Abstract—Haze reduces the visibility of image content and leads to failure in handling subsequent computer vision tasks. In this paper, we address the problem of single image dehazing by proposing a dehazing network named T-Net, which consists of a backbone network based on the U-Net architecture and a dual attention module. Multi-scale feature fusion can be achieved by using skip connections with a new fusion strategy. Furthermore, by repeatedly unfolding the plain T-Net, Stack T-Net is proposed to take advantage of the dependence of deep features across stages via a recursive strategy. To reduce network parameters, the intra-stage recursive computation of ResNet is adopted in our Stack T-Net. We take both the stage-wise result and the original hazy image as input to each T-Net and finally output the prediction of the clean image. Experimental results on both synthetic and real-world images demonstrate that our plain T-Net and the advanced Stack T-Net perform favorably against state-of-the-art dehazing algorithms and show that our Stack T-Net could further improve the dehazing effect, demonstrating the effectiveness of the recursive strategy.

Index Terms—Image dehazing, Multi-scale, Dual attention, Recurrent structure, Recursive strategy.

I. INTRODUCTION

Haze is a common atmospheric phenomenon that not only has a serious adverse effect on human visual perception, but also has a serious impact on the performance of modern computer vision systems for various visual tasks, such as image classification, object detection and video surveillance. Since hazy conditions such as fog, aerosols, sand and mist can scatter and adsorb light, images taken in a hazy environment suffer from many different degradations such as blurred visibility, color cast, reduced contrast and halo artifacts. Therefore, it is crucial to develop effective solutions to reduce the impact of image degeneration caused by environmental conditions through dehazing.

In the past ten years, single image dehazing has attracted widespread attention in the computer vision community. Its goal is to restore clean scenes from hazy images. According to the atmosphere scattering model [1], [2], [3], the hazy process can be approximated as,

\[ I(x) = J(x)t(x) + A(x)(1 - t(x)), \]

where \( I \) and \( J \) denote a hazy image and its clean scene, respectively. \( A \) represents the global atmospheric light, \( t \) describes the transmission map and \( x \) is the pixel location. The transmission \( t(x) \) of each pixel \( x \) is related to the depth \( d(x) \) of the pixel, and is represented as \( t(x) = e^{-\beta d(x)} \), where \( \beta \) is the atmospheric scattering coefficient and controls haze density. Many methods based on this physical model restore a clean scene \( J \) from a hazy image \( I \) by estimating the global atmospheric light and the transmission map of the hazy image.

Early prior-based methods attempt to estimate the transmission map by using the difference in statistical properties of hazy and clear images, such as contrast prior [4], dark channel prior (DCP) [5], color-lines [6], color attenuation prior [7], haze-lines [8] and color ellipsoidal prior [9]. However, these hypothesized priors are not suitable for all real-world images, and it is easy to obtain inaccurate approximations of the transmission map, which will in turn cause quality degradation of the restored images. To solve this problem, driven by the unprecedented success in recent years of deep learning in low level vision tasks [10], [11], [12], [13], [14], convolution neural networks (CNNs) have been adopted for the task of single image dehazing.

Earlier deep learning-based methods, including the DehazeNet [15], the multi-scale CNN (MSCNN) [16] and the residue learning technique [17], use deep networks to estimate the transmission map and exploit the conventional DCP’s [5] bright pixels method to estimate the atmospheric light. However, since the estimation of the atmospheric light is coarse, the results of these methods are not sufficiently satisfactory. Therefore, some methods try to use deep networks to estimate the transmission map and atmospheric light simultaneously or separately. For example, Li et al. [18] propose a network named AOD-Net to estimate a new variable that combines the transmission map with the atmospheric light, while Zhang et al. [19] use a two-stream network to estimate these two variables separately.
In addition to the methods based on the atmospheric scattering model, there are some works that use end-to-end deep neural networks to directly estimate clean images instead of explicitly estimating the transmission map and atmospheric light. Considering that the estimation of the transmission map and atmospheric light sometimes deviates from the real hazy image, directly estimating the clean image can avoid sub-optimal restoration. Inspired by the progresses of generative adversarial networks (GANs) in synthesizing realistic images, many methods use GANs as the framework to transform hazy images to clean images. For instance, Qu et al. [20] propose a GAN based on pix2pix to realize image-to-image translation for dehazing.

Image restoration methods based on deep learning are driven by data, requiring large-scale synthetic or real-world datasets for training. Since it is difficult to obtain paired hazy and clean images in the real world, synthetic datasets are generally used to train networks in experiments. These datasets are created by using the atmosphere scattering model to simulate the process of image degradation. Although proper guidance by the transmission map and atmospheric light is beneficial for effective dehazing under a wide range of haze densities and for dealing with color distortions caused by haze, the algorithms based on the estimation of the transmission map and atmospheric light also rely on the same physical model. These algorithms may suffer from inherent performance loss on real-world images due to the over-fitting of the algorithms on synthetic datasets. In comparison, the methods of direct mapping exhibit higher robustness on real-world hazy images.

For this reason, we design an end-to-end network named T-Net for single image dehazing, which directly predicts clean images without using the atmospheric scattering model. Our T-Net is a symmetrical T-shaped architecture that is composed of two components, a backbone module and a dual attention module. The backbone module is a network based on the U-Net [21] architecture, in which a residual dense block [22], upsampling block and downsampling block are used as basic blocks. To obtain distinctive features conducive to dehazing, a dual attention module [23], which can also enhance the robustness of the network, is embedded in the backbone module. Meanwhile, in view of the effectiveness of multi-scale features for image restoration [16], [24], [25], [26], [27], [28], [29], we use skip connections with a new fusion strategy [30] in T-Net to realize adaptive multi-scale feature fusion. Moreover, inspired by the recurrent structure of [25] and [31], we propose Stack T-Net by repeatedly unfolding our T-Net to further improve the performance of dehazing. Each stage of Stack T-Net uses the output of the previous stage and the original hazy image as input and outputs a clean image. Fig. 1 shows the dehazed results of our proposed T-Net and Stack T-Net on exemplar real-world images.

To summarize, our work has three-fold contributions as follows.

1) T-Net, a novel network with a dual attention module based on the U-Net architecture, is proposed to realize efficient information exchange across the multi-scale features from different levels to directly predict clean images.

2) Stack T-Net, which is created by repeatedly unfolding T-Net, can further improve the performance of dehazing by taking both the stage-wise result and the original hazy image as input to each stage.

3) Extensive experiments show that our plain T-Net and Stack T-Net perform favorably against state-of-the-art methods on both synthetic and real-world hazy images. Meanwhile, an ablation study is conducted to demonstrate the effects of different modules in the proposed network.

The rest of the paper is organized as follows. Section II discusses the related work of single image dehazing. Section III describes the structure of the plain network T-Net, including a backbone module and a dual attention module. Section IV explains the details of our overall dehazing network Stack T-Net and defines the loss function for training. Section V introduces the datasets used in the experiments, the results of the ablation study and the quantitative and qualitative comparisons between our approach and state-of-the-art methods on synthetic and real-world images. Section VI concludes the paper and discusses the direction of future work.

