Neurocomputing 437 (2021) 143–156

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Attentive U-recurrent encoder-decoder network for image dehazing

Shibai Yin [a], Yibin Wang [b,*], Yee-Hong Yang [c]

[a] Department of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China
[b] Department of Engineering, Sichuan Normal University, Chengdu, Sichuan 610066, China
[c] Department of Computing Science, University of Alberta, Edmonton T6G 2E8, Canada

ARTICLE INFO

Article history:
Received 10 October 2019
Revised 30 July 2020
Accepted 6 December 2020
Available online 14 January 2021
Communicated by Ruili Wang

Keywords:
Image dehazing
Attention mechanism
Recurrent network
Encoder-decoder
Convolutional neural network

ABSTRACT

Haze removal is an important pre-processing step in many computer vision tasks. Convolutional neural networks, especially the U-shaped networks, have shown to be effective in image dehazing. Nevertheless, these networks have three main limitations. First, the relevant haze information, e.g. concentration of haze, is totally ignored. Second, spatial inconsistency and information dilution usually occur when the networks refine the dehazed results with a coarse-to-fine strategy. Third, the receptive field of the network is not large enough to capture structural information. Motivated by these problems, a new attentive U-recurrent encoder-decoder dehazing network is presented, which consists of an attentive recurrent network and a U-recurrent encoder-decoder network. By assuming that haze layers with different depths can be detected by multiple stages, we use an attentive recurrent network to generate the haze attention map for guiding the U-recurrent encoder-decoder network with the concentration of haze to better estimate the clear image. Meanwhile, the features for dehazing are further enhanced and the dehazing results are refined in the U-recurrent encoder-decoder network. This design not only enables spatial consistency but also reduces information dilution with short recurrent pathways. Furthermore, a novel residual pyramid pooling module is also proposed and used in the U-recurrent encoder-decoder network, which provides the network with structural information and with an enlarged receptive field. The experimental results demonstrate that our method outperforms state-of-the-art dehazing algorithms on both synthetic and real hazy images.

1. Introduction

Image dehazing, which aims to remove the effect of foggy appearance in degraded images, has been widely used as an image pre-processing step in various vision tasks, e.g. industrial surveillance, remote sensing, restoration of photographs [1–3]. Due to the increasing interest from the research community, much progress has been made in this area [4–6]. However, it is still a challenging task to develop effective dehazing methods which can handle hazy images in complicated scenarios.

Conventional haze removal methods are based on the physical atmospheric scattering model and highly rely on hand-crafted priors. For example, the dark channel prior (DCP) regards that some pixels in a local patch have very low intensity values in at least one of the RGB channels and haze is removed by leveraging the DCP to compute the transmission map in the physical atmospheric scattering model [7]. In addition, the non-local prior method is based on the observation that colors in a hazy image form several color lines in RGB color space and these color lines are used to recover the clear image [8]. Although these hand-crafted prior-based methods have been demonstrated to be effective in many haze scenarios, they are not robust enough to remove haze when the priors are insufficient to describe the real world. Just as the DCP cannot handle white objects which have similar color as the atmospheric light.

Recently, Convolutional Neural Networks (CNNs) have shown their outstanding performance in many machine vision tasks, e.g. image restoration, image segmentation and object recognition. Meanwhile, recent works also demonstrate that U-shaped CNNs facilitate image dehazing by extracting hierarchical features in both bottom-up and top-down pathways [9,10,3]. Although these U-shaped CNNs have achieved promising performance by learning non-linear mappings between hazy images and their corresponding clear images, there are still factors limiting their performance. First, relevant haze information, which helps to remove haze in conventional methods, is totally ignored. Motivated by the success of injecting the rain drop information into an image deraining network [11], we believe that the performance of image dehazing can be similarly improved by incorporating the relevant haze

* Corresponding author.
E-mail address: yibeen.wong@gmail.com (Y. Wang).

https://doi.org/10.1016/j.neucom.2020.12.081
0925-2312/© 2021 Elsevier B.V. All rights reserved.

information into a dehazing network. Second, for a U-shaped network, the feed forward pattern hardly refines the dehazing results without adding more layers and hence, more parameters. Existing methods solve this problem by employing a recurrent unit, (e.g. LSTM) and by refining predicted results with the output of the network, leading to complicated computation and spatial inconsistency [12]13. In addition, informative features will be gradually diluted when features in the encoder are transmitted to the decoder. Third, the receptive field of a U-shaped network is not proportional to its depth. As reported in previous works [14], the empirical receptive fields of a U-shaped network are much smaller than that predicted in theory and are not large enough to capture useful structural information.

To solve the first problem, we adopt an Attentive Deraining Network (ADN) [11] proposed by Qian et al. as the backbone. Unlike the common U-shaped network, this network has an Attentive-Recurrent Network (ARN) in front of the encoder-decoder network for generating rain drop information. After injecting the rain drop information into the encoder-decoder network, the network is aware exactly where to focus on. Motivated by this idea, we use an ARN to generate a haze attention map which highlights haze concentration information. By repeatedly passing the input image and haze attention map through the ARN with the supervision of a transmission map, the haze layers with different depths can be detected iteratively.

To mitigate the second issue, we propose a U-recurrent encoder-decoder network to replace the existing encoder-decoder network in the backbone. Unlike most recurrent encoder-decoder networks that consider the whole network in one iterative step [12], we shorten the recurrent pathway using a novel recurrent mechanism for both the encoder and the decoder. Using the recurrent encoder as an example, by passing features through the recurrent encoder iteratively, the spatial consistency can be well maintained and the parameters of the encoder are shared across all iterations. To further maintain spatial–temporal dependencies and to relieve the problem of feature dilution, the rich information from an intermediate layer in the recurrent encoder are passed to the corresponding intermediate layer in the next iteration via recurrent convolution. Then, the captured high-level features are fed to the decoder which has a similar recurrent mechanism as that of the encoder.

To address the third issue, we design a novel Residual Pyramid Pooling Module (RPPM) and apply it to both the recurrent encoder and the recurrent decoder. By capturing the pooling features with varying downsampling rates and fusing them together with residual learning, the structural information is capable of merging with the local information at each iteration seamlessly. Because of the recurrent mechanism, the receptive field of the network is enlarged.

Fig. 1 shows the structure of the proposed network. The detailed discussion of this structure is given in Section 3. Because of using the attentive-recurrent network, the U-recurrent encoder-decoder network and the Residual Pyramid Pooling Module (RPPM), we refer to the proposed network as the Attentive U-recurrent Encoder-Decoder Dehazing Network (AUEDN). With the help of injected visual attention mechanism, we believe that our network may also work well in other domains, e.g. disaster management systems, industrial surveillance. For example, Muhammad et al. propose a light-weight network for smoke detection in foggy surveillance environments by classifying images into four classes: smoke, non-smoke, smoke with fog and non-smoke with fog [15]. To improve the accuracy of smoke detection, we could utilize our attentive-recurrent network to generate a smoke attention map based on the properties of smoke, e.g. color, density, shape, and inject this visual attention map into the subsequent network for identifying the smoke in foggy images. Similarly, the accuracy of the digital image authentication could be improved using our network to replace the common image hashing mechanism for feature extraction [1].

The contributions of our work are summarized as follows:

- We propose a Attentive U-Recurrent Encoder-Decoder Dehazing Network (AUEDN), consisting of an attentive recurrent network and a U-recurrent encoder-decoder network. By leveraging the attentive-recurrent network to generate the haze attention map with the supervision of the transmission map, the subsequent U-recurrent encoder-decoder network can remove haze guided by relevant haze information effectively.
- We propose a new U-recurrent encoder-decoder network using a new recurrent mechanism in both the encoder and the decoder. This design enables spatial consistency and reduces information dilution by shortening the recurrent pathways in both the encoder and the decoder. As well, the features for dehazing are enhanced and the dehazing results are refined by the proposed novel recurrent mechanism.
- We propose a novel residual pyramid pooling module, which is used in both the encoder and the decoder. Therefore, the receptive field of the network is enlarged by allowing local information to be merged with structural information seamlessly.

The rest of the paper is organized as follows: Section 2 discusses the related work in image dehazing. Section 3 explains the proposed network and its components e.g. an attentive recurrent network and a U-recurrent encoder-decoder network. Section 4 gives the loss functions. Section 5 and Section 6 present the experiments. Section 7 discusses the limitations. Section 8 concludes the paper.

2. Related works

Existing image dehazing methods can be classified into two categories: hand-crafted methods and CNNs based methods [16]. The hand-crafted methods remove haze based on the physical atmospheric scattering model and rely on hand-crafted priors to estimate the physical parameters in the model, e.g. the atmospheric light, the transmission map and the clear image. For example, He et al. propose the dark channel prior (DCP) to compute the transmission map based on the observation that in haze free images some pixels in a local patch have very low intensity values in at least one of the RGB channels [7]. Fattal et al. discover that the intensities of pixels in local image patches present a one-dimensional distribution in the RGB color space and leverage this prior to predict the transmission map [17]. Motivated by the above work, Berman et al. further present the non-local prior (NLP) which asserts that a small number of colors in a hazy image can be well approximated by corresponding color lines in the RGB color space [8]. Zhu et al. propose a linear color attenuation prior (CAP) based on the difference between the brightness and the saturation of pixels in a hazy image [18]. However, these priors are not applicable to all cases. To improve robustness, some fusion-based and optimization-based methods are proposed. For instance, Wang et al. propose a Constrained Total Variation Model (CTVM) which transforms the image dehazing problem to a constrained optimization problem and use alternating minimization scheme with a fast gradient projection algorithm to find the optimal solution [4]. Yu et al. present a Pixel-wise Alpha Blending method (PWAB) for dehazing nighttime images which obtains the transmission map by blending the results estimated by the dark channel prior and the bright channel prior [19]. Hernandez-Beltran et al. propose a dehazing method based on the Genetic Programming Estimator (GPE) for predicting the transmission function automatically [20]. Li et al. propose a Multi-scale Pyramid Fusion method (MPFM)

