Pyramid Global Context Network for Image Dehazing

Dong Zhao, Long Xu, Lin Ma, Jia Li, Yihua Yan

***Abstract*—Haze caused by atmospheric scattering and absorption would severely affect scene visibility of an image. Thus, image dehazing for haze removal has been widely studied in the literature. Within a hazy image, haze is not confined in a small local patch/position, while widely diffusing in a whole image. Under this circumstance, global context is a crucial factor in the success of dehazing, which was seldom investigated in existing dehazing algorithms. In the literature, the global context (GC) block has been designed to learn point-wise long-range dependencies of an image for global context modeling; however, patch-wise long-range dependencies were ignored. To image dehazing, patch-wise long-range dependencies should be highlighted to cooperate with patch-wise operations of image dehazing. In this paper, we first extend the point-wise GC into a Pyramid Global Context (PGC), which is a multi-scale GC, after undergoing the pyramid pooling. Thus, patch-wise long-range dependencies can be explored by the PGC. Then, the proposed PGC is plugged into a U-Net, getting an attentive U-Net. Further, the attentive U-Net is optimized by importing ResNet's shortcut connection and dilated convolution. Thus, the finalized dehazing model can explore both long-range and patch-wise context dependencies for global context modeling, which is crucial for image dehazing. The extensive experiments on synthetic databases and real-world hazy images demonstrate the superiority of our model over other representative state-of-the-art models from both quantitative and qualitative comparisons.**

***Index Terms*—Image dehazing, deep learning, global context modeling, pyramid global context (PGC), dilated residual U-Net (DRU).**

I. Introduction

Due to the existence of atmospheric particles, such as fog, haze, rain, dust, and fume in bad weather conditions, outdoor images capture often suffer from atmospheric absorption and scattering, leading to contrast loss, color degradation, and saturation attenuation. In practice, even in a sunny day, the atmosphere is not absolutely free of any floating particle. Consequently, the haze still exists in a captured image, especially for distant scenes [1]. The hazy image would lower the efficiency of high-level computer vision tasks and applications [2], [3], [4], such as object detection, semantic segmentation, video surveillance, and remote sensing. Therefore, over the past decades, removing haze from the image, namely image dehazing, has received significant attention in the computer vision community.

Dehazing, deraining [5], [6], desnowing [7] and raindrops removal [8] are closely connected. However, they are caused by different degradation processes under different weather conditions, so they follow different physical principles [9]. For image dehazing, scene depth is crucial for accurate transmission map estimation and, therefore, efficient image dehazing. To estimate the depth information, early methods resorted to auxiliary information, such as polarization [10], multiple images [11], [12], [13], and existing 3D geographic models [14]. They are, however, limited by the availability of auxiliary information. Thus, haze-relevant priors [1] were investigated, realizing haze removal from a single image without auxiliary information. Those hand-crafted haze-relevant priors estimate coarse transmission map t according to the well-received atmospheric scattering model [1]:

$$I_H(x) = I_D(x)t(x) + A(1 - t(x)), \tag{1}$$

where x represents a spatial location; I_H and I_D are the observed hazy image and clear scene radiance, respectively; t is transmission map; A is global atmospheric light. Of these priors, dark-channel prior (DCP) [1], color-lines prior [15], color attenuation prior (CAP) [16], difference-structure-preservation prior [17] and color ellipsoid prior (CEP) [18] are of patch-wise assumptions, assuming that the haze is constant in a local patch, while non-local prior (NLP) [19] estimates transmission map in a non-local manner, assuming that the degradation is different for every pixel. Although prior-based methods are usually simple and effective for some scenes, they may fail in practical scenarios as the prior does not hold.

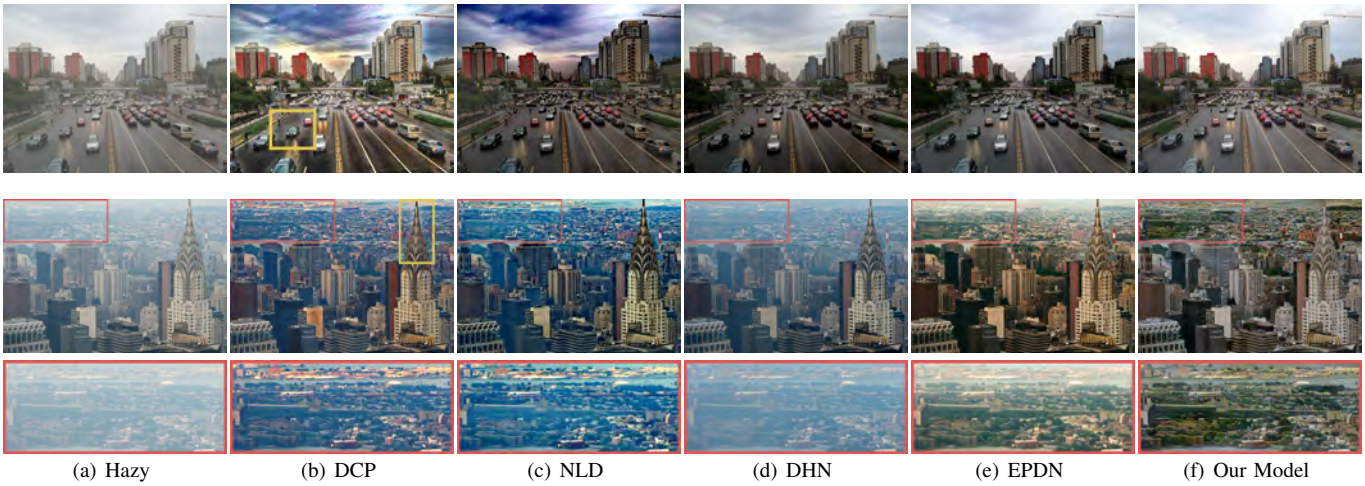
The primary deep learning-based methods, such as DehazeNet (DHN) [20], Multi-Scale CNN (MSCNN) [22] and All-in-One Dehazing (AOD) [23], have demonstrated great success of CNN for image dehazing. Trained on large-scale databases, these methods can extract effective features to estimate fine transmission directly or $\mathcal{K}$ [23] map. However, these methods still follow the conventional dehazing model as in Eq. (1) and mostly utilize low-level image features. Consequently, they are unable to effectively learn global context information of an image, compromising dehazing efficiency. Profiting from skip-connection or/and density connection, deep networks (such as Densely Connected Pyramid Dehazing Networks (DCPDN) [24], Gated Fusion Networks (GFN) [25] and Enhanced Pix2pix Dehazing Networks (EPDN) [21])

Corresponding author: Long Xu (email: lxu@nao.cas.cn).

D. Zhao, L. Xu and Y. Yan are with the Key Laboratory of Solar Activity, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China (e-mails: zhaodong@buaa.edu.cn, lxu@nao.cas.cn, yyh@nao.cas.cn).

L. Ma is with Meituan, Beijing, 100102, China (e-mail: forest.linma@gmail.com).

J. Li is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, 100191, China (e-mail: jiali@buaa.eud.cn).



(a) Hazy (b) DCP (c) NLD (d) DHN (e) EPDN (f) Our Model

Fig. 1. Dehazing for hazy sky regions (images in the first row) and distant scenes (images in the second row). The last row illustrates the corresponding zoom-in of the red rectangles labeled in the second row. (a): hazy images. (b)-(f): results of DCP [1], NLD [19], DHN [20], EPDN [21] and our model.

have been proposed by stacking more convolution layers. In this way, long-range dependencies can be explored to some extent since the receptive field of convolution becomes larger as the convolution layer goes deeper than primary learning-based methods. However, simply stacking convolution layers is not computationally efficient. Moreover, models that solely rely on convolutions exhibit limited ability in capturing long-range dependencies, partially due to difficulties of exchanging information between distant points [26].

Recent literatures [27], [26], [28], [29] have indicated that long-range dependencies can improve the performance of high-level vision tasks. However, they have not yet well investigated in existing dehazing methods. Fig.1 shows two cases of hazy images with sky regions (the first row) and distant scene regions (the second row). The DCP [1] and NLD [19] remove heavy haze from distant scenes successfully, as shown in the second row of Fig.1 (b) and (c). However, both of them yield over-saturation in the sky regions, as shown in the first row of Fig.1 (b) and (c). In the first row of Fig.1 (d) and (e), deep learning-based methods of DHN [20] and EPDN [21] are visually appealing in sky regions; however, they still contain a small amount of haze in distant scenes as shown in the second row of Fig.1 (d) and (e). The reason is that, without an effective long-range dependency modeling, large-scale context information about some cases (such as the sky and distant scenes) cannot be sufficiently mined from an input image.

In this paper, we exploit long-range dependencies among both points and patches for further profiting image dehazing. Unlike previous productions [30][31] that resorted to an additional network for this purpose, we characterize long-range dependencies by an effective and lightweight module/block of global context modeling, namely Global Context (GC) block [26]. In practice, GC block [26] is usually embedded into a deep network, such as ResNet [26][32], to realize global context modeling. Despite its powerful representative ability, GC block [26] only aggregates pixel-wise features together to form a global context feature, ignoring patch-wise dependencies. In fact, patch-wise context dependencies should be more appreciated in image dehazing. They provide patch-wise global context information that can be used in patch-wise dehazing. It has been found that patch-wise operation could allow image dehazing to avoid over-saturation caused by the pixel-wise operation of dehazing [33], [34]. Therefore, we think combining point-wise and patch-wise context modelings is better than using either of them individually. For this purpose, a new global context modeling block, namely Pyramid Global Context (PGC) block, is proposed. It is derived from the GC block [26] undergoing the spatial pyramid pooling [35], [36], [37], which can thereby learn multi-scale context dependencies.