II. RELATED WORK

In this section, we briefly discuss the single image dehazing methods, which can be roughly divided into two categories, prior-based methods and learning-based methods as mentioned above.

A. Prior-Based Methods

For the multiple image dehazing task, some additional information beneficial to dehazing can be obtained from several

![Fig. 1. Dehazed results on real-world hazy images. We show the results of T-Net and Stack T-Net, which both perform well on these images. Compared with T-Net, the results of Stack T-Net are cleaner and brighter, and exhibit finer details.](image-url)
hazy images of the same scene. For instance, some works [32], [33], [34] exploit the polarization properties of multiple hazy images taken from different angles with a polarizer to derive the transmission map. In contrast, single image dehazing is a highly ill-posed problem due to the lack of additional information from other sources. To address this problem, different priors or assumptions obtained through observations and statistics on a large amount of real data have been used in many dehazing methods. Except for the earliest methods [35], [36], [37], [38] based on varying image enhancement algorithms, most prior-based dehazing methods are proposed based on the atmosphere scattering model, obtaining clean images by estimating the transmission map and atmospheric light to invert (1). Representative works following this route include [4], [5], [6], [7], [8], [9], [39].

Fattal et al. [39] propose a dehazing algorithm by estimating the albedo under the assumption that the transmission map has no local correlation with the surface shading. Tan et al. [4] propose a local contrast-maximization method based on Markov random field (MRF), due to the observation that the contrast of hazy images is lower than that of the corresponding clear images. Observing that the intensity of at least one color channel in local regions of natural haze-free images is close to zero and that the pixel intensity increases as haze increases, He et al. [5] present the dark channel prior (DCP) as an approximation of haze distribution to estimate the transmission map. Meanwhile, this work also proposes a popular way to estimate the atmosphere light by averaging the top 0.1% of brightness pixels in hazy images. Many subsequent works make improvement on DCP to refine the estimation of the transmission map by using different edge-preserving smoothing filters [40], [41], [42], [43].

Discovering that haze causes the color-lines [44] to deviate from the origin, Fattal et al. [6] develop a new method to recover the transmission map, where color-lines are proposed by observing that pixels of small image patches are generally distributed as a 2D Gaussian along some lines in the RGB space. The color attenuation prior adopted in the linear model of [7] assumes that as haze increases, the brightness of images increases but saturation decreases. By assuming that the color of a clean image can be roughly decomposed into a combination of hundreds of different colors, Berman et al. [8] observe that the colors in hazy images form line-shaped clusters in the RGB space, where these lines are named haze-lines and pass through the coordinate value corresponding to the atmospheric light. Based on this observation, they propose a method to estimate the transmission map by calculating the haze-lines with the predicted atmospheric light, which is computed with the brightness pixels method of DCP.

Biu and Kim [9] construct color ellipsoids by statistically fitting hazy pixel clusters in the RGB space and then calculate a prior vector through color ellipsoid geometry to obtain the transmission map. With the restriction that the minimum channel of a hazy image is linearly related to that of its corresponding haze-free image, Wang et al. [45] introduce a new single image dehazing method by combining DCP [5] with linear transformation. To avoid the distortions in sky or other bright areas, the weakening strategies are introduced into the estimation process of the transmission map of DCP in this work. Moreover, the pixels for the calculation of the atmospheric light are better selected by locating the sky areas with an improved quad-tree subdivision method. Aimed at the problem that the atmospheric light estimation method of DCP [5] cannot reflect the local variation of illumination, Hu et al. [46] calculate the local atmospheric light of non-overlapping image patches and combine it with the global atmospheric light to generate an adaptive illumination map. The illuminance channel contains a large amount of information about the atmospheric light, whereas the chrominance channel contains little. Accordingly, Fang et al. [47] introduce a novel single image dehazing approach by reformulating the atmosphere scattering model in YUV color space. Despite some promising dehazing results that have been obtained, prior-based methods are limited by the hypothetical priors themselves. There is a certain gap between these priors and reality, which can cause the performance of these dehazing methods to be less than satisfactory.

B. Deep Learning-Based Methods

Recently, deep learning has achieved significant success in low-level vision tasks such as image super-resolution [12], [48], [49], deblurring [50], [51], [52], deraining [53], [54], [55] and desnowing [56], [57], which also include dehazing [15], [16], [18], [19], [58], [59], [60]. At present, there are two main ideas about dehazing methods based on deep learning. One is to estimate the transmission map and atmospheric light according to the atmospheric scattering model, and the other uses deep learning networks to directly predict clean images.

The deep learning dehazing methods according to the physical model use the same strategy as prior-based methods to restore clean images, but generally estimate the transmission map and atmospheric light by using specially designed CNNs instead of priors, so as to avoid the limitation on performance caused by the gap between hand-crafted priors and reality. Cai et al. [15] propose a dehazing model, DehazeNet, to estimate the transmission map. Ren et al. [16] design a multi-scale CNN (MSCNN) to estimate the transmission map with a coarse-to-fine strategy. Zhang et al. [19] employ a two-stream densely connected pyramid dehazing network (DPCD), which predicts the transmission map and atmospheric light separately. Li et al. [18] create a reformulation of the atmosphere scattering model by using a new variable to integrate the transmission map and the atmospheric light, and design an end-to-end neural network named AOD-Net to estimate this variable. Zhang et al. [61] introduce Famed-Net to estimate the same variable as [18]. Song et al. [58] present a novel ranking convolutional neural network to automatically learn robust haze pattern features with statistical and structural properties from large-scale haze datasets, where the learned features are used to restore the transmission map by using a random forest regression model. Dudhane et al. [62] propose PYF-Net, which consists of a YNet for the estimation of the transmission map in the RGB and YCbCr space and an FNet to fuse two transmission maps. Recently, some works combine the classical method DCP with deep learning. For example, Golts et al. [63] design a new loss based on DCP, which is used for the training of an unsupervised deep network to estimate the transmission map. Chen et al. [64] exploit a Patch Map Selection...
Network (PMS-Net) to adaptively set the patch size of the dark channel corresponding to each pixel. Explicitly or implicitly encoding image depth, the model-based methods can typically maintain the depth structure of images while dehazing and deal with depth-related color distortions. However, real haze scenes are more complicated than the atmospheric scattering model, which is only a coarse approximation of the haze model, leading to the unsatisfactory restoration of these methods on real images even though they perform well on synthetic images.

