*), G(z^*), B(z^*)]^\top. \tag{7}$$

There exist typical approaches to learn the mapping between the pairs of hazy and dehazed images, or hazy and transmission images. The approach building the haze-to-transmission mapping requires 1) estimating transmission maps from hazy images, and 2) refining the transmission maps to fit the spatial structure of hazy images. We, therefore, formulate the aforementioned model by an end-to-end model, which consists of the backbone net and the following differential guided filter.

B. Backbone of Dehaze Net

The estimator of transmission maps is derived from Pix2Pix [10], which consists of blocks of "convolution—instance normalization—leaky ReLU" in the encoder and blocks of "transposed convolution—instance normalization—Dropout (in the first three blocks)—ReLU" in the decoder. Skip connections are built between the encoder and decoder by concatenating feature maps with the same dimension, as is used in U-Net [44]. For brevity, we refer to the Pix2Pix backbone as G_0. Please refer to [10] for more details.

C. Revisit Feed-forward Propagation of Differentiable Guided Filter

As shown in Fig. 1, the differentiable guided filter follows the backbone net of Pix2Pix as a layer in our model, which enables refinement of the coarse estimations of transmission maps in a one-stage way. It is noted that the differentiable guided filter layer is derived from the guided filter [12] and the deep guided filter [14] but without any learnable parameters. Therefore, no corresponding gradients need to be given for this layer. The differentiable guided filter is notably different from its counterparts in [14] and [12]: the deep guided filter introduced trainable parameters into the guided filter. The vanilla guided filter is not differentiable and is usually employed as a postprocessing tool to refine the spatial structure. On the contrary, the differentiable guided filter in our model is appended to the backbone, which forms an end-to-end dehazing model and leads to dehazing in

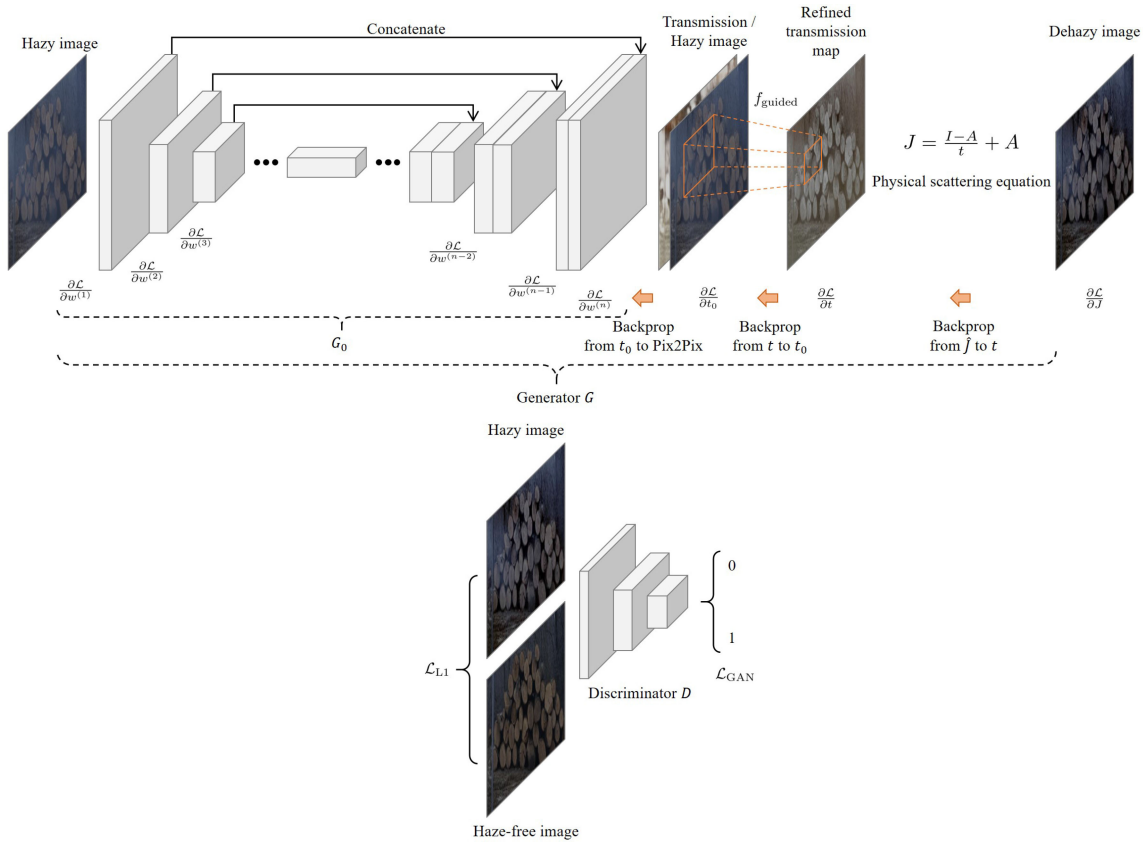

Fig. 1. Framework of Guided-Pix2Pix. G_0 and f_{guided} denote the Pix2Pix backbone and the differentiable guided filter layer, which will be fully described in Sections III-B, III-C, and III-D. In the feed-forward propagation of the training phase, our model first estimates the transmission maps with coarse spatial structures and then refines them with the differentiable guided filter layer. Dehazed images are later restored by the physical scattering equation. Joint L1 loss and adversarial loss between the dehazed and haze-free images are used to promote visual perception and genuineness of dehazing performance. In the backward propagation, the derivatives of the learnable parameters are propagated through the physical scattering equation, the differentiable guided filter layer, and reach the Pix2Pix backbone finally. In the inference phase, the transmission map is estimated, refined, and used in the scattering equation for dehazing.

a one-stage way. Experiments illustrate that our methods can effectively alleviate the halo artifacts between the edges of bright and dark pixels, in comparison against given common methods. Additionally, our dehazing model is quite tolerant to the adversarial training, demonstrated by the hyperparameter test. We briefly revisit the feed-forward propagation of the differentiable guided filter in this section.

We assume that given the input x and the guidance I, the output y should subject to a local linear model in a window ω_k centered at the pixel k

$$y_i = a_k I_i + b_k, \forall i \in \omega_k. \tag{8}$$

Besides, the output y can be obtained by x substracting some redundant components n

$$y_i = x_i - n_i. \tag{9}$$

Therefore, by minimizing the linear ridge regression model defined in (10), we can obtain the linear coefficients a_k and b_k in (11)–(15) as follows:

$$E(a_k, b_k) = \sum_{i \in \omega_k} \left((a_k I_i + b_k - x_i)^2 + \epsilon a_k^2 \right) \tag{10}$$

$$a_k = \frac{1}{\sigma_k^2 + \epsilon} \left\{ \frac{1}{|\omega_k|} \sum_{i \in \omega_k} I_i x_i - \mu_k \bar{x}_k \right\} \tag{11}$$

$$b_k = \bar{x}_k - a_k \mu_k \tag{12}$$

$$\mu_k = \frac{1}{|\omega_k|} \sum_{i \in \omega_k} I_i \tag{13}$$

$$\sigma_k^2 = \frac{1}{|\omega_k|} \sum_{i \in \omega_k} I_i^2 - \left\{ \frac{1}{|\omega_k|} \sum_{i \in \omega_k} I_i \right\}^2 \tag{14}$$

$$\bar{x}_k = \frac{1}{|\omega_k|} \sum_{i \in \omega_k} x_i. \tag{15}$$

Finally, the average coefficients a and b can be obtained by (16) and (17). The output y can be given by (18) as follows:

$$\bar{a}_i = \frac{1}{|\omega_i|} \sum_{k \in \omega_i} a_k \tag{16}$$

$$\bar{b}_i = \frac{1}{|\omega_i|} \sum_{k \in \omega_i} b_k \tag{17}$$

$$y_i = \bar{a}_i I_i + \bar{b}_i. \tag{18}$$

The feed-forward propagation of the differentiable guided filter is given by Algorithm 1, in which $.*$ and $./$ denote elementwise multiplication and division, and f_{mean} is the mean filter.

Algorithm 1: Forward of Differentiable Guided Filter

REQUIRE Input image to filter x, Guidance image I, Radius r, Regularization ϵ
ENSURE Filtered output y

$\text{mean}_I \leftarrow f_{\text{mean}}(I)$
$\text{mean}_x \leftarrow f_{\text{mean}}(x)$
$\text{corr}_I \leftarrow f_{\text{mean}}(I.*I)$
$\text{corr}_{Ix} \leftarrow f_{\text{mean}}(I.*x)$
$\text{var}_I \leftarrow \text{corr}_I - \text{mean}_I.*\text{mean}_I$
$\text{cov}_{Ix} \leftarrow \text{corr}_{Ix} - \text{mean}_I.*\text{mean}_x$
$a \leftarrow \text{cov}_{Ix}./(\text{var}_I + \epsilon)$
$b \leftarrow \text{mean}_x - a.*\text{mean}_I$
$\text{mean}_a \leftarrow f_{\text{mean}}(a)$
$\text{mean}_b \leftarrow f_{\text{mean}}(b)$
$y \leftarrow \text{mean}_a.*I + \text{mean}_b$
return y

D. Backward Propagation of Differentiable Guided Filter

The backward propagation of the differentiable guided filter is briefly introduced in this section. Naturally, the guided filter layer should be fully differentiable if the layer is embedded into the network to achieve the end-to-end training and inference, because it is necessary that the gradients with respect to trainable variables in the backbone net should be back-propagated through the guided filter layer, otherwise the model fails to train the parameters. Fortunately, because all computational operators are differentiable, including the mean filter, the elementwise multiplication, and division; our guided filter layer, consequently, is differentiable. We now give the mathematical partial derivative of the loss with respect to the input of the guided filter layer.

We first give the partial derivative of the scalar loss ℓ with respect to the input of the mean filter x in Proposition 1.

