Image and Vision Computing 146 (2024) 105014

Contents lists available at ScienceDirect

Image and Vision Computing

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

WaveletFormerNet: A Transformer-based wavelet network for real-world non-homogeneous and dense fog removal

Shengli Zhang [a], Zhiyong Tao [a,*,1], Sen Lin [b]

[a] School of Electronic and Information Engineering, Liaoning Technical University, Huludao, Liaoning, China
[b] School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang, Liaoning, China

ARTICLE INFO

Keywords:
Image dehazing
Convolution neural network
Wavelet transform
Swin transformer
Fog removal

ABSTRACT

Although deep convolutional neural networks have achieved remarkable success in removing synthetic fog, it is essential to be able to process images taken in complex foggy conditions, such as dense or non-homogeneous fog, in the real world. However, the haze distribution in the real world is complex, and downsampling can lead to color distortion or loss of detail in the output results as the resolution of a feature map or image resolution decreases. Moreover, the over-stacking of convolutional blocks might increase the model complexity. In addition to the challenges of obtaining sufficient training data, overfitting can also arise in deep learning techniques for foggy image processing, which can limit the generalization abilities of the model, posing challenges for its practical applications in real-world scenarios. Considering these issues, this paper proposes a Transformer-based wavelet network (WaveletFormerNet) for real-world foggy image recovery. We embed the discrete wavelet transform into the Vision Transformer by proposing the WaveletFormer and IWaveletFormer blocks, aiming to alleviate texture detail loss and color distortion in the image due to downsampling. We introduce parallel convolution in the Transformer block, which allows for the capture of multi-frequency information in a light-weight mechanism. Such a structure reduces computational expenses and improves the effectiveness of the network. Additionally, we have implemented a feature aggregation module (FAM) to maintain image resolution and enhance the feature extraction capacity of our model, further contributing to its impressive performance in real-world foggy image recovery tasks. Through extensive experiments on real-world fog datasets, we have demonstrated that our WaveletFormerNet achieves superior performance compared to state-of-the-art methods, as shown through quantitative and qualitative evaluations of minor model complexity. Additionally, our satisfactory results on real-world dust removal and application tests showcase the superior generalization ability and improved performance of WaveletFormerNet in computer vision-related applications compared to existing state-of-the-art methods, further confirming our proposed approach's effectiveness and robustness. Our code is available at https://github.com/shengli666666/WaveletFormerNet.

1. Introduction

Haze is a common atmospheric phenomenon that causes distortion and degradation of images. Image dehazing techniques are significant for many computer vision tasks, such as remote sensing processing [23,24] and video analysis and recognition [40,52]. In recent years, image dehazing has been a hot research topic in computer vision and image processing, serving as an essential low-level image recovery task and a preprocessing step for high-level vision tasks.

Many previous dehazing methods [21,60,39,58] have used the classical atmosphere scattering model (ASM) [34,37] to characterize the degradation process of hazy images by Eq. (1):

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

where $I(x)$ and $J(x)$ are the degraded images and the clear images, respectively. A represents global atmospheric light; $t(x) = e^{-\beta d(x)}$ is the transmission map, where β and $d(x)$ represent atmospheric scattering parameters and scene depth, respectively. The main idea of early prior-based dehazing methods is to estimate the medium transmission map

* Corresponding author.
E-mail address: taozhiyong@lntu.edu.cn (Z. Tao).

[1] Zhiyong Tao, Ph.D., a professor and doctor supervisor, has been working in the School of Electronic and Information Engineering at Liaoning Technical University since July 2000. His research interests include computer vision, deep learning, image processing, and pattern recognition.

https://doi.org/10.1016/j.imavis.2024.105014
Received 3 January 2024; Received in revised form 9 March 2024; Accepted 3 April 2024
Available online 10 April 2024

$t(x)$ and the global atmospheric light A by handcraft; these methods have made significant progress. However, these prior-based methods usually require time-consuming iterative optimization and manually designed priors; they need to be more consistent with the practice. As is well known, haze formation is related to natural factors, such as altitude, temperature, and humidity, but it is challenging to express hazy images with a simplistic model. Therefore, these prior-based methods may result in estimation errors when dealing with complex scenes such as the non-homogeneous and dense fog weather.

Nowadays, data-driven based methods have made significant progress in image dehazing. A handful of end-to-end models [17,25,46] have been proposed to mitigate the deep reliance on predefined prior information, but they typically have limitations in terms of interpretability. With the rapid development of deep learning techniques and the establishment of large-scale synthetic datasets [26], many data-driven approaches [53,58,49,29,56,1] use Convolution Neural Networks (CNNs) to extract features and build end-to-end dehazing networks to learn the transmission map between clear and hazy images.

Although the above data-driven approaches significantly improve the visual quality of image dehazing results, the following problems still exist for the current state-of-the-art (SOTA) CNNs: First, due to the complex distribution of haze in the real world, edge information is crucial to recovering the texture details of a clear image. Still, the points at the edges of the image may be ignored during the layer-by-layer convolution process, resulting in the loss of image details easily. Second, models trained on synthetic fog datasets do not work well for processing images in real-world non-homogeneous fog or dense fog, and the generalization ability and robustness of the network still need to be improved. Third, balancing the network's generalization ability with the model's complexity is a significant challenge for image dehazing tasks, especially handling challenging visual tasks.

Fig. 2 shows an example of images and their corresponding frequency information. The frequency information here represents edge detail. From Fig. 2, we can see that the clear image contains more edge details, while the hazy image reflects only some of the structural features and loses a lot of edge details. The low-frequency information of the image retains more structural information, such as color and target. In contrast, the high-frequency information of the image can specifically represent the edge details and textures of the image. Hence, the frequency information is essential for recovering the structure and texture of an image. Therefore, we exploit the discrete wavelet transform (DWT)

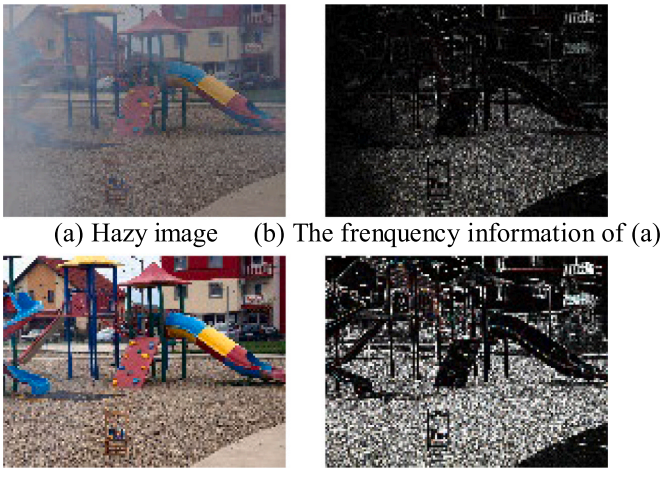

Fig. 2. Examples of clear and hazy images and their frequency information to demonstrate edge detection on clear and hazy images using the Laplacian function. Degraded images correspond to frequency information that reflects the structure and contours of the image. Still, clear images are also richer in frequency information after filtering, corresponding to more textural details.

and inverse discrete wavelet transform (IDWT) to decompose the RGB image into high and low-frequency information to guide the network for image recovery. The motivation is that DWT or IDWT can alleviate information loss and enlarge the receptive field with a better restoration performance. In addition, wavelet transform has been applied to denoising tasks [61,30,28] using traditional methods. Utilizing wavelet transform to incorporate multi-scale information gives the network frequency analysis capabilities. (See Fig. 1.)

