Multimedia Tools and Applications (2024) 83:11367–11385
https://doi.org/10.1007/s11042-023-15844-6

PSPAN:pyramid spatially weighted pixel attention network for image dehazing

YuBo Zhang[1] · **Tongxiang Xu**[1] · **Kang Tian**[1]

Received: 6 November 2022 / Revised: 1 March 2023 / Accepted: 15 May 2023 /
Published online: 28 June 2023

Abstract
Haze-free images are the prerequisites for many high-level visual tasks, and thus image dehazing has become an active topic in computer vision. However, the existing image dehazing algorithms are limited in face of unevenly distributed haze and dense haze in some scenes. In this paper, we propose a Pyramid Spatially Weighted Pixel Attention Network (PSPAN) for single image dehazing by leveraging complementarity among different levels of features in a pyramid manner with unique attention methods. The proposed PSPAN utilizes the feature pyramid as the core network and consists of three modules: an efficient Multi-scale Feature Extraction Attention module, a pyramid Spatially Weighted Pixel Attention module, and an image reconstruction module. Specifically, PSPAN preprocesses hazy images first before acquiring abundant shared features. After that, these features are sent to different branches. To effectively fuse useful information from these different branches and obtain better-dehazed results, we propose an efficient feature aggregation attention module. Finally, the image reconstruction module is used to restore clear images. Meanwhile, a loss function that combines a mean square error loss part, an edge loss part, and a perceptual loss part is employed in PSPAN which can better preserve image details. Experimental results demonstrate that the proposed PSPAN achieves superior performance to other existing state-of-the-art algorithms in terms of accuracy and visual effect.

Keywords Image dehazing · The feature pyramid · Multi feature extraction · Spatially weighted pixel attention

✉ Tongxiang Xu
xutongxiang521@gmail.com

YuBo Zhang
zhangyubo@nepu.edu.cn

Kang Tian
tiankangg@163.com

[1] School of Electrical Information Engineering, Northeast Petroleum University, 163319 Daqing, China

1 Introduction

Haze, fog or smoke usually affect visibility and obscure key information of the images. To deal with this issue, image dehazing has been widely studied in recent years which aims to recover clear images from their corresponding hazy images. The whole procedure can be formulated as:

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

Where $I(x)$ denotes the hazy image and $J(x)$ denotes the clear image, x denotes a pixel position in the image, A denotes the global atmospheric light, and $t(x)$ denotes the transmission map. In addition, the transmission map can be represented as $t(x) = e^{-\beta d(x)}$, where β and $d(x)$ represent individually the atmosphere scattering parameter and the scene depth.

Previous image dehazing approaches have focused more on restoring the clear image using priors such as dark-channel prior, contrast color lines, and haze-line prior. For example, He et al. [13]proposed a dark channel prior (DCP) based method for estimating the transmission map. Kansal et al. proposed a novel approach of image subsampling [17], which is used to construct the dark channel to improve the computational efficiency. Although these prior-based methods have achieved considerable success, their performances are limited because not all the images of real scenes are compatible with the predefined priors. Recently, deep learning has demonstrated effectiveness in various computer vision tasks. Various convolutional neural network (CNN) based methods have been proposed to estimate the transmission map and the atmospheric light. Once the transmission map and the atmospheric light are estimated, the dehazed image is restored through the atmosphere scattering model. Generally speaking, low-level features in a neural network refer partly to detailed information, and high-level features contain more semantic information. Both of them are critical for recovering a clear image, but most CNN-based methods usually use high-level features to achieve image dehazing. Moreover, these methods are based on the atmosphere scattering model. If the estimated transmission map and atmospheric light are not accurate, then the dehazed result will be of low quality.

Although the existing end-to-end dehazing algorithm has a better defogging effect than the one based on the physical model, it is easy to ignore the aggregation of multi-scale spatial information, resulting in the loss of image details, so there is still a problem of unsatisfactory dehazing effect. In addition, common attention mechanisms tend to adjust weights relying on a single dimension of information. To solve the above problems, we propose a novel end-to-end framework called the Pyramid Spatially Weighted Pixel Attention Network (PSPAN).

In this work, we propose a novel end-to-end framework called Pyramid Spatially Weighted Pixel Attention Network (PSPAN) for single image dehazing, which leverages complementarity among different level features in a pyramid manner with a unique attention mechanism. Specifically, PSPAN consists of three modules: a three-scale feature extraction attention module, a pyramid spatially weighted pixel attention module, and an image reconstruction module. To begin with, the three-scale feature extraction attention module extracts features at three different scales. At the same time, it integrates the efficient channel attention mechanism, which can expand the receptive field and extract different scale features through weighted screening for fusion. After that, these features are fed into the SWPAB block. The module extracts more significant attention features from the spatially weighted pixel attention blocks and then fuses these attention features into different levels. Finally, the image reconstruction module is used to restore a clear image based on the output of SWPAB. In addition, we introduce a training loss function consisting of three terms: the MSE loss, the Edge loss and

the perceptual loss. The MSE loss is utilized to measure the pixel-wise distance, while the Edge loss promotes generating a clean image with more details and the perceptual loss uses the pretrained model to extract the advanced perceptual features of the image in order to further repair the details. As shown in Fig. 1, the proposed PSPAN produces a more realistic image with more details.

The main contributions of the proposed image dehazing method are summarized as follows:

- We propose a novel end-to-end framework called Pyramid Spatially Weighted Pixel Attention Network (PSPAN) for single image dehazing, which can extract more informative features by the special attention block, and fuse the complementary features at different levels in a pyramid manner.
- The new proposed attention block (SWPAB) not only solves the problem that previous dehazing networks are difficult to focus on multi-dimension of information but also pays more attention to the feathers for dense hazy regions reconstruction.
- A loss function that combines a mean square error loss part, an edge loss part and a perceptual loss part is employed in PSPAN, which can better preserve image details.
- Extensive experiments on standard benchmark datasets demonstrate that the proposed PSPAN is better than the majority of existing methods.

2 Related work

2.1 Image dehazing

Previous image dehazing methods can be divided into prior-based methods and learning-based methods.

Fig. 1 Examples of image dehazing results. Top left: input hazy image. Top right and bottom left: restored haze-free images using DCP and AODNet respectively. Bottom right: dehazed image generated by the proposed method

Prior-based methods recover clear images by using prior statistics, such as the albedo of the scene in [8]. In the past few years, researchers have explored different priors for image dehazing [9, 13, 19, 35, 45]. Specifically, based on the observation that clear images have higher contrast than hazy images, Tan et al. [35] enhanced the visibility of hazy images by maximizing local contrast. He et al. [13] proposed dark channel prior (DCP) that the intensity of pixels in haze-free patches is very low in at least one color channel to achieve image dehazing. Furthermore, based on a general observation that small image patches typically exhibit a one-dimensional distribution in the RGB color space, Fattal [9] proposed an approach to recover the scene transmission using color lines. Zhu et al. [45] proposed color attenuation before obtaining the scene depth from the hazy image through supervised learning. To quickly and accurately estimate the transmission map, a sub-sampling based local minimum operation and fast gradient domain guided image filtering (GDGF) is applied on initial depth map [19].All the above methods heavily rely on hypothetical priors. However, those priors tend to lose effectiveness in complex scenes, leading to a performance drop.