Proposition 1: If $y = f_{\text{mean}}(x)$ and the partial derivative of the scalar loss ℓ with respect to the output y is given, then the partial derivative of ℓ with respect to the input x is

$$\frac{\partial \ell}{\partial x} = f_{\text{mean}}\left\{\frac{\partial \ell}{\partial y}\right\}. \tag{19}$$

Proof: In our case, the feed forward of the mean filter is

$$y_{m,n} = \sum_{\substack{m-r\le i\le m+r\\ n-r\le j\le n+r}} \frac{1}{|\omega|}\cdot x_{i,j}. \tag{20}$$

Elements in $\{y_{p,q}\}_{m-r\le p\le m+r, n-r\le q\le n+r}$ are accordingly associated with $x_{m,n}$. So the partial derivative of each $y_{p,q}$ with respect to $x_{m,n}$ is

$$\frac{\partial y_{p,q}}{\partial x_{m,n}} = \frac{1}{|\omega|}. \tag{21}$$

According to the chain rule, the partial derivative of the scalar loss ℓ with respect to each input $x_{m,n}$ is

$$\frac{\partial \ell}{\partial x_{m,n}} = \sum_{\substack{m-r\le p\le m+r\\ n-r\le q\le n+r}} \frac{\partial \ell}{\partial y_{p,q}}\frac{\partial y_{p,q}}{\partial x_{m,n}} = \sum_{\substack{m-r\le p\le m+r\\ n-r\le q\le n+r}} \frac{\partial \ell}{\partial y_{p,q}}\frac{1}{|\omega|}. \tag{22}$$

Finally, the partial derivative of the scalar loss ℓ with respect to the input of the mean filter x is

$$\frac{\partial \ell}{\partial x} = f_{\text{mean}}\left\{\frac{\partial \ell}{\partial y}\right\}. \tag{23}$$

■

Proposition 2: If $y = f_{\text{mean}}(x)$ and the partial derivative of the scalar loss ℓ with respect to the output y is given, then we have

$$\begin{aligned}\mathrm{d}\ell &= \mathrm{tr}\left\{\frac{\partial \ell}{\partial \mathrm{y}^{\mathrm{T}}}\mathrm{dy}\right\}\\ &= \mathrm{tr}\left\{\frac{\partial \ell}{\partial \mathrm{y}^{\mathrm{T}}}\mathrm{df_{mean}(x)}\right\}\\ &= \mathrm{tr}\left\{\mathrm{f}_{\text{mean}}^{\mathrm{T}}\left\{\frac{\partial \ell}{\partial \mathrm{y}}\right\}\mathrm{dx}\right\}.\end{aligned} \tag{24}$$

Proof: Using Proposition 1, (24) is confirmed. ■

Proposition 3: If $y = f_{\text{guided}}(I, x)$ in which I and x denote the guidance and the input, and the partial derivative of the scalar loss ℓ with respect to y is given, then the partial derivative of ℓ with respect to x is

$$\begin{aligned}\frac{\partial \ell}{\partial x} &= f_{\text{mean}}\left\{\frac{\partial \ell}{\partial a}.*\frac{1}{\sigma^2+\epsilon}\right\}.*I\\ &\quad + f_{\text{mean}}\left\{\frac{\partial \ell}{\partial b} - \frac{\partial \ell}{\partial a}.*\frac{1}{\sigma^2+\epsilon}.*f_{\text{mean}}(I)\right\}\end{aligned} \tag{25}$$

$$\frac{\partial \ell}{\partial b} = f_{\text{mean}}\left\{\frac{\partial \ell}{\partial y}\right\} \tag{26}$$

$$\frac{\partial \ell}{\partial a} = f_{\text{mean}}\left\{\frac{\partial \ell}{\partial y}.*I\right\} - \frac{\partial \ell}{\partial b}.*f_{\text{mean}}(I) \tag{27}$$

in which $.*$ and $./$ denote elementwise multiplication and division, and $1/(\cdot)$ is elementwise division, too.

Proof: The feed forward of the guided filter is given by

$$\sigma^2 = f(I.*I) - f(I).*f(I) \tag{28}$$

$$a = \{f(I.*x) - f(I).*f(x)\}./(\sigma^2+\epsilon) \tag{29}$$

$$b = f(x) - a.*f(I) \tag{30}$$

$$y = f(a).*I + f(b) \tag{31}$$

in which f denotes the mean filter for brevity.

The total differentials of ℓ and y satisfy

$$\mathrm{d}\ell = \mathrm{tr}\left\{\frac{\partial \ell}{\partial \mathrm{y}^{\mathrm{T}}}\mathrm{dy}\right\}. \tag{32}$$

Using (31), we have

$$\begin{aligned}\mathrm{d}\ell &= \mathrm{tr}\left\{\frac{\partial \ell}{\partial \mathrm{y}^{\mathrm{T}}}\mathrm{dy}\right\}\\ &= \mathrm{tr}\left\{\frac{\partial \ell}{\partial \mathrm{y}^{\mathrm{T}}}\mathrm{d}\left\{f(\mathrm{a}).*I + \mathrm{f}(b)\right\}\right\}\end{aligned}$$

$$= \mathrm{tr}\left\{\left\{\frac{\partial \ell}{\partial \mathrm{y}} .* I\right\}^{\mathrm{T}} \mathrm{df(a)} + \frac{\partial \ell}{\partial \mathrm{y}^{\mathrm{T}}} \mathrm{df(b)}\right\}$$
$$= \mathrm{tr}\left\{f^{\mathrm{T}}\left\{\frac{\partial \ell}{\partial \mathrm{y}} .* I\right\} \mathrm{da} + f^{\mathrm{T}}\left\{\frac{\partial \ell}{\partial \mathrm{y}}\right\} \mathrm{db}\right\}. \quad (33)$$

b is the function of a and, therefore, we have $\partial \ell / \partial b$

$$\frac{\partial \ell}{\partial b} = f\left\{\frac{\partial \ell}{\partial y}\right\}. \quad (34)$$

Substituting (34) into (33), we have

$$\begin{aligned}
\mathrm{d}\ell &= \mathrm{tr}\left\{f^{\mathrm{T}}\left\{\frac{\partial \ell}{\partial \mathrm{y}} .* I\right\} \mathrm{d}a + \frac{\partial \ell}{\partial \mathrm{b}^{\mathrm{T}}} \mathrm{db}\right\} \\
&= \mathrm{tr}\left\{f^{\mathrm{T}}\left\{\frac{\partial \ell}{\partial \mathrm{y}} .* I\right\} \mathrm{da} + \frac{\partial \ell}{\partial \mathrm{b}^{\mathrm{T}}} \mathrm{d}\left\{\mathrm{f(x)} - \mathrm{a} .* \mathrm{f(I)}\right\}\right\} \\
&= \mathrm{tr}\left\{f^{\mathrm{T}}\left\{\frac{\partial \ell}{\partial \mathrm{y}} .* I\right\} \mathrm{da} + \frac{\partial \ell}{\partial \mathrm{b}^{\mathrm{T}}} \mathrm{df(x)} - \left\{\frac{\partial \ell}{\partial \mathrm{b}} .* \mathrm{f(I)}\right\}^{\mathrm{T}} \mathrm{da}\right\} \\
&= \mathrm{tr}\left\{\left\{f\left\{\frac{\partial \ell}{\partial \mathrm{y}} .* I\right\} - \frac{\partial \ell}{\partial \mathrm{b}} .* \mathrm{f(I)}\right\}^{\mathrm{T}} \mathrm{da} + \frac{\partial \ell}{\partial \mathrm{b}^{\mathrm{T}}} \mathrm{df(x)}\right\}.
\end{aligned} \quad (35)$$

Therefore, we have $\partial \ell / \partial a$

$$\frac{\partial \ell}{\partial a} = f\left\{\frac{\partial \ell}{\partial y} .* I\right\} - \frac{\partial \ell}{\partial b} .* f(I). \quad (36)$$

Substituting (36) into (35), we have

$$\begin{aligned}
\mathrm{d}\ell &= \mathrm{tr}\left\{\frac{\partial \ell}{\partial \mathrm{a}^{\mathrm{T}}} \mathrm{da} + \frac{\partial \ell}{\partial \mathrm{b}^{\mathrm{T}}} \mathrm{df(x)}\right\} \\
&= \mathrm{tr}\left\{\begin{array}{l} \dfrac{\partial \ell}{\partial a^{\mathrm{T}}} \mathrm{d}\left\{(f(I .* x) - f(I) .* f(x)) ./ (\sigma^2 + \epsilon)\right\} \\ + \dfrac{\partial \ell}{\partial b^{\mathrm{T}}} \mathrm{d}f(x) \end{array}\right\} \\
&= \mathrm{tr}\left\{\begin{array}{l} \left\{\dfrac{\partial \ell}{\partial a} .* \dfrac{1}{\sigma^2 + \epsilon}\right\}^{\mathrm{T}} \mathrm{d}\left\{f(I .* x) - f(I) .* f(x)\right\} \\ + \dfrac{\partial \ell}{\partial b^{\mathrm{T}}} \mathrm{d}f(x) \end{array}\right\} \\
&= \mathrm{tr}\left\{\begin{array}{l} \left\{\dfrac{\partial \ell}{\partial a} .* \dfrac{1}{\sigma^2 + \epsilon}\right\}^{\mathrm{T}} \mathrm{d}f(I .* x) \\ + \left\{\dfrac{\partial \ell}{\partial b} - \dfrac{\partial \ell}{\partial a} .* \dfrac{1}{\sigma^2 + \epsilon} .* f(I)\right\}^{\mathrm{T}} \mathrm{d}f(x) \end{array}\right\} \\
&= \mathrm{tr}\left\{\left\{\begin{array}{l} f\left\{\dfrac{\partial \ell}{\partial a} .* \dfrac{1}{\sigma^2 + \epsilon}\right\} .* I \\ + f\left\{\dfrac{\partial \ell}{\partial b} - \dfrac{\partial \ell}{\partial a} .* \dfrac{1}{\sigma^2 + \epsilon} .* f(I)\right\} \end{array}\right\}^{\mathrm{T}} \mathrm{dx}\right\}
\end{aligned} \quad (37)$$

in which $1/(\sigma^2 + \epsilon)$ is also an elementwise operation.

Algorithm 2: Backward of Differentiable Guided Filter

REQUIRE Derivative of ℓ with respect to y ∂y,
Guidance image I, Radius r, Regularization ϵ
ENSURE Derivative of ℓ with respect to x ∂x

1: $\mathrm{mean}_I \leftarrow f_{\mathrm{mean}}(I)$
2: $\mathrm{corr}_I \leftarrow f_{\mathrm{mean}}(I .* I)$
3: $\mathrm{var}_I \leftarrow \mathrm{corr}_I - \mathrm{mean}_I .* \mathrm{mean}_I$
4: $\partial b \leftarrow f_{\mathrm{mean}}(\partial y)$
5: $\partial a \leftarrow f_{\mathrm{mean}}(\partial y .* I) - \partial b .* \mathrm{mean}_I$
6: $\partial x \leftarrow f_{\mathrm{mean}}(\partial a ./ (\mathrm{var}_I + \epsilon)) .* I + f_{\mathrm{mean}}(\partial b - \partial a ./ (\mathrm{var}_I + \epsilon) .* \mathrm{mean}_I)$
7: RETURN ∂x

Finally, we have $\partial \ell / \partial x$

$$\begin{aligned}
\frac{\partial \ell}{\partial x} &= f\left\{\frac{\partial \ell}{\partial a} .* \frac{1}{\sigma^2 + \epsilon}\right\} .* I \\
&\quad + f\left\{\frac{\partial \ell}{\partial b} - \frac{\partial \ell}{\partial a} .* \frac{1}{\sigma^2 + \epsilon} .* f(I)\right\}.
\end{aligned} \quad (38)$$

■

The backward propagation of the differentiable guided filter is given in Algorithm 2, in accordance with Propositions 1–3, in which $.*$ and $./$ denote elementwise multiplication and division, and f_{mean} is the mean filter.