Using a combined analysis of the above, we present WaveletFormerNet to mitigate the issue in this article. We have three main goals and design three key steps to address the complex distribution of haze characteristics in the real world. Specifically: **a)** To alleviate the image's detailed texture loss due to downsampling, we devise the WaveletFormer block and IWaveletFormer block to fully preserve the structure and texture details in the original image, combining the advantages of discrete wavelet transform (DWT) and inverse discrete wavelet transform (IDWT) with the Vision Transformer (ViT), respectively. **b)** To capture the multi-frequency signals in the lightweight mechanism, we introduce parallel convolution in ViT; this structural design also reduces the computational cost and model complexity. **c)** To maintain image resolution and enhance the receptive field of our network, we present the feature aggregation module (FAM) to process the association of information and the interaction of characteristics between different levels. Our key contributions can be summarized as follows:

We propose the WaveletFormer and IWaveletFormer blocks to alleviate texture detail loss and maintain image resolution. The parallel convolution in the Transformer blocks captures the multi-frequency information in the lightweight mechanism.

We present a feature aggregation module (FAM) to capture the long-range dependencies among information with different levels and further enhance the feature extraction capability of WaveletFormerNet.

We present WaveletFormerNet, an end-to-end wavelet reconstruction network guided by frequency information to tackle image dehazing problems under complex, hazy conditions in the real world. To validate the effectiveness of WaveletFormerNet, we conducted extensive experiments on both synthetic and real-world datasets. The results demonstrate that our method yields competitive dehazing performance in comparison to SOTA methods.

The remaining sections of this paper are structured as follows. Section 2 provides a concise overview of the existing research on image dehazing, wavelet transform, and ViT. In Section 3, we outline the proposed WaveletFormerNet framework, the proposed WaveletFormer and IWAveletFormer blocks, and our proposed FAM. In Section 4, we introduce the dataset details, implementation details, and the loss function we adopted. The evaluation of our proposed method is shown in Section 5, including a comparison to SOTA methods, generality analysis for WaveletFormerNet, application test, and ablation study. In Section 6, we summarize the conclusions and discussions about our study.

2. Related work

In this section, we focus on related work in image dehazing in Senction 2.1. We also introduce the related work of discrete wavelet transform in Section 2.2 and Vision Transformer in Section 2.3.

2.1. Image dehazing

The existing methods for image dehazing are broadly classified into two categories: traditional prior-based methods and learning-based methods.

2.1.1. Traditional methods

Most prior-based dehazing methods [21,60,46,59] use hazy and clear images to estimate the transmission map, then use ASM to recover

Fig. 1. Qualitative dehazing performance comparison among WaveletFormerNet and SOTA methods on four benchmark real-world hazy datasets (I-Haze, O-Haze, NH-Haze, and Dense-Haze datasets).

haze-free images. He et al. [21] proposed the dark channel prior (DCP), assuming that the image patches of haze-free outdoor images often have low-intensity values in at least one channel. To address the difference in brightness and saturation of hazy images, Zhu et al. [60] proposed color attenuation prior (CAP) to estimate the scene depth as solid prior knowledge. To eliminate the polarization effect of information. Shen et al. [46] proposed a globally nonuniform ambient light model to predict spatially varied ambient light and designed a bright pixel index to correct the transmission. With the predicted haze parameters, they reversed the atmospheric scattering model to restore visibility. Zhou et al. [59] proposed a robust polarization-based dehazing network. However, the specific scenario inherently limits the performance of these methods, and they may lead to undesirable color distortions when the scenario does not satisfy these priors. In contrast, WaveletFormerNet can reconstruct images with richer detail by leveraging the complementary advantages of prior- and deep learning-based methods.

2.1.2. *Deep learning methods*

Recently, deep learning techniques have been proposed to tackle the problem of underwater image dehazing. These techniques have shown promising results in the restoration of underwater images. They can be classified into three categories: (i) CNN-based methods, (ii) GAN-based methods, and (iii) Transformer-based methods.

2.1.2.1. CNN-based methods. A wide range of CNN-based methods [41,8,25,31,39,30,48,53,49,58] have dominated in recent years. Ren et al. [41] proposed MSCNN to estimate $t(x)$ using a coarse-scale network followed by local optimization. Li et al. [25] reiterated ASM and proposed AODNet to learn each hazy image and its $t(x)$. However, all of these methods rely on ASM, and the dehazing results are often color-distorted. To alleviate the bottleneck problem encountered in traditional multi-scale methods, Liu et al. [31] implemented an attention-based end-to-end dehazing network, GridDehazeNet. To enable more efficient dehazing network performance, Liu et al. [39] designed an FFANet with channel and spatial attention to obtain excellent dehazing performance. Zheng et al. [58] took FFANet as a baseline, proposing C^2PNet with a curricular contrastive regularization and the physics-aware dual-branch unit to enhance the network dehazing performance. However, behind the excellent performance achieved by these supervised methods, a large number of data pairs are required for the training; more importantly, these methods are almost trained on synthetic images [26,5,3,4,2], which can not be well generalized to real-world image dehazing.

2.1.2.2. GAN-based methods. Recently, some unsupervised data-driven methods have also made significant progress in image defogging. Ren et al. [42] first introduced Generative Adversarial Networks (GANs) to the field of image defogging that achieves mapping from foggy images to fog-free images by training a generator and a discriminator. Mehta et al. [35] propose SkyGAN for haze removal in aerial images, alleviating the degradation in image visibility. Dong et al. [15] propose a fully end-to-end GAN with a Fusion discriminator (FD-GAN) for image dehazing; this model can generate more natural and realistic dehazed images with less color distortion and fewer artifacts. Wang et al. [51] proposed a dual multiscale network, TMS-GAN, to alleviate the problem of limited domain transfer performance between trained synthetic blurred images and untrained real blurred images.

Li et al. [27] propose a novel dehazing algorithm by combining model-based and data-driven approaches. The proposed neural augmentation reduces the number of training data significantly, and the proposed neural augmentation framework converges faster than the corresponding data-driven approach [15,18,51,27]. However, unsupervised methods may be unstable during the training process, leading to problems such as the possibility of unstable results during enhancements.