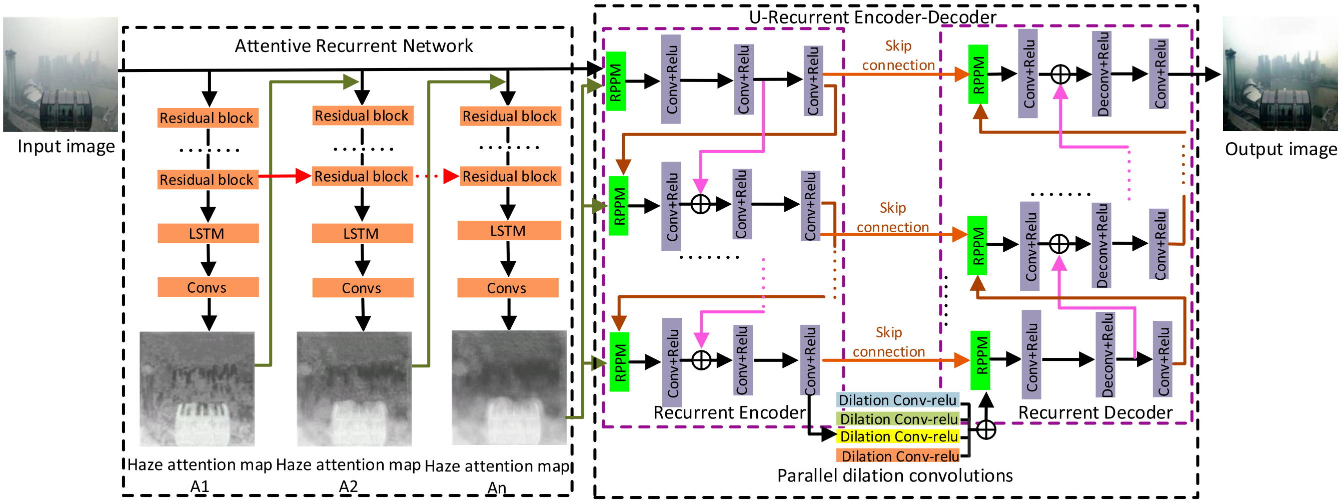


Fig. 1. The architecture of the proposed attentive U-recurrent encoder-decoder dehazing network.

which fuses the local detail enhancement and the global contrast enhancement in a multi-scale pyramid strategy [5]. Galdran et al. discover that clear images contain lower intensity values than that of hazy images and propose the Artificial Multiple-Exposure Image Fusion method (AMEF) by fusing information from multiple-exposure images [21]. Jiang et al. propose the bi-channel prior based on fusing the information provided by the dark channel prior (DCP) and the bright channel prior (BCP) [22]. Although these hand-crafted priors have been demonstrated to be effective in most situations, they are still not robust enough in handling complex scenarios.

Recently, CNNs based methods have shown their superior performance in many computer vision tasks including image dehazing. In one of the earlier works, a CNN is employed to estimate the transmission map, and then a conventional method is used to calculate the atmospheric light. For example, Cai et al. propose using Dehazenet to estimate the transmission map and use an empirical rule to predict the atmospheric light [23]. Ren et al. propose a Multi-Scale Convolutional Neural Network (MSCNN) to predict the tramsmission map in a coarse-to-fine strategy [13]. However, this approach is focused only on estimating the transmission map. If the transmission map was not accurately predicted, the final dehazing result would be impacted negatively. Therefore, recent research is focused on end-to-end networks which integrate all the intermediate processing into one framework. Li et al. propose the All-in-One Dehazing Network (AOD-N) to learn the mapping from a hazy image to a clear image [24]. Wu et al. propose the Densely Pyramid Residual Network (DPRN) by exploiting feature maps and depth at different scale [6]. Qu et al. propose the Enhanced Pix2pix Dehazing Network (EPDN) which removes haze by incorporating two well-designed enhancing blocks into a generative adversarial framework [25]. Li et al. present the GridDehazeNet (GridDN) to obtain a haze-free image by introducing a pre-processing module, a bottleneck and a post-processing module [26]. Yin et al. propose a novel image dehazing network with a parallel attention block [3]. Recent advances have shown that U-shaped CNNs combined with common dehazing schemes used in conventional methods achieve great success, e.g. embedding the physical atmospheric scattering model [9], using a fusion scheme [27], optimizing dehazing result by a coarse-to-fine strategy [13,12]. Specifically, Zhang et al. propose the Densely Connected Pyramid Dehazing Network (DCPDN) which employs an encoder-decoder network, a U-network and a Generative Adversarial Network (GAN) to learn the transmission map, the atmospheric light and the haze-free image based on the physical atmospheric scattering model [9]. Ren et al. build a Gated Fusion Network (GFN) based on the encoder-decoder architecture to generate confidence maps by fusing the white balanced image, the contrast enhanced image and the gamma corrected image effectively [27]. Meanwhile, the coarse dehazed result is fed back to the input for further optimization. In a similar spirit, Tao et al. propose a scale-recurrent network to optimize the coarse result by inserting a recurrent LSTM into the bottleneck of an encoder-decoder network [12]. However, there are still several factors which negatively impact the performance of the network. First, the relevant haze information, such as the distribution of haze, which can guide the removal of haze, is totally ignored. Second, the encoder-decoder network cannot enforce spatial consistency when it optimizes the predicted result becasue of the long recurrent pathway. As well, the informative features are gradually diluted when low-level features are transmitted to combine with high-level features in the bottom-up pathway and similarly, when high-level features are transmitted to combine with lower layers in the top-down pathway. Third, the receptive field of an encoder-decoder network is not proportional to its depth, because of the U-shaped structure. To solve the above problems, we propose a new Attentive U-Recurrent Encoder-Decoder Dehazing Network (AUEDN) based on the previous Attentive Deraining Network (ADN) [11]. By the supervision of the provided transmission map, the previous ARN in the ADN generates the haze attention map for guiding the subsequent encoder-decoder network of the haze distribution. As well, we replace the subsequent encoder-decoder network with our proposed U-recurrent encoder-decoder network. By using a novel recurrent mechanism in the encoder and decoder network, the resulting U-recurrent encoder-decoder network can maintain spatial consistency well. Furthermore, the intermediate information from the recurrent encoder/decoder is passed to the corresponding intermediate layer via recurrent convolution to reduce information dilution. Finally, we design a novel residual pyramid pooling module (RPPM), which is placed in both the encoder and decoder networks for enlarging the receptive field of the network.

3. Attentive U-recurrent encoder-decoder network for image dehazing

As displayed in Fig. 1, our proposed network, which is built based on the ADN [11], consists of two sub-networks: an Attentive-Recurrent Network (ARN) and a new U-Recurrent

Encoder-Decoder Network (UEDN). The ARN is the network adapted from the ADN. With the supervision of the transmission map, the purpose of the ARN in our network is to generate the haze attention map which makes the subsequent UEDN aware of the distribution of haze and to remove it effectively. The UEDN is our proposed novel network. With the help of the proposed recurrent mechanism and the embedding RPPM, the UEDN can maintain spatial consistency, reduce information dilution and enlarge the receptive field.

3.1. Attentive recurrent network

In the ADN, the ARN is used to generate a visual attention map, such as raindrop regions and their surrounding structures. This information can guide subsequent parts of the network to focus on those regions for restoration. For image dehazing, the haze concentration is related to scene depth d. Indeed, the transmission map t is related to the haze concentration as a function of depth d as $t = e^{-\beta d}$, where β is the scattering coefficient. Hence, haze at similar depth can be viewed as one layer and haze in the whole image can be regarded as an accumulation of multiple haze layers. Based on this assumption, the generation of the haze attention map can be decomposed into multiple stages. As can be seen in Fig. 1, the ARN, which contains residual blocks, a recurrent LSTM unit and convolutional layers, can be used to generate the haze attention map iteratively. By concatenating the initial attention map with the hazy input image and feeding them into the ARN, the network can generate more haze concentration information with increasing iteration. Besides, the hidden state in the LSTM unit provides haze concentration information which guides haze detection in subsequent iterations.

During training, we use the mean squared error (MSE) loss function between the output and the ground truth of the transmission map. Here, we set the values of the pixels in the initial attention map to 0.8 and the number of iterations to 3. Then, the loss function for the ARN is formulated as:

$$f_{ARN} = \sum_{n=1}^{3} \lambda f_{MSE}(A_n, T), \tag{1}$$

where f_{ARN} denotes the loss function of the ARN, and f_{MSE} represents the loss function based on the MSE. n denotes the iteration number. A_n is the haze attention map produced by the ARN at the n-th iteration. T is the ground truth of the transmission map. λ is a trade off factor which represents the contribution of each iteration and is set to a value of 1, which means that all iterations are of equal contribution.

We visualize the haze attention map after each iteration, illustrated in Fig. 2. Fig. 2(a) is the input hazy image and Fig. 2(b)–2(d) are the outputs of the ARN after the 1st, 2nd and 3rd iteration, respectively. From Fig. 2, we have two observations. First, the pixels in the close view have larger values, showing a lower haze concentration, while the pixels in distant view have smaller values, showing a higher haze concentration. Second, with increasing iteration, more and more haze concentration information can be captured and refined, which leads to a more accurate haze attention map.

3.2. U-recurrent encoder-decoder network

The previous encoder-decoder network in the ADN is a classic U-shaped architecture designed with 6 conv-relu blocks in the encoder and 4 conv-relu blocks and 2 deconv-relu block in the decoder, as displayed in Fig. 3(a). By extracting hierarchical information from the bottom-up and top-down pathways, the encoder-decoder network achieves success in image dehazing. However, one problem of this feed-forward network is that it is difficult to refine dehazing results by a coarse-to-fine scheme without adding more layers, and hence, more parameters. One possible solution is to employ a recurrent structure which feeds the output of the network back to its input and also to apply the recurrent unit, e.g. LSTM, in the network for connecting information between consecutive iterations. As shown in Fig. 3(b), Tao et al. insert an LSTM unit in the bottleneck layers of a recurrent encoder-decoder network to pass intermediate information from the last iteration to the current iteration [12]. The effect of this design is that the dehazing results can be refined with fewer parameters. However, there is still room for improvement. Due to the long recurrent pathway, spatial inconsistency and feature dilution occur when the features from the encoder go to the decoder and back to the input again. In addition, the LSTM unit with 4 inputs for the input gate, the forget gate, the cell state and the output gate requires many parameters. To ensure spatial consistency and to reduce feature dilution, we propose a new U-recurrent encoder-decoder to enhance extracted features and to refine dehazing results by shortening the effective recurrent pathways. The recurrent convolution which is adopted for passing intermediate status from the last iteration to the corresponding layer at the current iteration reduces the number of parameters. Besides, we further introduce a new residual pyramid pooling module (RPPM) and apply it in the recurrent encoder/decoder for enlarging the receptive field of the network. Finally, for passing more texture information, 4 parallel dilation convolutions with dilation factors 2,4,8 and 16, are used to connect the recurrent encoder and the recurrent decoder. The architecture of the proposed U-recurrent encoder-decoder network with 3 iterations is displayed in Fig. 3(c). A more condensed representation is shown in Fig. 1. In what follows, we explain the details of the U-recurrent encoder-decoder network.