We then embed the proposed PGC block into a U-Net, which is composed of four-stage contracting paths (Encoder) and expansive paths (Decoder). Each encoder/decoder block consists of a down-scale/up-scale convolutional layer and a dilated residual bottleneck (DRB) block. DRB is a small slice of the dilated residual network [38]. This network is built on a CNN, replacing conventional convolution by a dilated one, and also importing ResNet's shortcut connection [32], so it inherits the merits of both dilated convolution and ResNet. We abbreviate the final network as "PGC-UNet", dilated residual U-Net as "DRU" in the following content. In this PGC-UNet, U-Net infrastructure can retain localization accuracy with skip-connections, which concatenate feature maps of the Encoders and the Decoders; meanwhile, PGC can improve the ability of global context modeling. Additionally, dilated convolution can enlarge the receptive field of convolution.

The big difference between the proposed PGC-UNet and related works lies in that patch-wise haze-relevant feature, and therefore patch-wise operator of dehazing is learned by using a PGC block, which is consistent with the fact that haze is patch-wise instead of pixel-wise. In PGC block, long-range dependencies of multiple scales are explored for global context modeling. Contributions of this work can be summarised as follows:

- A PGC block is proposed, exploring long-range de-

pendencies among not only points but also patches. It provides an efficient solution to non-local/global context modeling.
- A new neural network is proposed for image dehazing, plugging PGC/GC blocks into a U-Net, getting an attentive U-Net, which is further optimized by the DRB block.
- Extensive experiments and ablation studies are performed on public databases to conclude the best solution of image dehazing, which could guide practical application and provide references to the peers.

II. Related Works

A. Deep Learning-based Single Image Dehazing

Most deep learning-based methods seek mapping relation between image features and parameters in (1). Earliest methods, including DHN [20], MSCNN [22], AOD [23], were designed as trainable end-to-end CNN-based architectures for medium transmission (or a $\mathcal{K}$ map in AOD [23]) estimation. Many works such as DCPDN [24], Iteration-Wise Priors (IWP) [39], and Dual-Path in Dual-Path Networks (DPDPN) [40], tended to strictly follow the physics-driven scattering model, by jointly estimating the transmission map t, atmospheric light A and clear scene radiance I_D. However, these methods heavily rely on training database and accurate estimation of t and A, which may be greatly limited in real-world scenarios. Unlike these methods, blind learning models disentangles image dehazing from the physical scattering model, no need to estimate t or/and A. For example, GFN [25] produces a haze-free image via fusing dehazed patches from three feature maps. EPDN [21] implements a pyramid pooling enhanced pix2pixHD [41] model, where the adversarial learning is repeated to contribute a powerful generator which can directly generate a haze-free image from a hazy one. GridDehazeNet (GridDN) [42] proposes a multi-scale estimation on a grid network [43], and embeds a channel-wise attention block into a network. However, all these methods do not perform well for capturing long-range dependencies, partially due to the difficulties of exchanging information among points far away from each other.

B. Global Context Modeling

1) Non-Local (NL) Block: To extract the global context of a visual scene, previous efforts mainly stack more convolution layers. However, increasing convolution layers is not computationally efficient and hard to optimize [27]. The NL block [27], as shown in Fig.2 (a), is proposed to model long-range dependencies using one layer, via self-attention mechanism [44]. Denote input and output features as $X \in \mathbb{R}^{h\times w\times c}$ and $Y \in \mathbb{R}^{h\times w\times c}$, where h, w, and c represent image height, width and channel dimensions. The query feature map Q, key feature map K and value feature map V are formed by three 1×1 convolutions W_q, W_k and W_v

$$Q = W_q(X), \quad K = W_k(X), \quad V = W_v(X); \tag{2}$$

where $Q \in \mathbb{R}^{h\times w\times c'}$, $K \in \mathbb{R}^{h\times w\times c'}$, $V \in \mathbb{R}^{h\times w\times c}$, c' is the number of channels of output feature maps. Then, c' feature maps are flattened to a vector of the size $n \times c'$, where $n = hw$ represents the total number of the spatial locations. The output of NL block [27] is

$$Y = W_z(\mathcal{N}_n(Q \times K^T) \times V) + X, \tag{3}$$

where W_z is a 1×1 convolutional layer. In [27], it chooses *softmax* as the normalizing function $\mathcal{N}_n$ which has been proved to work well in many tasks [27], [26]. Eq. (3) is applied to each pair of positions. There are totally n^2 pairs, resulting in $\mathcal{O}(n^2)$ memory complexity and $\mathcal{O}(c'n^2)$ computational complexity.

The memory and computational costs of NL attention grow quadratically with the resolution of the input. To overcome this drawback, Region-Level Non-Local (RLNL) [45][46] firstly divides input feature map into a grid of regions/patches. Then, it models long-range attention separately only within each region. Thus, this method cannot directly learn long-range dependencies across different regions. Asymmetric Non-Local Networks (APNL) [47] selects (1, 3, 6, 8) to be output anchor points of pyramid spatial pooling, ignoring point-wise context dependencies. It cannot give sufficient feature statistics about global context information to learn accurate haze-relevant features.

2) Global Context (GC) Block: Unlike non-local (NL) block [27] which performs global context modeling for each query position/point, surprisingly, GC block [26] found that global context modeled by NL block are almost the same for different query positions within an image, indicating only query-independent dependency is learned [26]. In addition, the 1×1 convolution W_v in NL block is removed and replaced by a bottleneck transform module in GC block as shown in Fig.2(b). This process can provide us a lightweight Squeeze-and-excitation (SE) block [48]. The bottleneck transform module consists of (1) a 1×1 convolution W_{v1} to compress the number of channels from c to $c/r(r > 1)$, (2) a layer

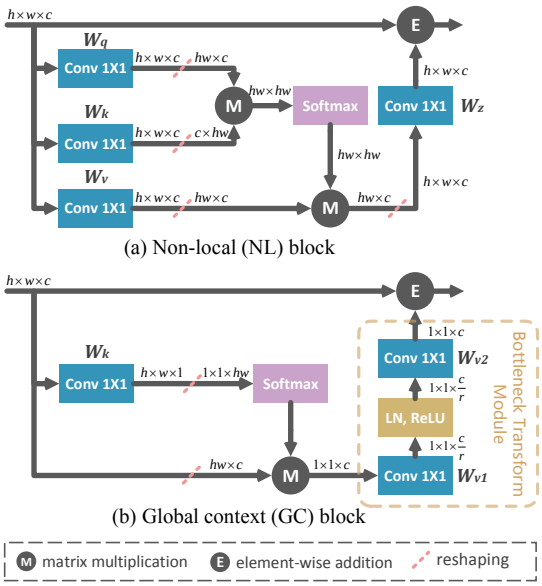


Fig. 2. Non-local (NL) block, global context (GC) block [26]. (LN is the short name of Layer Normalization, h, w and c are the height, width and channel of the input feature maps, respectively, and r is the bottleneck ratio.)

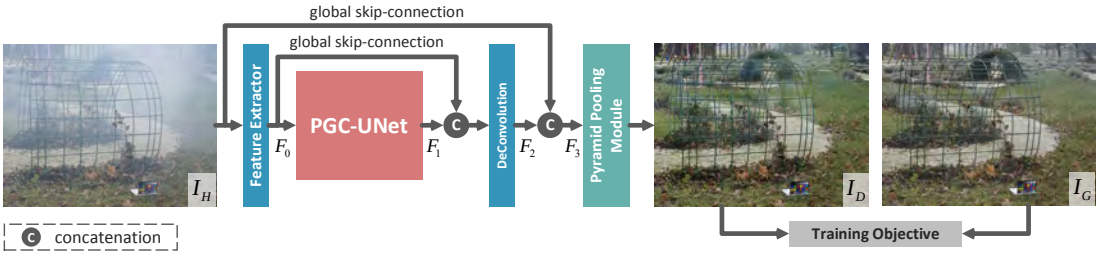


Fig. 3. Architecture of the proposed dehazing network. PGC: pyramid global context block. I_H, I_D and I_G represent hazy image, dehazed image and the ground truth, respectively.

normalization (LN), (3) a nonlinear function (ReLU) and (4) a 1×1 convolution W_{v2} to restore the number of channels from c/r to c. The output of GC block [26] is

$$Y = \mathcal{T}(\mathcal{N}_n(\bar{K})^T \times \bar{X}) + X. \tag{4}$$

where the reshaped key features are represented by $\bar{K} \in \mathbb{R}^{n \times 1}$, $\bar{X} \in \mathbb{R}^{n \times c}$ is the reshaped X, $\mathcal{T}$ is the bottleneck transform module (as shown in Fig.2 (b)), $\mathcal{N}_n$ is the normalizing function.

Despite good representation, the context modeling block in GC block [26] only aggregates the features of all positions together to form a global context feature. It is unable to learn the dependencies between patches effectively. Thus, to boost the abilities to learn patch-wise long-range dependencies on different scales, a simple yet effective global context modeling, namely PGC block, is proposed in this work. PGC inherits the merits of GC block [26] with lightweight and efficient point-wise long-range dependency modeling; additionally, it boosts patch-wise long-range dependency modeling.

III. Proposed Dehazing Method

A. Framework

Let I_H and I_D denote input hazy image and output dehazed image, respectively. Image height, width and channel are denoted by H, W and C ($C = 3$ for color image). The proposed network framework is shown in Fig. 3, which mainly consists of a feature extractor, a PGC aided U-Net module (PGC-UNet), a deconvolutional layer, and a pyramid pooling module.

he feature extractor aims to extract low-level feature F_0 from the hazy image. It contains one convolutional layer with kernel size is 7×7 and stride 1. F_0 is then fed into the PGC aided U-Net (PGC-UNet), and the output is F_1. The details of PGC-UNet will be introduced in subsection III-C.

We then use a deconvolutional layer, with kernel size is 7×7, to expand the output of the PGC-UNet to the original size of hazy image I_H and compress channels from 64 to 32. The output denoted as F_2, is further concatenated with hazy image I_H, and we obtain the feature maps F_3, as shown in Fig.3. Note that we use long-rang global skip-connections twice, i.e., $\mathcal{C}[F_1, F_0]$ and $\mathcal{C}[F_2, I_H]$, where $\mathcal{C}[\,]$ denotes the concatenation operation. These global skip-connections can transmit low-level features to F_3. In addition, they enable the mechanism of residual learning, facilitating gradient back-propagation of the deep network.