2.1.2.3. Transformer-based methods. Recently, Transformer [50] has gained increasing attention, image content and attention weights interact spatially as a result of spatially varying convolution. Guo et al. [19] proposed DeHamer to effectively integrate Transformer features and CNN features and bring the domain knowledge, such as task-specific prior, into Transformer for improving the performance.

Furthermore, Song et al. [49] proposed that DehazeFormer improves on Swin Transformer [32], which makes Transformer more useful for image dehazing. Compared to CNN-based networks, our approach can help the network pay more attention to attenuated color channels and spatial areas. In addition, since a GAN is coupled with a transformer, we can obtain a better performance with a relatively small number of parameters.

2.2. *Discrete wavelet transform*

As a traditional image processing technique, the discrete wavelet transform [12,55,13,30] is widely used for image analysis. Guo et al.

[20] proposed a DWSR combining the discrete wavelet transform with ResNet by predicting the residual wavelet subbands. Inspired by U-Net [44], Liu et al. [28] proposed MWCNN, which replaces pooling and non-pooling operations to reduce the number of parameters in the network. However, multiple uses of the discrete wavelet transform operations may result in redundant channels. Therefore, Yang et al. [54] proposed the Wavelet U-Net, which uses the discrete wavelet transform to extract edge features while applying the adaptive color transform that convolutional layers; this structure enhances the texture details in the image. Zou et al. [61] proposed SDWNet to obtain large sensory fields with a high spatial resolution and recover precise high-frequency texture details. Fu et al. [18] proposed a two-branch network DW-GAN to leverage the power of discrete wavelet transform in helping the network acquire more frequency domain information. These methods demonstrate the significant role of discrete wavelet transform in the image recovery process.

2.3. Vision Transformer

Attention mechanism [50] of deep learning has achieved a great process nowadays. Dosovitskiy et al. [16] proposed ViT with the direct application of the Transformer architecture, which projects images into token sequences via patch-wise linear embedding. The shortcomings of the ViT are its weak inductive bias and its quadratic computational cost. Until Liu et al. [32] proposed the Swin Transformer, they divided tokens into a window and performed self-attention within a window to maintain a linear computational cost. Guo et al. [19] proposed Dehamer modulate convolutional features via learning modulation matrices, which are conditioned on Transformer features instead of simple addition or concatenation of features. Song et al. [49] proposed the DehazeFormer, which can be viewed as a combination of Swin Transformer [32] and U-Net [44] with more comprehensive improvements in the normalization layer, nonlinear activation function, and spatial information aggregation scheme. DehazeFormer improves the network performance for single-image dehazing further. Although ViT enhances the image recovery performance, it may increase additional computational expense and ignore the haze distribution characteristics under complex conditions.

3. Methodology

In this section, We introduced the motivation and overview of our proposed WaveletFormerNet in Section 3.1; in Section 3.2, we present the detailed structure of the WaveletFormer and IWaveletFormer blocks. The detailed design of FAM, another module proposed in this paper, is placed in Section 3.3. At the end of the chapter, we introduce the ASPP Module we adopted in Section 3.4.

3.1. Overview

Fig. 3 illustrates the detailed structure of our WaveletFormerNet. Both encoding and decoding of WaveletFormerNet are based on the WaveletFormer and IWaveletFormer block, but the difference between the encoding and decoding segments is that downsampling and upsampling are replaced by DWT and IDWT, respectively. Although the WaveletFormer and IWaveletFormer block as the base block of the network mainly combines wavelet transform and Swin Transformer, we do not directly apply these two existing tools but improve them. We use the wavelet transform to transform the features to the frequency domain and use the frequency information to guide WaveletFormerNet to recover the structural and texture details of the image. In addition, our proposed parallel convolution also alleviates the receptive field caused by Swin Transformer. This structure of the proposed WaveletFormer and IWaveletFormer block also alleviates the details caused by downsampling loss and other problems. Furthermore, we propose a Feature Aggregation Module (FAM) to maintain image resolution and enhance the receptive field of our network, combining different levels of feature information. Finally, we adapt an atrous spatial pyramid pooling module (ASPP) in our network and adjust dilated convolution [10] with different expansion rates (rate = 3, 6, 9), obtaining features in different receptive fields. See Fig. 4 to visualize the quantitative comparison of our method real-world dehazing performance with other SOTA methods.

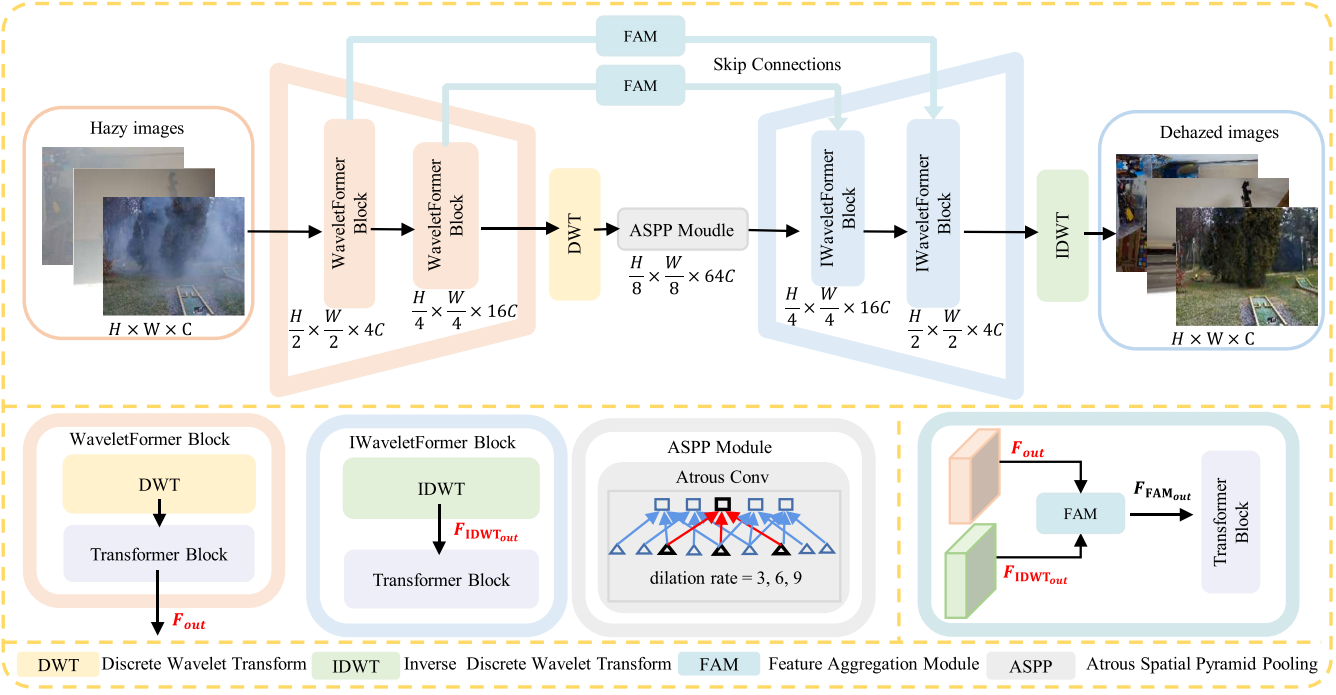

Fig. 3. The schematic illustration of the proposed WaveletFormerNet. WaveletFormer block and IWaveletFormer block consist of DWT and IDWT and Transformer block respectively, and IDWT is the reverse process of DWT.