Recurrent Encoder. The proposed recurrent encoder consists of 1 RPPM and 3 conv-relu blocks which doubles the number of channels of the previous layer and downsamples the feature maps to

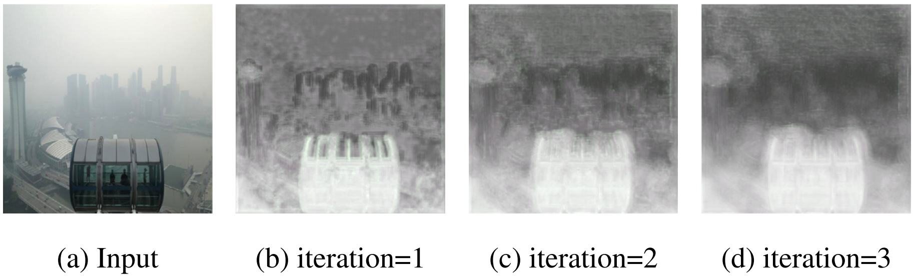
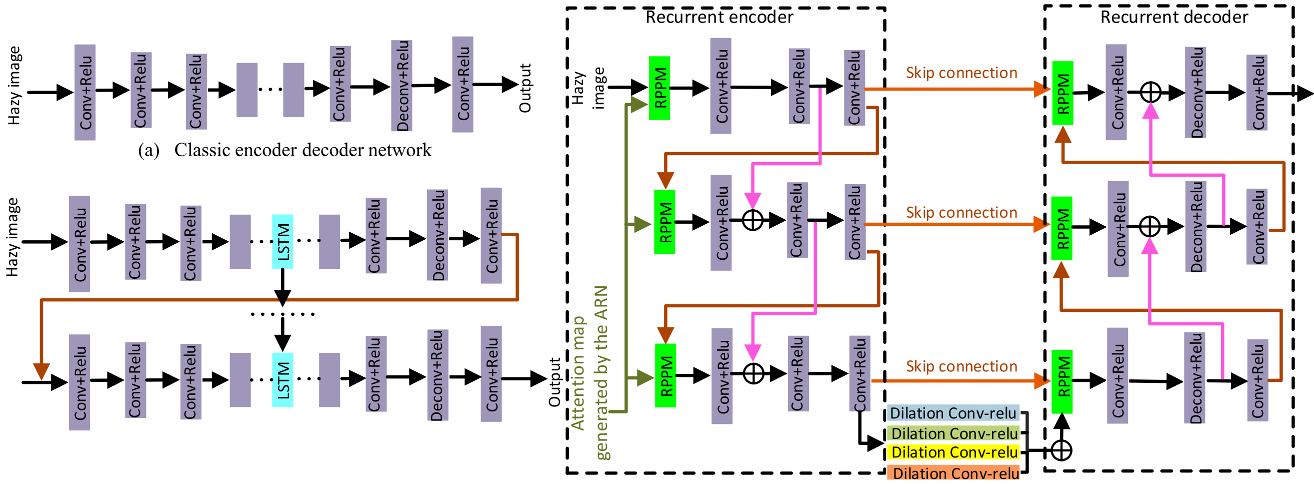

(a) Input (b) iteration=1 (c) iteration=2 (d) iteration=3

Fig. 2. Visualization of the haze attention map of the ARN after each iteration in testing. The pixel with lower value in the map represents a higher concentration of haze. With the increasing of iteration, ARN focuses more and more on hazy regions.



(a) Classic encoder decoder network

(b) Recurrent encoder decoder network with LSTM unit

(c) Our proposed U-recurrent encoder-decoder network with 3 iterations

Fig. 3. Comparison of different encoder-decoder models.

half size at each iteration. By passing the output feature maps through the encoder iteratively, the hierarchical features are extracted and enhanced by sharing parameters across all iteration steps. Hence, the function for the recurrent encoder is defined as

$$f^i = Net_{RE}([f^{i-1}, A^{i-1\downarrow}]; \theta_{RE}), i \geqslant 1 \tag{2}$$

where Net_{RE} is the recurrent encoder with parameter θ_{RE}. $[f^{i-1}, A^{i-1\downarrow}]$ refers to the concatenation of f^{i-1} and $A^{i-1\downarrow}$, where f^{i-1} is the output of the recurrent encoder at the $(i-1)th$ iteration and A is the haze attention map generated by the ARN. $(A)^{i-1\downarrow}$ represents the down-sampling operator to adapt the spatial size of A from the $(i-1)th$ iteration to the ith iteration. Then this concatenated feature also serves as the input of the encoder at the ith iteration. f^i is the output of the recurrent encoder at the ith iteration. It is noteworthy that f^0 is the original hazy image.

Although the features are enhanced by the recurrent mechanism as specified in Eq. (2), only the output of the encoder in the last iteration is considered for the recurrent inference, while the intermediate features containing rich information are not fully used at all. Besides, the receptive field of the encoder is not large enough to capture global features. To enlarge the receptive field of the encoder, we propose the RPPM and place it in the recurrent encoder as shown in Fig. 3(c). More details about the RPPM are given in the **Residual Pyramid Pooling Module** section. Meanwhile, to make full use of the intermediate features, we introduce a new recurrent convolution mechanism using residual connection.

Let E_RPPM with parameter θ_{E_RPPM} be the function of RPPM which is inserted in the recurrent encoder. E_CR1 with parameter θ_{E_CR1}, E_CR2 with parameter θ_{E_CR2} and E_CR3 with parameter θ_{E_CR3} denote as functions of the 1st, 2nd and 3rd conv-relu blocks in the encoder, respectively. The number of channels are doubled and the size of feature maps are downsampled by one half by passing features through the 2nd conv-relu block whose convolutional kernel size is 3 and stride is 2. The 1st and 2nd conv-relu blocks maintain the channel number and feature size by setting kernel size of 5 and 3 with stride of 1, respectively. The recurrent convolution is formulated as

$$\begin{aligned} f^i_{e\_rppm} &= E\_RPPM([f^{i-1}, A^{i-1\downarrow}]; \theta_{E\_RPPM}), \\ f^i_{e\_cr1} &= E\_CR1(f^i_{e\_rppm}; \theta_{E\_CR1}), \\ f^i_{e\_cr2} &= E\_CR2(f^i_{e\_cr1} + f^{i-1}_{e\_cr2}; \theta_{E\_CR2}), \\ f^i &= E\_CR3(f^i_{e\_cr2}; \theta_{E\_CR3}), i \geqslant 1 \end{aligned} \tag{3}$$

where $f^i_{e\_rppm}$ is the output of the RPPM in the recurrent encoder at the ith iteration. $f^i_{e\_cr1}$, $f^i_{e\_cr2}$ and f^i are the output features of the 1st, 2nd and 3rd conv-relu blocks at the ith iteration. $f^{i-1}_{e\_cr2}$ is the output feature of 2nd conv-relu block at the $(i-1)th$ iteration. By considering $f^{i-1}_{e\_cr2}$ as the hidden state, the recurrent convolution can be performed by residual connection which connects $f^{i-1}_{e\_cr2}$ with $f^i_{e\_cr1}$. Meanwhile, the combined feature $f^i_{e\_cr1} + f^{i-1}_{e\_cr2}$ further serves as the input to the 2nd conv-relu block at the ith iteration. Fig. 4(a) shows an example of this recurrent convolution operation. It is noteworthy that $f^0_{e\_cr2}$ represents a feature that has the same feature size as $f^1_{e\_cr1}$.

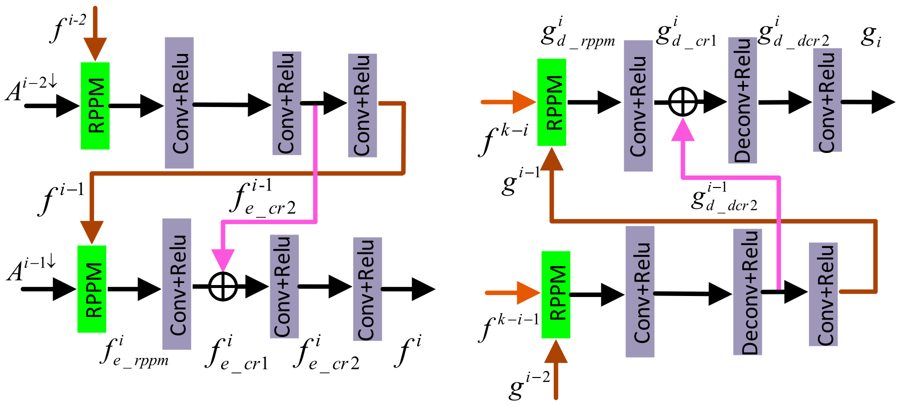
Compared to existing recurrent units, e.g. LSTM, the recurrent convolution operation can reduce recurrent parameters using residual connections. Besides, the recurrent convolution can also reduce feature dilution and enforce spatial dependencies effectively. The advantages of the recurrent convolution operation are demonstrated in our ablation study.

Parallel dilation convolutions. Before the output of the encoder enters the recurrent decoder, it first goes through 4 parallel dilation conv-relu blocks with varying dilation factors of 2,4,8 and 16. Unlike previous serial dilation conv-relu blocks in the backbone, this design can capture enhanced context information and can benefit dehazing results with higher accuracy. Its effectiveness is verified in our ablation study. The related operations are formulated as:

$$\begin{aligned} f_{dcr_2} &= DCR_2(f_{re}; \theta_{DCR_2}), \\ f_{dcr_4} &= DCR_4(f_{re}; \theta_{DCR_4}), \\ f_{dcr_8} &= DCR_8(f_{re}; \theta_{DCR_8}), \\ f_{dcr_{16}} &= DCR_{16}(f_{re}; \theta_{DCR_{16}}), \\ f_{dcr} &= \sigma(f_{dcr_2} + f_{dcr_4} + f_{dcr_8} + f_{dcr_{16}}), \end{aligned} \tag{4}$$

where f_{re} is the final output of the recurrent encoder; DCR_2 with parameter θ_{DCR_2}, DCR_4 with parameter θ_{DCR_4} , DCR_8 with parameter θ_{DCR_8} and DCR_{16} with parameter $\theta_{DCR_{16}}$ are 4 parallel dilation convolutional layers, where 2,4,8 and 16 are the corresponding dilation factors; σ denotes the nonlinear activation function ReLu. As shown in Eq. (4), the outputs of 4 parallel dilation convolutional layers, f_{dcr_2}, f_{dcr_4}, f_{dcr_8}, $f_{dcr_{16}}$, are further fused by the summation operation for enhancing context information f_{dcr}.