As opposed to the above methods, learning-based methods utilize convolutional neural networks to recover clear images from hazy images directly. These methods can be further divided into two sub-categories: physical-model-based methods and end-to-end methods. Due to the fact that prior-based methods are sensitive to changes in the environment, some physical-model-based methods utilize the feature extraction capabilities of CNNs to estimate various components of atmospheric scattering models. As an example, [25] used CNNs to estimate atmospheric light, [1] estimated transmission, and [22, 41] estimated both transmission and atmospheric light to identify haze-affected regions. And recently, end-to-end methods have shown a considerable improvement in performance for recovering areas affected by the haze in comparison with the above traditional methods. [21] proposed a lightweight network called AODNet, which can output images directly and is a real high-quality network. It was suggested by [31] to utilize an encoder-decoder formulation (GFN) to encode features from the hazy images, which are then extrapolated by a decoder to reconstruct the haze-free images. Mei et al. [26] described a Progressive Feature Fusion Network (PFFNet) that directly learns the nonlinear transformation function from observed hazy images to haze-free ones. The Enhanced Pix2Pix Dehazing Network (EPDN) [29] attempts to improve the dehazing performance by following the dehazing network with an enhancer. Dong et al. [6] proposed the Multi-Scale Boosted Dehazing Network (MSBDN), which incorporates the boosting strategy and the back-projection technique for image dehazing. In order to generate more visually pleasing dehazed images, [7] proposed a fusion of frequency priors with the image in an adversarial learning framework. And for the sake of better dehazing performance, [39] constructed a contrastive learning-driven autoencoder-like framework called AECRNet) based on the negative information.

2.2 Attention block

Usually, humans selectively pay attention to the targeted area with more useful information to obtain more detailed intelligence while suppressing other useless information. The attention mechanism in deep learning is similar to the selective visual attention mechanism in humans, and its purpose is to select and prioritize information more critical to the task goal. In recent years, the attention mechanism has been introduced into deep learning algorithms to handle a variety of computer vision tasks, including: [2, 15, 42, 44], and [24]. Mnih et al. first proposed the concept of the attention mechanism [27] and believed that it highlights the influence of a key input on the output by calculating the weight of the input data. According

to the relationship between feature channels, Hu et al. [15] propose a novel architectural element (SE) to establish inter-channel interdepend dencies, so as to adaptively reply to channel-level feature responses. By utilizing the channel attention mechanism to enhance the representational ability of a very deep residual network, [42] is able to adaptively extract informative, high-frequency, channel-attention features in the image. [10] propose the Dual Attention Network (DANet) based on the self-attention mechanism for the scene segmentation task. The proposed position attention module is designed to selectively learn the spatial interdependencies of features, while the channel attention module is utilized to emphasize channel interdependencies. Thus, precise segmentation results can be achieved with the two attention modules. Liu et al. [24] propose GridDehazeNet by integrating multi-scale estimation with the attention mechanism. As well as alleviating the bottleneck issue that occurs in some multi-scale networks, channel-wise attention is utilized to reconstruct features of diverse scales. Qin et al. [28] proposed the Attention-based Feature Fusion (FFA) structure which consists of two attention modules for dealing with feature information from channel and pixel spaces. Several of the above methods have demonstrated that attention mechanisms play a significant and powerful role in the image processing.

3 Method

In this paper, we propose a novel PSPAN network that combines the benefits of the attention mechanism and pyramid operations for image dehazing. Next we will first introduce the specific details of the proposed multi feature extract block (MFEB) and spatially weighted pixel attention block (SWPAB). Afterwards, we will describe the objective function used by the proposed network.

3.1 Overall framework

Figure 2 shows the overall architecture of the proposed network. The overall network consists of three modules, namely the multi-scale feature extraction attention module (denoted by the dotted red line), the pyramid spatially weighted pixel attention module (denoted by light green square in the dotted yellow line), and the image reconstruction module. At the beginning, we will pass the hazy image to be processed through a three-layer pyramid structure, and each pyramid block (MFEB) will transmit information of different scales to the next layer: the feature attention processing block (SWPAB). At this stage, the proposed attention block will process information according to different weights and then conduct upsampling to complete further fusion. This makes it possible to capture more crucial and informative features to predict better-dehazed results. At last, the feature information processed by the multi-layer attention block is then processed by the convolution recovery module and finally added to the original image to obtain the final output.

3.2 Multi feature extraction block

In order to get the features of different scales better, this paper designs a three-scale feature extraction attention module. In this module, three different scale convolutions are utilized to extract different information about the receptive field from the feature map to obtain feature maps of different scales. The extraction module of each scale is composed of a 3×3 convolution layer and an RREB (two Resblocks with ECA) layer. And the RREB layer