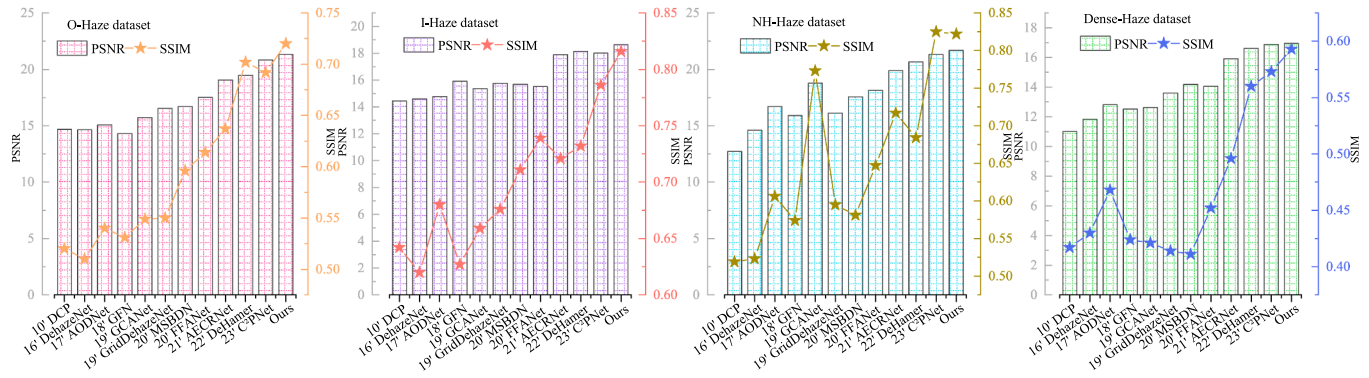

Fig. 4. Quantitative comparisons on referenced indicators (PSNR and SSIM) results among WaveletFormerNet and SOTA methods on four real-world datasets ((O-Haze, I-Haze, NH-Haze, and Dense-Haze)).

3.2. WaveletFormer and IWaveletFormer block

WaveletFormer and IWaveletFormer Blocks use DWT and IDWT to decompose the images from the frequency domain point of view, respectively, and the feature maps are used as inputs to the Transformer module with parallel convolution.

3.2.1. Frequency decomposition of images

Fig. 5 illustrates the detailed structure of the WaveletFormer block, adopting frequency information to guide the network in reconstructing a clear image. We can observe that the input image $F_{\mathrm{DWT_{in}}}$ can be divided into the low- and high-frequency details separated into four different frequency subbands: the low-frequency band F_{LL}, the horizontal subband F_{LH}, the vertical subband F_{HL} and the high-frequency subband F_{HH} on the diagonal edge of the original image. This mechanism alleviates detail and color loss and provides a better balance between network processing efficiency and image recovery performance. For the 2D discrete wavelet transform, we import the pytorch_wavelets package and use Daubechies wavelet basis functions.

3.2.2. Parallel convolution in Vision Transformer

According to the attention mechanism [50], given an input feature map $\mathbf{X} \in \mathbb{R}^{b\times h\times w\times c}$, we project $\mathbf{X}$ to $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ (query, key, value), and we compute the attention function for a set of queries simultaneously and pack them into a matrix $\mathbf{Q}$; so that the computed output matrix can be described as:

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{2}$$

And the Multi-Head Self-Attention (MHSA) [50] can be expressed as Eq. (3), where the projections are parameter matrices $\mathbf{W}_i^Q \in \mathbb{R}^{d_{\mathrm{model}}\times d_k}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_{\mathrm{model}}\times d_k}$, $\mathbf{W}_i^V \in \mathbb{R}^{d_{\mathrm{model}}\times d_v}$ and $\mathbf{W}^O \in \mathbb{R}^{hd_v\times d_{\mathrm{model}}}$.

$$\begin{aligned}\mathrm{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \mathrm{Concat}(\mathrm{head}_1,\cdots,\mathrm{head}_h)\mathbf{W}^O\\ \mathrm{head}_i &= \mathrm{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q,\mathbf{K}\mathbf{W}_i^K,\mathbf{V}\mathbf{W}_i^V\right)\end{aligned} \tag{3}$$

Fig. 5 illustrates the detailed structure of our WaveletFormer block. We employ $h = 8$ parallel attention layers, or heads, where $d_k = d_v = d_{\mathrm{model}}/h = 64$.

We perform an additional convolution of the features $F_{\mathrm{DWT_{out}}}$ from the DWT and then achieve a dynamic aggregation style of information with the production of MHSA in the spatial dimension, thus capturing the multi-frequency signals in the lightweight mechanism. Therefore, the output of the WaveletFormer block can be formulated as follows:

$$F_{\mathrm{out}} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} + \mathrm{Conv}(F_{\mathrm{DWT_{out}}}) \tag{4}$$

In the decoding, we use the IDWT in the IWaveletFormer block with the same filter as the DWT for image recovery. We present Fig. 6 to show a more detailed explanation or visual representation of the WaveletFormer and IWaveletFormer blocks. Our proposed WaveletFormerNet contains two WaveletFormer blocks and IWaveletFormer blocks in the encoding and decoding, respectively. As can be observed from the feature maps, the output feature maps of the WaveletFormer block are more salient in terms of frequency information as the encoding network deepens, and the edge details are sharpened. Comparing the output feature maps of the IWaveletFormer block in the two decoding stages, we can observe that the overall structure and texture details of the image are getting clearer, revealing the effectiveness of our proposed WaveletFormer and IWaveletFormer blocks.