The derivation of the guided filter in the matrix form is highly related to that in the scalar form given in [12]. We will investigate the relation in appendix.

E. Markovian Discriminator-Based Adversarial Training and Postprocessing

To effectively train the network, we combine the adversarial loss $\mathcal{L}_{\mathrm{GAN}}$ and the perceptual loss $\mathcal{L}_{\mathrm{L1}}$ as the total loss and update the network with the partial derivatives of the total loss with respect to the learnable parameters. The discriminator D is also derived from Pix2Pix [10], which is composed of blocks of "convolution—instance normalization—leaky ReLU". The output of D alternatively differentiates local patches of generated or real images, instead of a global identification. The hazy and dehazed/GT images are concatenated to feed the discriminator. See [10] for more details about D.

For brevity, the trainable Pix2Pix backbone is referred to as G_0, and the Guided-Pix2Pix (Pix2Pix backbone—guided filter layer—dehazing equation) as G. Using the network and the scattering model defined previously, the backbone G_0 takes the hazy image I as input and generates the estimation of the coarse transmission map t_0. Also, a coarse dehazed image $\hat{J}_0$ can be restored from the given factors

$$t_0 = G_0(I) \quad (39)$$
$$\hat{J}_0(z) = \frac{I(z) - A}{t_0(z)} + A. \quad (40)$$

The refinement of transmission and the dehazed image are given by

$$t = f_{\text{guided}}(I, t_0) \tag{41}$$

$$\hat{J}(z) = \frac{I(z) - A}{t(z)} + A. \tag{42}$$

The feed-forward propagation of our model estimates the refined transmission maps and the dehazed images, and the backward propagation calibrates the performance of dehazing. We now define the loss function in the backward propagation. Conventionally, the L1 loss is used as the perceptual loss as it captures and gathers the differences between pixels of the dehazed and the clear image, given by

$$\mathcal{L}_{\text{L1}} = \mathbb{E}_{J\sim p_{\text{data}}, \hat{J}\sim p_G}\left[||J - \hat{J}||_1\right]. \tag{43}$$

Adversarial loss guides G to generate the estimation of a haze-free image that can be hardly differentiated by discriminator D, which is derived from GAN. The objective function is given by

$$V(G, D) = \min_G \max_D \mathbb{E}_{J\sim p_{\text{data}}}\left[\log D(J)\right] + \mathbb{E}_{\hat{J}\sim p_G}\left[\log(1 - D(\hat{J})\right]. \tag{44}$$

Optimized by gradient descent methods, the trainable backbone G_0 can be updated by back propagation of the GAN training strategy [25].

The inference and refinement of transmission are given by (45) and the image dehazing is given by (46), after the parameters of G_0 are secured

$$t = f_{\text{guided}}(I, G_0(I)) \tag{45}$$

$$\hat{J}(z) = \frac{I(z) - A}{t(z)} + A. \tag{46}$$

Besides, contrast-sensitive Retinex postprocessing is employed for better visualization: first, the multiscale Retinex enhances the details of the dehazed image, using

$$r(z) = \sum_k w_k \left\{\log \hat{J}(z) - \log\left[F_k(z) * \hat{J}(z)\right]\right\} \tag{47}$$

where $F_k(z) * \hat{J}(z)$ denotes Gaussian blur. Conventially, three scales, $\sigma_{\text{Gauss}} = 15, 80, 250$, and $w_k = 1/3$ are used in our multiscale Retinex.

Then, given a dynamic control parameter d_r, the mean μ_r, the standard deviation σ_r, and the desired data range of $r(z)$ are computed

$$\min_r = \mu_r - d_r \cdot \sigma_r \tag{48}$$

$$\max_r = \mu_r + d_r \cdot \sigma_r. \tag{49}$$

Finally, given the minimum and the maximum, the dehazed image is rescaled and clipped into $0 \ldots 255$

$$\hat{J}(z) = \frac{r(z) - \min_r}{\max_r - \min_r} \cdot 255. \tag{50}$$

IV. Experiments and Discussion

We evaluate our dehazing model by conducting a variety of experiments, including comparisons with other methods, ablation study with respect to the refinement, and hyperparameter test with respect to r and ϵ.

A. Implementation Details

Our experiments are conducted with TensorFlow [45] framework. The input hazy images are resized to 512×512, and the refined transmission maps and dehazed images share the same size. The radius and ϵ of the differentiable guided filter are empirically assigned to 60 and 10^{-8}, respectively. The guidance images are the grayscale of the input images. The ADAM [46] optimizer is applied to the optimization of the network and its initial learning rate, β_1, and β_2 are 10^{-4}, 0.5, and 0.999, respectively. The learning rate decays by 0.9 per 100 iteration steps. These hyperparameters are empirically given but the key parameters will be thoroughly investigated in the following experiments.

We collect several synthetic and real-world datasets to evaluate our model, including RESIDE-SOTS indoor/outdoor [47], RICE [48], and I/O-HAZE [49], [50] datasets. REalistic Single Image DEhazing (RESIDE) dataset [47] is composed of both synthetic and real-world hazy/haze-free three-channel natural images, and its synthetic objective test set (SOTS) collects 500 indoor/outdoor pairs of images, employed as the comparative dataset. Remote sensing Image Cloud rEmoving (RICE) dataset [48] contains 500 pairs of cloudy/cloud-free 3-band images, collected from Google Earth. O-HAZE [50] is an outdoor-scene hazy/haze-free dataset, while I-HAZE [49] is an indoor dataset, and three-channel hazy images of both of them are generated by dedicated haze machines. For dehazing, the 500 pairs of RESIDE-SOTS outdoor [47] hazy/clear images are contaminated by mild haze, while the haze in the I/O-HAZE [49], [50] dataset is severe, nonhomogeneous, and more challenging. On the other hand, the 500 pairs of homogeneously hazy/clear images in RICE [48] are generated from Google Earth by switching the display of the cloud layer. We train our model only on the half dataset of I-HAZE [49] and refer to it as the natural image dehazing model, and train on RICE [48] as the remote sensing dehazing model. For natural images, we test and evaluate our model on the RESIDE-SOTS indoor/outdoor [47], half I-HAZE [49], and O-HAZE [50] datasets, to demonstrate the effectiveness and generalization. For remote sensing images, we train on the half RICE [48] and evaluate on the others.

B. Comparison Against State-of-the-Art Methods

We first compare our method against representative and state-of-the-art natural image dehazing methods, including DCP [1], [2], AOD-Net [23], DCPDN [5], EPDN [7], DehRet [51], KTDN [52], and the deep guided filter (DGF) [14]. Particularly, DCP [1], [2], and DCPDN [5] estimate the transmission map to restore a clear image, while AOD-Net [23], EPDN [7], DehRet [51], KTDN [52], and DGF [14] remove haze in an image-to-image translation way. GP2P-G denotes the outputs of our Guided-Pix2Pix, while GP2P-Rescale and GP2P-Retinex

TABLE I
AVERAGE PSNR AND SSIM OF COMPARISONS ON SYNTHETIC DEHAZING DATASETS (AVERAGE ± STANDARD DEVIATION)

Dataset	DCP [1, 2]	AOD-Net [23]	DCPDN [5]	EPDN [7]	DGF [14]	DehRet [51]	KTDN [52]	GP2P-*G*	GP2P-Rescale	GP2P-Retinex
Indoor										
PSNR	12.24 ± 5.19	18.73 ± 3.22	13.74 ± 3.53	25.02 ± 2.67	17.16 ± 2.40	15.40 ± 2.63	13.82 ± 2.64	16.71 ± 3.59	14.48 ± 3.53	12.32 ± 2.34
SSIM	0.6678 ± 0.1741	0.8290 ± 0.0823	0.7330 ± 0.1102	0.9186 ± 0.0406	0.7698 ± 0.0704	0.6982 ± 0.1094	0.7106 ± 0.0881	0.7833 ± 0.0815	0.6910 ± 0.0810	0.6456 ± 0.1042
Outdoor										
PSNR	15.53 ± 5.19	19.50 ± 2.25	19.19 ± 4.30	18.33 ± 3.36	15.94 ± 3.88	17.14 ± 1.86	15.96 ± 2.86	21.42 ± 2.47	17.17 ± 2.40	13.58 ± 1.71
SSIM	0.7877 ± 0.1633	0.8549 ± 0.0485	0.8421 ± 0.0757	0.8037 ± 0.0558	0.7740 ± 0.0715	0.7370 ± 0.0719	0.7618 ± 0.0679	0.8448 ± 0.0673	0.7588 ± 0.0545	0.6564 ± 0.0835
I-HAZE-15										
PSNR	11.56 ± 4.85	14.47 ± 2.67	14.96 ± 3.35	15.12 ± 2.82	13.42 ± 1.88	15.72 ± 2.75	16.70 ± 2.62	16.93 ± 3.01	13.13 ± 2.45	12.90 ± 1.96
SSIM	0.5990 ± 0.1528	0.6225 ± 0.1070	0.6553 ± 0.1243	0.6411 ± 0.0767	0.6094 ± 0.0748	0.6496 ± 0.1032	0.6923 ± 0.0910	0.6998 ± 0.1025	0.6002 ± 0.0891	0.5449 ± 0.0657
O-HAZE										
PSNR	13.18 ± 5.12	14.83 ± 1.53	14.80 ± 3.20	16.33 ± 2.38	15.15 ± 1.69	12.04 ± 2.55	18.52 ± 2.11	16.05 ± 3.18	10.38 ± 2.12	10.99 ± 1.83
SSIM	0.5274 ± 0.1596	0.4670 ± 0.1045	0.4911 ± 0.1200	0.5448 ± 0.1016	0.5562 ± 0.0755	0.4779 ± 0.1096	0.6461 ± 0.076	0.5654 ± 0.1301	0.4590 ± 0.0998	0.4438 ± 0.0901

[1]We reproduce dehazing results using the pretrained models.

denote the aforementioned rescaled and Retinex-enhanced results. I/O-HAZE [49], [50], RESIDE-SOTS indoor and outdoor [47] datasets are used as benchmark datasets. The quantitative comparisons between dehazing methods are listed in Table I and the visual assessments are shown in Figs. 3–6.