Finally, inspired by [36], [24] and [21], F_3 is fed into a pyramid pooling module to make sure the details of features from different scales are embedded in the final dehazed image I_D. Feature maps on different scales provide different receptive fields, which helps to reconstruct an image on various scales.

B. Pyramid Global Context (PGC) Block

From our investigation, we found that global context modeling is crucial for haze-relevant features learning. Original GC block [26] simplifies non-local block [27] by explicitly using a query-independent attention map for all query positions/points. It is point-wise, so patch-wise dependencies among patches are ignored. We think that this point-wise spatial attention cannot accord with the well-known patch-wise prior in previous works [1]. The pixel-wise prior based methods, such as Non-local prior [19], commonly lead to over-saturation in recovered image. The main reason is that, in an image patch where the depth/transmission is constant, pixel-wise approaches estimate transmission per pixel, leading to over-estimation of haze from its actual value [33].

Therefore, in this paper, we aim to enhance the GC block [26] to learn both point-wise and patch-wise long-range haze-relevant attentions. We propose to utilize pyramid spatial pooling to sample sparse anchor point for each region/patch with pooling sizes $\ell_i = 1/2^i$, $i \in \{0, 1, 2, 3\}$, as shown in Fig.4. The pooled feature maps X_{ℓ_i} is expressed as:

$$X_{\ell_i} = \mathcal{P}_{\ell_i}(X) \in \mathbb{R}^{h_i, w_i, c}, \quad i \in \{0, 1, 2, 3\}, \tag{5}$$

where $\mathcal{P}_{\ell_i}(\cdot)$ is the pyramid spatial pooling operator with pooling size ℓ_i; h_i, w_i and c represent height, width and channel dimensions of X_{ℓ_i}, and we have $h_i = h\ell_i$ and $w_i = w\ell_i$.

Then, for each pooled feature maps X_{ℓ_i}, we use the same context modeling module in GC block [26] (see Fig.4) to mine long-range dependencies between the anchor points. The output of GC is denoted by $K_{\ell_i} = W_{k\ell_i}(X_{\ell_i})$ at scale i. Given a set of weights $\{\alpha_{\ell_i}, i \in \{0, 1, 2, 3\}\}$, the outputs of GC at four scales are concatenated to form the output attention map of PGC:

$$\bar{X}_{att} = \mathcal{N}_n(\mathcal{C}[\bar{K}_{\ell_i}; i \in \{0, 1, 2, 3\}]^T) \times \mathcal{C}[\alpha_{\ell_i} \bar{X}_{\ell_i}; i \in \{0, 1, 2, 3\}], \tag{6}$$

where $\bar{K}_{\ell_i} \in \mathbb{R}^{n_i \times 1}$ (n_i=$h_i \times w_i$), $W_{k\ell_i}$ is the 1×1 convolution at scale i, $\bar{X}_{\ell_i} \in \mathbb{R}^{n_i \times c}$ is reshaped from X_{ℓ_i}, $\mathcal{N}_n$ is the

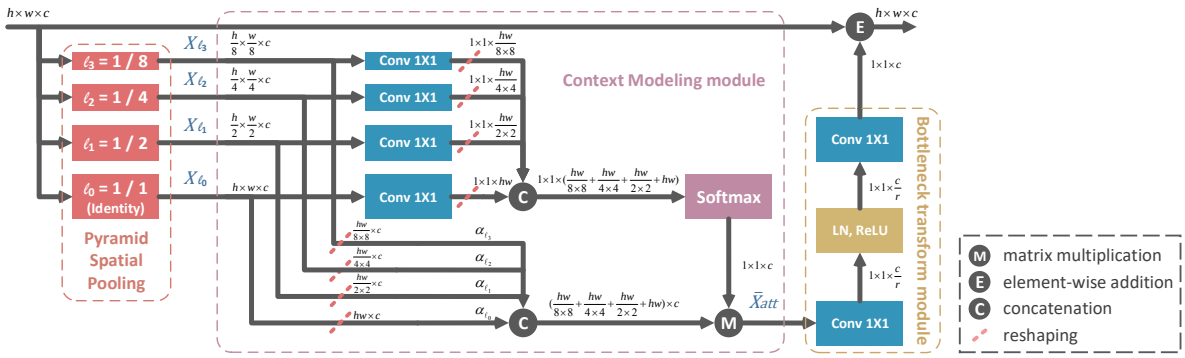
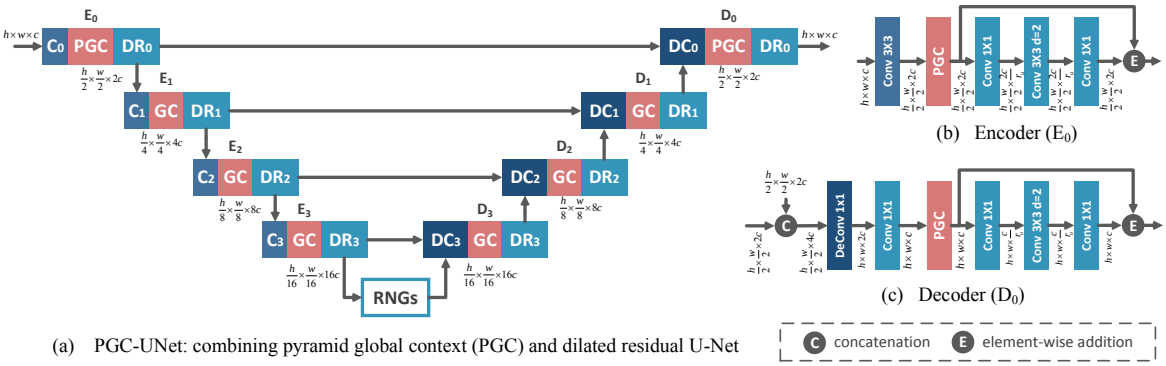


Fig. 4. The pyramid global context (PGC) block. LN:Layer Normalization. X_{ℓ_i} and α_{ℓ_i} are the pooled feature and concatenated weight of scale ℓ_i, respectively. X_{att} is the output attention map of context modeling module. h, w and c are the height, width and channel of the input feature maps, respectively. r is the bottleneck ratio.



(a) PGC-UNet: combining pyramid global context (PGC) and dilated residual U-Net

(b) Encoder (E_0)

(c) Decoder (D_0)

Fig. 5. Framework of the PGC aided U-Net (PGC-UNet). We adopt bottleneck layers both in encoder and decoder, with the ratio is r_u. h, w and c are the height, width and channel number of input features. C_j/DC_j: the down-/up-scale convolutional layer in E_j/D_j. DR_j: Dilated ResNet bottleneck. $RNGs$: ResNet Groups. $d = 2$ represents the delation=2.

normalizing function, which is *softmax* function here. $\mathcal{C}[\ ]$ denotes the operation of concatenation. α_{ℓ_i} is weight at scale i, and in our work, it is defined as:

$$\alpha_i = (\frac{1}{\ell_i})^2,\ \ i \in \{0, 1, 2, 3\}. \tag{7}$$

Note that the pooling size of ℓ_0=1/1 is the same as the basic GC [26], while other scales learn patch-wise long-range dependencies of different scales.

The final output of the PGC block is:

$$Y = \mathcal{T}(\bar{X}_{att}) \ + \ X \tag{8}$$

where $\mathcal{T}$ represents a bottleneck transform module as shown in Fig.4. PGC block is lightweight and effective for long-range dependency modeling. For dehazing purpose, it is then plugged into a U-Net with dilated convolution and ResNet's shortcut connection to form the final model, namely PGC-UNet, which will be introduced in subsection III-C.

C. PGC aided U-Net (PGC-UNet)

U-Net [49] is equipped with skip-connection, which directly propagates features of a lower layer to a higher layer, skipping the layers between them, which endows it with a good capability of image restoration. Thus, U-Net is selected to be the backbone of the proposed PGC-UNet. In this paper, we first explore PGC for global context modeling. Then, the PGC block is plugged into a U-Net, getting the proposed PGC-UNet. It should be pointed that here the U-Net is a dilated residual optimized one [38]. The framework of the proposed PGC-UNet is shown in Fig.5, which mainly consists of three parts: contracting path (encoder), ResNet Groups (RNGs), and expansive path (decoder).

Contracting path. The contracting path (encoder path) consists of four stages, i.e. $E_j(j$=$0, 1, 2, 3)$ shown in Fig.5(b). Each encoder block consists a down-scale layer ($C_j(j$=$0, 1, 2, 3)$, downscaling spatial resolution with stride s=2, and doubling the number of feature maps) followed by a basic Dilated ResNet Bottleneck (DRB) [38] ($DR_j(j$=$0, 1, 2, 3)$). In addition, the proposed PGC module is embedded between them for global context modeling. Here, DRB block is used because dilated convolution [50] can provide enlarging receptive field without need of downscaling image spatial resolution. We use the same dilated factors (d=2) for all encoder blocks.

ResNet groups (RNGs). As shown in Fig.5(a), the feature transformation module in the proposed network consists of several basic ResNet [32] Groups (RNGs). RNGs are equipped with shortcut connections of ResNet, so it allows us to train a

deeper network which has the better capability of representing very complex functions.

Expansive path. The expansive path (decoder path) consists of four stages, i.e. $D_j(j{=}0,1,2,3)$ shown in Fig.5 (c). Each decoder block is similar to the corresponding encoder block, i.e., a PGC block followed by a DRB block. The difference lies in that each decoder has two layers to merge the aggregated features: the first layer is a transposed convolution used to expand the resolution of the concatenated features and halve the number of channels; the second 1×1 convolution further halves the number of channels, so that the output has the same number of channels with the one of the encoders in up one stage.