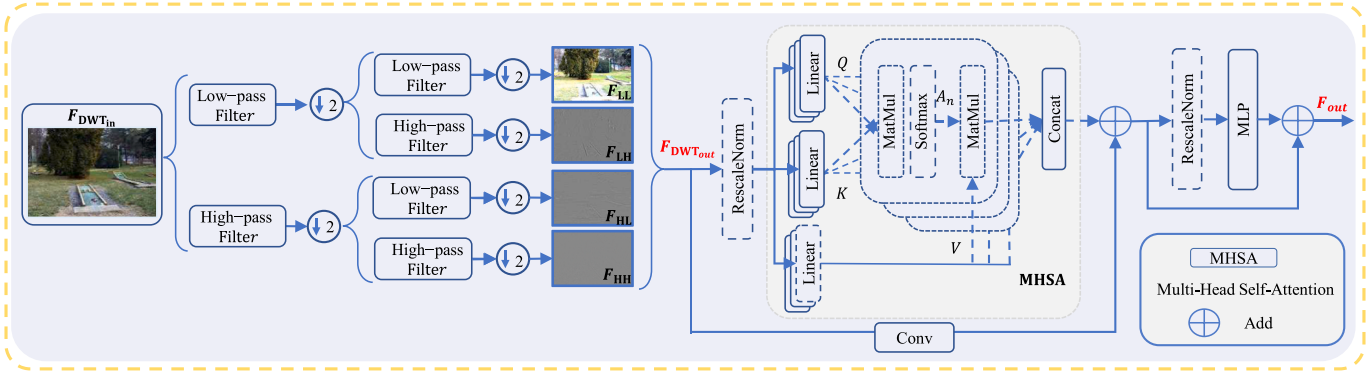

Fig. 5. The architecture of proposed WaveletFormer and IWaveletFormer blocks. Note: The WaveletFormer block and the IWaveletFormer block have the same structure: they utilize DWT and IDWT to substitute downsampling and upsampling, respectively.

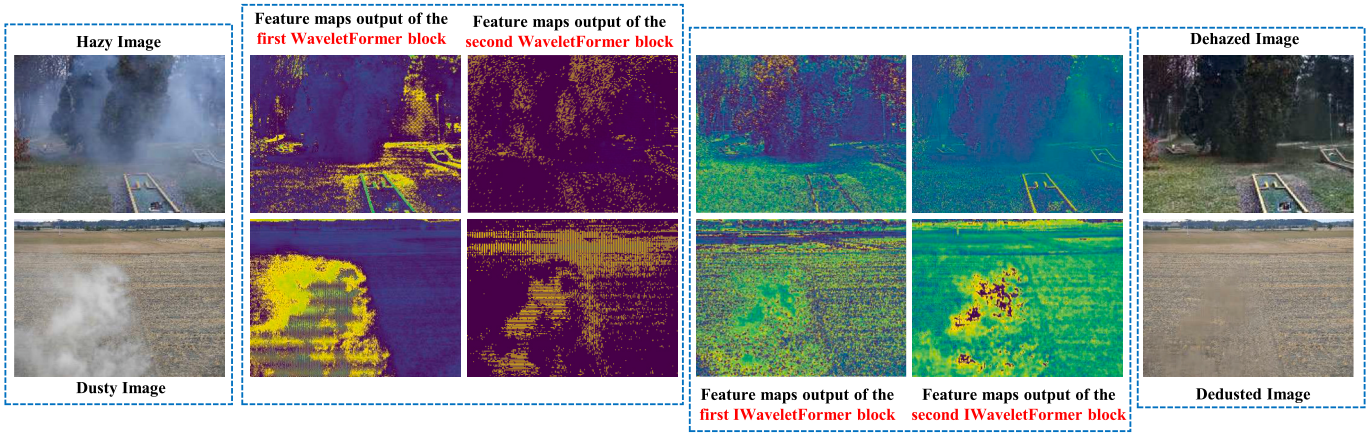

Fig. 6. Visual results of the features in our proposed WaveletFormer block and IWaveletFormer block with degraded images (hazy and dusty images). Our proposed WaveletFormerNet contains two WaveletFormer blocks and IWaveletFormer blocks in the encoding and decoding, respectively. As can be observed from the feature maps, the output feature maps of the WaveletFormer block are more salient in terms of frequency information as the encoding network deepens, and the edge details are sharpened. Comparing the output feature maps of the IWaveletFormer block in the two decoding stages, we can observe that the overall structure and texture details of the image are getting clearer.

3.3. Feature aggregation module

Fig. 7 illustrates the detailed structure of the proposed FAM. As the key component of the FAM, the Multi-Head Cross-Attention (MHCA) [38] introduces the feature map F_{out} of the WaveletFormer block and the feature map $F_{\text{IDWT}_{\text{out}}}$ from IWaveletFormer block into the MHSA [50] for processing, the computed weight values $\boldsymbol{Y}$ to be rescaled by the sigmoid activation function. The resulting feature tensor $\boldsymbol{Z}$ will be summed with the feature map F_{out} to obtain the high-level feature tensor. In addition, the feature tensor $\boldsymbol{X}$ produced by $F_{\text{IDWT}_{\text{out}}}$ is also obtained at a high-level feature tensor after operations such as ReLu. Finally, we concatenate these two high-level feature tensors as the $F_{\text{MHCA}_{\text{out}}}$. Therefore, $F_{\text{FAM}_{\text{out}}}$ can be expressed as:

$$F_{\text{FAM}_{\text{out}}} = \text{MLP}(\text{MHCA}(F_{\text{out}}, F_{\text{IDWT}_{\text{out}}})) + F_{\text{out}} \tag{5}$$

The proposed feature aggregation module (FAM) introduces the feature map F_{out} of the WaveletFormer block and the feature map $F_{\text{IDWT}_{\text{out}}}$ from IWaveletFormer block for processing. Fig. 8 illustrates that FAM is a link between the encoding and decoding stages, guiding our WaveletFormerNet to generate images with more crisp textures and rich