First, we evaluate these methods from the visual perspective. DCP [1], [2] is always referred to as a baseline of dehazing due to its visual effectiveness and simple implementation, and it is able to achieve a wonderful dehazing performance, provided that delicate parameter configuration is given. Its drawbacks, however, have been fully presented and improved, especially for the halo artifacts surrounding buildings. DehRet [51] has been proved to be an equivalent to DCP, so they generate similar results. These halo artifacts are also shown in the results of EPDN [7]. Visually, DGF [14] reduces the color intensity but introduces color cast as well. KTDN [52] achieves the best results on I/O-HAZE datasets; however, it brings artifacts on the low-frequent areas. Nevertheless, our method can reduce the halo artifacts surrounding buildings, which should be attributed to the merits of the physical model-based fashion and transmission map refinement: the embedded guided filter significantly refines the spatial structure of the transmission maps, which results in the improvement of halo artifacts and brightness consistency, as shown in Fig. 2. Furthermore, our model simplifies the backbone net compared to DCPDN [5] when they achieve similar visual performance. The details in the dehazed image, in addition, can be enhanced by the rescale method or Retinex postprocessing.

Our method has its intrinsic drawbacks as well: it dehazes more slightly compared to those state-of-the-art methods, which seems to yield an underdehazing visual performance. Besides, our methods can also hardly remove the haze close to the atmospheric light, which may appear at the deepest pixels within the image.

The quantitative assessment is listed in Table I. Interestingly, EPDN [7] achieves the best quantitative performance on the indoor dataset but our method outperforms others on the outdoor dataset, in terms of PSNR and SSIM. On the O-HAZE dataset, KTDN [52] quantitatively performs better than other methods. We attribute this to their high-relevant customization and training on the corresponding datasets: EPDN [7] is sufficiently trained on the indoor dataset, and KTDN [52] fits the O-HAZE dataset as well. Similarly, the visual check can account for the assessment: the intrinsic drawback of our method degrade the indoor results but slightly dehazing fits the outdoor dataset

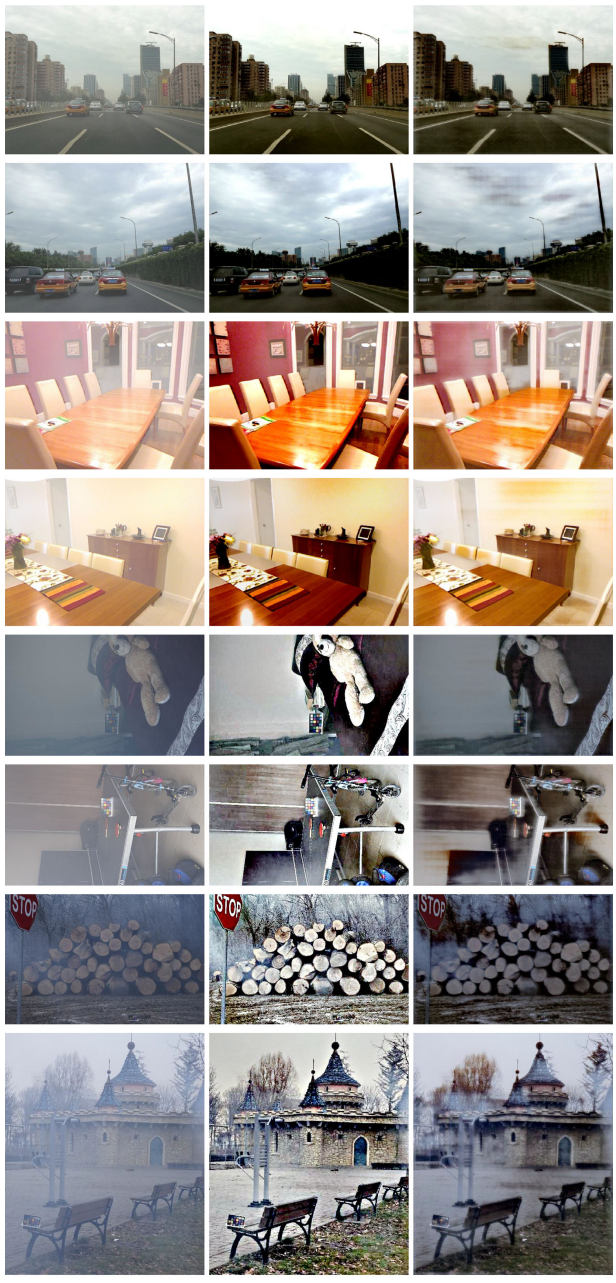

Fig. 2. Effect of the embedded guided filter. The transmission maps from the Pix2Pix backbone produce severe halo artifacts, while they can be dramatically refined and improved by the guided filter layer.

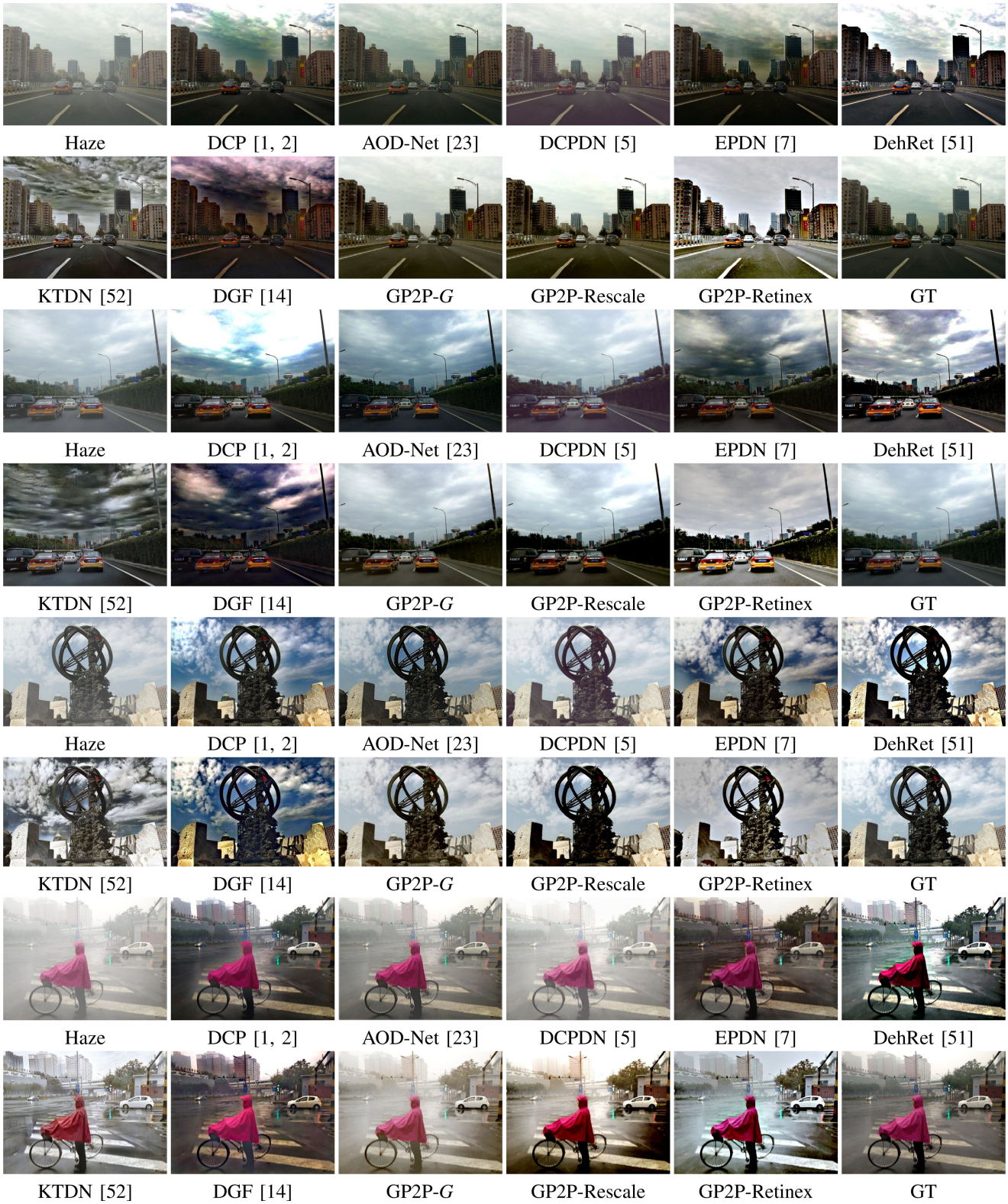

Fig. 3. Visual comparisons on RESIDE-SOTS outdoor dataset. We compare our methods with DCP [1], [2], AOD-Net [23], DCPDN [5], EPDN [7], DehRet [51], KTDN [52], and DGF [14]. Note that DCP and EPDN degrade the dehazing performance in terms of the halo artifacts surrounding the building and the armillary sphere, AOD-Net may overestimate the haze, DGF visually manipulates the color intensity, and KTDN introduces artifacts in the low-frequent areas. But, our method can refine dehazing, prevent halo artifacts, and generate clear images.