It should be pointed that we only plug the PGC block into the first stage of encoder/decoder, i.e., E_0 and D_0, as shown in Fig.5 (a). This is mainly because the PGC module is most efficient at the largest resolution. Its efficiency declines gradually going along with the increase of stage since feature maps are accordingly down-scaled smaller and smaller. The GC block [26] can be seen as a special form of the PGC with a single pooling size of $1/2$.

D. Training Objective

The optimization objective of the proposed network is defined as following:

$$\mathcal{L} = \lambda_{gan}\mathcal{L}_{gan} + \lambda_{fm}\mathcal{L}_{fm} + \lambda_{ssim}\mathcal{L}_{ssim} + \lambda_{l1}\mathcal{L}_{l1}, \quad (9)$$

where $\mathcal{L}_{gan}$ is the adversarial loss, $\mathcal{L}_{fm}$ is the feature matching loss, $\mathcal{L}_{ssim}$ is the SSIM loss, and $\mathcal{L}_{l1}$ is the L_1 loss.

Adversarial Loss: Denote the ground truth by I_G. In our work, the adversarial loss of conditional GAN is defined as below:

$$\begin{aligned}\mathcal{L}_{gan} = \arg\min_{\boldsymbol{G}}\{\max_{\boldsymbol{D}_1,\boldsymbol{D}_2}\{\sum_{k=1,2}(\mathbb{E}_{I_H,I_G}[\log \boldsymbol{D}_k(I_H,I_G)] \\ +\mathbb{E}_{I_H}[\log(1-\boldsymbol{D}_k(I_H,I_D))])\}\},\end{aligned} \quad (10)$$

where $\boldsymbol{G}$ is our dehazing model, and we have $I_D = \boldsymbol{G}(I_H)$; $\boldsymbol{D}_k$ (k=1,2) are the discriminators. $\mathbb{E}$ represents the mean operation on a batch of training samples. It should be pointed out that a two-scale discriminator is used here as shown in Fig. 6, where each input image and its half sampling version respectively undergo two different discriminators with the same network structure, but the different size outputs. It has been proved multi-scale discriminator has good property for guiding the generator to generate images from coarse to fine granularity [21]. In addition, the discriminator adopts the concept of convolutional "PatchGAN [51] which means that each image is divided into small patches (e.g., 8×8) rather than a whole for discriminating real/false (real: positive or false: negative). The two scales of the discriminator have the same patch size, so they output a $H/8\times W/8$ and a $H/16\times W/16$ binary matrices where "0" and "1" indicate the probability of each patch being real or false. From Fig. 6, two mean square errors (MSEs) are computed for the two scales independently, and then combining them together gives the final loss of the two-scale discriminator.

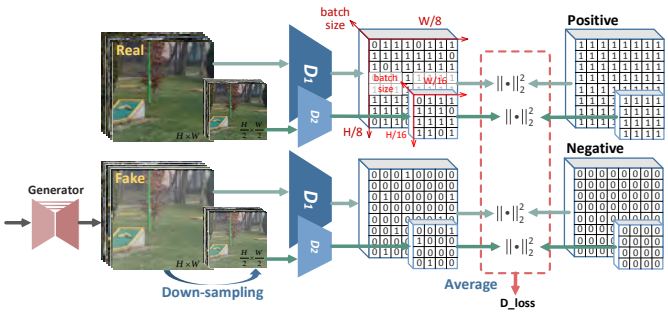


Fig. 6. Architectures of multi-scale discriminators. Note that, we omit the hazy inputs of the discriminators $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ to simplify the illustration.

Feature Matching Loss: Feature matching addresses the instability of GAN by specifying a new objective for the generator that prevents it from overfitting on the current discriminator.

$$\mathcal{L}_{fm} = \min_{\boldsymbol{G}}\mathbb{E}_{I_D,I_G}\sum_{k=1,2}\sum_{z=1}^{Z}\frac{1}{N_z}(\parallel \boldsymbol{D}_k^z(I_D) - \boldsymbol{D}_k^z(I_G)\parallel_1), \quad (11)$$

where Z is the total number of layers used for feature extraction, N_z is a number of elements in each layer, $\boldsymbol{D}_k^z$ is the operator of the feature extraction of the z-th layer in $\boldsymbol{D}_k$. The adversarial loss together with feature matching loss is used to make the GAN module learn global information and recover image structure by using multi-scale features.

SSIM Loss: Since human eyes are the final judger for evaluating the efficiency of a dehazing algorithm, measuring image quality from perspective of human visual system (HVS) is desirable in image processing tasks. Therefore, SSIM, which has good correlation with HVS, with simple formulas and easy implementation, is incorporated into loss function as:

$$\mathcal{L}_{ssim} = 1 - \mathbb{E}_{I_D,I_G}SSIM(I_D,I_G), \quad (12)$$

where $SSIM(\cdot)$ is the SSIM of the paired images.

L_1 **Loss**: It facilitates the feature selection in the model optimization, and it is defined as:

$$\mathcal{L}_{l1} = \mathbb{E}_{I_D,I_G}\parallel I_D - I_G\parallel_1 . \quad (13)$$

IV. Experiments

A. Database

For training our model, we collect the samples from the databases of O-HAZE [52], I-HAZE [53], and RESIDE [54] to form a mixed training database, as demonstrated in Table I. Besides, images from O-HAZE [52] and I-HAZE [53] are divided into small patches and resized to the same size (512×512) to be training samples, as shown in Fig.7. Since the O-HAZE [52] and I-HAZE [53] provide real hazy images and their haze-free images, samples from these two databases account for a larger proportion in our training database. RESIDE [54] is widely adopted as the benchmark database in many dehazing works [4], [21], [42] due to its large scale and diverse image contents. The RESIDE [54] contains synthesized hazy images in Indoor Training Set (ITS) and Outdoor Training Set (OTS)

TABLE I
MIXED TRAINING AND TESTING DATABASES.

	Training Database				Testing Database					
	O-HAZE	I-HAZE	BESIDE-OTS		O-HAZE	I-HAZE	SOTS(outdoor)	Middlebury	HazeRD	Real-World
Number	5110	3982	5000	Number	587	300	350	23	70	109

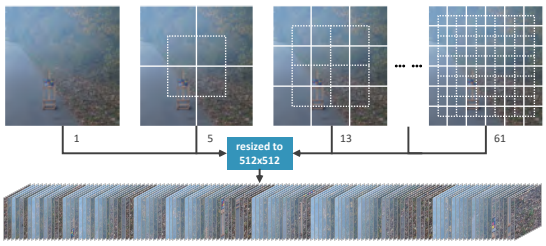


Fig. 7. Patch-based training strategy on O-HAZE [52] and I-HAZE [53] databases. Each sample is equally divided into 1, 5, 13, 25, 41 and 61 patches, and further resized to 512×512.

databases. In our mixed database, 5000 samples from the OTS database are included.

To demonstrate generalization ability of our model, we test it on several databases which have no overlap with the training database, including O/I-HAZE [52][53], SOTS (outdoor) [54], Middlebury [55], HazeRD [56] and real-world hazy images. HazeRD [56] contains 14 haze-free images of real outdoor scene and corresponding depth maps. For each image, we synthesize 5 hazy images with fixed atmospheric light $\boldsymbol{A} = [0.76, 0.76, 0.76]$ and different scattering coefficients $\beta = \{0.05, 0.1, 0.2, 0.5, 1\}$. Moreover, for comparisons on real-world images, we collect 109 hazy images mainly from [25] and [57].

B. Implementation Details

The proposed model is developed on PyTorch deep learning package. It is trained on our training database with 14,092 512×512 images from O-HAZE [52], I-HAZE [53] and OTS [54] databases. During training, we adopt ADAM [58] as the optimization algorithm with a batch size of 6. The initial learning rate is set to 0.0002 for both generator and discriminator, where the exponential decay rates are set as (β_1, β_2)=$(0.6; 0.999)$. Following [59], [60], [61], we use instance normalization layer [62] instead of batch normalization [63] in our network except GC block [26] and PGC block. According to (7), for all of the comparison experiments, we set the weight α=$[1, 4, 16, 64]$ for levels (ℓ_0=1, ℓ_1=$1/2$, ℓ_2=$1/4$, ℓ_3=$1/8$), respectively (i.e. the PGC-UNet$-\alpha 64$ model mentioned in the subsection IV-E). The parameters of our hybrid loss function are set as (λ_{gan}, λ_{fm}, λ_{ssim}, λ_{l1})=(0.001, 0.15, 1, 1). For more details, please access the source codec via https://github.com/phoenixtreesky7/PGC-DN.

C. Performance Comparisons with State-of-the-art Methods

In this subsection, extensive experiments are conducted to verify the effectiveness of our proposed model, compared with state-of-the-art methods of prior-based and deep learning-based methods, including DCP [1], Multi-Scale Fusion (MSF) [64], CAP [16], NLP [19], DHN [20], MSCNN [22], GFN [25], EPDN [21] and GridDN [42].

1) Evaluations on Synthetic Databases:

First, we evaluate the proposed model on synthetic hazy images which are selected from O/I-HAZE [52], [53], SOTS (outdoor) [54], Middlebury [55], HazeRD [56] databases, as shown in Table I. We strictly follow the author's recommendations in their papers for performance comparisons. The learning-based methods pre-trained testing models are downloaded from their open sources. To quantitatively evaluate our model, we employ the peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [65], and feature similarity index for color image (FSIMc) [66]. Since the FSIMc uses the phase congruency, gradient magnitude, and chromatic features to represent complementary aspects of image visual quality beyond the SSIM. We also employ the Color Naturalness Index (CNI) [67][68][69][70] to evaluate the color quality of the dehazed image. In CNI, the naturalness of the reproduced image is defined as the "degree of correspondence between human perception and reality world" [68][69]. In our experiments, we follow the same set of CNI in [68].