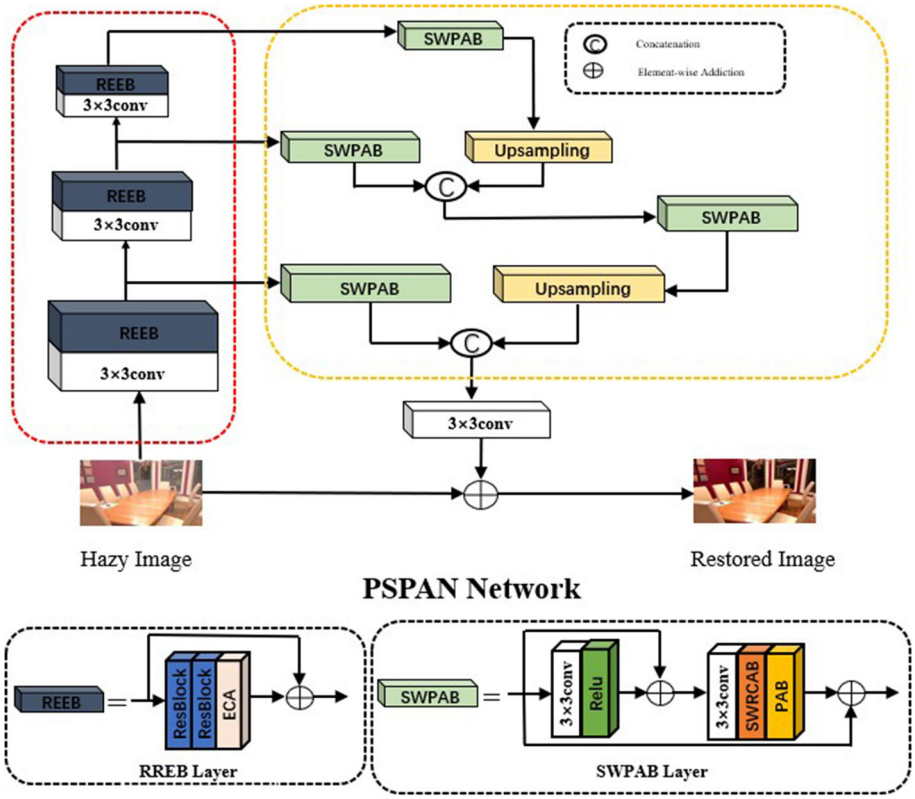

Fig. 2 Overall architecture of PSPAN: (1) Extract multi-scale features using the proposed three-scale feature extraction module. Every feature extraction stage of the module consists of two components, namely a 3×3 convolution layer and an RREB layer; (2) The three-scale features generated by the feature extraction module are then fed into the proposed pyramid feature attention module. Three attention blocks are used to process the features at different scales in a top-down pyramid fashion; (3) The image reconstruction module, including a convolution operation and a simple element-wise addition operation, is adopted to restore the dehazed single image

contains two ResBlocks [12] and an ECAblock, its overall block is presented in Fig. 3 (in the green line) and they collectively form a new residual network. In the first MFEB layer, the depth (the number of channels) of feature maps is increased to 32 and the following two layers increase the depth of the feature maps to 64 and 128 while reducing the resolution of the feature maps by half, respectively. Unlike previous works that only used the output features of the third stage, all the outputs of the three stages are fed into the pyramid feature attention module.

Inspired by the RRB (Residual Block with SE) module [5], we designed the residual network called RREB, which utilizes the efficient channel attention mechanism (ECA). Considering that skip connections can provide long-range information compensation and enable residual learning, we combine the ECA operation with residual blocks in the dehaze residual network. Spatial contextual information has been shown to be effective in single image dehazing. Nevertheless, the different feature channels in the same layer are independent of one another, and had little correlation during the previous convolution operation. In light of the fact that ECA can model a correlation between different feature channels, we can

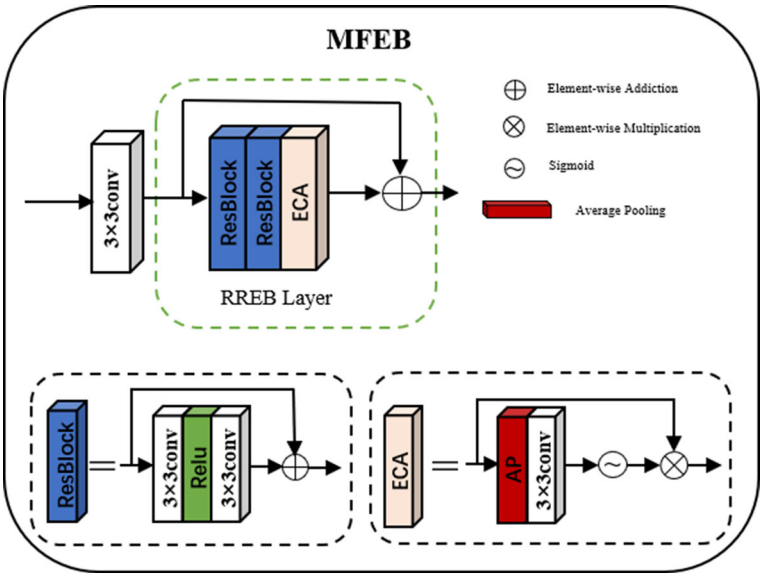

Fig. 3 Detailed structure of multi feature extraction block

intensify the feature channel that has more context information by giving it a larger weight. Conversely, feature channels that have less spatial contextual information will just receive a small weight. As shown in the blue square in Fig.3, ResBlock adopts the jump connection method to improve the learning ability of the network.To further extract features of the current scale, conventional convolution CB_i is followed by two ResBlocks (RB_{1i} and RB_{2i}) in the RREB layer to ensure the integrity of current scale feature extraction. The overall MFEAB module is expressed as:

$$MFEAB_i(x) = ECA_i(RB_{2i}(RB_{1i}(CB_i(F_{i-1}(x))))) + CB_i(F_{i-1}(x)) \quad (2)$$

Where $CB_i(\cdot)$ denotes the convolution function, $ECA_i(x)$ denotes the ECAblock operation, and $F_{i-1}(x)$ represents the currently entered feature.

The efficient channel attention mechanism [38] is illustrated by the light pink square in Fig.3. Firstly, we will carry out global average pooling without dimension reduction; Secondly, the kernel size of the convolution layer will be determined adaptively to facilitate cross-channel information interaction; Then we will use the sigmoid function to determine the weight value of the feature map; Finally, the weight value of the feature map will be used to adjust the input feature map and output the weighted feature map. In each feature extraction module (MFEAB), an efficient channel attention mechanism is used to filter the salient features of the current scale by weighting instead of the original features, which improves the efficiency and performance of the network. The efficient channel attention mechanism $ECA_i(x)$ is expressed as:

$$ECA_i(x) = \delta(Convk_i(g_i(F_{i-1}(x)))) \otimes F_{i-1}(x) \quad (3)$$

Where $g_i(\cdot)$ denotes global average pooling function; $Convk_i(\cdot)$ represents convolution functions with kernel size $k \times k$; δ is the sigmoid function.

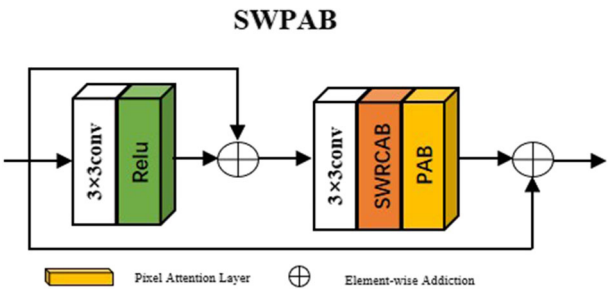

Fig. 4 The architecture of SWPAB

3.3 Spatially weighted pixel attention block

Considering that the attention mechanism [36, 37, 40] has been widely incorporated into the design of neural networks, it has played a significant role in the performance of networks. Inspired by the work [28], we further design a novel feature attention (SWPAB) module. The SWPAB module combines spatially weighted residual channel attention (SWRCA) and pixel attention (PA) into channel-wise and pixel-wise features, respectively. And then we use a new structure to link the two features. As SWPAB treats different features and pixels unequally, it can provide additional flexibility in dealing with different types of information. In other words, to ensure that the network captures more informative features, the new attention block called spatially weighted pixel attention block is employed to explore the interdependencies among features in channels, spatial and pixel.