Recurrent decoder. The recurrent decoder, which consists of 1 RPPM, 2 conv-relu blocks and 1 deconv-relu block, further maps f_{dcr} to the clear image and refines the predicted result by an



(a) Recurrent convolution in encoder (b) Recurrent convolution in decoder

Fig. 4. Unfolded architecture of the recurrent convolution.

iterative scheme. A similar recurrent architecture in the recurrent encoder is also used in the decoder. As well, a skip connection between iterations in the decoder is also incorporated. Let the number of iterations for both the encoder and the decoder be k, the function of the recurrent decoder is defined as

$$g^i = Net_{RD}([g^{i-1}, f^{k-i}]; \theta_{RD}), i \geqslant 1 \tag{5}$$

where Net_{RD} is the recurrent decoder with parameter θ_{RD}; $[g^{i-1}, f^{k-i}]$ refers to the concatenation of g^{i-1} and f^{k-i}. Here, g^{i-1} and f^{k-i} are the outputs of the recurrent encoder and the recurrent decoder at the $(i-1)th$ and $(k-i)th$ iteration, respectively. From that, we find that the skip connections between f^{k-i} and g^{i-1} both with the same spatial size can be used to accelerate network convergence. It is noteworthy that g^0 is equal to f_{re}.

Based on Eq. (5), the recurrent convolution operations in the recurrent decoder are expressed as

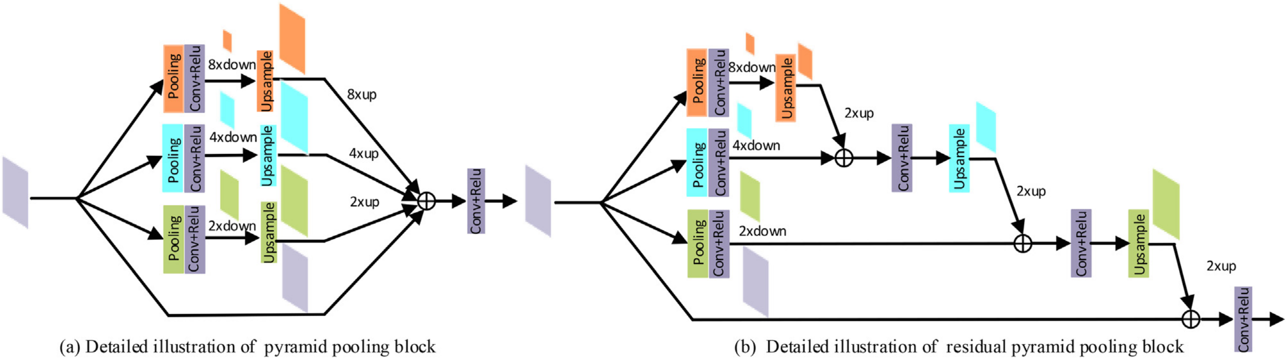
$$\begin{aligned} g^i_{d\_rppm} &= D\_RPPM([g^{i-1}, f^{k-i}]; \theta_{D\_RPPM}), \\ g^i_{d\_cr1} &= D\_CR1(g^i_{d\_rppm}; \theta_{D\_CR1}), \\ g^i_{d\_dcr2} &= D\_DCR2(g^i_{d\_dcr1} + g^{i-1}_{d\_dcr2}; \theta_{D\_CR2}), \\ g^i &= D\_CR3(g^i_{d\_cr2}; \theta_{D\_CR3}), i \geqslant 1 \end{aligned} \tag{6}$$

where D_RPPM is the function of the $RPPM$ in the decoder; $g^i_{d\_rppm}$ is the output of the RPPM at the ith iteration. D_CR1, D_DCR2 and D_CR3 are the 1st conv-relu block, the 2nd deconv-relu block, and the 3rd conv-relu block with parameters $\theta_{D\_CR1}, \theta_{D\_DCR2}$ and θ_{D_CR3}, respectively. $g^i_{d\_cr1}, g^i_{d\_cr2}$ and g^i are the outputs of the above mentioned blocks at the ith iteration. $g^{i-1}_{d\_dcr2}$ is the output of the deconv-relu block at the $(i-1)th$ iteration. All the kernel sizes are set to 3 and strides are set to 1, except for the deconvolution in the deconv-relu block whose stride is 2. Then this deconv-relu block can double the spatial size of features and reduce the number of channels by one half. Fig. 4(b) shows an example of the recurrent convolution operation in the decoder.

Residual Pyramid Pooling Module. The recurrent mechanisms in the encoder and the decoder enable spatial consistency and reduce feature dilution, but they cannot enlarge the receptive field of the whole network. As mentioned in [28], if the local features at each level are aware of structural information at different scale spaces, the receptive field of the encoder-decoder network can be enlarged effectively. Take the simple pooling-based network as an example, by inserting the pyramid pooling block into each module in the top-down pathway, not only the features at different levels of the decoder are merged with structural information seamlessly but also the receptive field of the network is enlarged. The detailed illustration of this pyramid pooling block is displayed in Fig. 5(a). As can be seen, features are converted to different scale spaces by feeding them to 3 average pooling layers with downsampling rates of 2,4 and 8. Then the output features from these 3 branches are upsampled to the resolution of original input and are fused with the input by a summation operation, followed by a convolutional layer. However, one potential drawback of this design is the information lost, especially when features with a larger downsampling rate, e.g. 8, is upsampled to the same spatial size of the input directly for feature fusion. Hence, it is necessary to bridge this gap by upsampling features and fusing them gradually. To address this issue, we propose to add a residual pyramid pooling module (RPPM) in the recurrent encoder, and at each iteration, make feature maps aware of the global structural information. The design of the RPPM is displayed in Fig. 5(b). The RPPM has 4 branches. One branch delivers the original input, while the other branches generate the downsampled features with downsampling rates 2,4 and 8 in scale space, respectively, using an average pooling layer and a convolutional layer. In the next step, these upsampled structural features are fused gradually using residual learning. In particular, the downsampled features generated by the largest downsampling rate, i.e. 8, are upsampled by a factor 2 and fused with features generated by a downsampling rate of 4 by the summation operation, followed with a convolutional layer which serves as a weighting layer for the summation fusion. Similarly, the fused features are further upsampled by a factor of 2 and fused with the features generated by a downsampling rate of 2 in a similar scheme. These operations are repeated until all the features generated by the 4 branches are fused completely.

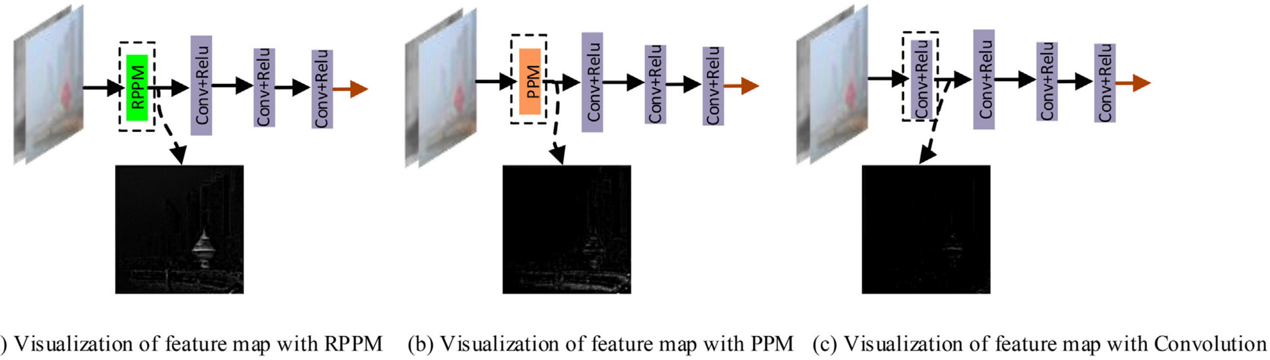
Our RPPM has two main advantages. First, by fusing the features generated by two adjacent downsampling rates each time, the downsampled features from one branch only need to be upsampled by a factor 2. Compared with upsampling these features to the original resolution of the input directly, the proposed design reduces the problem of information lost effectively. Besides, the fusion strategy using summation and convolutional can facilitate gradient propagation and allow the convolutional layer to learn to accommodate the importance of each branch.

To explore the effect of the RPPM, we visualize the features with RPPM, pyramid pooling module(PPM) and conv-relu block in the network, respectively. Since the pyramid pooling module has been demonstrated to enlarge the receptive field of a decoder [14], we only visualize feature near the RPPM, the PPM and the conv-relu block in the encoder. By comparing features after the RPPM (see Fig. 6(a)), the alternative PPM (see Fig. 6)) and the alternative



(a) Detailed illustration of pyramid pooling block (b) Detailed illustration of residual pyramid pooling block

Fig. 5. Comparison of an existing pyramid pooling module and our proposed residual pyramid pooling module.



(a) Visualization of feature map with RPPM (b) Visualization of feature map with PPM (c) Visualization of feature map with Convolution

Fig. 6. Visualizing feature maps with the proposed residual pyramid pooling module (RPPM), the pyramid pooling module and the conv-relu block, respectively.

conv-relu block (see Fig. 6(c)) at the 1st iteration, it can be easily seen that the RPPM allows the network capture more structural features with sharp edges. While the PPM captures inferior structrual features with few sharp edges and the conv-relu block captures minimal structural information. Therefore, the design of RPPM provides local features at each iteration with structural information and enlarges the receptive field of the whole network. The advantages of RPPM are demonstrated in our ablation study.

4. Loss functions

Inspired by the success of hybrid loss functions in training the dehazing network, we empirically find that the standard L_2 loss combined with the L_G gradient loss are effective to train our network. By adopting L_2 and L_G together, we can recover the clear image with more detailed information. With the loss function L_{ARN} designed for the ARN in Section 3.1, the final loss function of our network can be written as

$$L = \lambda L_2 + \gamma L_G + L_{ARN}, \tag{7}$$

where λ and γ are the weighting coefficients, representing the influence of L_2 and of L_G, respectively. In this work, we set the value of 1 for both λ and γ, which means that L_2 and L_G contribute equally.

L_2 is defined as

$$L_2 = \frac{1}{N}\sum_{t=0}^{N}\|I_t - J_t\|_2, \tag{8}$$

where $I_t, t = 1, 2 \ldots, N$ and $J_t, t = 1, 2, \ldots, N$ represent the set of clear images predicted by our network and the corresponding ground truths, respectively.