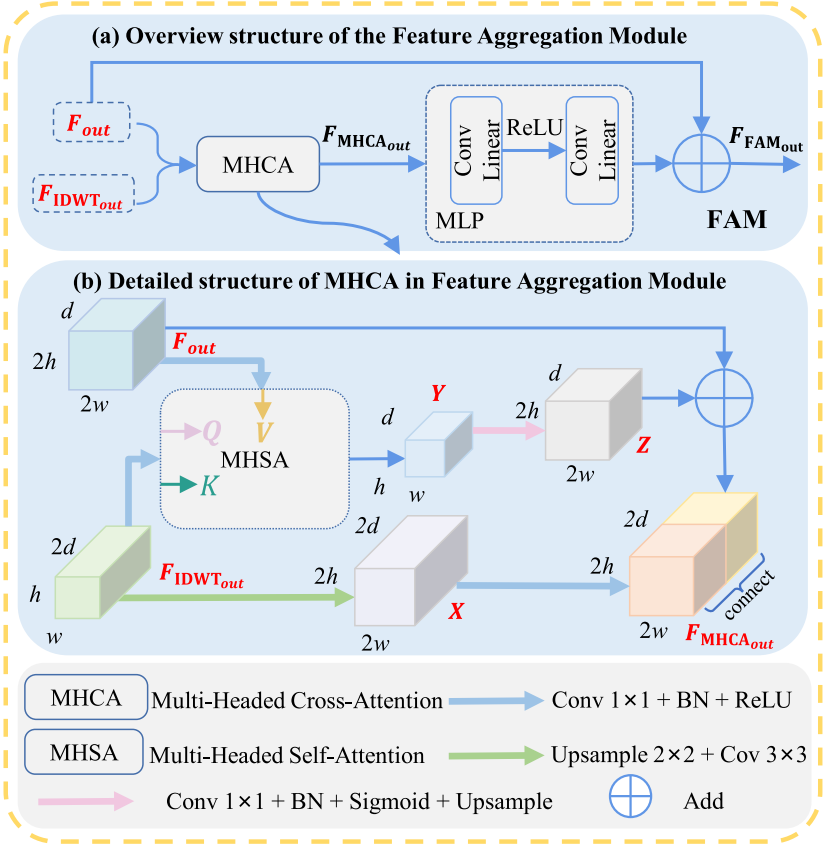

Fig. 7. Architecture of proposed feature aggregation module employed in the proposed WaveletFormerNet.

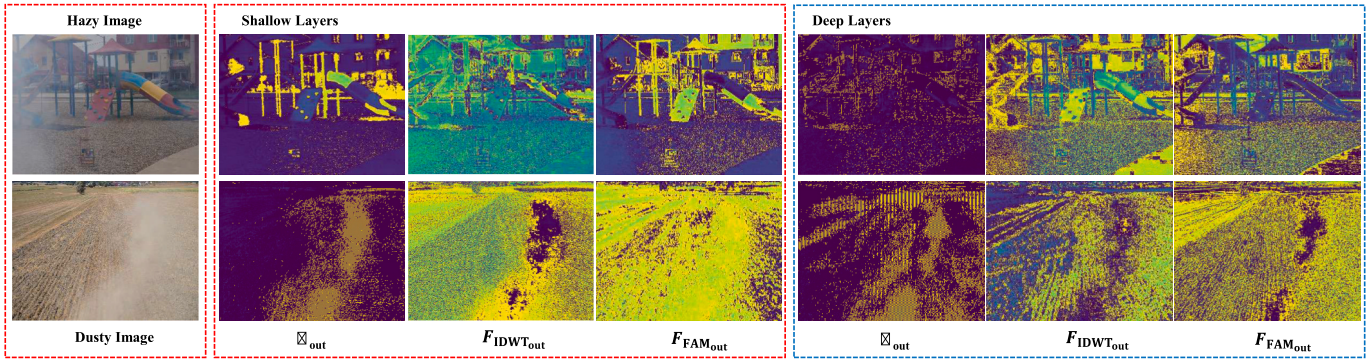

Fig. 8. Visual results of the intermediate features in our proposed feature aggregation module with degraded images (hazy and dusty images). The corresponding modulated features $F_{\text{FAM}_{\text{out}}}$ are also presented. The features F_{out} in the WaveletFormer block have long-range attention but coarse textures, while the features $F_{\text{IDWT}_{\text{out}}}$ in the IWaveletFormer block are with precise details. The modulated features produced by the feature aggregation module inherit the characteristics of both Transformer features and frequency information, i.e., long-range dependencies and clear textures.

details. The proposed FAM removes irrelevant or noisy areas from skip connections and highlights those important areas, capturing the long-range relationship among different receptive field features and improving decoding efficiency.

3.4. Atrous spatial pyramid pooling module

We adopt the atrous spatial pyramid pooling module (ASPP module) in our network. Unlike previous dehazing networks that use repeated upsampling and downsampling to obtain large receptive domains, we use dilated convolution to obtain features in different receptive fields. Fig. 9 illustrates the principle of the ASPP module used in our paper. The input $F_{\text{ASPP}_{\text{in}}}$ is sampled in parallel with a convolution of holes at different expansion rates (rate = 3, 6, 9); the results are then concatenated together to expand the number of channels, and the channels of the output $F_{\text{ASPP}_{\text{out}}}$ are reduced by 1×1 convolution.

4. Experiment setup

This section introduces the training and test datasets in Section 4.1. Our loss function and training details are in Section 4.2. Finally, our comparison methods and evaluation metrics are embodied in Section 4.3.

4.1. Training datasets

We extensively and comprehensively evaluated our model and compared SOTA methods on real-world and synthetic datasets in the same experiment setting.

4.1.1. Real-world datasets

We use the following four datasets to evaluate our experiments: the NTIRE 2018 image dehazing dataset (I-Haze), the outdoor NTIRE 2018 image dehazing dataset (O-Haze), a benchmark for image dehazing with dense-haze and haze-free images (Dense-Haze), and the NTIRE 2020 dataset for non-homogeneous dehazing challenge (NH-Haze).

4.1.1.1. I-Haze [5] and O-Haze [3]. They contain 25 and 35 hazy images (size 2833 × 4657 pixels) respectively for training. Both datasets contain 5 hazy images for validation along with their corresponding ground truth images. We used training data for training and validation data for the test.

4.1.1.2. Dense-Haze [2]. It contains 45 hazy images (size 1200 × 1600 pixels) for training 5 hazy images for validation and 5 more for testing with their corresponding ground truth images. We have performed training on training data and tested our model with test data.

4.1.1.3. NH-Haze [4]. It contains 45 hazy images (size 1200 × 1600 pixels) for training. We selected 40 pairs of data for training and the rest for testing.

4.1.1.4. Real-world datasets expansion. We use the same training strategy and dataset expansion process for all four real-world datasets. Specifically, we randomly cropped the original images into square patches of 512 × 512 pixels; these patches are not identical for every epoch. To augment the training data, we implemented random rotations (90, 180, or 270 degrees) and random horizontal flips when processing

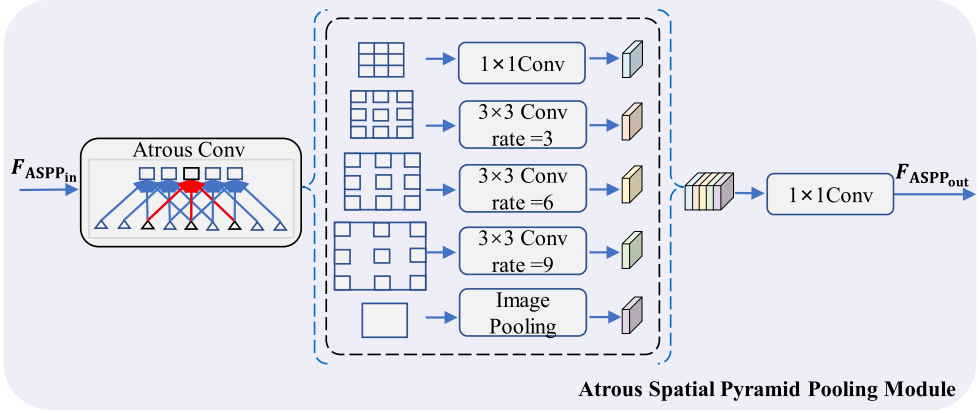

Fig. 9. The architecture of our adopted Atrous Spatial Pyramid Pooling Module (ASPP Module).

the training data. This step allows these small real-world datasets to be expanded into larger datasets that are more efficient and more suitable for data-driven methods of training experiments.