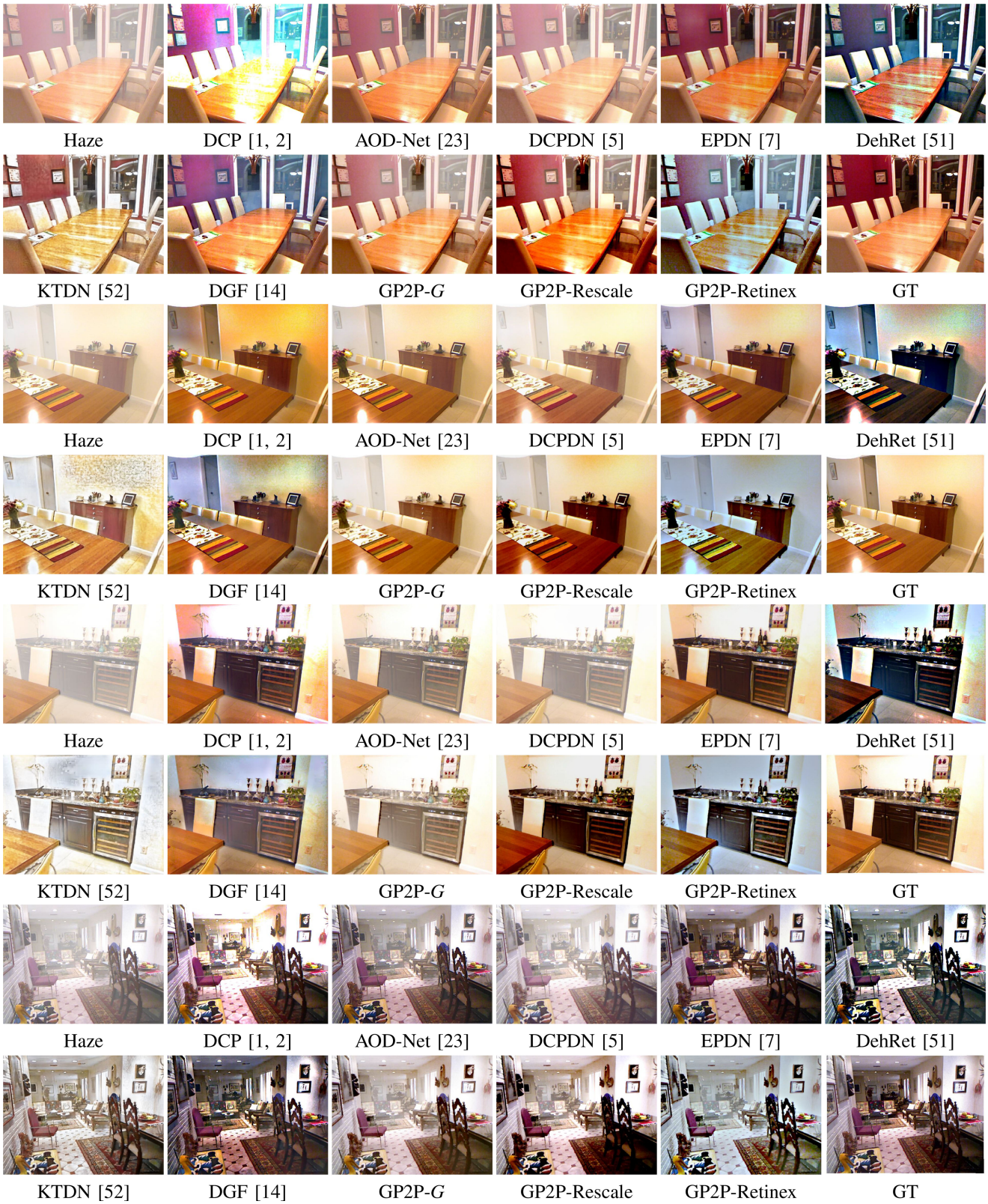

Fig. 4. Visual comparisons on RESIDE-SOTS indoor dataset. We compare our methods with DCP [1], [2], AOD-Net [23], DCPDN [5], EPDN [7], DehRet [51], KTDN [52], and DGF [14]. Similarly, DCP and EPDN degrade the dehazing performance in terms of the halo artifacts, AOD-Net may overestimate the haze, DGF visually manipulates the color intensity, and KTDN introduces artifacts in the low-frequent areas. Our method can also refine dehazing, prevent halo artifacts, and generate clear images.

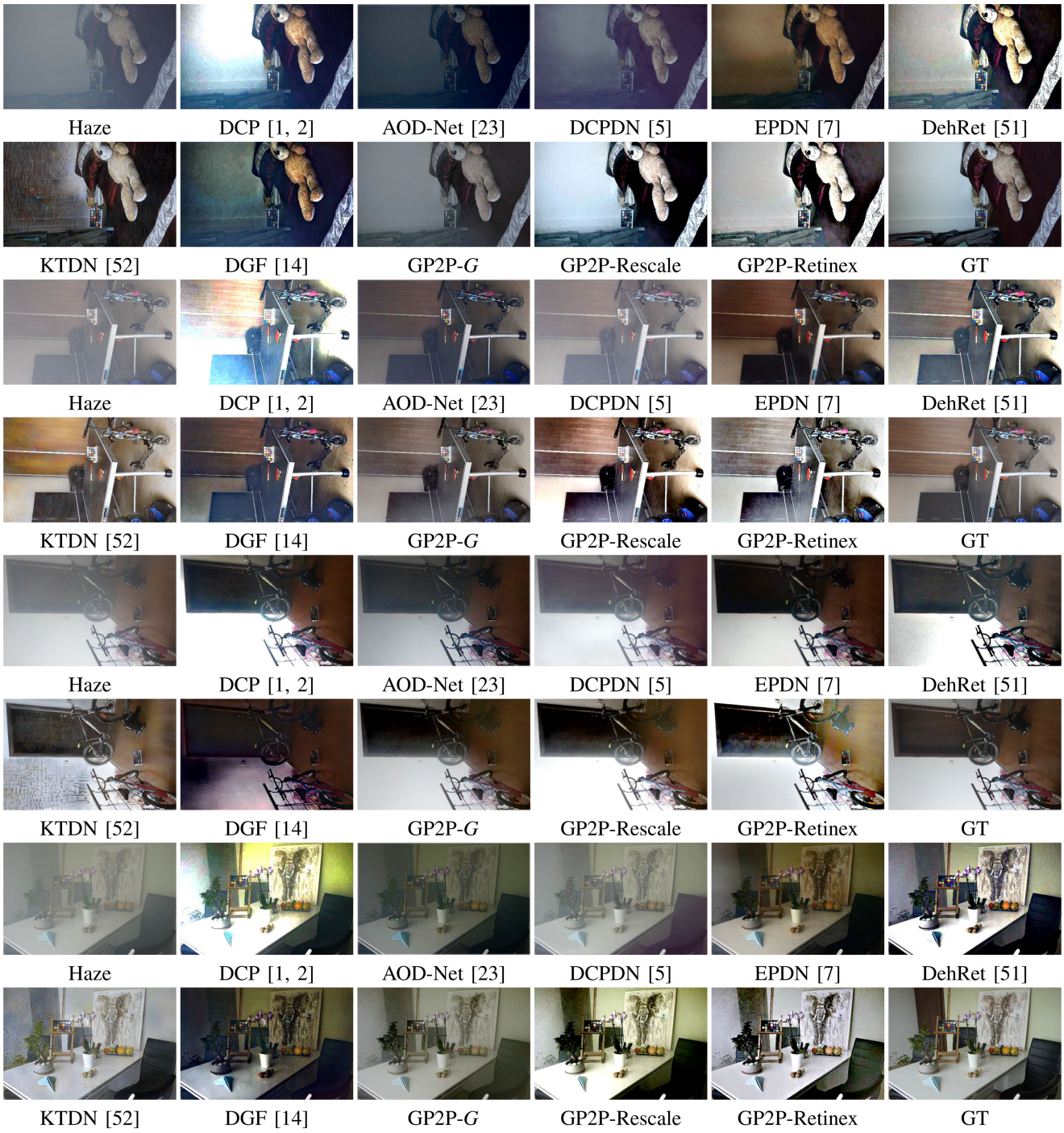

Fig. 5. Visual comparisons on I-HAZE dataset. We compare our methods with DCP [1], [2], AOD-Net [23], DCPDN [5], EPDN [7], DehRet [51], KTDN [52], and DGF [14]. Similarly, DCP and EPDN degrade the dehazing performance in terms of the halo artifacts, AOD-Net may overestimate the haze, DGF visually manipulates the color intensity, and KTDN introduces artifacts in the low-frequent areas. Our method can also refine dehazing, prevent halo artifacts, and generate clear images.

well. We will improve its intensity by introducing a specific controlling knob.

The generalization of our method is confirmed on the O-HAZE [50], RESIDE-SOTS indoor, and outdoor datasets. The dehazing model has been trained on the half I-HAZE dataset and transferred to dehaze on the aforementioned datasets without fine-tuning. As seen in Figs. 3, 4, and 6, we find that our method can also achieve acceptable dehazing performance without retraining, which demonstrates the generalization of our method. The area around the selection of atmospheric light, however, cannot be fully restored, and the nonhomogeneous haze cannot be thoroughly detected and removed, which should be an improvement in future work.

C. Ablation Study With Respect to Differentiable Guided Filter

We investigate the effect of the differentiable guided filter in the ablation study. RICE [48] dataset is employed to test the