As is shown in Fig. 4, we adopt the idea of skip connection and the attention mechanism and design a basic block consisting of multiple local residual learning skip connections and feature attention. For one thing, the local residual learning allows the information of the thin haze region and low-frequency information to be bypassed through multiple local residual learning, making the main network learn more useful information. And spatially weighted residual channel attention and pixel attention further improve the capability of SWPAB. In this structure, shallow information can be retained and passed on to deeper layers. Most importantly, the SWPAB gives different weights to different level features before feeding all features to the feature fusion module, the weight is obtained by adaptive learning of this module. The SWPAB module can be described as:

$$\begin{aligned} SWPAB_i\left(x\right) = & PA_i\left(SWRCAB_i\left(CB_i\left(F_{i-1}\left(x\right)+\sigma\left(CB_i\left(F_{i-1}\left(x\right)\right)\right)\right)\right)\right) \\ & + F_{i-1}\left(x\right) \end{aligned} \tag{4}$$

Where $SWRCAB_i\left(x\right)$ and $PA_i\left(x\right)$ represent the SWRCAB operation and PA operation respectively; σ denotes relu function.

Squeeze and excitation residual blocks (SEResBlock) [15] have been widely used as a common residual network. However, SEResBlock employs a global average pool operation to learn the weight of each channel that equally aggregates all input features, ignoring the inconsistent concentration of haze. As a way to pay more attention to seriously degraded regions and informative channels, the Spatially Weighted Residual Channel Attention Block (SWRCAB) [14] was presented to focus more attention on content-aware channel level contact. As is depicted in Fig. 5, SWRCAB first learns spatial weights of input features through a convolutional layer followed by a sigmoid layer; then it obtains the spatial weights via element-wise multiplication; and finally, it gets each channel's attention by applying a

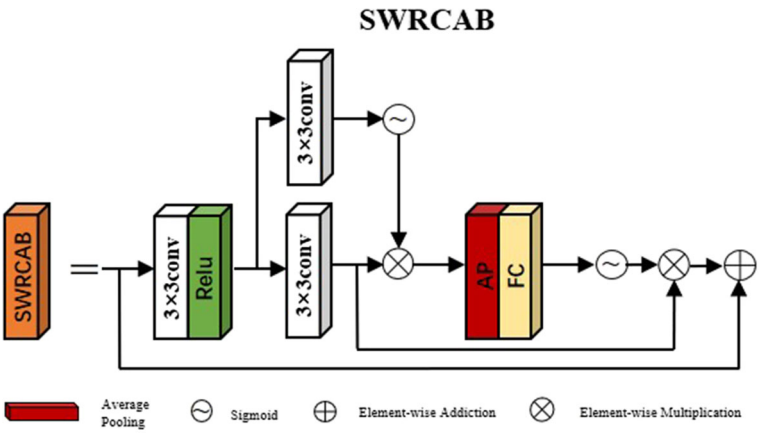

Fig. 5 The structure of spatially weighted residual channel attention block

global average pooling layer which is followed by a linear transformer layer and a sigmoid activation layer.

3.4 Loss function

To optimize the proposed network, three loss functions are utilized, namely the MSE loss $\mathscr{L}mse$, the Edge loss $\mathscr{L}edge$, and the Perceptual loss $\mathscr{L}per$.

MSE loss To measure the differences between the clear image and the output dehazed image on a pixel-wise basis, Mean Square Error (MSE) is used. The MSE loss can be defined as follows:

$$\mathscr{L}_{mse} = \frac{1}{C\ W\ H} \sum_{c=1}^{C} \sum_{i=1}^{W} \sum_{j=1}^{H} \left(I_{c,i,j}^{\text{clear}} - \tilde{I}_{c,i,j}^{\text{dehazed}} \right)^2 \tag{5}$$

Where C , W, and H represent the channel number, width, and height of an image, respectively. $I_{c.i,j}^{clear}$ is the value of ground truth at the location (i, j) of the channel c, and $\tilde{I}_{c,i,j}^{dehazed}$ corresponds to the value of the dehazed image generated by PSPAN.

Edge loss The Edge loss function is introduced to the network in order to recover a clear image with more detail. First, the convolution operation $Conv$ with Laplace operator [11] is used to obtain the edge images of the clear and dehazed images. Then, the $Tahn$ activation function is used to map the values of edge images to [0, 1]. Finally, the pixel-wise distance (L_1 Norm) is used to measure the differences between clear and dehazed edge images. The Edge loss function is given by:

$$\begin{aligned} \mathscr{L}_{edge} = & || Tahn \left(Conv \left(I^{clear}, k_{laplace} \right) \right) \\ & - Tahn \left(Conv \left(\tilde{I}^{dehazed}, k_{laplace} \right) \right) ||_1. \end{aligned} \tag{6}$$

Perceptual loss The concept of perceptual loss has been widely applied to image reconstruction since it was first proposed. By measuring the gap between the high-level feature representations extracted from a pre-trained deep neural network, the perceptual loss can

calculate the visual difference between the dehazed image and the ground truth quantifiably. To fully extract the potential information from high-level feature space, we apply a perceptual loss based on the VGG-16 network [34] to construct fine details in this work. In particular, this VGG-16 network is pre-trained on ImageNet. Perceptual loss is described as:

$$\mathscr{L}_{per} = \frac{1}{3}\sum_{r}\frac{|\phi_r(J) - \phi_r(\widetilde{J})|}{N_k} \tag{7}$$

Where ϕ_r denotes the output of $r-th$ layer in VGG-16. In this work, we set $r \in \{relu1\_2, relu2\_2, relu3\_3\}$. N_r represents the output size of the $r-th$ layer. Different from the traditional perceptual loss [16], we replace L_2 loss with L_1 loss for better dehazing effect.

Total loss For further exploring the performance of the proposed method, the total loss of is a multi-faceted loss function in the training stage which can think about the pixel-level similarity, the edge similarity and the perceptual similarity. And it is given by:

$$\mathscr{L} = \mathscr{L}_{mse} + \alpha \cdot \mathscr{L}_{edge} + \beta \cdot \mathscr{L}_{per} \tag{8}$$

Where α, β is a hyper-parameter that is used to yield the final loss . In this work, α is set to 0.01 and β is set to 0.01.

4 Experiments

In this section, extensive experiments are conducted on both synthetic and real-world datasets to demonstrate the effectiveness of the proposed network. We evaluate the performance of the proposed approach (PSPAN) by comparing its dehazing results quantitatively and subjectively with those of DCP [13], AODNet [21], DehazeNet [1], DCPDN [43], MSCNN [30] , MSBDN [6], EPDN [29], GirdDehazeNet [24], GCANet [3], GFN [31], DRN [14], FD-GAN [7] and [4, 18, 20], etc. The implementation codes provided by the respective authors of the above state-of-the-art approaches are used and the best results in each of the following quantitative comparisons are highlighted in bold. In addition, two ablation studies are conducted to verify the effectiveness of the used loss and the new proposed module.