The deep learning dehazing methods of direct mapping regard dehazing as an image-to-image translation problem. In recent years, generative adversarial networks (GANs) have shown great reconstruction ability in image generation and translation [12], [65], [66], and are subsequently used in the field of dehazing. Qu et al. [20] propose a pix2pix GAN with two enhancing blocks to predict clean images and enhance the detail and color. Raj et al. [67] design a conditional GAN based on the U-Net architecture for dehazing. Engin et al. [68] design an end-to-end dehazing network based on CycleGAN with no need for paired hazy and corresponding clean images for training. Du et al. [69] design a GAN with an adaptive loss to facilitate end-to-end perceptual optimization and propose a new post-processing method for halo artifact removal using guide filters. Shao et al. [70] apply an end-to-end network that is made up of two sub-networks, a bidirectional translation network based on CycleGAN to bridge the gap between the synthetic and real domains, and a dehazing network to restore clean images from the hazy images before and after translation. In addition to GAN, some works are proposed based on other architectures. Ren et al. [71] introduce an end-to-end network based on the encoder-decoder architecture and adopt a novel fusion-based strategy that derives three inputs by using three pre-processing methods. Liu et al. [28] propose a grid network GridDehazeNet for single image dehazing, which consists of three modules, pre-processing, backbone and post-processing. Li et al. [59] introduce a level-aware progressive network (LAP-Net), where the sub-network of each stage predicts the transmission with different haze-level supervision, and the final output is yielded with an adaptive integration strategy. Li et al. [60] present a two-stage single image dehazing network called PDR-Net, of which the first stage is a haze removal subnetwork that generates the coarse haze-free image through the perceptual loss [48], and the second stage is a refinement subnetwork that further enhances the visual effect of the haze-free result. Dong et al. [72] introduce a Multi-Scale Boosted Dehazing Network (MSBDN), which is built based on the U-Net architecture with two strategies designed for feature propagation and fusion, boosting and error feedback. Shin et al. [73] exhibit a triple convolutional network (TCN) with dual supervision for single image dehazing, which simulates the mismatch problem based on the region-model to enhance dehazed images. The direct-mapping methods avoid sub-optimal restoration by skipping the estimation of the atmospheric light and transmission map. However, due to the lack of the atmosphere scattering model constraints, many methods will lose depth information during the dehazing process, which causes the depth structure of clean images to be damaged and makes the depth-related color distortions difficult to address.

Our method estimates clean images by direct mapping, and fully considers how to propagate and maintain depth information in the design of the network structure, effectively solving the problem of the loss of depth information in the direct-mapping methods.

III. T-Net

In this section, we describe the proposed T-Net, a symmetrical T-shaped network consisting of two sub-modules, a backbone module based on the U-Net architecture and a dual attention module. Fig. 2 illustrates the architecture of T-Net, and the details are given in the following.

A. Backbone Module

The backbone module of T-Net is based on the U-Net architecture. We adopt a new fusion strategy to make use of the skip connections. As shown in Fig. 2, the network mainly includes three types of basic blocks, residual dense block (RDB) [22], upsampling block and downsampling block. The detailed structure of these three types of blocks is shown in Fig. 3.

Extensive research has demonstrated that the use of multi-scale features is beneficial to various image understanding tasks. High-level features are of the downsampled spatial resolution but compress more semantic contextual information that is necessary for scene understanding of images. In contrast, low-level features are of higher resolution to help localize objects but contain less semantic contextual information. Therefore, the feature fusion of different levels can simultaneously preserve spatial information from low-level features and exploit the semantic contextual information from high-level features for image understanding. Image restoration includes the process of image understanding to determine what should be kept on the images and what should be removed. To make full use of the information from different levels of features, multi-scale feature fusion...
has been applied to many image restoration tasks, such as image deraining [25], [26], image deblurring [74], [75] and image dehazing [27], [28], [72], which significantly improves the performance of the algorithms.

The U-Net [21] architecture is originally proposed for semantic segmentation, consisting of a contracting path to capture contextual cues, and a symmetric expanding path for precise localization, as well as multiple lateral connections between the contracting path and its symmetric expanding path, which are called skip connections and designed for multi-scale feature fusion. Moreover, we consider that in addition to haze density being related to image depth, image scene is also closely related to image depth. There is considerable spatial information retained in low-level features, which contain the depth information of images. The skip connections of U-Net enable the depth information in low-level features to propagate forwards and be combined with the semantic contextual information in high-level features, which allows the depth structure to be restored in the dehazed images. For these reasons, we design our T-Net based on this architecture and make several improvements for dehazing.

As shown in Fig. 2, the backbone is a symmetric network, which mainly includes three pairs of RDB blocks and four pairs of upsampling and downsampling blocks in the trunk road (the red path in Fig. 2), and three RDB blocks in the lateral connections. In addition, there is a convolutional layer without an activation function at the beginning of the network, which generates 16 linear feature maps as the learned input from a hazy image. There is another convolutional layer at the end, which is symmetrical to the beginning and used to generate high-quality dehazed images.

The RDB [22] block, which keeps the number of feature maps unchanged, can extract abundant local features via densely connected convolutional layers. Therefore, we choose it as the basic block for feature generation. As shown in Fig. 3, we use five convolutional layers in each RDB and set the growth rate to 16. Every convolutional layer takes the concatenation of the output features of all convolutional layers before it as input, and the final output of each RDB block is the combination of the output of the last layer and the input of the RDB block through channel-wise addition. The first four convolutional layers are used to extract features, and the last layer (kernel size = 1, 1 × 1 convolution) is used to keep the number of channels of the output feature the same as that of the input feature. To reduce information loss, we use convolutional layers to realize upsampling and downsampling; the detailed structure is shown in Fig. 3. The upsampling block and the downsampling block both consist of two convolutional layers, of which the first layer adjusts the size of the feature maps by setting different convolution kernel sizes, and the other layer uses a 1 × 1 convolution to change the number of channels. In each upsampling block, the number of feature maps decreases by half as the size of the feature maps increases by one time, which is the reverse in each downsampling block.

We use a new fusion strategy to realize skip connections, considering that features from different scales may not be equally important. Motivated by [30], we set two trainable fusion parameters for each skip connection, where every fusion parameter is a n-dimensional vector (n is the number of channels of the features before fusion). Moreover, RDB blocks instead of 1 × 1 convolutional layers are used in the lateral connection to obtain more feature combinations, which improves the possibility of obtaining more effective information for dehazing. The backbone includes four pairs of upsampling and downsampling blocks, so the feature from feature fusion can be expressed as

\[
\tilde{F}_i^j = \alpha f_u(F_i^j) + \beta f_a(F_{i+1}^j),
\]

\[
F_{i+1}^j = f_d(F_i^j),
\]

\[
f_x(\cdot) = \begin{cases} f_u(\cdot), & i < m - 1, \\ f_a(\cdot), & i = m - 1, \end{cases}
\]

\[
i = 0, 1, \ldots, m - 1; \quad j = 0, 1, \ldots, n,
\]

(2)

where \(\alpha\) and \(\beta\) are the fusion parameters for the two features to be fused, \(f_u(\cdot), f_a(\cdot), f_d(\cdot)\) and \(f_x(\cdot)\) stand for the functions of RDB, upsampling, downsampling and dual attention, respectively, \(F_i^j\) represents the j-th feature channel after the i-th downsampling, \(\tilde{F}_i^j\) represents the fused feature symmetric with \(F_{i+1}^j\), \(m\) is the number of upsampling and downsampling block pairs, and \(n\) is the number of channels of the features.

B. Dual Attention Module

As shown in Fig. 2, in addition to the backbone module, we use a dual attention module in T-Net, which is embedded in the middle of the backbone module and includes two blocks, the position attention block and the channel attention block. Fig. 4 is an overview of the dual attention module.