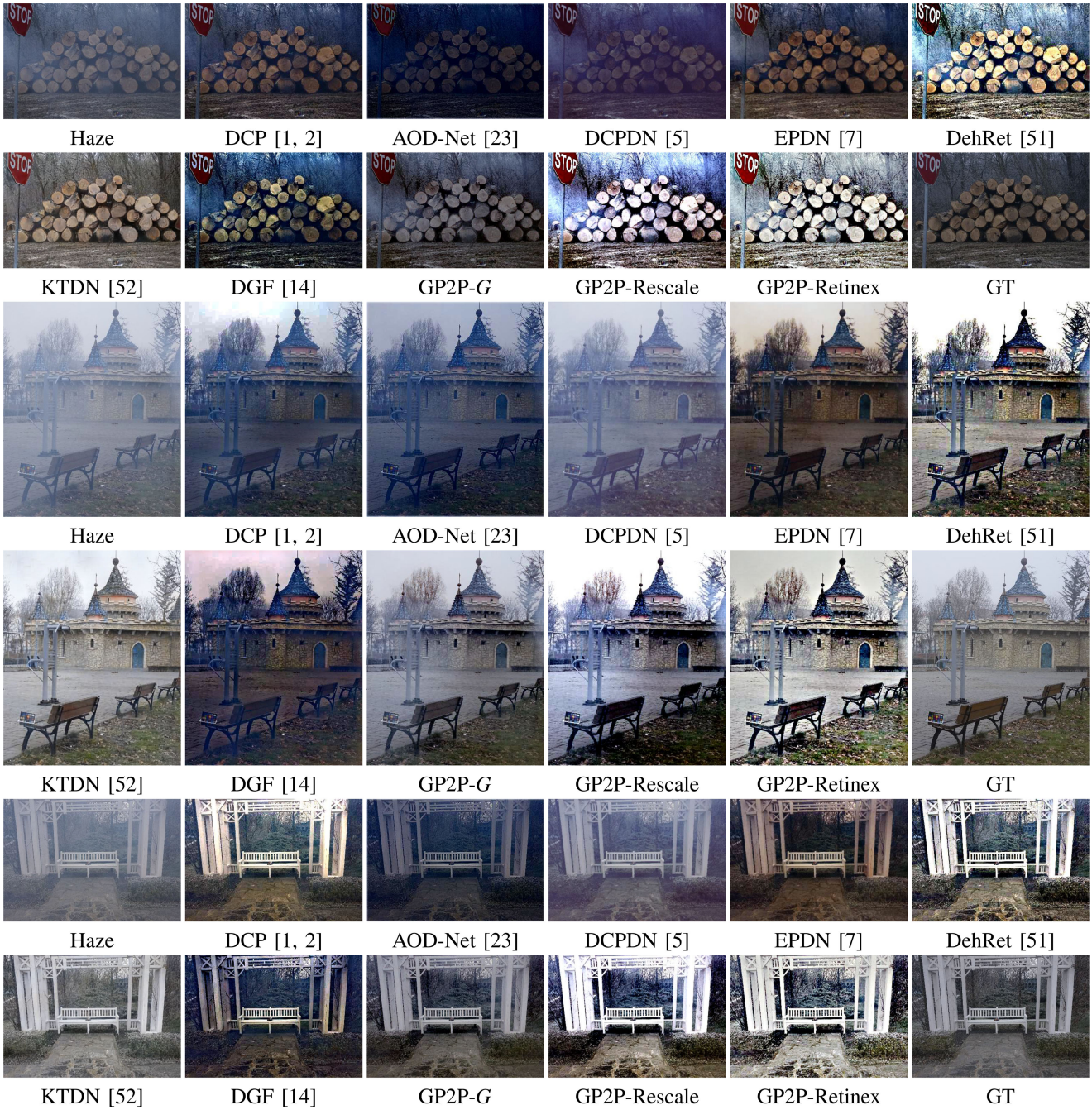

Fig. 6. Visual comparisons on O-HAZE dataset. We compare our methods with DCP [1], [2], AOD-Net [23], DCPDN [5], EPDN [7], DehRet [51], KTDN [52], and DGF [14]. Similarly, DCP and EPDN degrade the dehazing performance in terms of the halo artifacts, AOD-Net may overestimate the haze, DGF visually manipulates the color intensity, and KTDN introduces artifacts in the low-frequent areas. Our method can also refine dehazing, prevent halo artifacts, and generate clear images.

dehazing effect on remote sensing images. The effect of each component should be clarified first: the backbone should be able to estimate the coarse distribution of existing haze within an image and form the potential transmission map, while the adversarial training (GAN architecture) and the differentiable guided filter are in charge of different dehazing services: the adversarial training ensures realistic outputs, which has been demonstrated in plenty of relevant literature, while the guided filter refines the spatial structure of transmission maps. Consequently, they improve the quality of the dehazed images in a joint way. Accordingly, we focus on investigating the effect of the guided filter rather than that of the adversarial training.

As shown in Fig. 7, the guided filter refines the spatial structure of the potential transmission maps, yielding homogeneous results without oversaturated artifacts: the filter follows the backbone network and refines the transmission map jointly. The prediction of the backbone network can coarsely capture the spatial structure and the following differentiable guided filter can refine the TME to retain the spatial features of the image. Furthermore, the results also show the dehazing performance

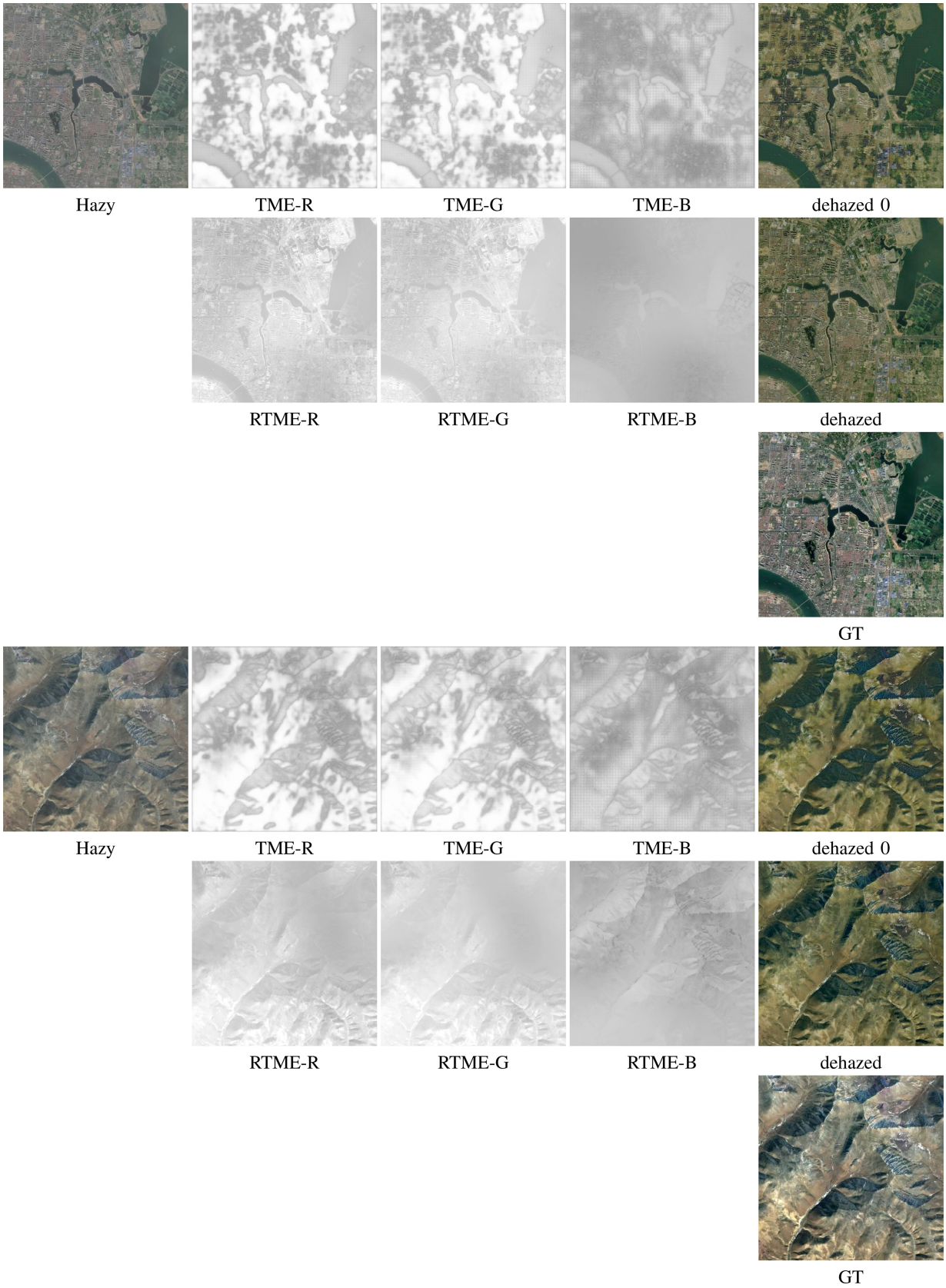

Fig. 7. Ablation study with respect to the differentiable guided filter. The backbone network coarsely captures the spatial structures while the following filter refines and yields fine transmission maps.

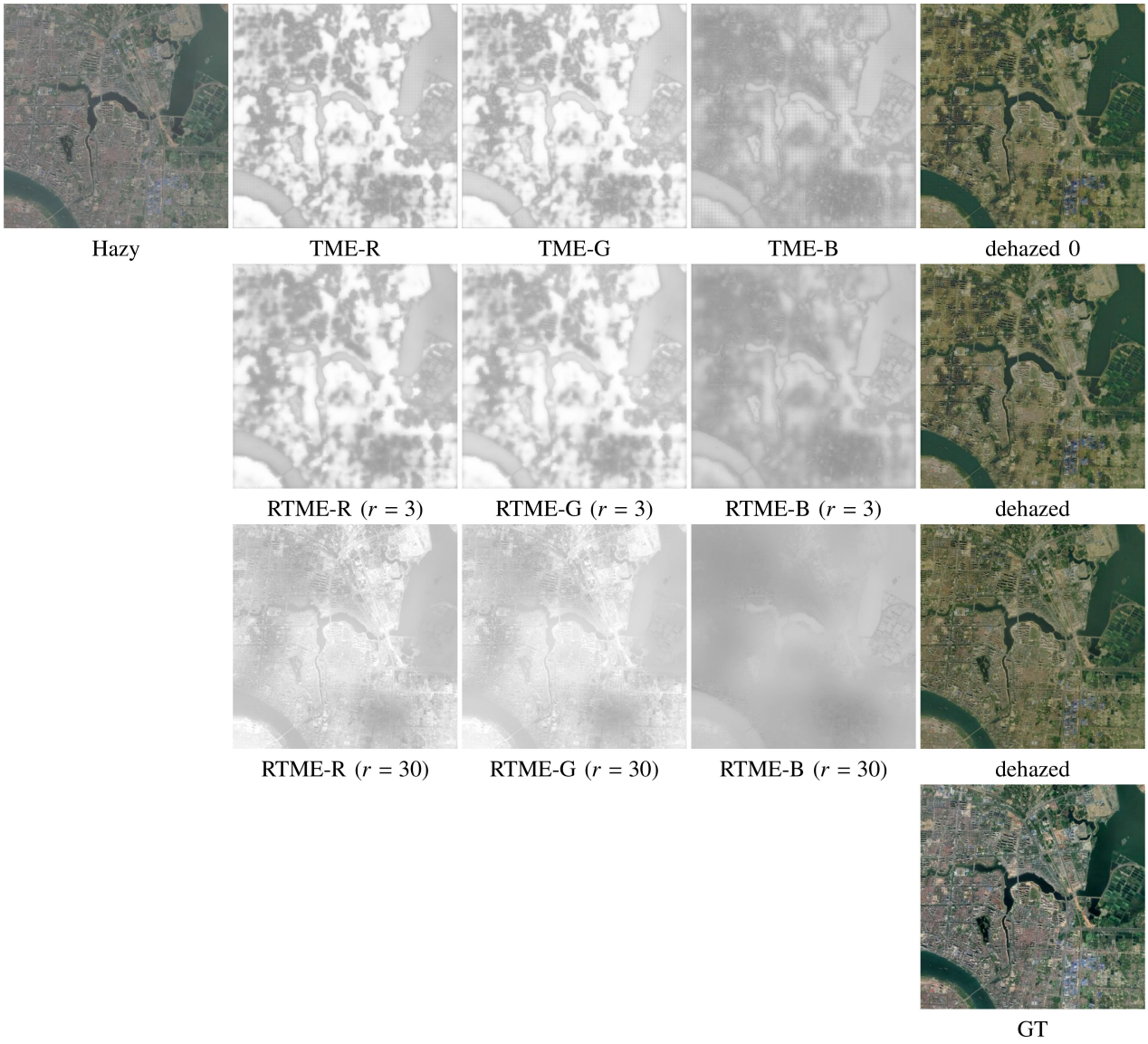

Fig. 8. Hyperparameter sensitivity with respect to the radius of the guided filter (1). The radius varies from 3 to 120. A higher radius yields smooth results and retains more spatial features.

on the remote sensing image: The estimation of the refined transmission map, referred to as RTME, is three-channel because three channels are contaminated by haze in different degrees. Note that the TME of the blue channel is contaminated more severely than that of the red and green channels, which has been also shown in Fig. 7.