L_G is defined using gradient operations in the horizontal and the vertical directions:

$$L_G = \frac{1}{N}\sum_{t=0}^{N}\|(G_v(I_t) - G_v(J_t)\|_2 + \frac{1}{N}\sum_{t=0}^{N}\|(G_h(I_t) - G_h(J_t)\|_2, \tag{9}$$

where G_v and G_h are the vertical and horizontal gradient operators, respectively. Such a loss function allows us to preserve fine details and to remove artifacts.

5. Experimental results

In this section, we first give the implementation details and discuss the datasets. Then, we conduct a series of ablation studies for analyzing the effectiveness of each component of our network. Finally, we demonstrate the performance of the proposed network by comparing it with existing state-of-the-art methods on both synthetic and real hazy images.

5.1. Implementation details

The proposed network is implemented using PyTorch and is trained on a PC with 4 NVIDIA RTX 2080 Ti GPUs. For fair comparisons, the parameters setting are based on values reported by existing image dehazing methods [9,10], e.g. all the input images are resized to 512 × 512, the batch size is set to 2 and the optimization algorithm ADAM is used with $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The default settings of the ARN in the ADN are also used for our ARN. Besides, the strides are set to 1 for the deconvolution and the dilation convolution in the U-recurrent encoder-decoder network, except for the 2nd convolution in the recurrent encoder and the 2nd deconvolution in the recurrent decoder with the strides set to 2. All the kernel sizes are set to 3 except that of the 1st convolution layer of the recurrent encoder with the kernel size set to 5. The number of iterations k is set to 2 for the recurrent encoder and the recurrent decoder and the network is trained for 2,400,000 iterations.

5.2. Datasets

Although there are several existing datasets, the number of synthetic hazy images contained in them is large. For example, the public **RESIDE** dataset contains 313,950 synthetic outdoor images [29]. Directly training our model using existing datasets would cost too much training time and it would not be fair to compare our model with other dehazing models which are trained with only 4,000∼10,000 synthetic images. Hence, similar to existing dehazing methods, we create our training dataset based on existing public datasets. First, 4000 synthetic images provided by [9] are selected as indoor images which are used to synthesize images via the physical atmospheric scattering model, with 1000 depth images, random atmospheric light $A \in \{0.5, 1\}$ and scattering coefficient $\beta \in \{0.4, 0.6\}$. Then, another 4000 synthetic images from the RESIDE dataset with $A \in \{0.8, 0.85, 0.9, 1\}$ and $\beta \in \{0.04, 0.06, 0.08, 0.1, 0.12, 0.16, 0.2\}$ are randomly selected as outdoor images. Moreover, another 400 indoor images from [9] denoted as **Test A** and 400 outdoor images from RESIDE denoted as **Test B** form our testing dateset, denoted as **Test C**. Overall, we create 8000 training images and 800 testing images in total. Besides, the public testing dataset **SOTS** in **RESIDE** [29] which contains 500 indoor images and 500 outdoor images with different haze concentration is adopted for testing the robustness of our algorithm.

5.3. Ablation studies

In this subsection, we first test the effect of the number of iterations for the U-recurrent encoder-decoder network and the combined loss function. Then, we further demonstrate the effectiveness of the haze attention map, the recurrent encoder/decoder, the RPPM and the parallel dilation conv-relu blocks. Finally, we evaluate different configurations of the recurrent encoder/decoder and the RPPM.

5.3.1. Analysis on the number of iterations for U-recurrent encoder-decoder network

In order to verify the effect of different number of iterations and to select the optimal number of iterations for the recurrent encoder/decoder, we vary the number of iterations k from 1 to 4 for both the recurrent encoder and the recurrent decoder. The performance of the AUEDN with different number of iterations are reported in Table 1. As can be seen, the performance improves when k is increased from 1 to 2 but declines when $k > 2$. Hence, k is set to 2 in all subsequent experiments.

5.3.2. Analysis on the number of iterations for training

Fig. 7 shows the PSNR and the SSIM of our model with epoch = 400. As can be seen, the curves of PSNR and SSIM converge when epoch = 300 and they stabilize after 300. Hence, we set the number of epochs to 300 in our experiments for training. Because our training dataset contains 8000 images, 300 epochs is equivalent to 2,400,000 iterations.

5.3.3. Analysis on the combined loss function

Based on the AUEDN, we further study the effect of the proposed combined loss function with two different settings. We remove L_2 or L_G from Eq. (7) while keeping the other items unchanged. From Table 1 we find that both changes degrade the performance compared to the combined loss function as defined in Eq. (7). Hence, training with the combined loss function leads to higher PSNR and SSIM than training with either single loss function.

5.3.4. Analysis on the effectiveness of haze attention map

To confirm that the haze attention map generated by the ARN can make subsequent subnetwork aware of the haze concentration, we conduct experiments based on the backbone ADN with two different settings, namely, ADN without supervision of the transmission map (ADN w/o TM) and ADN with supervision of the transmission map (ADN w TM). ADN w TM denotes the haze attention map is generated by the ARN in the ADN with guidance of the loss function f_{ARN}. On the contrary, ADN w/o TM represents the haze attention map generated without using the loss function f_{ARN} and the whole network is trained using the combined loss L_2+L_G. The related results are displayed in Table 2. As can been seen, ADN w TM outperforms ADN w/o TM significantly, which demonstrates that the haze attention map obtained by the loss function f_{ARN} for the ARN does make subsequent sub-network aware of the haze concentration.

5.3.5. Analysis on the effectiveness of the recurrent encoder/decoder

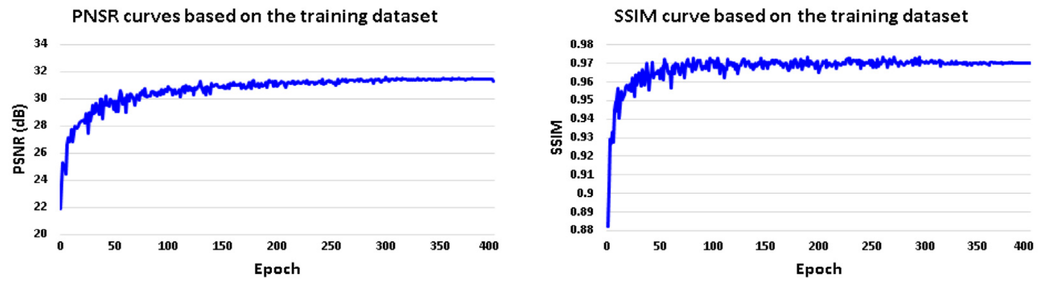
To test the effectiveness of the recurrent encoder/decoder, we compare the performance of the AUEDN w/o RPPM and AUEDN w RNN. The AUEDN w/o RPPM denotes the AUEDN without the RPPM in the recurrent encoder and the recurrent decoder, respectively. Based on the AUEDN w/o RPPM, the AUEDN w RNN further applies the recurrent mechanism to the whole encode-decoder network rather than applying the recurrent mechanism to the encoder and to the decoder. For fair comparison, the number of iterations is set to 2 in both cases. The results shown in the 3rd, 4th and 5th columns of Table 2 confirm that the AUEDN w/o RPPM gives performance gain over the backbone ADN w/o TM and also outperforms the AUEDN w RNN, which demonstrates that our model does benefit from the proposed recurrent encoder and recurrent decoder. Due to enhancing features and refining results with short recurrent pathways, the proposed recurrent mechanism enables spatial consistency and reduces the problem of information dilution effectively.

5.3.6. Analysis on the effectiveness of RPPM

We compare the performance of the AUEDN and the AUEDN w/o RPPM for verifying whether or not the RPPM provides the local features at each iteration with structural information at different scale spaces. The result displayed in the 6th row of Table 2 indicates that the AUEDN is qualitatively better than the AUEDN w/o RPPM (the fourth column in Table 2). Hence, the RPPM is crucial to the recurrent encoder and recurrent decoder for enlarging receptive field of the network.

Table 1
Ablation analysis for the number of iterations and the loss function on **SOTS** testing dataset. Best results are shown in bold.

	Number of iteration				Loss function		
	k = 1	k = 2	k = 3	k = 4	$L_2 + L_{ARN}$	$L_G + L_{ARN}$	$L_2 + L_G + L_{ARN}$
PSNR	29.01	**31.40**	27.74	26.84	30.57	30.45	**31.57**
SSIM	0.965	**0.9734**	0.9562	0.9490	0.9729	0.9718	**0.9733**



(a) (b)

Fig. 7. Convergence analysis for the number of iterations in terms of PSNR and SSIM values. (a) The PSNR curve based on the training dataset; (b) The SSIM curve based on the training dataset.

Table 2
Ablation studies for AUEDN variants on **SOTS** testing dataset. Best results are shown in bold.

Model	ADN w/o TM	ADN w TM	AUEDN w/o RPPM	AUEDN w RNN	AUEDN	AUEDN w SD	AUEDN w/o RC	AUEDN w PPM
PSNR	28.74	30.12	30.77	29.12	**31.40**	28.97	30.42	30.81
SSIM	0.9652	0.9678	0.9710	0.9705	**0.9734**	0.9619	0.9714	0.9719

5.3.7. Analysis on the effectiveness of parallel dilation conv-relu blocks

We use the serial dilation conv-relu blocks to replace the proposed parallel dilation conv-relu blocks in the AUEDN to evaluate the effectiveness of the parallel dilation conv-relu blocks. Here we denote this variant model as AUEDN w SD. Comparing the results of the AUEDN (the 6th column of Table 2) and the AUEDN w SD (the 7th column of Table 2), we can see that the performance of the AUEDN w SD declines noticeably. On the other hand, we also observe that, by using the RPPM, the recurrent encoder/decoder and the parallel dilation conv-relu blocks together in AUEDN, the performance are further enhanced compared to other variant models (the 2nd, 3rd, 4th, 5th and 7th columns in Table 2). Hence, we conclude that the three components of our proposed network all contribute to improving results.

5.3.8. Analysis on the configuration of the recurrent encoder/decoder

To better understand the configuration of the recurrent encoder/decoder, we remove the recurrent convolutional operation in the encoder and the decoder, and conduct the recurrent operation by taking the output of the encoder/decoder back to its input again. The results of the related variant model AUEDN w/o RC shown in the 8th column in Table 2 demonstrate that in the absence of recurrent operations in the proposed network, the performance declines (the 6th column in Table 2). This supports the statement that the recurrent convolutional operation plays an important role in the proposed recurrent encoder/decoder.