4.1.2. Synthetic datasets

For the objectivity of the experimental results, we evaluated our and SOTA methods under the same experimental conditions. We train our WaveletFormerNet on two training sets of RESIDE: indoor and outdoor. The indoor training set (ITS) has 13,990 hazy images, and the outdoor training set (OTS) has 296,695 hazy images. We selected SOTS as our testing set, the SOTS is from the RESIDE dataset, including 500 indoor and 500 outdoor hazy images.

4.2. Loss function and training details

Refer to previous work [57], to balance both visual perception and quantitative assessments, we combine ℓ_1 loss $\mathscr{L}_{\ell_1}$, multiscale structural similarity (MS-SSIM) loss $\mathscr{L}_{\mathbf{MS-SSIM}}$ and $\mathscr{L}_{\text{per}}$ loss linearly.

Concretely, the ℓ_1 loss retains color and brightness and converges quickly, providing a wider and more stable gradient, which can be described as follows:

$$\mathscr{L}_{\ell_1} = \frac{1}{N}\sum_{i=1}^{N} \| I_i - GT_i \| \tag{6}$$

where I_i is the dehazing image processed by WaveletFormerNet and GT_i refers to the ground truth.

The $\mathscr{L}_{\mathbf{MS-SSIM}}$ loss integrates the variations of resolution and visualization conditions to consider structural differences, compared to other loss functions, the $L_{\text{MS-SSIM}}$ loss preserves the contrast in the high-frequency region, it can be described as follow:

$$\mathscr{L}_{\mathbf{MS-SSIM}}(I_i, GT_i) = 1 - \text{MS} - \text{SSIM}(I_i, GT_i) \tag{7}$$

Inspired by the current hot research in the field of image dehazing, we adopt L_{per} [47] to promote the perceptual similarity of dimensional spatial features and perceive the image from a high-dimension. The L_{per} can be expressed by the following equation:

$$L_{\text{per}} = \sum_{i=1}^{N} \|\omega(I_{\text{out}_i}) - \omega(GT_i)\| \tag{8}$$

where $\omega(\cdot)$ denotes the extraction of two groups of feature maps from 2nd and 5th pooling layers of VGG16 [47] (which has been pre-trained from ImageNet [45]).

Therefore, our loss function can be expressed as:

$$\mathscr{L}_{\text{total}} = \lambda_1 \mathscr{L}_{\text{MS-SSIM}} + \lambda_2 \mathscr{L}_{\ell_1} + \lambda_3 \mathscr{L}_{\text{per}} \tag{9}$$

where λ_1, λ_2, and λ_3 are empirically set to 1, 2, and 1 to balance the scales of multiple losses, respectively.

We conducted a comparative experiment on GeForce RTX 4090 using PyTorch 1.11.0. Adam optimizer is adopted, the initial learning rate is set to 0.0001, betas = (0.9, 0.999), the batch size is 16, the crop size is 256 × 256, and the total number of epochs is 150. We adopt the cosine annealing strategy [22] to adjust the learning rate η_t from the initial value to 0. Assuming T is the total number of batches, η is the initial learning rate, at the batch t, η_t can be expressed by following the cosine function:

$$\eta_t = \frac{1}{2}\left(1 + cos\left(\frac{t\pi}{T}\right)\right)\eta \tag{10}$$

4.3. Comparison methods and evaluation metric

We conduct a comprehensive comparison with SOTA methods on synthetic and real-world datasets, and we mainly select some SOTA methods for comparison according to the three main categories of image dehazing methods, including prior-based methods, ASM-based methods, and hazy to clear image translation-based methods.

We adopt two objective quantity evaluation metrics peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). In addition, we also adopt three no-reference metrics Entropy [6], natural image quality evaluator (NIQE [36]), and fog aware density evaluator (FADE [11]) to evaluate all the above methods. Specifically, a higher Entropy score indicates that the image presents more detail; a lower NIQE score indicates better image quality and a lower FADE score indicates better visibility.

5. Experiment results

We show the results of our proposed WaveletFormerNet compared to the SOTA method on real-world and synthetic datasets in Section 5.1 and Section 5.2, respectively. In Section 5.3, we show the parameters and runtime analysis. We display a generality analysis for WaveletFormerNet on the RB-Dust agricultural dust dataset in Section 5.4. We also conduct the application test in Section 5.5 and the ablation study in Section 5.6.

5.1. Results analysis on the real-world datasets

We show the comparative results of qualitative effects on the four datasets in Fig. 10, Fig. 11, Fig. 12, and Fig. 13. We can observe that DCP produces bluer results in real-world datasets, with severe color bias due to the complex hazy conditions where the pure prior theory is not applicable. Furthermore, the output results of AODNet, FFANet, and AECRNet show severe color distortion and incomplete haze removal; Dehamer and C^{2}PNet outperform the above three methods, removing haze very well and providing enjoyable visual effects. However, our results are closer to the ground truth than these two SOTA methods, and WaveletFormerNet produces visually pleasing dehazing images, which can retain richer texture details.

Table 1 shows the quantitative comparison of WaveletFormerNet with SOTA methods on referenced indicators (PSNR and SSIM). We can observe that methods such as AECRNet, Dehamer, and C^{2}PNet have demonstrated exemplary performance. However, for the challenging tasks of non-homogeneous fog and dense fog, our WaveletFormerNet improves the PSNR metrics by 0.36 dB and 0.07 dB each compared to the second-best on the NH-Haze and Dense-Haze datasets.

Moreover, Table 2 and Fig. 15 show the quantitative comparison of WaveletFormerNet with SOTA methods on non-referenced indicators (Entropy, NIQE, and FADE) on four real-world datasets (O-Haze, I-Haze, NH-Haze, and Dense-Haze). The results of the quantitative comparison showed the better image quality and visibility produced by our WaveletFormerNet.

To compare the dehazing performance of each algorithm more comprehensively, we selected a selection of natural hazy images [17] taken in the real world (including cones, house, and train), which often appear in the dehazing literature, and a selection of images from the RTTS dataset, to perform a qualitative comparison in Fig. 16. In addition, we select three no-reference evaluation metrics, Entropy, NIQE, and FADE, and show the quantitative comparison between our method and the SOTA methods on natural hazy images and RTTS datasets in Table 4. From Fig. 16, we can observe that each algorithm has different characteristics in terms of the results of processing unpaired foggy datasets: some of them deepen the color of the output image, some enhance the contrast of the image, some sharpen the details in the image, some increase the brightness value of the image, etc. The quantitative evaluation metrics comparison results in Table 4 show that our method produces a competitive comparison with DeHamer and C^{2}PNet; however, compared to these methods, our output results have richer textures and more natural colors.