D. Hyperparameter Effect on Dehazing

We explore the effect of two key hyperparameters in the guided-pix2pix: the radius of the differentiable guided filter r and the initial learning rate. The radius controls the transfer performance of spatial structures from original hazy images, while the initial learning rate affects the training efficacy. The visual assessments on the RICE dataset are shown in Figs. 8–10.

Figs. 8 and 9 show the effect of the radius of the guided filter. The radius controls the transfer performance of spatial structures from original hazy images: a higher radius yields a finer transmission map while a lower radius keeps the refinement as the output from the backbone network. As can be seen in Figs. 8 and 9, the radius of 3 seldom modify the transmission map, but the radius of 30 significantly refine it. On the other hand, the transmission maps of the radius over 60 vary merely compared to that of the radius of 60. Consequently, the radius is recommended to over 60 to achieve acceptable performance for images of 512×512.

The initial learning rate is a key parameter to control training efficacy: a higher α leads to a spatial structure corruption while a lower α discourages training. We test the initial learning rate because the differentiable guided filter is observed to significantly modify the estimation of the transmission map even if the filter has no learnable parameters. The visual results are shown in Fig. 10.

Fig. 10 shows the effect of the initial learning rate α. The initial learning rate α affects training efficacy: a higher α makes

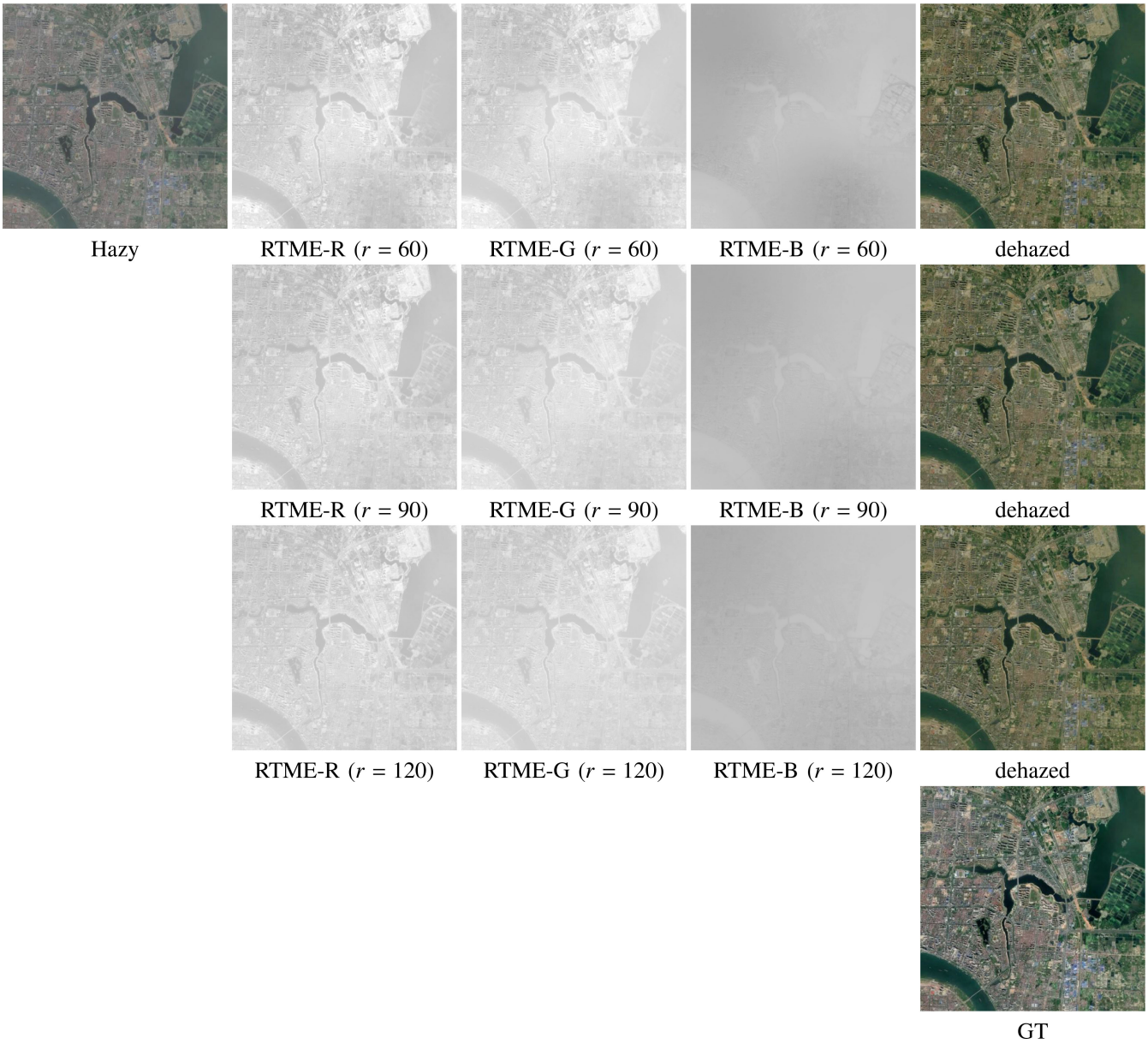

Fig. 9. Hyperparameter sensitivity with respect to the radius of the guided filter (2). The radius varies from 3 to 120. A higher radius yields smooth results and retains more spatial features.

training corrupted in learning spatial structures and features, while a lower α punishes training hard and yields inappropriate results. We test α varying from 10^{-5} to 10^{-3} and observe that α of 10^{-4} is more appropriate compared to the others. Fortunately, the following guided filter layer enables the structural transfer from the hazy images and yields acceptable dehazing results.

V. Conclusion

In this article, we propose an end-to-end learnable dehazing network, which is referred to as Guided-Pix2Pix, to jointly estimate and refine the transmission map and further dehaze images by the physical scattering equation. Instead of a two-stage model of predicting and postprocessing, Guided-Pix2Pix enables predicting refined transmission maps in one feed-forward step, and then it substitutes these potential refinements into the physical scattering equation to restore dehazed images. Specifically, the Pix2Pix backbone is employed to achieve the coarse prediction of the transmission, and then the embedded differentiable guided filter refines its spatial structure to alleviate the potential halo artifacts. In the training phase, perceptual and adversarial losses are jointly employed to guide the dehazing network to generate haze-free and realistic images, which is our distinct contribution against the deep guided filter [14]. Also, the differentiable guided filter ensures the forward propagation of the feature maps and back-propagation of gradients, which guarantees the inference in a one-stage way. Experiments show that our model tends to improve the visibility of hazy images and alleviate the halo artifacts along edges, in terms of the visual assessment in the experiments. Besides, the guided filter layer significantly promotes

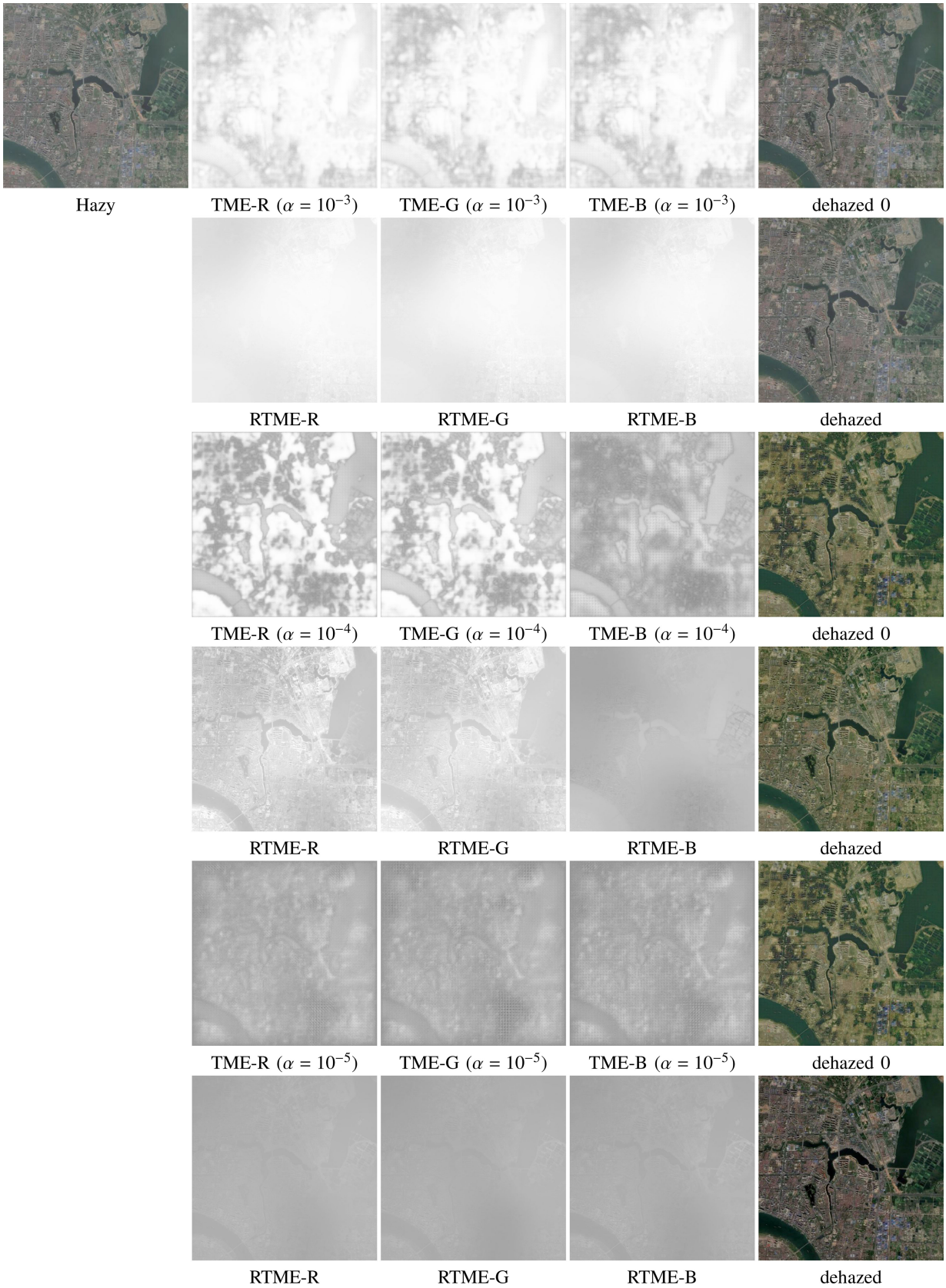

Fig. 10. Hyperparameter sensitivity with respect to the initial learning rate α. The initial learning rate varies from 10^{-5} to 10^{-3}. A higher α causes the learning corruption of spatial features, while a lower α leads to the laziness of training. Fortunately, the following guided filter enables transferring the spatial structures of hazy images.