5.3.9. Analysis on the configuration of the RPPM

To evaluate the effectiveness of the configuration of the proposed RPPM, we use the pyramid pooling module (PPM) (see Fig. 6(a)) to replace the RPPM in our AUEDN and denote this model as AUEDN w PPM. The quantitative result shown in Table 2 indicates that the RPPM with progressive fusion strategy by summation works better than the PPM with direct fusion.

5.4. Evaluation on the synthetic dataset

In this section, the proposed AUEDN is compared with 6 state-of-the-art methods, including two traditional dehazing methods, e.g. DCP [7] and NLP [8], three CNNs-based dehazing methods, e.g. Dehazenet [23], AOD-N [24] and DCPDN [9], and one CNNs-based deraining method ADN [11]. Meanwhile, Test A, Test B, Test C and SOTS are used for evaluating the performances of the above methods. For fair comparison, the dehazing results of the compared methods are generated by the original code provided by the authors. As for the ADN, because this network is designed for image deraining, we apply the loss f_{ARN} to the ARN and re-train the ADN with our 8000 synthetic outdoor and indoor images for obtaining the dehazing model, which is actually the model ADN w TM discussed in Section 5.3.4.

Quantitative results are reported in Table 3. We observe that our model surpasses all state-of-the-art methods using Test A, Test B and Test C datasets. Besides, we further test the performance of different methods on the public SOTS dataset which includes 500 indoor images and 500 outdoor images with different haze concentration. With the data collected from [19,21,20], Table 4 reports the quantitative results of our method and 5 hand-crafted methods. Among them, PWAB [19], AMEF [21] and GPE [20] are recently released methods which have favorable performance on dehazing. As can be seen, our method not only ranks first, but also outperforms all of these methods with a large margin. In addition, Table 5 shows the performance of our method and 7 latest CNN-based methods. As can be seen, GridDN [26] ranks first and our AUEDN ranks second on the indoor images of the SOTS dataset. The competitive performance of GridDN can be attributed to its preprocessing module and post-precessing module. In contrast, our model without extra pre or post processing still has competitive performance and the difference of SSIM's between our method and that of GridDN is small, only 0.02. However, our method ranks first on the outdoor images of SOTS and outperforms all state-of-the-art methods.

We also show the visual results in Fig. 8. Fig. 8(a) displays the input images, Fig. 8(b)-Fig. 8(j) display the dehazing results of 9 methods, including the best and the second best hand-crafted methods in Table 4, i.e. DCP [7] and NLP [8] and 7 CNN-based methods in Table 5, i.e. Dehazenet [23], AOD-N [24], DCPCN [9], GFN [27], GridDN [26], EPDN [25], ADN [11] and our AUEDN. It is noteworthy that we skip GFN because the recent released methods, e.g. GridDN and EPDN, have demonstrated that they outperform GFN in visual effect by a large margin. Fig. 8(k) shows the ground truth. For a close-up examination, the zoom-in details of the regions enclosed in red rectangles are shown in the 2nd, 4th

Table 3
Quantitative comparisons of existing methods on the synthetic testing datasets in terms of PSNR/SSIM. ▷ means the model is re-trained on synthetic 8000 images. Best results are shown in bold.

Method		DCP [7]	NLP [8]	Dehazenet [23]	AOD-N [24]	DCPDN [9]	ADN [11]▷	AUEDN
Test A	PSNR	13.95	17.44	20.19	17.83	29.22	29.75	**29.87**
	SSIM	0.8824	0.7959	0.8773	0.8842	0.9560	0.9703	**0.9768**
Test B	PSNR	13.95	16.59	22.30	18.54	28.12	27.30	**27.37**
	SSIM	0.8664	0.7736	0.9159	0.8520	0.9416	0.9714	**0.9716**
Test C	PSNR	13.77	17.01	21.24	18.19	28.67	28.52	**28.62**
	SSIM	0.8753	0.7847	0.8966	0.8681	0.9488	0.9701	**0.9742**

Table 4
Quantitative comparisons of our method and hand-crafted methods on the public SOTS dataset. Best results are shown in bold.

Method		DCP [7]	NLP [8]	GPE [20]	PWAB [19]	AMEF [21]	AUEDN
Indoor	PSNR	16.62	16.69	11.97	15.96	16.01	**27.80**
	SSIM	0.8179	0.7767	0.6301	0.7415	0.7573	**0.9621**
Outdoor	PSNR	19.13	18.85	15.91	12.33	17.62	**35.00**
	SSIM	0.8148	0.8476	0.7297	0.6759	0.8201	**0.9848**

Table 5
Quantitative comparisons of our method and CNN-based methods on the public SOTS dataset. ▷ means the model is re-trained on synthetic 8000 images. Mixture represents indoor images and outdoor images. Red and blue indicate the best and second best performance, respectively.

Method		Dehazenet [23]	AOD-N [24]	DCPCN [9]	GFN [27]	GridDN [26]	EPDN [25]	ADN ▷ [11]	AUEDN
Indoor	PSNR	21.14	19.06	15.85	22.30	32.16	25.06	27.64	27.80
	SSIM	0.8472	0.8504	0.8175	0.8800	0.9836	0.9232	0.9580	0.9621
Outdoor	PSNR	20.46	20.29	19.93	21.55	30.86	22.57	32.47	35.00
	SSIM	0.8514	0.8765	0.8449	0.8444	0.9819	0.8630	0.9727	0.9848
Mixture	PSNR	20.8	19.67	17.89	21.92	31.51	23.81	30.05	31.40
	SSIM	0.8493	0.8634	0.8312	0.8622	0.9827	0.8931	0.9653	0.9734

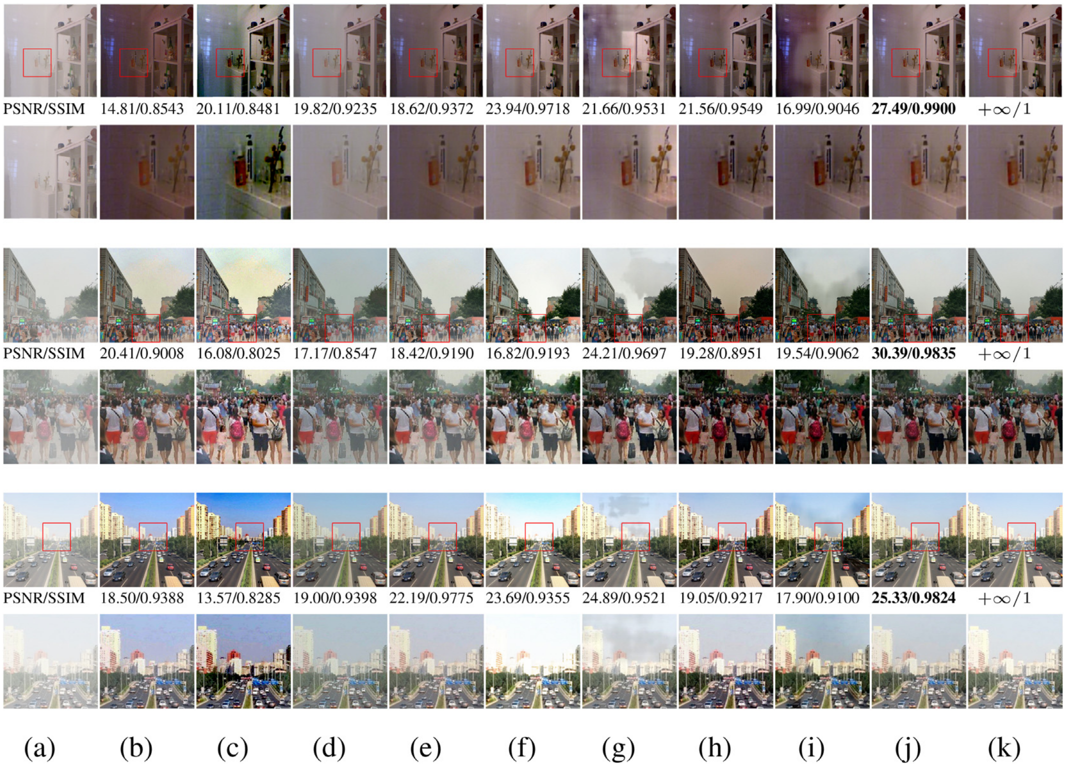


Fig. 8. Dehazing results on the synthetic dataset. (a) Input; (b) DCP [7]; (c) NLP [8]; (d) AOD-N [24]; (e) Dehazenet [23]; (f) DCPDN [9]; (g) ADN [11]; (h) EPDN [25]; (i) GridDN [26]; (j) Ours; (k) Ground Truth.

and 6th rows of Fig. 8. As can be seen, although the compared methods remove haze in most cases, their results are either over dehazed (too dark) or under dehazed (too bright). As an example, the results of DCP and NLP (see Fig. 8 and Fig. 8(c)) have color distortion and are darker than that of the ground truth (see Fig. 8(k)). To be specific, the visually dark corner in the bathroom and color distorted sky regions can be observed in the 2nd, 4th and 6th rows of Fig. 8(k)) and Fig. 8(c). Besides, there are still residual haze and artifacts in the results of AOD-N and Dehazenet, e.g. the zoom-in details displayed in the 2nd and 4th rows of Fig. 8(c) and Fig. 8 (e), respectively. Although the dehazed results shown in the 5th row of Fig. 8(e), and Fig. 8(e) are close to the corresponding result shown in the ground truth (see Fig. 8(k)), upon detailed inspection on zoom-in details in the 6th row of Fig. 8(k)), and Fig. 8(e), one can observe that the sky regions are darker than that of the ground truth. On the other hand, the DCPDN over dehazed images by making the results brighter. For example, the sky regions in the 3rd and the 5th rows of Fig. 8(f) are much brighter than that of the ground truth (see Fig. 8(k)). The ADN removes haze in most scenes, except generating artifacts in some regions, such as the color distorted sky regions in the 3rd and the 5th rows of Fig. 8(g). The results by EPDN and GridDN seems to be better visually than that of other CNN-based methods. However, there is still an issue of over dehaze problem. For instance, Fig. 8(g) and Fig. 8(i) are much darker that the ground truth Fig. 8(k), especially the wall and sky regions in the 1st and 3rd rows of Fig. 8(k), and Fig. 8(i). In contrast, our method (see Fig. 8(j)) is able to remove haze with much less color distortion and the results are visually closest to the ground truth. The quantitative PSNR/SSIM shown under each image confirm the superior performance of the proposed method.