Comparisons on the O-HAZE database. Fig.8 (1) shows the qualitative comparisons on O-HAZE [52] database. DCP [1] and NLD [19] tend to cause blur and color distortions, e.g., grassland in Figs.8 (1-b) and (1-c). AOD [23] and EPDN [21] with fine details, appear more visually pleasant than those of NLD [19], MSCNN [22] and GridDN [42]. However, the color is distorted in Figs.8 (1-f) and (1-g) compared to ground truth. Our model achieves the best visual quality on outdoor database O-HAZE [52], as shown in Fig.8 (1-i). Table II also reveals that the best PSNR, SSIM, FSIMc and second best CNI on O-HAZE database are achieved by our model.

Comparisons on the SOTS database. From Fig.8 (2), it can be observed that all algorithms have good visual perception on SOTS [54] database. However, MSCNN [22], AOD [23] and GridDN [42] remains haze at distant scenes. The proposed model has pretty good visual perception as shown in Fig. 8 (2-i). From Table II, DHN [20], EPDN [21], GridDN [42] and our model obtains the top four highest scores of the four metrics. The GridDN [42] achieves the best SSIM score, which is 0.014/1.4% SSIM larger than our model. With respect to PSNR, our model outperforms the GridDN [42] by 0.49 dB. Regarding CNI metric, our model is also the best among all compared algorithms.

Comparisons on the HazeRD database. To further validate the robustness of our model to different intensity of haze, we test it on HazeRD [56] database. Figs.8 (3) and (4) illustrate

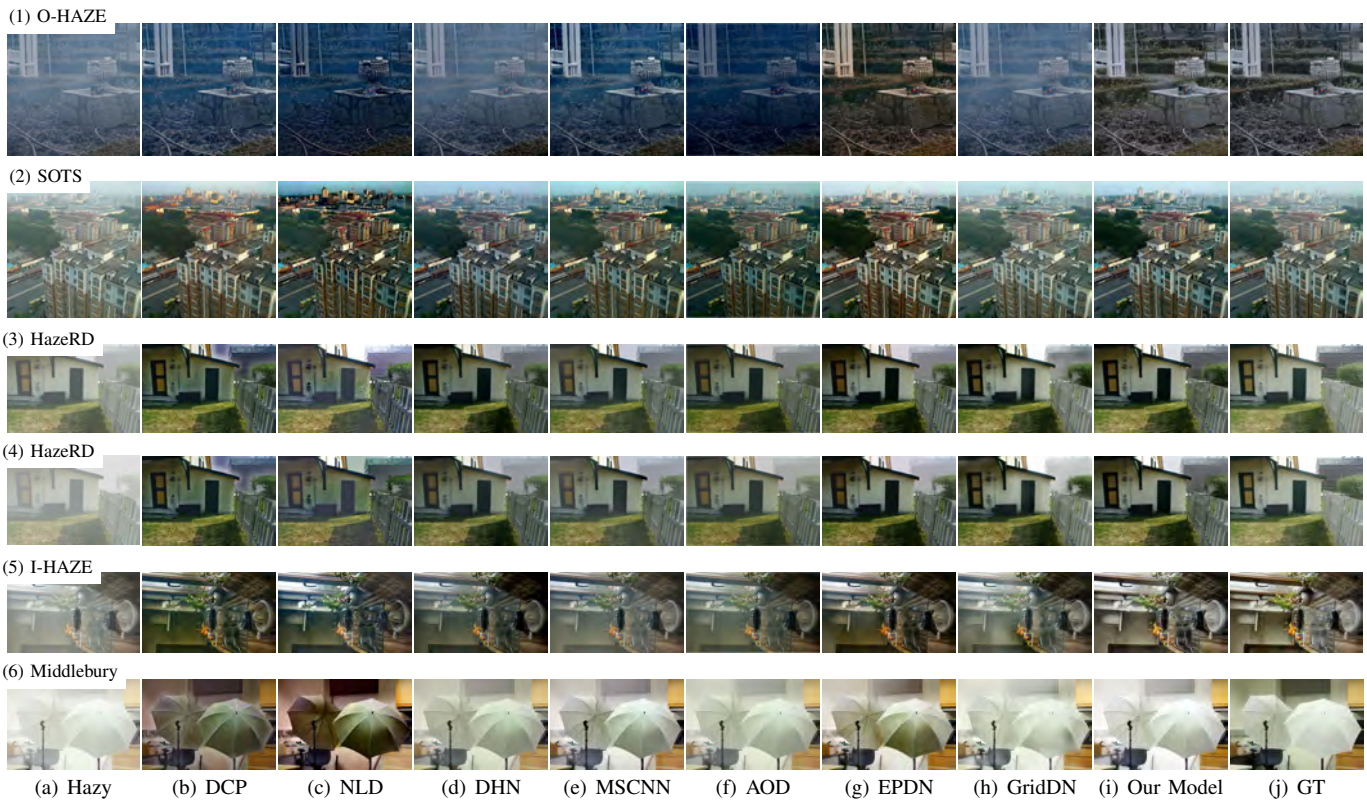
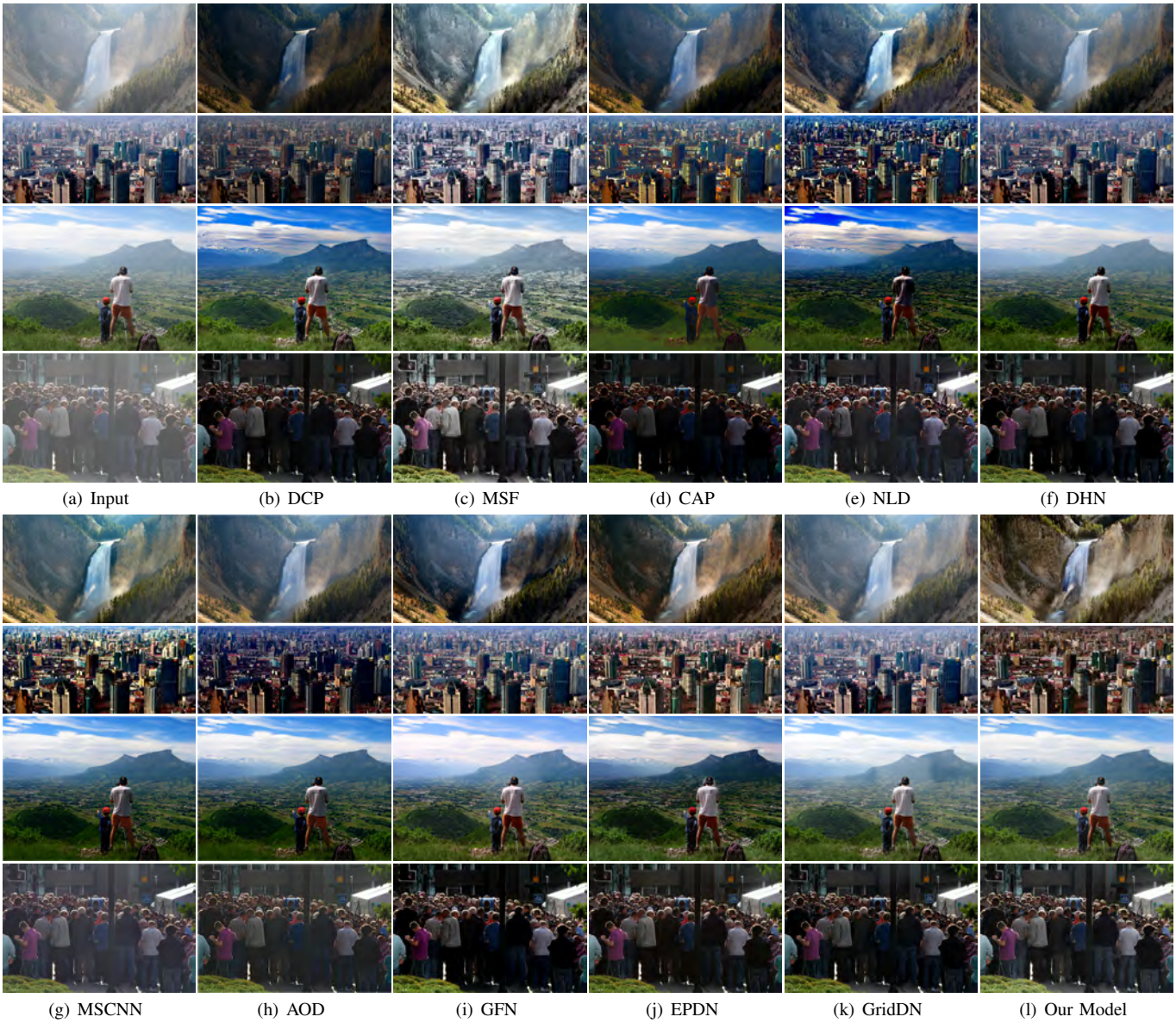


Fig. 8. Qualitative results of synthetic image dehazing. (a) is the hazy image. (b)-(i) are the results of DCP [1], NLD [19], DHN [20], MSCNN [22], AOD [23], EPDN [21], GridDN [42] and our model. (j) is the ground truth.