Discriminant feature representations are the key for scene understanding and can be obtained by capturing long-range contextual information. The dual attention module, which can adaptively integrate local features with their global dependencies, is
stands for the operation of concatenation, \( x \) separately represent the input and the output \( y \) is the original hazy image and \( f_0 = x \) and \( y = f_0 \), and \( (3) = y \in \{1, \ldots, K\} \) is the number of channel attention modules. First, a spatial attention matrix is generated by a softmaxed product of two identical original hazy images as input. We choose the consistent with other stages, the first stage uses the concatenation of the output of the previous stage as the input of the next stage, while the performance of deraining is limited by the deraining result of each stage. Therefore, we use the strategy of [31], using the concatenation of the output of the previous stage and the original hazy image as input, where the original hazy image can supplement the information loss caused by each dehazing stage. As shown in Fig. 5, we use T-Net with the same structure in each stage. The inference of Stack T-Net at the \( k \)-th stage can be formulated as

\[
    x^k = f_{in}(x^0, y^{k-1}),
\]

\[
    y^k = f_k^T(x^k),
\]

where \( x^k \) and \( y^k \) separately represent the input and the output of the \( k \)-th stage, \( x^0 \) is the original hazy image and \( y^0 = x^0 \), \( f_{in} \) stands for the operation of concatenation, \( f_k^T \) denotes the mapping of T-Net at the \( k \)-th stage, and \( k \in \{1, \ldots, K\} \). T-Net of each stage plays the role of a dehazing sub-network, which directly learns the mapping from hazy images to dehazy results. Thereby, the input of each stage is a six-channel image, which is the concatenation of two three-channel RGB images, and the output is a three-channel RGB image. Meanwhile, to be consistent with other stages, the first stage uses the concatenation of two identical original hazy images as input. We choose the output of the last stage as the final dehazing result.

In this section, we introduce the overall dehazing network Stack T-Net constructed by repeatedly unfolding the plain T-Net, as well as the loss function used in our method. The first subsection below explains the details of the architecture of Stack T-Net, and Fig. 5 is the illustration of a \( K \)-stage Stack T-Net, where sub-figure (a) exhibits the detailed structure, and sub-figure (b) shows the information flow and the feature resolution change among different stages. The second subsection introduces the loss function we used for training the full network.

A. Details of the Architecture

Inspired by [25] and [31], we introduce the recurrent structure to our method. In [25], Yang et al. use a recurrent network for deraining but use the deraining result of the previous stage as the input of the next stage, while the performance of deraining is limited by the deraining result of each stage. Therefore, we use the strategy of [31], using the concatenation of the output of the previous stage and the original hazy image as input, where the original hazy image can supplement the information loss caused by each dehazing stage. As shown in Fig. 5, we use T-Net with the same structure in each stage. The inference of Stack T-Net at the \( k \)-th stage can be formulated as

\[
    x^k = f_{in}(x^0, y^{k-1}),
\]

\[
    y^k = f_k^T(x^k),
\]

where \( x^k \) and \( y^k \) separately represent the input and the output of the \( k \)-th stage, \( x^0 \) is the original hazy image and \( y^0 = x^0 \), \( f_{in} \) stands for the operation of concatenation, \( f_k^T \) denotes the mapping of T-Net at the \( k \)-th stage, and \( k \in \{1, \ldots, K\} \). T-Net of each stage plays the role of a dehazing sub-network, which directly learns the mapping from hazy images to dehazy results. Thereby, the input of each stage is a six-channel image, which is the concatenation of two three-channel RGB images, and the output is a three-channel RGB image. Meanwhile, to be consistent with other stages, the first stage uses the concatenation of two identical original hazy images as input. We choose the output of the last stage as the final dehazing result.

However, the network of recurrent structures has a considerably large number of parameters to learn, which requires a great deal of memory and decreases the efficiency of dehazing. The more complex the structure of the network is, the more likely it leads to over-fitting. Witnessing the role of recursive
Fig. 5. The architecture of Stack T-Net. (a) The detailed structure. (b) The information flow and the feature resolution change among different stages. We use T-Net as the sub-network to build Stack T-Net by repeatedly unfolding it several times. Each dotted line of (b) represents a different resolution, and the resolution of each dotted line is 1/4 of that of the above one. Since we use T-Net with four pairs of upsampling and downsampling blocks as sub-network, there are five dotted lines in (b) (containing the original resolution).

computation in image restoration [76], [77], [78], we apply it to our method. In our practice, we utilize the inter-stage recursive computation instead of repeatedly unfolding the plain network. Since inter-stage recursive computation requires each stage to share the same network parameters, each stage can be expressed as

\[ f^k_T(\cdot) = f_T(\cdot). \] (4)

Experimental results verify that this recursive strategy can improve the dehazing effectiveness without designing a deeper and more complex network.

B. Loss Function

The loss function of our method measures the error of the dehazing results at each stage and is used to train the proposed network. The MSE loss has a smooth function curve, which is convenient for the use of the gradient descent algorithm. However, it is more sensitive to outliers and more easily causes gradient explosions than the \( L_1 \) loss. Therefore, we employ the smooth \( L_1 \) loss that combines the MSE loss and the \( L_1 \) loss to enhance the robustness of the network.

Let \( I_c(x) \) denote the intensity of the \( c \)-th color channel at pixel \( x \) in the ground truth and let \( J_c(x)^k \) represent the intensity of the same position of the dehazed image in the \( k \)-th stage. \( K \) and \( N \) denote the number of stages and pixels, respectively. The smooth \( L_1 \) loss of the \( k \)-th stage (\( L^k_{SL_1} \)) can be defined as

\[ L^k_{SL_1} = \frac{1}{N} \sum_{x=1}^{N} \sum_{c=1}^{3} f^k_{SL_1}(|J_c^k(x) - I_c(x)|), \] (5)

where

\[ f^k_{SL_1}(e) = \begin{cases} 0.5e^2, & 0 \leq e < 1, \\ e - 0.5, & e \geq 1. \end{cases} \] (6)

When the error between the dehazed result and the ground truth exceeds the limited value, the \( L_1 \) norm is used instead of the \( L_2 \) norm to reduce the influence of outliers on the network parameters.

The perceptual loss [48] measures image visual similarities between the dehazed image and the ground truth more effectively than the pixel-wise loss, for which we take advantage of the perceptual loss to strengthen the finer details of the dehazed images. By leveraging multi-scale features extracted from VGG16 [79] pre-trained on ImageNet [80], the perceptual loss of the \( k \)-th dehazed stage can be defined as

\[ L^k_p = \frac{1}{C_j H_j W_j} \left\| F_j(J^k) - F_j(I) \right\|_2^2, \] (7)

where \( J^k \) represents the dehazed result of the \( k \)-th stage, \( I \) represents the ground truth, each \( F_j(\cdot) (j = \{1, 2, 3\}) \) separately denotes the output features of the relu1-2, relu2-2 and relu3-3 layers of VGG-16, and \( C_j, H_j \) and \( W_j \) specify the dimensions of these features.