4.1 Experiments setup

Dataset We adopt the RESIDE dataset to train and test the proposed method, which is a large-scale synthetic hazy image dataset proposed in [23]. RESIDE is divided into five different subsets: Indoor Training Set (ITS), Outdoor Training Set (OTS), Synthetic Objective Testing Set (SOTS), Real-World Task-Driven Testing Set (RTTS), and Hybrid Subjective Testing Set (HSTS). ITS, OTS, and SOTS are synthetic datasets, images in RTTS are from real scenes, and HSTS contains both synthetic and real-world images. The training set of RESIDE contains 13,990 hazy images which are synthesized using 1,399 clear images from the NYU Depth Dataset V2 [33] and the Middlebury stereo [32]. The testing set, named Synthetic Objective Testing Set (SOTS), selects 500 indoor images and 500 outdoor ones from the NYU Depth Dataset V2 to synthesize hazy images. Here we name them RESIDE-Indoor and RESIDE-Outdoor, respectively. In this work, ITS and SOTS are used as training set and testing set, respectively. At the same time, in order to test the dehazing effect on the real hazy images, we use RTTS (Unannotated Real Hazy Images) as the test dataset.

Aside from the RESIDE dataset, we also use the LIVE Image Defogging Database from Choi's databaset [4] as a supplement to verify the generality of the dehazing effect.

Implementation details The PSPAN-Net is implemented by PyTorch 1.5.0 with one NVIDIA GTX1080TI GPU. The models are trained using Adam optimizer with a batch size of 1 is adopted, where the exponential decay rates $\beta 1$ and $\beta 2$ equal to 0.5 and 0.999, respectively. The initial learning rate is set to 0.0001 and drops to 60% of the original uniformly every twenty epochs.

Evaluation metric In this paper, PSNR, SSIM and LPIPS are selected as indicators for evaluating synthetic image datasets. PSNR, also known as peak signal-to-noise ratio, is based on the error between corresponding pixels, that is, based on error-sensitive image quality evaluation. The larger the value, the smaller the image distortion. SSIM, also known as structural similarity, is a measure of the similarity between two images. Its value range is [0, 1], and the closer the value is to 1, the more similar the images are. Learned perceptual image patch similarity (LPIPS) is also used to measure the difference between two images and is more in line with human perception than above traditional methods. The lower the value of LPIPS, the more similar the two images are.

In order to evaluate and compare the proposed model with previous methods from a more comprehensive perspective, except for the above two most commonly used reference subjective evaluation metrics, we also selected two additional evaluation metrics: natural image quality evaluator (NIQE) and color naturalness index (CNI). The design idea of NIQE is to construct a series of features to measure image quality and use these features to fit a multivariate Gaussian model. These features are extracted from some simple and highly regular natural landscapes. The smaller the value of NIQE, the more the characteristics of the image conform to the natural image with high rules, which means that its quality is better. The CNI is a measure of whether an image scene is real and natural based on human vision. The value ranges from 0 to 1, and the closer the CNI is to 1, the more natural the image is. In this paper, these two metrics are tested on realistic dataset.

4.2 Comparison with state of the art

4.2.1 Results on synthetic dataset

Synthetic datasets provide access to extremely diverse characteristics such as scene setting, differing camera properties and illumination conditions, which are covered in large amounts of paired datasets, making them indispensable. We compare the proposed method with previous state-of-the-art image dehazing methods both quantitatively and qualitatively. In this process, we carry out these experiments on two datasets: RESIDE-Indoor and RESIDE-Outdoor.

Quantitative evaluation Table 1 shows the quantitative comparisons of different methods on the RESIDE-Indoor and RESIDE-Outdoor datasets, in which the digital values are the results from the SOTS database in terms of average PSNR and SSIM. Higher values of PSNR and SSIM represent better performance. As shown in Table 1, the proposed method achieves the second best performance with 33.91 dB PSNR but the best performance with 0.99 SSIM meanwhile compared with the other methods on the Indoor dataset. Although KDDN achieves the best performance in PSNR on RESIDE-Indoor, we perform better in SSIM than it. Meanwhile, we achieve the best PSNR and SSIM on RESIDE-Outdoor. Lower

Table 1 PSNR↑ and SSIM↑ comparisons for different methods on the RESIDE dataset

Method	Indoor PSNR	SSIM	Outdoor PSNR	SSIM
DCP	16.62	0.82	19.13	0.81
DehazeNet	21.14	0.85	22.46	0.85
MSCNN	19.84	0.83	22.06	0.90
AODNet	19.06	0.85	20.29	0.85
DCPDN	19.00	0.84	19.71	0.83
GFN	22.30	0.88	21.55	0.84
EPDN	25.06	0.92	16.22	0.76
GirdDehaze	32.16	0.98	22.57	0.86
FD-GAN	22.14	0.90	23.35	0.92
MSBDN	32.79	0.98	23.16	0.94
KDDN	**34.72**	0.98	–	–
DRN	33.01	0.98	24.44	0.94
OKDNet	30.92	0.99	23.38	0.94
Proposed	33.91	**0.99**	**25.41**	**0.95**

values of LPIPS represent better performance. It can be seen that the proposed method outperforms most the other dehazing methods in terms of LPIPS metric from Table 2. Only on the RSEIDE-Outdoor, PSPAN is slightly inferior to GirdDehaze and MSBDN in terms whose gaps are just 0.022 and 0.017, respectively. As mentioned above, the proposed method outperforms most of the previous methods on the RESIDE dataset in terms of PSNR ,SSIM and LPIPS metrics.

Visual evaluation Figure 6 shows the qualitative comparisons of the visual effect on the Indoor and Outdoor datasets of SOTS. DCP tends to produce darker images compared with the ground truth, as this method often fails to accurately estimate the haze thickness of images. Additionally, DCP suffers from the problem of color distortion, which degrades the quality of their recovered images. GCANet suffers from the same problems as DCP, where the details of distant image fog are blurry and shiny, leading to color distortion problems. It is observed that there remain lots of haze residuals and renders in the dehazed images

Table 2 LPIPS↓ comparisons for different methods on the RESIDE dataset

Methods	Indoor	Outdoor
DCP	0.099	0.105
DehazeNet	0.071	0.075
AODNet	0.303	0.074
DCPDN	0.129	0.073
GFN	0.065	0.105
EPDN	0.053	0.119
GirdDehaze	0.012	**0.017**
FD-GAN	0.082	0.075
MSBDN	0.014	0.022
Proposed	**0.009**	0.039