TABLE II
PERFORMANCE COMPARISONS IN TERMS OF AVERAGE PSNR, SSIM, FSIMC, CNI EVALUATIONS ON FIVE SYNTHETIC DATABASES.[1]

Database	Metric	DCP	NLD	DHN	MSCNN	AOD	EPDN	GridDN	Our Model[2]	Our Model-RESIDE[3]
O-HAZE	PSNR	16.520	15.148	16.270	17.142	14.954	18.351	17.105	**24.907**	18.540
	SSIM	0.332	0.335	0.383	0.374	0.282	0.505	0.377	**0.773**	0.513
	FSIMc	0.979	0.976	0.979	0.982	0.974	0.982	0.971	**0.990**	0.983
	CNI	0.782	0.814	0.818	0.811	0.787	**0.901**	0.813	0.892	0.850
SOTS	PSNR	18.051	16.853	23.335	19.133	20.198	22.524	28.281	**28.780**	28.613
	SSIM	0.604	0.901	0.901	0.852	0.887	0.874	**0.970**	0.956	0.951
	FSIMc	0.992	0.985	0.997	0.994	0.990	0.996	0.998	**0.999**	**0.999**
	CNI	0.880	0.855	0.875	0.880	0.850	0.885	0.853	**0.890**	0.876
HazeRD	PSNR	15.602	15.066	15.672	15.773	15.629	15.792	15.364	**17.037**	16.477
	SSIM	0.656	0.607	0.622	0.637	0.614	0.598	0.675	**0.696**	0.649
	FSIMc	0.978	0.973	0.978	0.975	0.973	0.978	0.970	**0.980**	**0.980**
	CNI	0.860	0.809	0.860	0.864	**0.902**	0.900	0.836	0.889	0.864
I-HAZE	PSNR	14.835	14.929	16.572	17.061	15.009	16.325	16.057	**26.985**	18.328
	SSIM	0.429	0.567	0.572	0.547	0.582	0.622	0.629	**0.889**	0.657
	FSIMc	0.968	0.968	0.974	0.974	0.961	0.974	0.963	**0.985**	0.978
	CNI	0.850	0.857	0.853	0.853	0.839	0.882	0.849	**0.929**	0.876
Middlebury	PSNR	13.978	13.873	**15.965**	14.203	13.321	14.493	12.523	13.904	13.259
	SSIM	0.695	0.743	**0.798**	0.774	0.755	0.701	0.761	0.714	0.745
	FSIMc	0.969	0.968	0.970	**0.971**	0.967	0.970	0.962	0.968	0.967
	CNI	0.902	0.906	**0.927**	0.914	0.925	0.921	0.915	0.920	0.913
Total (1330 samples)	PSNR	16.473	15.521	18.161	17.525	16.354	18.791	19.639	**25.791**	20.943
	SSIM	0.449	0.558	0.582	0.559	0.535	0.637	0.612	**0.842**	0.672
	FSIMc	0.980	0.976	0.983	0.983	0.975	0.983	0.976	**0.990**	0.986
	CNI	0.829	0.836	0.845	0.843	0.823	0.893	0.835	**0.900**	0.865

[1] The **bold** and underline values indicate the best and second-best results.
[2] Our model trained on mixed databases (as shown in Table I).
[3] Our model trained only on the RESIDE-OTS database.



(a) Input (b) DCP (c) MSF (d) CAP (e) NLD (f) DHN

(g) MSCNN (h) AOD (i) GFN (j) EPDN (k) GridDN (l) Our Model

Fig. 9. Qualitative results of real-world image dehazing. (a): hazy image. (b)-(k): results of DCP [1], MSF [64], CAP [16], NLD [19], DHN [20] and MSCNN [22], AOD [23], GFN[25], EPDN [21] and GridDN [42]. (l): restored images of our model.

two input images that are coming from the same scene but with different haze-levels. DCP [1] generates halo artifacts on the wall of the house. DCP [1] and NLD [19] suffer from color distortion on the house. Other learning-based methods could keep accurate color and structure, as shown in Figs.8 (3-d) to (3-h). However, it can be found that the distant houses are not restored clearly by checking Figs.8 (4-d) to (4-h). Our model performs almost the same over these two different haze-levels, as shown in Figs. 8 (3-i) and (4-i), indicating the robustness to different haze-level. This conclusion can also be verified by Table II, where our method performs the best on the metrics of PSNR, SSIM, and FSIMc.

Comparisons on the I-HAZE database. As shown in Fig.8 (5-i), our model achieves the best visual quality on I-HAZE [53] database. MSCNN [22], AOD [23] and GridDN [42] fail to remove haze effectively from Figs.8 (5-e), (5-f) and (5-h). The other four methods of DCP [1], NLD [19], DHN [20] and EPDN [21] yield overly-enhanced glass and blackboard. The quantitative evaluations on I-HAZE [53] shown in Table II demonstrate that, our method surpass the second best method (except for our model trained on the RESIDE database) by a large margin of 9.924 dB, 0.26 SSIM, 0.011 FSIMc, and 0.047 CNI scores.

Comparisons on the Middlebury database. Fig. 8 (6) shows the comparisons of all algorithms on Middlebury [55] database. We can find that DCP [1] and NLD [19] are vulnerable to white objects. This flaw is commonly associated with prior-based methods (including DCP and NLD), which cannot well distinguish between white objects and real hazes. As shown in Figs.8 (6-b) and (6-c), the white wall and umbrella are over-enhanced, resulting in unreal color. This problem is alleviated by learning-based methods [20], [22],

TABLE III
PERFORMANCE COMPARISONS IN TERMS OF AVERAGE BCEA AND CNI ON REAL WORLD OUTDOOR IMAGES (**BOLD** AND UNDERLINE NUMBERS INDICATE THE BEST AND SECOND-BEST RESULTS.)

Indicators	DCP	MSF	CAP	NLD	DHN	MSCNN	AOD	GFN	EPDN	GridDN	Our Model
BCEA - e_r	1.025	1.018	0.724	**1.250**	0.713	0.757	0.833	0.661	1.026	0.639	<u>1.070</u>
BCEA - $\overline{g}$	1.720	**1.905**	0.995	<u>1.845</u>	1.192	1.447	1.401	1.494	1.672	1.239	1.807
BCEA - s_r	0.008	0.004	0.028	0.024	0.015	0.011	0.007	0.004	**0.000**	<u>0.001</u>	**0.000**
CNI	0.817	0.820	0.807	0.793	0.797	0.802	0.776	0.839	<u>0.853</u>	0.808	**0.861**

[25], [21], [42]. Our method also successfully refrains from over-saturation on white objects, as shown in Fig.8 (6-i). From Table II, DHN [20], MSCNN [22] and AOD [23] perform better than our model with respect to PSNR, SSIM and FSIMc. The reason is that they all were trained on indoor NYU2 [71] and/or Middlebury [55] databases where hazy images are artificially synthesized from atmospheric scattering model (1), while our training database does not contain indoor samples.

Quantitative comparisons on all synthetic databases. In Table II, the last row gives performance comparisons overall testing images. It can be observed that our model performs the best in terms of PSRN, SSIM, and FSIMc metrics, with 6.152dB PSNR improvement beyond the second best algorithm, namely GridDN [42], 0.205 SSIM and 0.007 FSIMc improvements beyond the second best algorithm, namely EPDN [21]. Additionally, our model obtains the highest CNI scores, indicating the method yield more natural and vivid colors.

2) Our Model Trained on RESIDE Database:

For fair comparisons, we further retrained our model only on the RESIDE database, which was used in EPDN [21] and GridDN [42]. The testing results are reported in the column of "Our Model-RESIDE" of Table II. It can be found that our model gets the second-best results of PSNR, SSIM, and FSIMc on both O-HAZE and I-HAZE databases, better than EPDN [21] and GridDN [42]. Experiments on SOTS [54] and HazeRD [56] databases demonstrate that our model outperforms EPDN and GridDN with respect to the PSNR and FSIMc, while the SSIM and CNI are comparable, with less than 2% SSIM and 4% CNI drops on both databases. Finally, on the dataset containing all of these five synthetic databases, our model is better than EPDN and GridDN in terms of all metrics except CNI.

3) Evaluation on Real-world Database:

Quantitative comparisons on all synthetic databases. In Table II, the last row gives performance comparisons overall testing images. It can be observed that our model performs the best in terms of PSRN, SSIM, and FSIMc metrics, with 6.152dB PSNR improvement beyond the second best algorithm, namely GridDN [42], 0.205 SSIM and 0.007 FSIMc improvements beyond the second best algorithm, namely EPDN [21]. Additionally, our model obtains the highest CNI scores, indicating the method yield more natural and vivid colors.

Qualitative Comparisons. Fig. 9 shows comparisons among all tested algorithms on real-world hazy images. The prior-based methods, DCP [1] and CAP [16], remove haze effectively. However, the dehazed images of DCP [1] suffer from over-enhanced visual artifacts in the sky regions, as shown in the third images of Fig.9 (b); the CAP generally yields over-saturation as shown in Fig.9 (c). The fusion method MSF [64] generates images of bright color, yet remains haze at distant scenes. NLD [19] achieves better visual quality than the previous methods [1], [64], [16], with plausible details. However, it leads to over-saturation, e.g., the clothes and sky regions in the third image of Fig.9 (e).

We also compare our model with the widely developed deep learning-based dehazing methods. Generally, learning-based methods outperform prior-based methods. AOD [23] and GridDN [42] avoid over-enhanced problem generally. However, they remain a small amount of haze in some regions. Although MSCNN [22], DHN [20] and GFN [25] have better visual quality than AOD [23] and GridDN [42], they have slight color distortion, e.g., the clothes of the boy in the third image of Fig.9 (g) and the crowds in the fourth images of Figs.9 (f) and (k). EPDN [21] performs better than previous methods, however, like most of deep learning-based methods, it produces insufficient dehazing in distant scenes as shown in the second image of Fig.9 (j). By contrast, our model recovers richer and more saturated colors, and effectively remove the haze in distant scenes, as shown in Fig.9 (l).

Quantitative Comparisons. We further use a bind contrast enhancement assessment (BCEA) [72] to evaluate the effectiveness of our model quantitatively. The BCEA consists of three indicators: the rate of edges newly visible e_r, the geometric mean of normalized gradients $\overline{g}$, and the rate of saturated (black or white) pixels s_r. The first two indicators evaluate the ability of edge and contrast restoration, respectively. The s_r is indispensable to evaluate the degree of over-enhanced. A good dehazing algorithm should get higher e_r and $\overline{g}$, and lower s_r at the same time.

The average BCEA indexes are listed in Table III. It can be observed that prior-based methods achieve better (larger) e_r and $\overline{g}$. However, their s_r are commonly higher (worse) than 0.008 (except MSF [64] with s_r=0.004). For example, the NLD [19] gets the best e_r and the second best $\overline{g}$, its s_r is the second worst, indicating that the dehazed results are highly over-enhanced. The CNN-based methods, DHN [20], MSCNN [22] and AOD [23] are inferior to other methods with respect to BCEA. The reason may lie in that these three methods are trained on indoor images, leading to low efficiency on real-world images. By contrast, our model achieves better visual quality and the top three scores of the BCEA, thanks to global context modeling. GFN [25], EPDN [21], GridDH [42] successfully refrain from over-saturation and halo artifacts, so

(a) hazy (b) RU (c) DRU (d) GC-UNet (e) PGC-UNet-α1 (f) PGC-UNet-α64

Fig. 10. Ablation studies: qualitative comparisons of RU, DRU, GC-UNet, PGC-UNet-α1 and PGC-UNet-α64 at 50*th* epoch.