5.5. Evaluation on Real Dataset

We compare our method with state-of-the-art dehazing methods on real hazy images which are downloaded from the internet. Since the ground truths of real images are not available, we only display visual comparisons in Fig. 9. Similar to the comparison on the synthetic dataset, the zoom-in details are shown in the next row for close examination. As can be seen, the DCP, NLP and EPDN tend to darken the images and produce color distortion, e.g., the color of sky regions in the 1st and 3rd rows of Fig. 9(b), and (h) turns to yellowish color. Meanwhile, unrealistic color can be observed clearly in the zoom-in details shown in the 2nd and 4th rows of Fig. 9(h) and (c). The AOD-N and Dehazenet leave the haze in the results (see Fig. 9(c) and (e)). Although the ADN tends to remove haze without residual haze, there are still artifacts in some regions, such as the dark shadow in the 1st and 3rd rows of Fig. 9 (g). The DCPCN not only generates unrealistic color in the sky region (see the 1st row of Fig. 9(f)), but also leaves haze in distant view (see the 4th row of Fig. 9(f)). The GridDN also produces a halo effect and residual haze in the images to some extent, e.g. the obvious halo effect in the sky region shown in the 1st row and 3rd row of Fig. 9(f)) and the residual haze also can be found in the 2nd row of Fig. 9(i). Our results displayed in Fig. 9(j) have much less color distortion and show sharper contours. The zoom-in details shown in the 2nd and the 4th rows of Fig. 9(h) show clear visual results.

5.5.1. Comparison with optimization-based and fusion-based methods

In this section, we further verify the robustness of our method by comparing it with optimization-based and fusion-based methods, e.g. AMEF [21], CTVM [4], MPFM [5] and PWAB [19], on a real image. Among them, CTVM is an optimization-based method, and AMEF, MPFM and PWAB are fusion-based methods. Thanks to the available results and codes provided by their corresponding authors [4,21,5], Fig. 10 shows the visual results. As can be seen, CTVM produces the natural effect and removes haze effectively, except that details are lost. The zoom-in image shown in the second row of Fig. 10(b) shows that sharp edges are not preserved. This is because CTVM assumes that the surface radiance is piecewise smooth and utilizes an alternating minimization scheme to solve the constrained optimization problem, but cannot handle fine details. Since PWAB cannot estimate the transmission map accurately, obvious color distortion is introduced in the buildings (see the first row of Fig. 10(e)). Besides, the color of detail shown

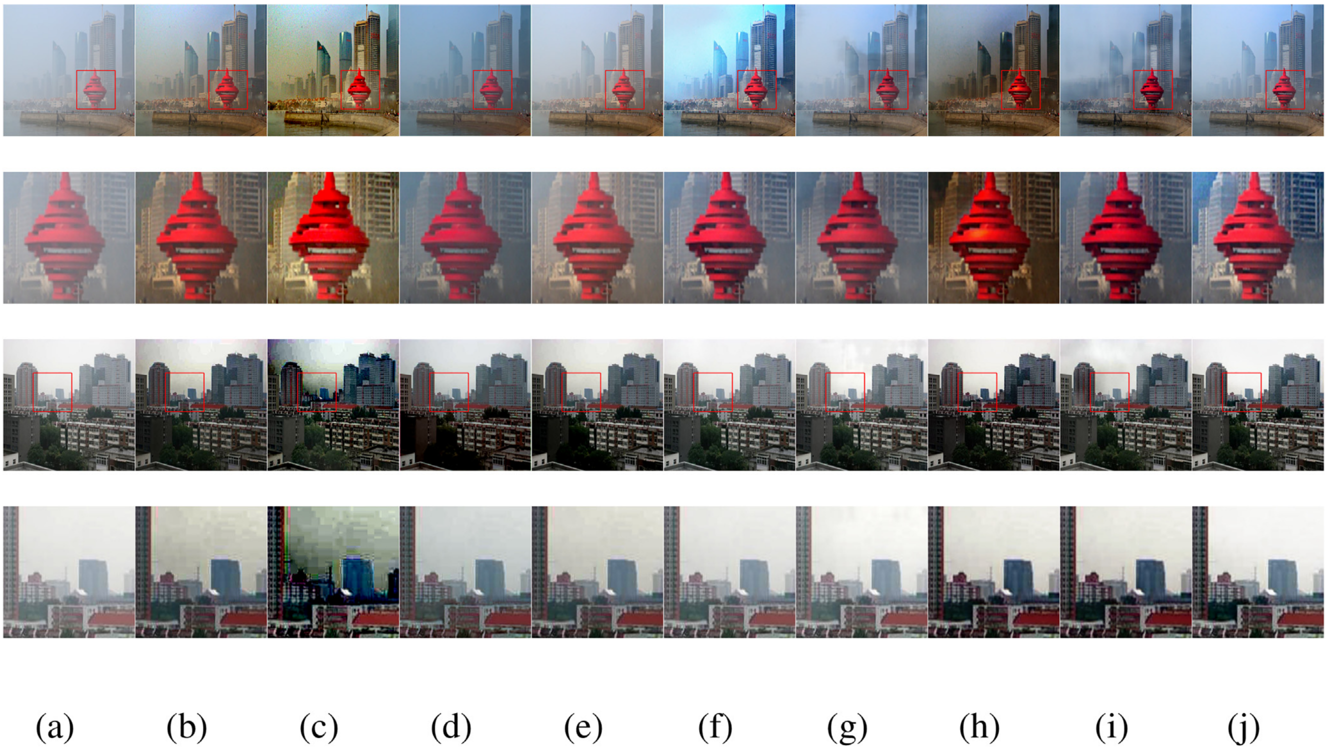


Fig. 9. Dehazing results on real images. (a) Input; (b) DCP [7]; (c) NLP [8]; (d) AOD-N [24]; (e) Dehazenet [23]; (f) DCPDN [9]; (g) ADN [11]; (h) EPDN [25]; (i) GridDN [26]; (j) Ours.

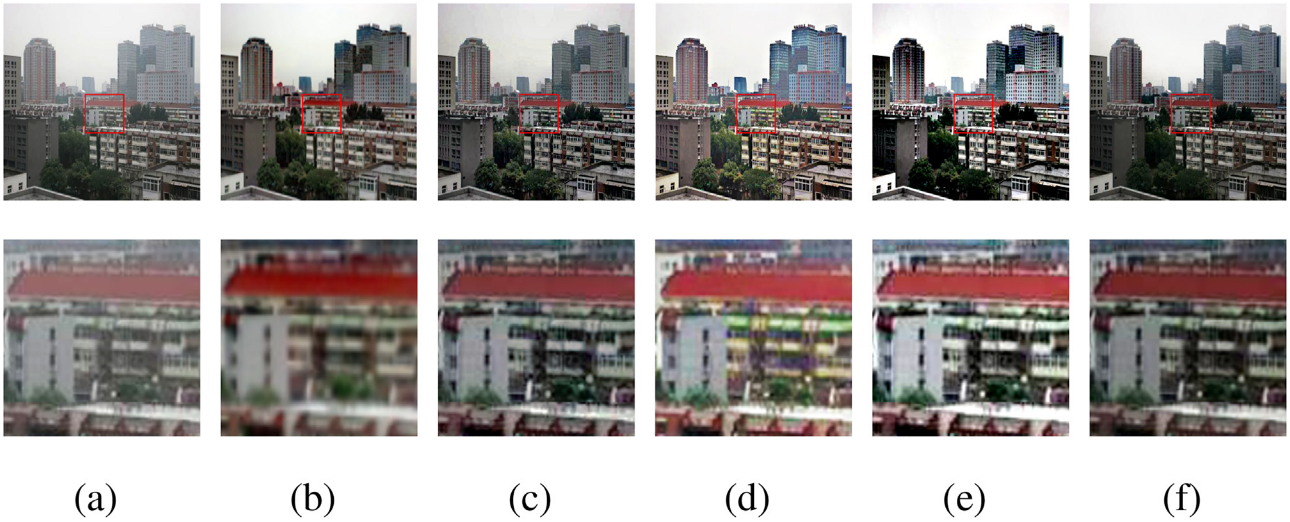


Fig. 10. Comparison of real image between our method and fusion-based and optimization-based methods. (a) Input; (b) CTVM [4]; (c) AMEF [21]; (d) MPFM [5]; (e) PWAB [19]; (f) Ours.

in the second row of Fig. 10(e) is too bright. The MPFM also produces unrealistic image in Fig. 10(d), because the fused weight maps calculated by haze-relevant features cannot describe all cases. Although the results by AMEF have better visual effect, there is still residual haze in the zoom-in detail in the second row of Fig. 10(c). In contrast, our method gives more realistic results with sharp details (see Fig. 10(f)).

5.5.2. Comparison with multi-scale CNN-based methods

Multi-scale information is helpful to obtain global information and recover details. Hence, current CNN-based methods usually incorporate multi-scale information into the framework, e.g. DRPN [6], MSCNN [13] and EPDN [25]. In this section, we further compare our method with these 3 CNN-based methods. Among them, DRPN not only employs the densely connected construction to extract different features, but also adopts the multi-scale pyramid pooling layer to capture the global information for final estimation. MSCNN uses a coarse-scale network to predict the transmission map, and follows with a fine-scale network to refine the coarse result with details. EPDN also employs the pyramid pooling layer in the enhancing block for embedding details of features from different scales into the clear image. The visual results of different methods are displayed in Fig. 11. As can be seen, the multi-scale information helps CNN-based methods recover details very well, e.g. the result by EPDN (see Fig. 11(d)). However, the river in the distance view has an unrealistic color. The result by DRPN has residual haze (see Fig. 11(b)). The result by MSCNN has color distortion, e.g. the color of Fig. 11(c) appears yellowish. In contrast, our method recovers the real image with realistic color and rich details (see Fig. 11(e)).

6. Run time

As our model contains two networks: a recurrent attentive network and a U-recurrent encoder-decoder network, a natural question to ask is how fast can the proposed method dehaze an image? In this section, we compare the average run time of our method with 5 state-of-the-art methods based on the public SOTS dataset via a PC with 4 NVIDIA RTX 2080 Ti GPUs. The related results are reported in Table 6. From that, we observe that the proposed AUEDN ranks second in run time, only second to AOD-N [24]. The high efficiency of AOD-N can be attributed to its light-weight architecture which contains 5 convolutional layers only. However, this design also leads to poor performance (see Table 3). Our method, which uses the recurrent architecture to share parameters across time stages, obtains promising performance with competitive run time. From Table 6, we observe that our method are more efficient than recent released EPDN [25], GridDN [21] and PIDN [30].