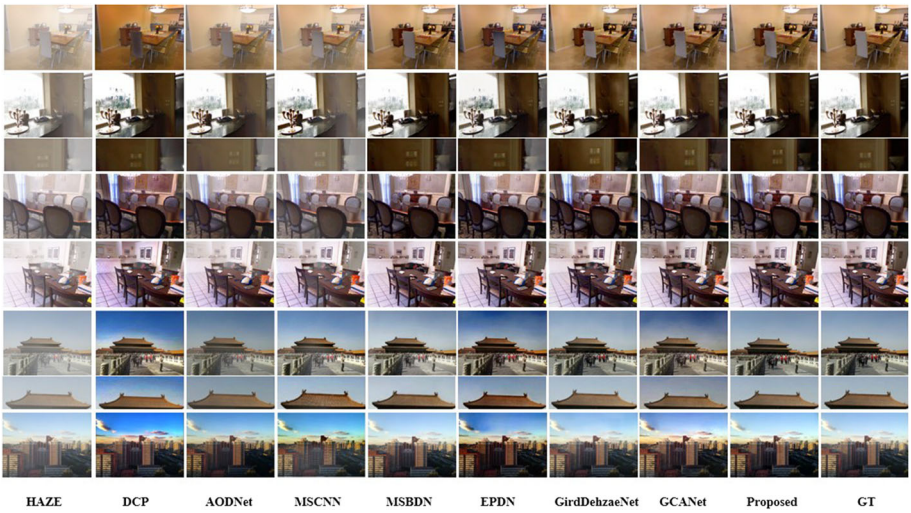

Fig. 6 Visual comparison results on the SOTS dataset. The first column presents the hazy images. The results of seven representative state-of-the-art single-image dehazing methods are illustrated separately. The dehazed results of the proposed method and the ground truth images are shown in the last two columns. The upper five rows show the results of the indoor subset, while the last three rows are dehazed images of the outdoor subset

of AODNet and MSCNN. Although EPDN achieves better results, there is a local gap with ground truth because of its exposure, and it is also darker than the ground truth in some cases. MSBDN and GridDehazeNet achieve the restored images with higher quality. However, they still generate some gray-mottled artifacts as shown in Fig. 6 and cannot completely remove the haze in some regions. Furthermore, the image in the third row is a magnified display of the door position on the right of the image in the second row and only the proposed method and GirdDehazeNet correctly handle the residual haze in this area. In conclusion, the proposed method achieves the best performance in terms of haze removal and it can generate more

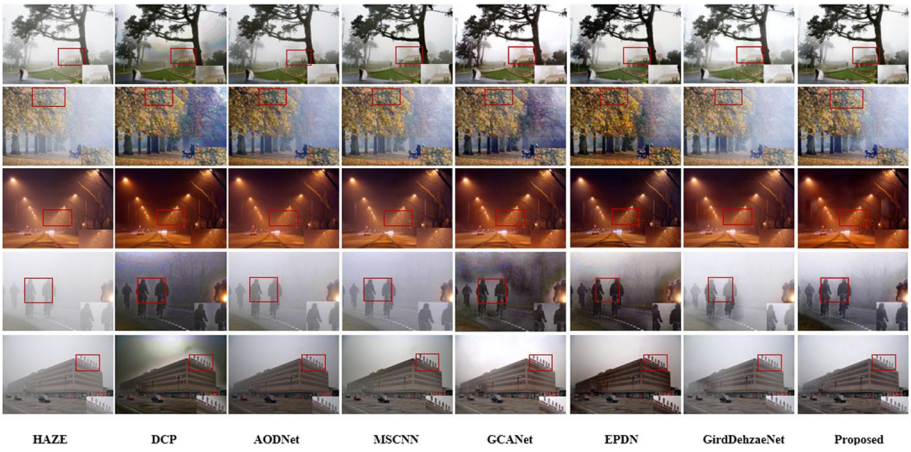

Fig. 7 Visual comparison with state-of-the-art dehazing methods on the RTTS dataset (Pictures are named Img1-5 from top to bottom)

Fig. 8 Some images from the choi's dataset(Hazy ones represent the original pictures and clear ones represent the images processed by the proposed method)

natural dehazed images with more realistic brightness and color fidelity. At the same time, the dehazed images produced by the proposed method are free of major artifacts and are able to preserve more detailed information.

4.2.2 Results on realistic dataset

Test on the RTTS dataset Recent learning-based dehazing methods tend to present insufficient generalization ability and poor dehazing effect on real-world images since they are trained on synthetic datasets. Hence, several real-world hazy images from the RTTS dataset are selected to verify the performance of the proposed method when applied in real scenes. As shown in Fig. 8, it can be observed that DCP suffers from serious color distortions for real hazy images. (e.g., the sky in Fig. 7). GCANet also produces the color distortion problem. For GirdDehazeNet, AODNet, MSCNN, haze removal is incomplete in a dense haze situation, we can find a lot of unremoved haze. The results of EPDN look more natural. This is because we do not use the same training method as EPDN which is trained with a generative adversarial scheme. With the help of adversarial learning, it recovers more realistic images from the real-world dataset. However, EPDN makes local areas dark in some cases. In general, the

Table 3 NIQE↓ comparisons for different methods on the RTTS dataset

Methods	Img1	Img2	Img3	Img4	Img5	Average
DCP	2.6547	3.3088	3.0130	4.8625	3.2342	3.4092
AODNet	2.9939	2.8322	3.1806	2.8932	10.6827	4.5165
MSCNN	3.5433	3.1512	2.8928	3.1022	3.2857	3.1950
GCANet	2.1980	3.6747	2.9225	4.0718	3.5245	3.2783
EPDN	2.0840	3.4109	2.9504	3.5532	3.1946	3.0386
GirdDehazeNet	1.7465	3.1064	2.9249	3.2646	3.4985	2.9082
Proposed	1.8302	3.1418	2.9907	3.4709	3.0326	**2.8932**

Table 4 CNI↑ comparisons for different methods on the choi's dataset

Methods	Img78	Img88	Img90	Img91	Img93	Img97	Img99	Img100	Average
DCP	0.7929	0.5906	0.6755	0.8010	0.8662	0.8466	0.4603	0.6307	0.7080
MSCNN	0.6587	0.5857	0.7345	0.9570	0.5656	0.6327	0.8065	0.7265	0.7084
choi et al.	0.9543	0.5124	0.9420	0.8790	0.5971	0.7604	0.2529	0.6276	0.6907
Isha et al.	0.9728	0.6005	0.7519	0.8917	0.9806	0.9028	0.4379	0.7168	0.7819
AODNet	0.6751	0.8241	0.8247	0.9080	0.8656	0.6389	0.9898	0.9773	0.8379
GCANet	0.6230	0.6740	0.5261	0.9406	0.6090	0.9503	0.8800	0.8210	0.7530
EPDN	0.6659	0.7108	0.7517	0.9830	0.5969	0.9118	0.8421	0.8021	0.7830
GridDehazeNet	0.7069	0.7388	0.4701	0.9285	0.6957	0.7505	0.9831	0.9809	0.7818
Proposed	0.6395	0.9936	0.7297	0.9371	0.6189	0.7059	0.9861	0.9779	**0.8497**

Table 5 The ablation experiments by considering different configurations of the proposed network on RESIDE Indoor (only retain MSE Loss)

Name	PSNR	SSIM
Base	27.62	0.9225
Base+RRB	27.07	0.9250
Base+RREB	30.90	0.9575
Base+RREB+SWPAB1	30.72	0.9597
Base+RREB+SWPAB2	31.05	0.9624
Base+RREB+SWPAB3	32.44	0.9765
Base+RREB+SWPAB4	33.52	0.9838
Base+RREB+SWRCAB4	32.84	0.9778

proposed PSPAN is more effective than existing methods in removing haze and preserving texture details on the RTTS dataset. Also from Table 3, we reach the best performance and surpass the second place 0.015 in NIQE as the average. This shows that the dehazed images processed by the proposed technology offer better image quality.