TABLE IV
SUBJECTIVE IMAGE QUALITY ASSESSMENT.

Indexes	AOD	EPDN	GridDN	Our Model
naturalness	3.360	3.923	3.373	**4.030**
contrast	3.315	3.870	3.298	**3.965**
colorful	3.118	3.870	3.255	**3.950**

TABLE V
ABLATION STUDIES IN TERMS OF PSNR, SSIM AND CNI.

Module of E/D	Weight of PGC	PSNR	SSIM	CNI
RU		24.525	0.929	0.870
DRU		24.621	0.929	0.874
GC-UNet		24.771	0.944	0.889
PGC-UNet-α1	[1, 1, 1, 1]	24.817	0.946	0.897
PGC-UNet-α64	[1, 4, 16, 64]	**25.064**	**0.952**	**0.902**

they are with very small s_r, however, their e_r and $\overline{g}$ are less (worse) than ours.

We further calculate the Color Naturalness Index (CNI) [68] for real-world dehazing results. Table III gives the average CNIs of different methods tested on 109 real-world samples. As we can find, our model obtains the highest CNI, indicating that our results possess more faithful color and richer details.

D. Subjective Image Quality Assessment

We also conduct a subjective image quality assessment (IQA) to evaluate the proposed model, comparing to other deep learning-based methods, including AOD [23], EPDN [21], and GridDN [42]. We randomly selected 5, 5, 15, 3, 3, and 19 samples from O/I-HAZE [52], [53], SOTS [54], Middlebury [55], HazeRD [56], and real-world testing databases, respectively, composing of a dataset of 50 samples for subjective IQA. Eighteen observers were asked to give their opinion about overall quality of each image with respect to the following three aspects:

- "***naturalness***": image looks natural after haze is effectively removed;
- "***contrast***": image content and edges are clear with good contrast;
- "***colorful***": image looks visually pleasant without oversaturation.

The evaluation is reported on the five-point scale: "*bad*": 1; "*poor*": 2; "*fair*": 3; "*good*": 4; and "*excellent*": 5. A single-stimulate method (ACR) [**?**] is used in subjective IQA, in which the observers were asked to rate image quality on one of the five scales. All of the test images are displayed randomly. Average scores on overall image quality, "naturalness", "contrast" and "colorful" are summarized in Table IV. It can be observed that the proposed model performs the best among all compared methods, consistent with the aforementioned objective evaluation, and verifying the superiority of the proposed model with respect to human visual perception.

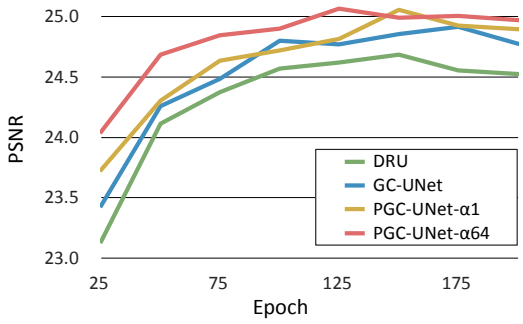


Fig. 11. Average PSNRs of DRU, GC-UNet, PGC-UNet-α1 and PGC-UNet-α64 at 25*th*, 50*th*, 75*th*, 100*th*, 125*th*, 150*th*, 175*th* and 200*th* epoch.

E. Ablation Study

For checking the contribution of each module in our model, we conduct an ablation study on the following five different configurations: 1) RU: U-Net with a residual block (shortcut connection) but without dilated convolution; 2) DRU: U-Net with dilated residual bottleneck block; 3) GC-UNet: global context aided U-Net with dilated residual bottleneck block; 4) PGC-UNet-α1: pyramid global context aided U-Net with dilated residual bottleneck block, using uniform weights, α1=[1, 1, 1, 1], for each stage pooling; 5) PGC-UNet-α64: pyramid

TABLE VI
AVERAGE RUNNING TIMES OF DIFFERENT METHODS TESTED ON 852 SAMPLES.

Average Pixel Number	Average Running Time (Matlab/CPU)						Average Running Time (Python/GPU)			
$(\sum_{i=1}^{852} H_i \times W_i)/852$	DCP	MSF	CAP	NLD	NHN	MSCNN	AOD	EPDN	GridDN	Our Model
$\approx$ 423,875	2.022	**0.591**	1.539	2.940	2.898	2.346	**0.188**	0.293	0.309	0.315

global context aided U-Net with dilated residual bottleneck block, with non-uniform weights, α64=[1, 4, 16, 64] for different scale pooling.

The qualitative comparisons are demonstrated in Fig. 10. They are further analyzed as follows.

RU & DRU: Fig.10 (b) gives the result of RU (configuration 1)). It can be observed that the color of the woods in the red rectangle tends to be distorted. By contrast, as shown in Fig.10(c), DRU (configuration 2)) gives us better visual experience. The reason is that dilated convolution renders DRU enlarging receptive field with an increase of convolutional layers. This mechanism could explore global context information to some extent so that DRU can aggregate global context information for learning haze relevant features.

GC block: In GC-UNet (configuration 3)), features of convolutional layers are further enhanced by GC block [26] before they are fed into the following dilated residual bottleneck block, so that global context features are modeled and propagated to next stage, i.e., from E_i to E_{i+1} and from D_{i+1} to D_i. Comparing Figs.10 (c) and (d), GC-UNet yields better dehazing than DRU around the park bench, which verifies that GC block is more efficient than DRU for exploring global context information. In addition, the success of the GC block also indicates that a global context is really crucial for haze-relevant features learning.

PGC block: GC block [26] acquires point-wise long-range dependencies, while PGC block exploits both point-wise and patch-wise long-range dependencies. Hence, the processing of PGC is more in line with the real situation of almost constant haze within a local patch. Comparing Fig.10 (e) with Fig. 10 (d), PGC-UNet-α1 (configuration 4)) is superior to GC-UNet, indicating superior performance of the proposed PGC module. From PGC-UNet-α1 to PGC-UNet-α64 (configuration 5)), patch-wise long-range dependencies are further highlighted for global context modeling. The result of PGC-UNet-α64 in Fig. 10 (f) is visually more pleasing than the one of PGC-UNet-α1.

Table V lists average PNSR, SSIM and CNI for five different configurations. It can be observed that GC-UNet outperforms DRU, indicating that GC block contributes a better representation of the hazy image and increases the dehazing ability of the network. PGC-UNet-α1 performs better than GC-UNet, but only a small gap since uniform weights contribute less patch-wise global context. PGC-UNet-α64 is the best with respect to all of the three metrics since it enables a patch-wise global context with the largest weight. We also demonstrate PSNR curves (shown in Fig.11) along with training epoch for all compared algorithms. It can be found that PGC-UNet-α64 converges faster and keeps more stable than others.

F. Computational Complexity

We gather time consumptions of all compared algorithms over randomly selected 852 images with different resolutions for statistics of computational complexites. DCP [1], MSF [64], CAP [16], NLD [19], DHN [20] and MSCNN [22] are realized by MATLAB codes, implemented on a Windows10 PC with an Inter(R) Core(TM) i7-8700 CPU @ 3.20 GHz processor with 16GB RAM. Other deep learning-based methods, including AOD [23], EPDN [21], GridDN [42], and our proposed one, are realized by Python, and implemented over a NVIDIA Tesla P100 GPU.

Average running times are listed in Table VI. Among Matlab codes mentioned above, MSF [64] is with the least time consumption since it avoids to estimate the transmission map and atmospheric light. Among deep learning-based methods, AOD [23] costs the least in time consumption. The proposed model is comparable to GridDN [42] with respect to time consumption. In spite of a little bit of overhead of computational complexity, the amount of dehazing efficiency achieved by the proposed one is deserved compared to other methods.

V. CONCLUSION

It was found that global context modeling was crucial for image dehazing. This paper further points out that patch-wise long-range dependencies should be stressed for global context modeling. For this purpose, we first propose a PGC block that could explore patch-wise long-range dependencies of different scales. An end-to-end dehazing network is then proposed by plugging the PGC block into a U-Net, which is further enhanced by a dilated residual bottleneck (DRB) block. Extensive experiments and ablation studies demonstrate that the proposed PGC contributes the backbone network, i.e., U-Net, more efficiency for image dehazing, which also verifies our ideas concerning patch-wise long-range dependencies and global context modeling for benefits of image dehazing. The proposed PGC is lightweight and computationally efficient. It can be easily plugged into any other networks in case of need global context modeling.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 61811530282, 61872429, 11790301 and 11790305.

REFERENCES

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

[2] Q. Liu, X. Gao, L. He, and W. Lu, "Single image dehazing with depth-aware non-local total variation regularization," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5178–5191, 2018.

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

[4] J. Zhang and D. Tao, "Famed-net: A fast and accurate multi-scale end-to-end dehazing network," *IEEE Transactions on Image Processing*, vol. 29, pp. 72–84, 2020.

[5] R. Li, L.-F. Cheong, and R. T. Tan, "Heavy rain image restoration: Integrating physics model and conditional adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1633–1642.

[6] W. Yang, R. T. Tan, J. Feng, Z. Guo, S. Yan, and J. Liu, "Joint rain detection and removal from a single image with contextualized deep networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 6, pp. 1377–1393, 2019.

[7] D.-W. Jaw, S.-C. Huang, and S.-Y. Kuo, "Desnowgan: An efficient single image snow removal framework using cross-resolution lateral connection and gans," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[8] H. Lin, C. Jing, Y. Huang, and X. Ding, "a^2net: Adjacent aggregation networks for image raindrop removal," *IEEE Access*, vol. 8, pp. 60 769–60 779, 2020.