7. Limitation

As we all know that most CNN-based dehazing methods highly rely on synthesized datasets to train their models. Although

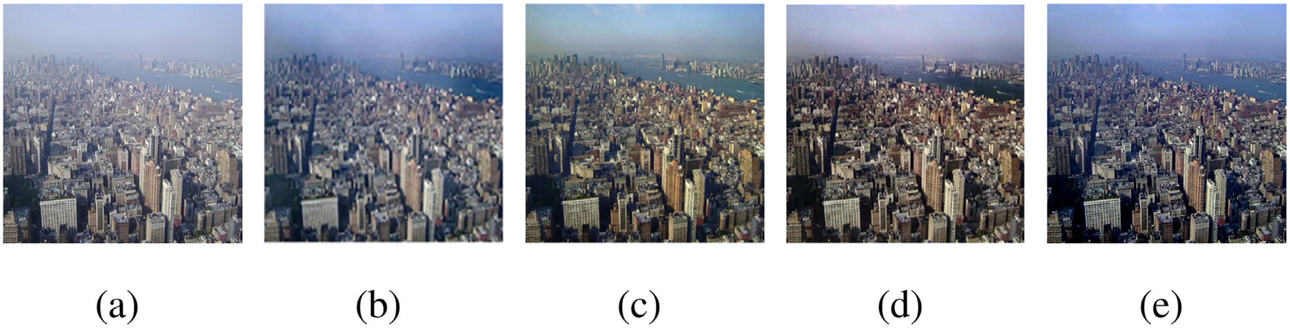


Fig. 11. Comparison of real image between our method and fusion-based and optimization-based methods. (a) Input; (b) DRPN [6]; (c) MSCNN [13]; (d) EPDN [25]; (e) Ours.

Table 6
Comparison of our proposed AUEDN with other art-of-the-state methods in terms of run time (second). Red and blue indicate the fastest and second fastest results, respectively.

	EPDN [25]	GridDN [21]	PIDN [30]	AOD-N [24]	DCPDN [9]	AUEDN
Platform	Pytorch	Pytorch	Pytorch	Pytorch	Pytorch	Pytorch
Run time	0.121	0.245	0.115	0.002	0.056	0.054

synthetic parameters tend to describe physical characteristics of haze by various parameters, e.g. atmospheric light and scattering coefficients, these datasets still cannot capture the large range of real images and hence, models training using these datasets cannot be generalized to all scenes. For example, most dehazing models cannot handle night-time images, due to the lack of such images in the training dataset. Hence, our method also shares the same limitation. Obviously, it is desirable to create a dataset using real-world images, rather than synthetic images. Recently, Wang et al. propose a semi-automatic method to generate real rain images by incorporating temporal priors and human supervision into a mathematical model [31]. Cai et al. build a real-world super-resolution dataset by adjusting the focal length of the camera [32]. Inspired by such an initiative, we plan to create a dataset with real hazy images and use it to train our model in the future.

8. Conclusion

In this paper, we propose a new attentive U-recurrent encoder-decoder network for image dehazing. By assuming that haze at different depths can be detected by multiple stages, the proposed method leverages the recurrent attentive network to generate the haze attention map for guiding the subsequent U-recurrent encoder-decoder network to dehaze and to estimate the clear image. Our U-recurrent encoder-decoder network then enhances the features for dehazing and refines the dehazing results using a new recurrent mechanism in the encoder and the decoder. This design not only enables spatial consistency but also reduces information dilution by shortening the recurrent pathways. To further enlarge the receptive field of the whole network, a novel residual pyramid pooling module is proposed and used in the U-recurrent encoder-decoder network, providing local features with structural information effectively.

Extensive ablation studies have verified the effectiveness of each proposed component in our network. Experiments conducted on synthetic images and real images also indicate that our method outperforms state-of-the-art methods quantitatively and qualitatively.

CRediT authorship contribution statement

Shibai Yin: Conceptualization, Methodology, Software, Writing - original draft. **Yibin Wang:** Data curation, Resources, Investigation, Validation. **Yee-Hong Yang:** Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 61502396), the Education Department Foundation of Sichuan Province (No. 18ZB0484), the Fundamental Research Funds for the Central Universities (No. JBK2002026) and the Natural Sciences and Engineering Research Council of Canada. In addition, this work is also supported by the Computer Graphics Laboratory of the University of Alberta, Key Laboratory of Financial Intelligence and Financial Engineering of Sichuan Province and Collaborative Innovation Center for the Innovation and Regulation of Internet-Based Finance.

References

[1] M. Sajjad, I.U. Haq, J. Lloret, W. Ding, K. Muhammad, Robust image hashing based efficient authentication for smart industrial environment, IEEE Trans. Industr. Inf. 15 (2019) 6541–6550.
[2] Z. Shao, W. Zhou, X. Deng, M. Zhang, Q. Cheng, Multilabel remote sensing image retrieval based on fully convolutional network, IEEE J. Selected Topics Appl. Earth Observations Remote Sensing (2020).
[3] S. Yin, Y. Wang, Y.-H. Yang, A novel image-dehazing network with a parallel attention block, Pattern Recogn. 107255 (2020).
[4] W. Wang, C. He, X.-G. Xia, A constrained total variation model for single image dehazing, Pattern Recogn. 80 (2018) 196–209.
[5] Y. Li, Q. Miao, R. Liu, J. Song, Y. Quan, Y. Huang, A multi-scale fusion scheme based on haze-relevant features for single image dehazing, Neurocomputing 283 (2018) 73–86.
[6] Y. Wu, Y. Qin, Z. Wang, X. Ma, Z. Cao, Densely pyramidal residual network for uav-based railway images dehazing, Neurocomputing 371 (2020) 124–136.
[7] K. He, J. Sun, X. Tang, Single image haze removal using dark channel prior, IEEE Trans Pattern. Anal. Mach. Intell. 33 (2010) 2341–2353.
[8] D. Berman, T. Treibitz, S. Avidan, Non-local image dehazing, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1674–1682.
[9] H. Zhang, V.M. Patel, Densely connected pyramid dehazing network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3194–3203.
[10] H. Zhang, V. Sindagi, V.M. Patel, Multi-scale single image dehazing using perceptual pyramid deep network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 902–911.
[11] R. Qian, R.T. Tan, W. Yang, J. Su, J. Liu, Attentive generative adversarial network for raindrop removal from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2482–2491.
[12] X. Tao, H. Gao, X. Shen, J. Wang, J. Jia, Scale-recurrent network for deep image deblurring, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8174–8182.
[13] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, M.-H. Yang, Single image dehazing via multi-scale convolutional neural networks, in: European conference on computer vision, Springer, 2016, pp. 154–169.
[14] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, J. Jiang, A simple pooling-based design for real-time salient object detection, arXiv preprint arXiv:1904.09569 (2019)..
[15] K. Muhammad, S. Khan, V. Palade, I. Mehmood, V.H.C. De Albuquerque, Edge intelligence-assisted smoke detection in foggy surveillance environments, IEEE Trans. Industr. Inf. (2019).
[16] S. Shaw, R. Gupta, S. Roy, A review on different image de-hazing methods, in: Emerging Technology in Modelling and Graphics, Springer, 2020, pp. 533–540.
[17] R. Fattal, Dehazing using color-lines, ACM Trans. Graphics (TOG) 34 (2014) 13.
[18] Q. Zhu, J. Mai, L. Shao, A fast single image haze removal algorithm using color attenuation prior, IEEE Trans. Image Process. 24 (2015) 3522–3533.
[19] T. Yu, K. Song, P. Miao, G. Yang, H. Yang, C. Chen, Nighttime single image dehazing via pixel-wise alpha blending, IEEE Access 7 (2019) 114619–114630.
[20] J.E. Hernandez-Beltran, V.H. Diaz-Ramirez, L. Trujillo, P. Legrand, Design of estimators for restoration of images degraded by haze using genetic programming, Swarm Evol. Comput. 44 (2019) 49–63.
[21] A. Galdran, Image dehazing by artificial multiple-exposure image fusion, Signal Processing 149 (2018) 135–147.
[22] Y. Jiang, C. Sun, Y. Zhao, L. Yang, Image dehazing using adaptive bi-channel priors on superpixels, Comput. Vis. Image Underst. 165 (2017) 17–32.
[23] B. Cai, X. Xu, K. Jia, C. Qing, D. Tao, Dehazenet: An end-to-end system for single image haze removal, IEEE Trans. Image Process. 25 (2016) 5187–5198.
[24] B. Li, X. Peng, Z. Wang, J. Xu, D. Feng, Aod-net: All-in-one dehazing network, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4770–4778.
[25] Y. Qu, Y. Chen, J. Huang, Y. Xie, Enhanced pix2pix dehazing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8160–8168.

[26] X. Liu, Y. Ma, Z. Shi, J. Chen, Griddehazenet: Attention-based multi-scale network for image dehazing, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7314–7323.
[27] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, M.-H. Yang, Gated fusion network for single image dehazing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3253–3261.
[28] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.
[29] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, Z. Wang, Benchmarking single-image dehazing and beyond, IEEE Trans. Image Process. 28 (2018) 492–505.
[30] D. Ren, W. Zuo, Q. Hu, P. Zhu, D. Meng, Progressive image deraining networks: a better and simpler baseline, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3937–3946.
[31] T. Wang, X. Yang, K. Xu, S. Chen, Q. Zhang, R.W. Lau, Spatial attentive single-image deraining with a high quality real rain dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12270–12279.
[32] J. Cai, H. Zeng, H. Yong, Z. Cao, L. Zhang, Toward real-world single image super-resolution: A new benchmark and a new model, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3086–3095.

Shibai Yin obtained her Ph.D. degree from the Chang'an University, Xi'an, China, in 2013. Since December, 2016, She is an associate professor in the Department of Economic Information Engineering, Southwestern University of Finance and Economics. Her research interests include image processing and machine vision.

Yibin Wang obtained his Ph.D. degree from the Northwestern Polytechnical University, Xi'an, China, in 2016. He is currently serving as a lecturer in Engineering Department, Sichuan Normal University. His research interests include image processing and machine vision.

Yee-Hong Yang received the Ph.D. from the University of Pittsburgh. He was on the faculty in the Department of Computer Science, the University of Saskatchewan from 1983 to 2001 and served as Graduate Chair from 1999 to 2001. Since July, 2001, he is professor in the Department of Computing Science, the University of Alberta. He is senior member of the IEEE and serves on the editorial board of the numerous international journals and on the program committees of many international conferences. His research interests cover a wide range of topics from computer graphics (e.g., physics-based modelling, texture analysis and synthesis, and static and dynamic image-based modeling and rendering) to computer vision (e.g., stereo and multi-view stereo, underwater imaging, and medical imaging). He has published many technical papers in international journals such as TPAMI, IJCV, TIP, and PR, and conferences such as ICCV, CVPR and ECCV.