problematic adversarial training and is of great help to yield acceptable results. Its generalization capability is also confirmed on a challenging outdoor nonhomogeneous hazy dataset. Failure of dehazing near the brightest part may be caused by the arbitrary atmospheric light estimation (the pixel of highest color intensity), which should be a valuable future work. In addition, we will further explore more sophisticated approaches to removing nonhomogeneous haze within images, which remains challenging for both prior-based and learning-based methods.

APPENDIX

Relation to The Kernel of Guided Filter

The derivation of the guided filter in the matrix form is highly related to that in the scalar form given in [12], so we investigate the relation in this section. The kernel of the guided filter is given in [12]

$$W_{ij} = \frac{1}{|\omega|^2} \sum_{k:(i,j)\in\omega_k} \left\{ 1 + \frac{(I_i - \mu_k)(I_j - \mu_k)}{\sigma_k^2 + \epsilon} \right\}. \tag{51}$$

Now, we investigate the relationship between the kernel and the given partial derivative of the output y_i with respect to the input x_j in the scalar form, which can cast another insight into understanding the differentiability of the guided filter layer.

The partial derivative of the scalar loss ℓ with respect to the input x_j is given by (52) according to the chain rule

$$\frac{\partial \ell}{\partial x_j} = \sum_{i\in\omega_j} \frac{\partial \ell}{\partial y_i} \frac{\partial y_i}{\partial x_j}. \tag{52}$$

We recall the partial derivative of ℓ with respect to x in

$$\begin{aligned} \frac{\partial \ell}{\partial x} =& f\left\{ \frac{\partial \ell}{\partial a} .* \frac{1}{\sigma^2 + \epsilon} \right\} .* I \\ &+ f\left\{ \frac{\partial \ell}{\partial b} - \frac{\partial \ell}{\partial a} .* \frac{1}{\sigma^2 + \epsilon} .* f(I) \right\} \end{aligned} \tag{53}$$

$$\frac{\partial \ell}{\partial a} = f\left\{ \frac{\partial \ell}{\partial y} .* I \right\} - \frac{\partial \ell}{\partial b} .* f(I) \tag{54}$$

$$\frac{\partial \ell}{\partial b} = f\left\{ \frac{\partial \ell}{\partial y} \right\}. \tag{55}$$

We consider the partial derivative of ℓ with respect to x_j and substitute (54) and (55) into the term with respect to a and b, and then it yields

$$\begin{aligned} \frac{\partial \ell}{\partial x_j} &= \frac{1}{|\omega|} \sum_{k\in\omega_j} \left\{ \frac{\partial \ell}{\partial a_k} * \frac{1}{\sigma_k^2 + \epsilon} \right\} * I_j \\ &\quad + \frac{1}{|\omega|} \sum_{k\in\omega_j} \left\{ \frac{\partial \ell}{\partial b_k} - \frac{\partial \ell}{\partial a_k} * \frac{\mu_k}{\sigma_k^2 + \epsilon} \right\} \\ &= \frac{1}{|\omega|} \sum_{k\in\omega_j} \left\{ \frac{\partial \ell}{\partial a_k} * \frac{I_j - \mu_k}{\sigma_k^2 + \epsilon} + \frac{\partial \ell}{\partial b_k} \right\} \\ &= \frac{1}{|\omega|} \sum_{k\in\omega_j} \left\{ \begin{array}{l} \dfrac{1}{|\omega|} \displaystyle\sum_{i\in\omega_j} \left\{ \frac{\partial \ell}{\partial y_i} * (I_i - \mu_k) \right\} * \frac{I_j - \mu_k}{\sigma_k^2 + \epsilon} \\ + \dfrac{1}{|\omega|} \displaystyle\sum_{i\in\omega_j} \frac{\partial \ell}{\partial y_i} \end{array} \right\} \\ &= \sum_{i\in\omega_j} \frac{\partial \ell}{\partial y_i} \left\{ \frac{1}{|\omega|^2} \sum_{k\in\omega_j} \left\{ 1 + \frac{(I_i - \mu_k)(I_j - \mu_k)}{\sigma_k^2 + \epsilon} \right\} \right\}. \end{aligned} \tag{56}$$

According to the chain rule and (52), we get the partial derivative of the output y_i with respect to the input x_j

$$\frac{\partial y_i}{\partial x_j} = \frac{1}{|\omega|^2} \sum_{(i,k)\in\omega_j} \left\{ 1 + \frac{(I_i - \mu_k)(I_j - \mu_k)}{\sigma_k^2 + \epsilon} \right\}. \tag{57}$$

Equation (57) is the expression of the filter kernel $W_{i,j}$.

REFERENCES

[1] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1956–1963.

[2] H. He and W.-C. Siu, "Single image super-resolution using Gaussian process regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 449–456.

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

[4] R. Fattal, "Single image dehazing," *Acm Trans. Graph.*, vol. 27, no. 3, pp. 1–9, 2008.

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

[6] Y. Li, Q. Miao, W. Ouyang, Z. Ma, and Y. Quan, "LAP-Net: Level-aware progressive network for image dehazing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3275–3284.

[7] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8152–8160.

[8] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[10] 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.*, 2017, pp. 5967–5976.

[11] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[12] 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.

[13] 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.

[14] H. Wu, S. Zheng, J. Zhang, and K. Huang, "Fast end-to-end trainable guided filter," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1838–1847.

[15] 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.

[16] L. Kratz and K. Nishino, "Factorizing scene albedo and depth from a single foggy image," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1701–1708.

[17] R. Fattal, "Dehazing using color-lines," *ACM Trans. Graph.*, vol. 34, pp. 1–14, 2014.

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

[19] 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.

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

[21] 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.
[22] X. Yang, Z. Xu, and J. Luo, "Towards perceptual image dehazing by physics-based disentanglement and adversarial training," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7485–7492.
[23] 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.*, 2017, pp. 4780–4788.
[24] D. Yang and J. Sun, "Proximal dehaze-net: A prior learning-based deep network for single image dehazing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 729–746.
[25] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[26] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.
[27] 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.*, 2014, pp. 184–199.
[28] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.
[29] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics," in *Proc. Int. Conf. Learn. Representations*, 2016.
[30] 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.
[31] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[32] 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.*, 2016, pp. 2536–2544.
[33] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6721–6729.
[34] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.
[35] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4471–4480.
[36] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2813–2821.
[37] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*.
[38] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," in *Proc. 5th Int. Conf. Learn. Representations*, 2017.
[39] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GAN," vol. 30, pp. 5767–5777, 2017.
[40] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. Int. Conf. Learn. Representations*, 2018.
[41] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Representations*, 2018.
[42] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Representations*, 2019.
[43] R. Szeliski *et al.*, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1068–1080, Jun. 2008.
[44] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Prof. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[45] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: https://www.tensorflow.org/
[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.
[47] B. Li *et al.*, "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2019.
[48] D. Lin, G. Xu, X. Wang, Y. Wang, X. Sun, and K. Fu, "A remote sensing image dataset for cloud removal," 2019, *arXiv:1901.00600*.
[49] C. O. Ancuti, C. Ancuti, R. Timofte, and C. D. Vleeschouwer, "I-Haze: A dehazing benchmark with real hazy and haze-free indoor images," in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst.*, 2018, pp. 620–631.
[50] C. O. Ancuti, C. Ancuti, R. Timofte, and C. D. Vleeschouwer, "O-haze: A dehazing benchmark with real hazy and haze-free outdoor images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 867–8678.
[51] A. Galdran, A. Alvarez-Gila, A. Bria, J. Vazquez-Corral, and M. Bertalmío, "On the duality between retinex and image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8212–8221.
[52] H. Wu, J. Liu, Y. Xie, Y. Qu, and L. Ma, "Knowledge transfer dehazing network for nonhomogeneous dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 1975–1983.

Libin Jiao received the B.S. degree in computer science and technology and the Ph.D. degree in computer software and theory from Beijing Normal University, Beijing, China, in 2014 and 2019, respectively.

He is currently a Postdoctoral Research Fellow with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include machine learning, data mining, and computer vision.

Changmiao Hu received the B.S. degree in mathematics from North China Electric Power University, Beijing, China, in 2006, and the Ph.D. degree in signal and information processing from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2012.

From 2012 to 2018, he was an Assistant Researcher with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences. He is currently with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include remote sensing image processing, geometric correction, and atmospheric correction.

Lianzhi Huo received the B.S. degree in geographic information system from Huazhong Agricultural University, Wuhan, China, in 2007, and the Ph.D. degree in signal and information processing from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2012.

From 2012 to 2015, he was an Assistant Researcher with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences. From 2015 to 2018, he was a Postdoctoral Fellow with the University of Idaho, USA, working on a NASA project. He is currently with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include image classification and machine learning.

Dr. Huo is a Reviewer for more than 15 international journals, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, IEEE Geoscience and Remote Sensing Letters, and *Neurocomputing*.

Ping Tang received the Ph.D. degree in mathematics from Beijing Normal University, Beijing, China, in 1996.

She spent two year as a Postdoctoral Researcher with the Institute of Geophysics, Chinese Academy of Sciences, Beijing, China. She is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. Her research interests include remote sensing image processing and applications, pattern classification, and big data analytics.