[9] R. Li, R. T. Tan, and L.-F. Cheong, "All in one bad weather removal using architectural search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3175–3185.

[10] Y. Y. Schechner, S. Narasimhan, and S. Nayar, "Instant dehazing of images using polarization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2001, pp. I–I.

[11] S. G. Narasimhan, "Chromatic framework for vision in bad weather," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2000, pp. 598–605.

[12] S. Nayar and S. Narasimhan, "Vision in bad weather," in *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 1999, pp. 820–827.

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

[14] J. Kopf, B. Neubert, B. Chen, M. Cohen, D. Cohenor, O. Deussen, M. T. Uyttendaele, and D. Lischinski, "Deep photo: model-based photograph enhancement and viewing," *International Conference on Computer Graphics and Interactive Techniques*, vol. 27, no. 5, p. 116, 2008.

[15] R. Fattal, "Dehazing using color-lines," *ACM Transactions on Graphics (TOG)*, vol. 34, 2014.

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

[17] L. He, J. Zhao, N. Zheng, and D. Bi, "Haze removal using the difference-structure-preservation prior," *IEEE Transactions on Image Processing*, vol. PP, no. 99, pp. 1–1, 2017.

[18] T. M. Bui and W. Kim, "Single image dehazing using color ellipsoid prior," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 999–1009, 2018.

[19] D. Berman, S. Avidan *et al.*, "Non-local image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1674–1682.

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

[21] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8160–8168.

[22] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 154–169.

[23] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4770–4778.

[24] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3194–3203.

[25] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang, "Gated fusion network for single image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3253–3261.

[26] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2019, pp. 0–0.

[27] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.

[28] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 433–442.

[29] K. Yue, M. Sun, Y. Yuan, F. Zhou, E. Ding, and F. Xu, "Compact generalized non-local network," in *Advances in Neural Information Processing Systems*, 2018, pp. 6510–6519.

[30] Z. Cheng, S. You, V. Ila, and H. Li, "Semantic single-image dehazing," *arXiv preprint arXiv:1804.05624*, 2018.

[31] W. Ren, J. Zhang, X. Xu, L. Ma, X. Cao, G. Meng, and W. Liu, "Deep video dehazing with semantic segmentation," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1895–1908, 2018.

[32] K. He, X. Zhang, S. Ren, and S. Jian, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[33] D. Zhao, L. Xu, Y. Yan, J. Chen, and L.-Y. Duan, "Multi-scale optimal fusion model for single image dehazing," *Signal Processing: Image Communication*, vol. 74, pp. 253–265, 2019.

[34] L. Xu, D. Zhao, Y. Yan, S. Kwong, J. Chen, and L.-Y. Duan, "Iders: Iterative dehazing method for single remote sensing image," *Information Sciences*, vol. 489, pp. 50–62, 2019.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[36] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[37] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 2169–2178.

[38] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 472–480.

[39] Y. Liu, J. Pan, J. Ren, and Z. Su, "Learning deep priors for image dehazing," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 2492–2500.

[40] A. Yang, H. Wang, Z. Ji, Y. Pang, and L. Shao, "Dual-path in dual-path network for single image dehazing," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 4627–4634.

[41] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8798–8807.

[42] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 7314–7323.

[43] D. Fourure, R. Emonet, É. Fromont, D. Muselet, A. Trémeau, and C. Wolf, "Residual conv-deconv grid network for semantic segmentation," *arXiv preprint arXiv:1707.07958*, 2017.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[45] G. Li, X. He, W. Zhang, H. Chang, L. Dong, and L. Lin, "Non-locally enhanced encoder-decoder network for single image de-raining," *arXiv preprint arXiv:1808.01491*, 2018.

[46] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 065–11 074.

[47] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 593–602.

[48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.
[49] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.
[50] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.
[51] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134, 2017.
[52] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-haze: a dehazing benchmark with real hazy and haze-free outdoor images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 754–762.
[53] C. Ancuti, C. O. Ancuti, R. Timofte, and C. De Vleeschouwer, "I-haze: a dehazing benchmark with real hazy and haze-free indoor images," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2018, pp. 620–631.
[54] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2018.
[55] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German Conference on pattern recognition*. Springer, 2014, pp. 31–42.
[56] Y. Zhang, L. Ding, and G. Sharma, "Hazerd: an outdoor scene dataset and benchmark for single image dehazing," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3205–3209.
[57] R. Fattal, "Single image dehazing," *ACM transactions on graphics (TOG)*, vol. 27, no. 3, p. 72, 2008.
[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[59] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8183–8192.
[60] Z. Xu, X. Yang, X. Li, X. Sun, and P. Harbin, "Strong baseline for single image dehazing with deep features and instance normalization." in *BMVC*, vol. 2, no. 3, 2018, p. 5.
[61] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1501–1510.
[62] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.
[63] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[64] C. O. Ancuti and C. Ancuti, "Single image dehazing by multi-scale fusion," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3271–3282, 2013.
[65] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
[66] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.
[67] S. N. Yendrikhovski, F. J. J. Blommaert, and H. de Ridder, "Perceptually optimal color reproduction," in *Human Vision and Electronic Imaging III*, B. E. Rogowitz and T. N. Pappas, Eds., vol. 3299, International Society for Optics and Photonics. SPIE, 1998, pp. 274 – 281.
[68] K. Huang, Q. Wang, and Z. Wu, "Natural color image enhancement and evaluation algorithm based on human visual system," *Computer Vision and Image Understanding*, vol. 103, no. 1, pp. 52–63, 2006.
[69] C. Z. GUO Fan, TANG Jin, "Objective measurement for image defogging algorithms," *Journal of Central South University*, no. 1, pp. 272–286, 2014.
[70] F. Guo, G. Lan, X. Xiao, and B. Zou, *Parameter Selection of Image Fog Removal Using Artificial Fish Swarm Algorithm*. Intelligent Computing Theories and Application. Springer, Cham., 2018.
[71] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012, pp. 746–760.
[72] N. Hautiere, J.-P. Tarel, D. Aubert, and E. Dumont, "Blind contrast enhancement assessment by gradient ratioing at visible edges," *Image Analysis & Stereology*, vol. 27, no. 2, pp. 87–95, 2008.

Dong Zhao received the M.S. degree in Key Laboratory of Complex System Intelligent Control and Decision, Department of Automation, Beijing Institute of Technology, Beijing, China, in 2016, and the Ph.D. degree from National Astronomical Observatories, Chinese Academy of Sciences , Beijing, China, in 2020. His research interests include computer vision and image processing.

Long Xu (M'12) received his M.S. degree in applied mathematics from Xidian University, Xi'an, China, in 2002, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He was a Postdoc with the Department of Computer Science, City University of Hong Kong, the Department of Electronic Engineering, Chinese University of Hong Kong, from July Aug. 2009 to Dec. 2012. From Jan. 2013 to March 2014, he was a Postdoc with the School of Computer Engineering, Nanyang Technological University, Singapore. Currently, he is with the Key Laboratory of Solar Activity, National Astronomical Observatories, Chinese Academy of Sciences. His research interests include image/video processing, solar radio astronomy, wavelet, machine learning, and computer vision.

He was selected into the 100-Talents Plan, Chinese Academy of Sciences, 2014.

Lin Ma (M'13) received the B.E. and M.E. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively, and the Ph.D. degree from the Department of Electronic Engineering, The Chinese University of Hong Kong, in 2013. He was a Researcher with the Huawei Noah's Ark Laboratory, Hong Kong, from 2013 to 2016. He was a Principal Researcher with the Tencent AI Laboratory, Shenzhen, China, from 2016 to 2020. He is a currently a Researcher with the Meituan, Beijing, China. His current research interests lie in the areas of computer vision, multimodal deep learning, specifically for image and language, image/video understanding, and quality assessment. Dr. Ma received the Best Paper Award from the Pacific-Rim Conference on Multimedia in 2008. He was a recipient of the Microsoft Research Asia Fellowship in 2011. He was a finalist in HKIS Young Scientist Award in engineering science in 2012.

Jia Li (M'12-SM'15) received the B.E. degree from Tsinghua University in Jul. 2005 and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in Jan. 2011. He is currently a Full Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. Before he joined Beihang University in Jun. 2014, he used to conduct research in Nanyang Technological University, Peking University and Shanda Innovations. He is the author or coauthor of over 80 technical articles in refereed journals and conferences such as TPAMI, IJCV, TIP, CVPR and ICCV. His research interests include computer vision and multimedia analysis, especially the perception and understanding of visual contents in extreme environments. He is supported by the Research Funds for Excellent Young Researchers from National Nature Science Foundation of China since 2019. He was also selected into the Beijing Nova Program (2017) and received the Second-grade Science Award of Chinese Institute of Electronics (2018), two Excellent Doctoral Thesis Award from Chinese Academy of Sciences (2012) and the Beijing Municipal Education Commission (2012), and the First-Grade Science-Technology Progress Award from Ministry of Education, China (2010). He is a senior member of IEEE, CIE and CCF. More information can be found at http://cvteam.net.

Yihua Yan is currently a professor and chief scientist of the solar radio research, and the director of the Key Laboratory of Solar Activity and the Solar Physics Division, National Astronomical Observatories, Science Academy of Sciences. He received his B. E degree and M. E degree from Northwestern Polytechnical University, Xi'an, China in 1982 and 1985, respectively, and the Ph.D. degree from Dalian University of Technology in 1990. He was a Foreign Research Fellow at National Astronomical Observatory of Japan from 1995 to 1996, and an Alexander von Humboldt Fellow at Astronomical Institute, Wurzburg University, Germany from 1996 to 1997. He is currently the president of International Astronomical Union (IAU) Division E: Sun & Heliosphere for the period from 2015 to 2018. His research interests include solar magnetic fields, solar radio astronomy, space solar physics and radio astronomical methods.