By combining the smooth \( L_1 \) loss and the perceptual loss, the total loss of all stages is defined as

\[ L = \sum_{k=1}^{K} L^k_{SL_1} + \lambda \sum_{k=1}^{K} L^k_p, \] (8)
that is,
\[ L = L_{SL} + \lambda L_P, \]
where \( \lambda \) is set to 0.04 to control the relative weights on the two loss components.

V. EXPERIMENTS

In this section, quantitative and qualitative experimental results are shown to demonstrate the effectiveness of the proposed method. First, we carry out an ablation study to demonstrate the effectiveness of each module of our network. Then, we conduct a series of experiments to compare the performance of our methods with state-of-the-art methods on both synthetic and real-world datasets. At last, we perform a runtime analysis to evaluate the time complexity of our proposed method.

A. Datasets

In the training phase, we use the datasets of [70] as the training set. This dataset contains 6000 synthetic hazy images from the RESIDE dataset [81], where the RESIDE dataset [81] contains both synthesized and real-world hazy/clean image pairs of indoor and outdoor scenes, and is split into five subsets, namely, ITS (Indoor Training Set), OTS (Outdoor Training Set), SOTS (Synthetic Object Testing Set), URHI (Unannotated Real Hazy Images) and RTTS (Real Task-driven Testing Set). Among the 6000 images, 3000 are chosen from the ITS, and the rest are from the OTS. All the images are randomly cropped to \( 256 \times 256 \) and randomly flipped for data augmentation. Then the pixel values are normalized to \([-1,1]\). Since our training set is a subset of ITS and OTS, we use SOTS as the total testing set in the test phase to compare the performance of our proposed method with other methods on synthetic data. Furthermore, the URHI dataset, which contains 1000 real-world hazy images of different sizes, is used to further evaluate the generalization ability of these methods in real-world scenarios.

B. Network Implementation

We implement our framework with Pytorch [82] and use the Adam optimizer [83] with a batch size 14 to train the network, where the momenta \( \beta_1 \) and \( \beta_2 \) adopt the default values of 0.9 and 0.999, respectively. We train every model for 2000 epochs in total. The learning rate is initially set to 0.001, reduced by half every 20 epochs, and kept fixed at 0.0001 from the 80-th epoch. Our proposed method is evaluated against the following state-of-the-art approaches: DCP [5], MSCNN [16], DehazeNet[15], NLD [8], AOD-Net [18], GFN [71], DCPDN [19], EPDN [20] and DA_dehaze [70]. For the synthetic dataset, we adopt the peak-signal-to-noise ratio (PSNR) and structural similarity measure (SSIM) as quantitative evaluation metrics, where PSNR can measure the pixel-by-pixel difference and SSIM can quantify the structural difference between a dehazed output image and its original haze-free scene. Although the hazy images of the real-world dataset have no corresponding ground truth, we choose two no-reference image evaluation metrics, the blind/referenceless image spatial quality evaluator (BRISQUE) [84] and perception-based image quality evaluator (PIQUE) [85] to further evaluate the image quality of the generated haze-free results. BRISQUE quantifies the possible loss of “naturalness” in a distorted image based on a statistical model of spatial natural scenes. PIQUE gathers the distortion scores predicted from the statistical features of different local patches to obtain a general quality score. Unlike PSNR and SSIM, lower values of BRISQUE and PIQUE represent better image quality. The results of all experiments are shown in the following.

C. Ablation Study

To evaluate the effectiveness of several key modules in our network, we perform ablation studies with the following three strategies.

The first ablation study is conducted to determine the configurations of the backbone module of the proposed T-Net, that is, the number of upsampling and downsampling block pairs and the number of RDB block pairs in the trunk road. RDB blocks are used as feature generators in our network. Simply put, the more RDB block pairs we use, the deeper the network and the deeper the features we can extract. The position attention block contains a matrix multiplication of \((N, C) \times (C, N)\), where \( N \) is the number of pixels in each channel of the feature. Thus, it is easy to cause out of memory if the input feature size of this block is too large. We set the initial pairs of upsampling and downsampling blocks as 2 to avoid this problem. Table I shows the performance of T-Net with different configurations on the SOTS dataset, where \( m \) and \( n \) represent the number of upsampling and downsampling block pairs and the number of RDB block pairs in the trunk road, respectively.

There is a tendency shown in Table I that the average PSNR and SSIM values increase as \( m \) and \( n \) increase. The exception occurs when \( n = 1 \), i.e., the performance of T-Net does not improve as \( m \) increases. The reason is that, the representation ability of features is influenced by the number of RDB block pairs in the trunk road. When \( n \) is set to 1, the features extracted by our network contain insufficient deep discriminant information, which makes the fitting ability of the neural network poor. Moreover, when \( n > 1 \), the performance of adding a pair of upsampling and downsampling blocks is better than adding a pair

| Configuration | SOTS |
|--------------|------|
| \( m \) | \( n \) | PSNR | SSIM |
| 2 | 1 | 22.90 | 0.8954 |
| 2 | 2 | 26.86 | 0.9451 |
| 3 | 1 | 27.43 | 0.9473 |
| 3 | 2 | 27.49 | 0.9520 |
| 4 | 1 | 28.13 | 0.9536 |
| 4 | 2 | 28.30 | 0.9535 |
| 3 | | **28.55** | **0.9543** |
of RDB blocks (see, e.g., (3,2) versus (2,3), (4,2) versus (3,3) in the form of (m,n)), which verifies the effectiveness of multi-scale features. Compared with single-scale features, multi-scale features can provide more discriminant information that is helpful for image understanding. As shown in Table I, the average PSNR and SSIM values are the highest when \( m = 4, n = 3 \). Therefore, we use a T-Net with four pairs of upsampling and downsampling blocks and three pairs of RDB blocks in the trunk road as our network in the following experiments.

The second ablation study is conducted to verify the effectiveness of each module by comparing the performance of several variants of T-Net. We set the trunk road without the dual attention module (where we use a RDB block to replace the dual attention module) as the basic model. Starting from this basic model, we create other variants by gradually injecting our modifications, which include usual skip connections realized by 1×1 convolutions, usual skip connections realized by RDB blocks, our skip connections, as well as our skip connections and dual attention module (T-Net). Table II shows the performance of each variant of T-Net on the SOTS dataset.

According to Table II, we can see that the performance is further improved every time a new component is added to the basic model, which justifies the overall design. The comparison among the four variants other than T-Net demonstrates the effectiveness of our skip connections with a new fusion strategy. Specifically, the RDB blocks used in the lateral connection instead of 1×1 convolution can extract more complex semantic information for dehazing, and features of different scales have different importances in feature fusion. Furthermore, the addition of the dual attention module has the greatest effect on the dehazing performance improvement, which effectively enhances the robustness and generalization ability of the network. This ablation study demonstrates the contribution of each component in our T-Net.

The third ablation study is conducted to evaluate our proposed Stack T-Net with different recursive stage number \( K \) and demonstrates the effectiveness of our recursive strategy. Limited by physical memory, we set \( K = 1, 2, 3 \). Table III shows the performance on the SOTS dataset of Stack T-Net with different recursive stage numbers.