Test on the choi's dataset Figure 8 shows some selected images from choi's dataset. To further show the generalization ability of the proposed method, we process these selected images with different dehazing methods. And then we measure the CNI parameters to further conduct comparative experiments. In Table 4, it is observed that the proposed method is superior to the others, achieving the best performance with 0.8497 CNI. The comparison results further validate that the proposed method can more effectively restore dehazed image with natural color and good visual effect.

4.3 Ablation study

To further explore the effectiveness of the proposed PSPAN, two ablation studies (Tables 5 and 6) are conducted to verify whether specific module parts of the proposed PSPAN and various losses are effective.

In the previous article, we introduced the two most important modules of the proposed network framework, namely RREB and SWPAB. In Table 5, we will test these two proposed points. Each pyramid block (MFEB) will transmit information of different scales to the feature attention processing block (SWPAB), and considering that the sampled information of the first two sizes has also been processed by the SWPAB, so there are actually four SWPAB blocks involved. At the same time, in order to reduce the interference of other factors, we only retain MSE loss. The following network variants are constructed: (1) Base: the traditional convolution is closely followed by two ordinary ResBlock while removing four SWPAB modules. (2) Base+RRB: use the RRB module to replace the common residual structure.

Table 6 Comparison of loss functions used to train the proposed model on Indoor dataset

MSE Loss	✓	✓	✓
Edge Loss		✓	✓
Perceptual Loss			✓
PSNR	33.52	33.50	33.91
SSIM	0.9838	0.9846	0.9868

(3) Base+RREB: use the RREB module composed of ECAblock and residual network to replace the common residual structure. (4)(5)(6)(7) Base+RREB+SWPAB (number increases in turn): add SWPAB modules to PSPAN network in turn. (8) Base+RREB+SWRCAB4: replace four SWPAB modules with the same quantity SWRCAB modules.

The base network achieved the worst results in terms of PSNR and SSIM in the previous table. The performances of the Base+RREB and the Base+RREB+SWPAB (number) are improved by adding the RREB block and SWPAB blocks. By comparing the results of (1) and (2), the proposed RREB module is superior to the original RRB. At the same time, the experiment (7)(8) also proves that the result of SWPAB module is better than that of SWRCAB. In a word, the full scheme of Base+REEB+SWPAB4 outperforms other architectures in the test dataset, which certifies that RREB and SWPAB are essential to detail-recovery image dehazing. It can also be seen that both considering low-level and high-level features is important for image dehazing.

And beyond that, we perform the ablation experiments to validate the necessity of the loss functions. From the results given in Table 6, we can see that the edge loss contributes to 0.0008 SSIM. The perceptual loss further boosts the performance by 0.41 dB PSNR and 0.0022 SSIM. We prove the effectiveness of the two added loss functions added and the combination of loss functions ensures the effectiveness of haze removal.

5 Conclusion

In this work, we introduce a novel end-to-end dehazing network called Pyramid Spatially Weighted Pixel Attention Network (PSPAN) to tackle the challenging single image dehazing problem. PSPAN is composed of a three-scale extraction module, a pyramid feature attention module, and an image reconstruction module. PSPAN is able to efficiently restore the haze-free image directly. In addition, we propose a novel loss set that combines edge loss and perceptual loss with mse loss to help the network learn more detailed information. Moreover, qualitative and quantitative experiments indicate that the proposed method outperforms most of the state-of-the-art learning-based and traditional approaches in terms of removing the haze and recovering image details.

Acknowledgements This work is partially supported by Heilongjiang Province Natural Science Foundation (LH2022F005) and Northeast Petroleum University Guiding Innovation Fund (No.15071202202).

Data Availability Statements The datasets analysed during the current study are available in the public RESIDE Dataset and public LIVE Image Defogging Database. And the different algorithms' results which performed in datasets during the current study are available from the public paper or the corresponding author on reasonable request.

Declarations

Conflicts of interest The authors declare that they have no conflict of interests.

References

1. Cai B, Xu X, Jia K, Qing C, Tao D (2016) Dehazenet: An end-to-end system for single image haze removal. IEEE Transactions on Image Processing 25(11):5187–5198. https://doi.org/10.1109/TIP.2016.2598681