As shown in Table III, Stack T-Net usually achieves higher average PSNR and SSIM values as the recursive stage number increases. However, the SSIM value of Stack T-Net with three stages is slightly lower than the value of Stack T-Net with two stages, because the best model is saved according to the highest PSNR value instead of the SSIM value. Actually, the highest SSIM value of Stack T-Net with three stages is higher than that of Stack T-Net with two stages in the experiment. This ablation study demonstrates that our recursive strategy is effective.

### D. Performance Comparison on the Synthetic Dataset

The proposed method is tested on the same synthetic dataset SOTS to qualitatively and quantitatively compare with the state-of-the-art methods, including DCP [5], NLD [8], MSCNN [16], DehazeNet [15], AOD-Net [18], GFN [71], DCPDN [19], EPDN [20] and DA_dehaze [70]. Apart from DCP and NLD which are prior-based methods, the others are deep learning-based methods. Moreover, MSCNN, DehazeNet, AOD-Net and DCPDN are based on the atmosphere scattering model, and the other three methods are end-to-end networks. We test two models of our algorithm in the comparison experiment, namely T-Net and Stack T-Net with three stages.

Table IV presents the quantitative comparison results of different dehazed methods on the SOTS dataset in terms of the average PSNR and SSIM values. Except for the last two rows, the data in this table are all quoted from the work of Shao [70]. As shown in Table IV, our work outperforms the state-of-the-art methods by a wide margin. Not only Stack T-Net but also our plain T-Net outperform the state-of-the-art DA_dehaze [70] in terms of both the average PSNR and SSIM values on the SOTS dataset.

Fig. 6 shows the qualitative comparison results among different dehazing methods, where the synthetic images are chosen from the SOTS datasets. The haze densities and scenes of the shown images in Fig. 6 are different, of which four images are selected from the indoor subsets of SOTS and the rest are from the outdoor subset. We note that the first three methods [8], [15], [18] perform poorly on most images and easily
cause under-dehazing, color oversaturation and other distortion problems. The dehazing results of DCPDN [19] and EPDN [20] seem better, but DCPDN [19] tends to make the brighter areas in images overexposed and EPDN [20] tends to reduce the brightness of the darker areas in images, which both cause the loss of many details in the images (see, e.g., the first and third rows in Fig. 6(e) and (f)). DA_dehaze [70] performs well on most images, but tends to cause color distortions (see, e.g., the first, second, fifth and sixth rows in Fig. 6(g), of which the color is slightly different from the ground truth images) and sometimes causes halo artifacts (see, e.g., the sky area of the third row in Fig. 6(g)), where other methods except our methods have the same problem.

Compared with the state-of-the-art methods, our T-Net and Stack T-Net have the best performance in terms of haze removal and are effective in suppressing halo artifacts and color distortions. The dehazed images are visually most similar to their ground truth images (see, e.g., Fig. 6(h) and (i)). In addition, we note that Stack T-Net can further eliminate hazy areas on the basis of T-Net (see, e.g., the first row in Fig. 6(h), in which the haze in the middle is eliminated in the same row of (i)), demonstrating that the recursive strategy is helpful for image dehazing.

E. Performance Comparison on Real-World Hazy Images

We further compare our methods with the same state-of-the-art approaches on real-world images to evaluate the generalization ability of our methods in real-world scenarios, where the real-world images used in the experiment are chosen from the URHI dataset. In addition, we calculate the average BRISQUE and PIQUE values of different methods on the URHI dataset to evaluate the image quality of the dehazed images. Since the clean ground truth images to the real-world hazy images are not available, we focus mainly on the visual comparison of different approaches and regard the values of no-reference evaluation metrics as the secondary support for our algorithms to outperform other methods. The qualitative and quantitative comparisons are separately shown in Fig. 7 and Table V.

According to Fig. 7, we note that the image distortions on real-world data are primarily the same as those on synthetic data, and some problems are even more obvious on real-world data. The dehazed images of NLD suffer from severe color distortions and halo artifacts (see, e.g., Fig. 7(b)), as well as other problems such as overexposure (see, e.g., the second and fifth rows in Fig. 7(b)). DehazeNet and AOD-Net tend to under-dehaze images (see, e.g., the second, third, fifth and last rows in Fig. 7(c), (d)). DCPDN
Fig. 7. Visual comparisons on the URHI dataset. We show the visual results of different methods on real-world data. The first column shows the hazy images, and the other columns represent the dehazed results of different methods. The results of our proposed T-Net and Stack T-Net are separately shown in the last two columns.

TABLE V
QUANTITATIVE COMPARISON ON THE URHI DATASET BETWEEN THE STATE-OF-THE-ART DEHAZING METHODS

| Methods     | URHI  |
|-------------|-------|
|             | BRISQUE ↓ | PIQUE ↓ |
| Origin      | 32.73 | 17.77 |
| DCP [5]     | 31.12 | 18.93 |
| NLD [8]     | 31.87 | 17.64 |
| MSCNN [16]  | 30.49 | 18.30 |
| DehazeNet [15] | 30.86 | 18.30 |
| AOD-Net [18] | 42.84 | 19.67 |
| DCPDN [19]  | 45.98 | 35.02 |
| EPDN [20]   | 30.23 | 19.93 |
| DA_dehaze [70] | 37.46 | 30.20 |
| T-Net       | 22.76 | 12.38 |
| Stack T-Net | **21.98** | **12.04** |

achieves better dehazing performance but tends to cause light halo artifacts (see, e.g., the sky area of the first, second, third, fourth and sixth rows in Fig. 7(e)) and overexposure (see, e.g., the fifth, seventh rows in Fig. 7(e)). The direct-mapping methods EPDN and DA_dehaze seem to remove more haze but cause damage to the depth structure of the dehazed images, which is reflected in the dehazed results as a large area of halo artifacts (see, e.g., the sky area of the first, second, third, fourth and sixth rows, the edge of the airplane of the third row, and the edge of the people of the fifth row in Fig. 7(f) and (g)) and color distortions (see, e.g., the first, second, fourth, fifth and sixth rows Fig. 7(f), (g)).

Compared with the state-of-the-art methods, our methods have the best visual effect on real-world hazy images. As shown in Fig. 7, in addition to effective haze removal, T-Net and Stack T-Net can also restore the depth structure well and suppress halo artifacts and color distortions. There are almost no artifacts or distortions in our dehazed images, of which the sky area and the object edges are clean and smooth (see, e.g., the first four rows in Fig. 7(h) and (i)). Moreover, our methods can restore the color of the image better than other methods (see, e.g., the first, fourth and sixth rows in Fig. 7, where the color of our methods is most similar to the hazy image). Although T-Net and Stack T-Net both have good performance on real-world images, Stack T-Net can further improve the quality of the images (see, e.g., the fifth row in Fig. 7(h) and (i), where the image of (i) has higher brightness and looks cleaner than the same image of (g)). In particular, even for images with severe haze and deep depth (see, e.g., the second row in Fig. 7), our methods can remove a certain amount of haze while maintaining the authenticity of the images, without causing image distortion due to halo artifacts, as in other methods.

The average BRISQUE and PIQUE values in Table V further quantify the performance of these methods on real-world data, where the first row shows the initial values of the URHI dataset. In this table, down arrow indicates that the lower the
value, the better the performance is. It is obvious that our Stack T-Net has the best performance in terms of the two no-reference metrics, followed by T-Net. Comparing this table with Table IV, we observe that some methods perform significantly differently on real-world and synthetic data, such as DA_dehaze [70]. However, this is basically consistent with the visual performance of these methods shown in Fig. 7. BRISQUE tends to quantify the statistical distribution difference between the test image and real-world natural scenes, while PIQUE focuses on the degree of distortions on local image patches. From the above analysis, we know that DA_dehaze is prone to over-dehazing, resulting in the destruction of the depth structure and the introduction of additional noise. Therefore, the distribution of the dehazed images of DA_dehaze has a large difference from real-world natural images, and the distortions of local patches are also obvious. This is the reason why its BRISQUE and PIQUE values are both higher than the initial values. Moreover, most dehazing methods can effectively reduce the BRISQUE value but have difficulty in reducing the PIQUE value, since they are unable to address distortions (like halo artifacts) on local patches. However, since our method considers the preservation of image depth information during the construction process, it can acquire clean dehazed images of high authenticity with little additional noise. These images are close to the statistical distribution of real natural images and have less distortion, which is confirmed by the visual effects of our method on URHI, as shown in Fig. 7. Overall, our methods achieve a good balance between dehazing and maintaining image authenticity.

F. Runtime Analysis

To evaluate the time complexity of our proposed method, we perform a runtime analysis to compare the time of dehazing one image from SOTS on average. As shown in Table VI, our subnetwork T-Net takes an average of 0.16 s to dehaze one image, which is close to the time taken by lightweight networks such as AOD-Net, EPDN and DehazeNet. However, the performance of T-Net is much better than these methods. Our proposed Stack T-Net takes an average of 0.12 s more per image than T-Net to dehaze, but the time it takes is only approximately half of the previous best method DA_dehaze. The study indicates that our methods are both effective and efficient.

| Methods          | Runtime(s) |
|------------------|------------|
| DCP [5]          | 0.08       |
| MSCNN [16]       | 0.11       |
| AOD-Net [18]     | 0.13       |
| T-Net            | 0.16       |
| EPDN [20]        | 0.17       |
| DehazeNet        | 0.18       |
| Stack T-Net      | 0.28       |
| DCPDN [19]       | 0.30       |
| DA_dehaze [70]   | 0.59       |
| NLD [8]          | 20.83      |

VI. Conclusion

In this work, we propose a new end-to-end dehazing network named T-Net, which is based on the U-Net architecture and contains a backbone module and a dual attention module. Inspired by the recursive strategy, we further propose Stack T-Net by repeatedly unfolding the plain T-Net. Through ablation studies, we verify that the overall design of T-Net is effective and that the recursive strategy is helpful for dehazing. Experimental results on both synthetic and real-world images demonstrate that our T-Net and Stack T-Net perform favorably against state-of-the-art dehazing algorithms, and that our Stack T-Net can further improve dehazing performance.

Real-world images under severe haze lose a lot of texture details and color information, which causes the performance degradation of state-of-the-art dehazing methods or makes the color and details of the dehazing results deviate from the original images. Our future work will focus on this problem, and further study the inherent latent features of invariance.

REFERENCES

[1] E. J. Mccartney, *Optics of the atmosphere: Scattering by molecules and particles* New York, NY, USA: John Wiley and Sons, Inc., 1976, Art. no. 421.
[2] S. G. Narasimhan and S. K. Nayar, “Vision and the atmosphere,” *Int. J. Comput. Vis.*, vol. 48, no. 3, pp. 233–254, 2002.
[3] Y. Li, S. You, M. S. Brown, and R. T. Tan, “Haze visibility enhancement: A survey and quantitative benchmarking,” *Comput. Vis. Image Understanding*, vol. 165, no. Dec., pp. 1–16, 2016.
[4] R. T. Tan, “Visibility in bad weather from a single image,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[5] K. He, J. Sun, and X. Tang, “Single image haze removal using dark channel prior,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.
[6] R. Fattal, “Dehazing using color-lines,” *ACM Trans. Graph.*, vol. 34, no. 1, Dec. 2015.
[7] Q. Zhu, J. Mai, and L. Shao, “A fast single image haze removal algorithm using color attenuation prior,” *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3522–3533, Nov. 2015.
[8] D. Berman, T. treibitz, and S. Avidan, “Non-local image dehazing,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.
[9] T. M. Bui and W. Kim, “Single image dehazing using color ellipsoid prior,” *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 999–1009, Feb. 2018.
[10] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.
[11] J. Kim, J. K. Lee, and K. M. Lee, “Deeply-recursive convolutional network for image super-resolution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.
[12] C. Ledig et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.
[13] Y. Tai, J. Yang, and X. Liu, “Image super-resolution via deep recursive residual network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3147–3155.
[14] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising,” *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.
[15] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, “Dehazenet: An end-to-end system for single image haze removal,” *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.
[16] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M. H. Yang, “Single image dehazing via multi-scale convolutional neural networks,” in *Proc. Eur. Conf. Comp. Vis.*, 2016, pp. 154–169.
[17] Y. Huang, Y. Wang, and Z. Su, “Single image dehazing via a joint deep modeling,” in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 2840–2844.
[18] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, “Aod-net: All-in-one dehazing network,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4780–4788.
I. Omer and M. Werman, “Color lines: Image specific color representation,” in *IEEE Trans. Image Process.*, vol. 24, no. 7, pp. 2808–2817, 2015.

Z. Li and J.-Y. Zhu, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

Z. Li and J. Zheng, “Edge-preserving decomposition-based single image haze removal,” *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5417–5431, Dec. 2015.

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

I. Omer and M. Werman, “Color lines: Image specific color representation,” in *Proc. IEEE Trans. Image Process. Conf. Comput. Soc. Comput. Vis. Pattern Recognit.*, 2004, vol. 2, pp. II–II.

W. Wang, X. Yuan, W. Yu, and Y. Liu, “Fast image dehazing method based on linear transformation,” *IEEE Trans. Multimedia*, vol. 19, pp. 1142–1155, 2017.

H.-M. Hu, H. Zhang, Z. Zhao, B. Li, and J. Zheng, “Adaptive single image dehazing using joint local-global illumination adjustment,” *IEEE Trans. Multimedia*, vol. 22, pp. 1485–1495, 2020.

F. Fang, T. Wang, Y. Yang, T. Zeng, and G. Zhang, “Variational single image dehazing for enhanced visualization,” *IEEE Trans. Multimedia*, vol. 22, pp. 2537–2550, 2020.

J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *Proc. Euro. Conf. Comput. Vis.*, Springer, 2016, pp. 694–711.

X. Yang et al., “DRFN: Deep recurrent fusion network for single-image super-resolution with large factors,” *IEEE Trans. Multimedia*, vol. 21, pp. 328–337, 2019.

K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu, and H. Li, “Adversarial spatio-temporal learning for video deblurring,” *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 291–310, Mar. 2018.

C.-H. Liang, Y.-A. Chen, Y.-C. Liu, and W. Hsa, “Raw image deblurring,” *IEEE Trans. Multimedia*, vol. 24, pp. 61–72, 2022.