2. Kansal, I, Kasana, SS (2018) Minimum preserving subsampling-based fast image de-fogging. Journal of Modern Optics, **65**(18):2103–2123. https://doi.org/10.1080/09500340.2018.1499976
3. Fattal, R (2008) Single image dehazing. ACM transactions on graphics (TOG) **27**(3):1–9. https://doi.org/10.1145/1360612.1360671
4. Tan, RT (2008) Visibility in bad weather from a single image. In: 2008 IEEE conference on computer vision and pattern recognition, pp 1–8. https://doi.org/10.1109/CVPR.2008.4587643. IEEE
5. Fattal, R (2014) Dehazing using color-lines. ACM transactions on graphics (TOG), **34**(1):1–14. https://doi.org/10.1145/2651362
6. Zhu, Q, Mai, J, Shao, L (2015) A fast single image haze removal algorithm using color attenuation prior. IEEE transactions on image processing, **24**(11):3522–3533 . https://doi.org/10.1109/TIP.2015.2446191
7. Dong Y, Liu Y, Zhang H, Chen S, Qiao Y (2020) Fd-gan: Generative adversarial networks with fusion-discriminator for single image dehazing. Proceedings of the AAAI Conference on Artificial Intelligence 34:10729–10736. https://doi.org/10.1609/aaai.v34i07.6701
8. Fattal R (2008) Single image dehazing. ACM transactions on graphics (TOG) 27(3):1–9. https://doi.org/10.1145/1360612.1360671
9. Fattal R (2014) Dehazing using color-lines. ACM transactions on graphics (TOG) 34(1):1–14. https://doi.org/10.1145/2651362
10. Yang, H, Pan, J, Yan, Q, Sun, W, Ren, J, Tai, Y-W (2017) Image dehazing using bilinear composition loss function. https://doi.org/10.48550/arXiv.1710.00279
11. Li, C, Guo, J, Porikli, F, Fu, H, Pang, Y (2018) A cascaded convolutional neural network for single image dehazing. IEEE Access, **6**:24877–24887. https://doi.org/10.1109/ACCESS.2018.2818882
12. Li, B, Peng, X, Wang, Z, Xu, J, Feng, D (2017) Aod-net: All-in-one dehazing network. In: Proceedings of the IEEE international conference on computer vision, pp 4770–4778
13. He K, Sun J, Tang X (2011) Single image haze removal using dark channel prior. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(12):2341–2353. https://doi.org/10.1109/TPAMI.2010.168
14. Mei, K, Jiang, A, Li, J, Wang, M (2018) Progressive feature fusion network for realistic image dehazing. In: Asian Conference on Computer Vision, pp 203–215. https://doi.org/10.1007/978-3-030-20887-5_13. Springer
15. Qu, Y, Chen, Y, Huang, J, Xie, Y (2019) Enhanced pix2pix dehazing network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8160–8168
16. Dong, H, Pan, J, Xiang, L, Hu, Z, Zhang, X, Wang, F, Yang, M-H (2020) Multi-scale boosted dehazing network with dense feature fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2157–2167. https://doi.org/10.48550/arXiv.2004.1338
17. Kansal I, Kasana SS (2018) Minimum preserving subsampling-based fast image de-fogging. J Modern Optics 65(18):2103–2123. https://doi.org/10.1080/09500340.2018.1499976
18. Kansal I, Kasana SS (2018) Fusion-based image de-fogging using dual tree complex wavelet transform. Int J Wavelets Multiresolution Inf Process 16(06):1850054. https://doi.org/10.1142/S0219691318500546
19. Kansal I, Kasana SS (2020) Improved color attenuation prior based image de-fogging technique. Multimed Tools Appl 79(17–18):12069–12091. https://doi.org/10.1007/s11042-019-08240-6
20. Lan Y, Cui Z, Su Y, Wang N, Li A, Zhang W, Li Q, Zhong X (2022) Online knowledge distillation network for single image dehazing. Scientific Reports 12(1):1–13. https://doi.org/10.1038/s41598-022-19132-5
21. Hu, J, Shen, L, Sun, G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). https://doi.org/10.48550/arXiv.1709.01507
22. Li C, Guo J, Porikli F, Fu H, Pang Y (2018) A cascaded convolutional neural network for single image dehazing. IEEE Access 6:24877–24887. https://doi.org/10.1109/ACCESS.2018.2818882
23. Li B, Ren W, Fu D, Tao D, Feng D, Zeng W, Wang Z (2018) Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing 28(1):492–505. https://doi.org/10.1109/TIP.2018.2867951
24. Mnih, V, Heess, N, Graves, A, et al (2014) Recurrent models of visual attention. Advances in neural information processing systems, vol 27
25. Lu H, Li Y, Nakashima S, Serikawa S (2016) Single image dehazing through improved atmospheric light estimation. Multimed Tools Appl 75(24):17081–17096. https://doi.org/10.1007/s11042-015-2977-7
26. Qin, X, Wang, Z, Bai, Y, Xie, X, Jia, H (2020) Ffa-net: Feature fusion attention network for single image dehazing. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 11908–11915. https://doi.org/10.1609/aaai.v34i07.6865
27. He, K, Zhang, X, Ren, S, Sun, J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
28. Qin X, Wang Z, Bai Y, Xie X, Jia H (2020) Ffa-net: Feature fusion attention network for single image dehazing. Proceedings of the AAAI conference on artificial intelligence 34:11908–11915. https://doi.org/10.1609/aaai.v34i07.6865

29. Wang, Q, Wu, B, Zhu, P, Li, P, Zuo, W, Hu, Q (2020) Supplementary material for 'eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the 2020 IEEE/CVF conference on computer vision and pattern recognition, IEEE, Seattle, WA, USA, pp 13–19
30. Xu, K, Ba, J, Kiros, R, Cho, K, Courville, A, Salakhudinov, R, Zemel, R, Bengio, Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp 2048–2057. PMLR
31. Vaswani, A, Shazeer, N, Parmar, N, Uszkoreit, J, Jones, L, Gomez, A.N, Kaiser, Ł, Polosukhin, I (2017) Attention is all you need. Advances in neural information processing systems, **30**
32. Wang, X, Girshick, R, Gupta, A, He, K (2018) Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7794–7803. https://doi.org/10.48550/arXiv.1711.07971
33. Hong, M, Xie, Y, Li, C, Qu, Y (2020) Distilling image dehazing with heterogeneous task imitation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3462–3471
34. Gilbarg, D, Trudinger, NS (1983) Elliptic partial differential equations of second order, grundlehren der mathematischen wissenschaften. Berlin Heidelberg New York ed
35. Simonyan, K, Zisserman, A (2014) Very deep convolutional networks for large-scale image recognition. https://doi.org/10.48550/arXiv:1409.1556
36. Johnson, J, Alahi, A, Fei-Fei, L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision, pp 694–711. https://doi.org/10.1007/978-3-319-46475-6_43. Springer
37. Zhang, H, Patel, VM (2018) Densely connected pyramid dehazing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3194–3203. https://doi.org/10.48550/arXiv.1803.08396
38. Ren, W, Liu, S, Zhang, H, Pan, J, Cao, X, Yang, M-H (2016) Single image dehazing via multi-scale convolutional neural networks. In: European conference on computer vision, pp 154–169. Springer
39. Chen, D, He, M, Fan, Q, Liao, J, Zhang, L, Hou, D, Yuan, L, Hua, G (2019) Gated context aggregation network for image dehazing and deraining. In: 2019 IEEE winter conference on applications of computer vision (WACV):pp 1375–1383. https://doi.org/10.1109/WACV.2019.00151
40. Lan, Y, Cui, Z, Su, Y, Wang, N, Li, A, Zhang, W, Li, Q, Zhong, X (2022) Online knowledge distillation network for single image dehazing. Scientific Reports, **12**(1):1–13. https://doi.org/10.1038/s41598-022-19132-5
41. Kansal, I, Kasana, SS (2018) Fusion-based image de-fogging using dual tree complex wavelet transform. International Journal of Wavelets, Multiresolution and Information Processing, **16**(06):1850054. https://doi.org/10.1142/S0219691318500546
42. Choi, LK, A.C.B. You, J (2015) Referenceless prediction of perceptual fog density and perceptual image defogging. Image Process, **24**(11). https://doi.org/10.1109/TIP.2015.2456502
43. Li, B, Ren, W, Fu, D, Tao, D, Feng, D, Zeng, W, Wang, Z (2018) Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing **28**(1):492–505. https://doi.org/10.1109/TIP.2018.2867951
44. Zhang X, Jiang R, Wang T, Huang P, Zhao L (2021) Attention-based interpolation network for video deblurring. Neurocomputing 453:865–875. https://doi.org/10.1016/j.neucom.2020.04.147
45. Zhu Q, Mai J, Shao L (2015) A fast single image haze removal algorithm using color attenuation prior. IEEE transactions on image processing 24(11):3522–3533. https://doi.org/10.1109/TIP.2015.2446191