K. Zhang et al., “Deblurring by realistic blurring,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2737–2746.

K. Zhang et al., “Beyond monocular deblurring: Stereo image deblurring via semantic understanding,” in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 71–89.

Y. Que, S. Li, and H. J. Lee, “Attentive composite residual network for robust rain removal from single images,” *IEEE Trans. Multimedia*, vol. 23, pp. 3059–3072, 2021.

K. Zhang et al., “Beyond monocular deblurring: Parallel stereo deblurring network via semantic prior,” Int. J. Comput. Vis., pp. 1–16, 2022.

Y.-F. Liu, D.-W. Jaw, S.-C. Huang, and J.-N. Hwang, “DesnoutNet: Context-aware deep network for snow removal,” *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3064–3073, Jun. 2018.

K. Zhang, R. Li, Y. Yu, W. Luo, and C. Li, “Deep dense multi-scale network for snow removal using semantic and depth priors,” *IEEE Trans. Image Process.*, vol. 30, pp. 7419–7431, 2021.

Y. Song, J. Li, X. Wang, and X. Chen, “Single image dehazing using ranking convolutional neural network,” *IEEE Trans. Multimedia*, vol. 20, pp. 1548–1560, 2018.

Y. Li et al., “Lap-net: Level-aware progressive network for image dehazing,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3276–3285.

C. Li, C. Guo, J. Guo, P. Han, H. Fu, and R. Cong, “Pdr-net: Perception-inspired single image dehazing network with refinement,” *IEEE Trans. Multimedia*, vol. 22, pp. 704–716, 2020.

J. Zhang and D. Tao, “Fated-net: A fast and accurate multi-scale end-to-end dehazing network,” *IEEE Trans. Image Process.*, vol. 29, pp. 72–84, 2020.

A. Dudahe and S. Murala, “Ryf-net: Deep fusion network for single image haze removal,” *IEEE Trans. Image Process.*, vol. 29, pp. 628–640, 2020.

A. Cherkas, D. Freedman, and M. Elad, “Unsupervised single image dehazing using dark channel prior loss,” *IEEE Trans. Image Process.*, vol. 29, pp. 2692–2701, 2020.

W. Chen, J. Ding, and S. Kuo, “PMS-net: Robust haze removal based on patch map for single images,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11 681–11 689.

P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proc. IEEE Trans. Pattern Recognit.*, 2017, pp. 1125–1134.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

N. Bharath Raj and N. Venketeswaran, “Single image haze removal using a generative adversarial network,” in *Proc. Int. Conf. Wireless Commun. Signal Process. Neww.*, 2020, pp. 37–42.

D. Engin, A. Genc, and H. K. Ekenel, “Cycle-dehaze: Enhanced cyclegan image dehazing,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1805–1809.

Y. Shao, L. Li, W. Ren, C. Gao, and N. Sang, “Domain adaptation for image dehazing,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2808–2817.

W. Ren et al., “Gated fusion network for single image dehazing,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3253–3261.

H. Dong et al., “Multi-scale boosted dehazing network with dense feature fusion,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2157–2167.
[73] J. Shin, H. Park, and J. Paik, “Region-based dehazing via dual-supervised triple-convolutional network,” IEEE Trans. Multimedia, vol. 24, pp. 245–260, 2022.

[74] O. Kapyn, T. Martyniuk, J. Wu, and Z. Wang, “Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 8878–8887.

[75] S. Nah, T. Hyun Kim, and K. Mu Lee, “Deep multi-scale convolutional neural network for dynamic scene deblurring,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 3883–3891.

[76] J. Kim, J. K. Lee, and K. M. Lee, “Deeply-recursive convolutional network for image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1637–1645.

[77] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha, “Recurrent squeeze-and-excitation context aggregation net for single image deraining,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 254–269.

[78] Y. Tai, J. Yang, and X. Liu, “Image super-resolution via deep recursive residual network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 3147–3155.

[79] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. 3rd Int. Conf. Learn. Representations, San Diego, CA, USA, May 7–9, 2015, pp. 1–14.

[80] O. Russakovsky et al., “Imagenet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015.

[81] B. Li et al., “Benchmarking single-image dehazing and beyond,” IEEE Trans. Image Process., vol. 28, no. 1, pp. 492–505, Jan. 2019.

[82] A. Paszke et al., “Automatic differentiation in pytorch,” in Proc. NIPS Autodiff Workshop: Future Gradient-based Mach. Learn. Softw. Techn., 2017, pp. 1–4.

[83] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. 3rd Int. Conf. Learn. Representations, San Diego, CA, USA, May 7–9, 2015, pp. 1–14.

[84] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” IEEE Trans. Image Process., vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[85] V. N., P. D., M. C. Bh, S. S. Channappayya, and S. S. Medasani, “Blind image quality evaluation using perception based features,” in Proc. 21st Nat. Conf. Commun., 2015, pp. 1–6.

Lirong Zheng received the B.E. degree in 2019 from the School of Electronic and Information Engineering, Shenzhen University, Shenzhen, China, where she is currently working toward the Ph.D. degree. She is also Member of the ATR National Key Laboratory of Defense Technology, Shenzhen University. Her research interests include intelligent information processing, video processing, and pattern recognition.

Yanshan Li received the M.Sc. degree from the Zhejiang University of Technology, Hangzhou, China, in 2005, and the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 2015. He is currently an Associate Professor with the ATR National Key Laboratory of Defense Technology, Shenzhen University, Shenzhen, China. His research interests cover computer vision, machine learning, and image analysis.

Kaihao Zhang received the Ph.D. degree from the College of Engineering and Computer Science, The Australian National University, Canberra, ACT, Australia. His research interests include computer vision and deep learning. He has more than 30 refereed publications in international conferences and journals, including CVPR, ICCV, ECCV, NeurIPS, AAAI, ACMMM, TPAMI, IJCV, TIP, TMM, etc.

Wenhan Luo (Member, IEEE) received the B.E. degree from the Huazhong University of Science and Technology, China, in 2009, the M.E. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, China, in 2012, and the Ph.D. degree from Imperial College London, London, U.K., in 2016. He is currently an Associate Professor with Sun Yat-sen University, Guangzhou, China. Before being a Faculty Member with University, he was an applied Research Scientist with Tencent, Shenzhen, China, solving real-world problems using computer vision and machine learning techniques. Prior to Tencent, he was with Amazon (A9), Palo Alto, CA, USA, where he developed deep models for better visual search experience. Before that, he was a Research Scientist with Tencent AI Lab. He has authored or coauthored more than 60 peer-reviewed papers, including more than 40 of them published in top-tier conferences and journals, like ICML, CVPR, ICCV, ECCV, AAAI, ACL, ACMMM, ICLR, TPAMI, AI, IJCV, TIP, and two of them are ESI highly cited papers. His research interests include several topics in computer vision and machine learning, such as motion analysis (especially object tracking), image/video quality restoration, image/video synthesis, reinforcement learning. His work was the recipient of the CVPR 2019 Best Paper Nominee. He was also the recipient of the 2022 ACM China Rising Star Award (Guangzhou Chapter).