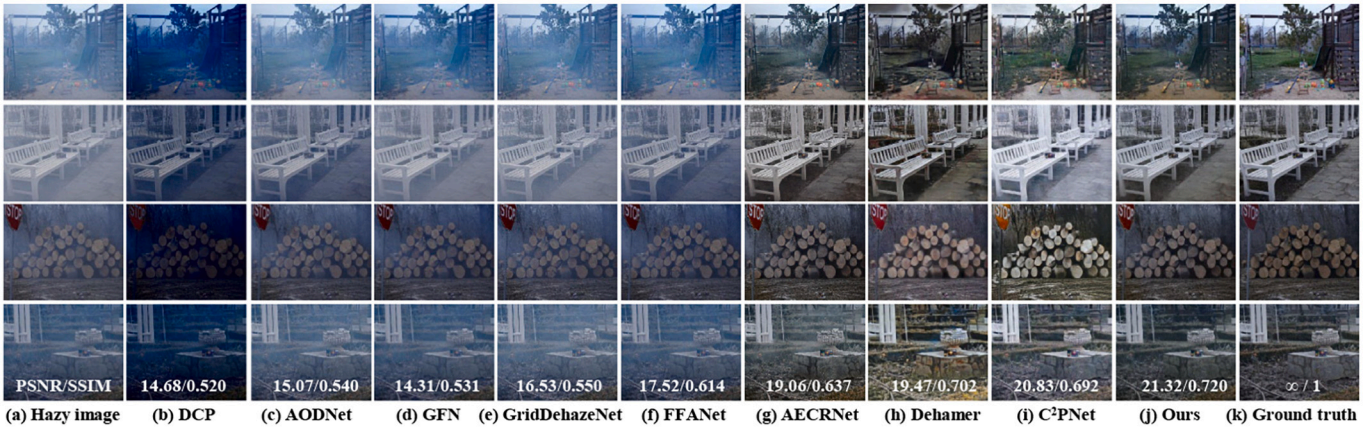

Fig. 10. Qualitative comparison results among WaveletFormerNet and SOTA methods on the real-world fog O-Haze dataset.

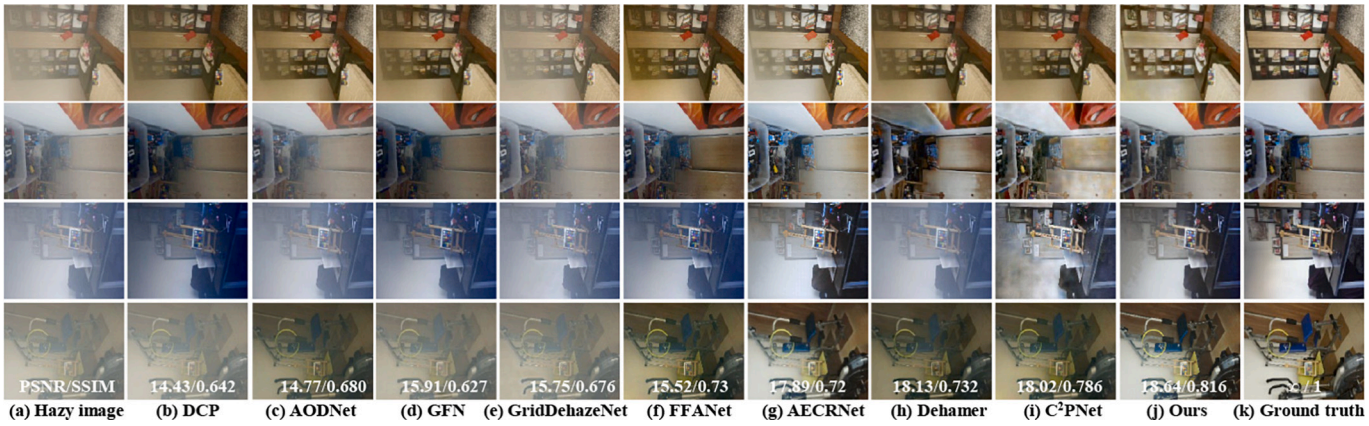

Fig. 11. Qualitative comparison results among WaveletFormerNet and SOTA methods on the real-world fog I-Haze dataset.

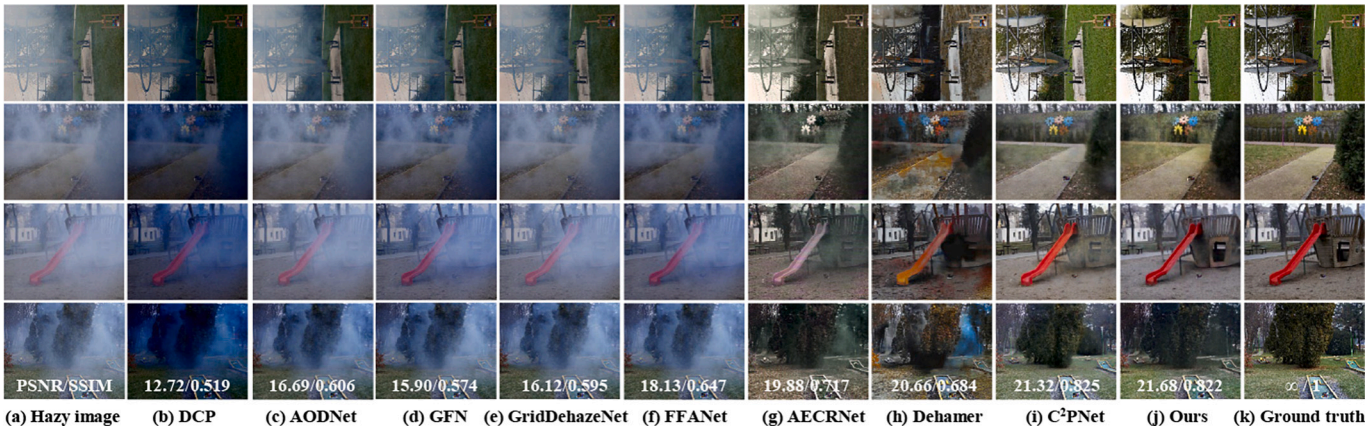

Fig. 12. Qualitative comparison results among WaveletFormerNet and SOTA methods on the real non-homogeneous fog NH-Haze dataset.

5.2. Results analysis on the synthetic dataset

Table 3 and Fig. 14 show that our proposed WaveletFormerNet produces competitive comparison results with SOTA methods on SOTS datasets. Compared to FFANet, DCP and AODNet do not remove fog very thoroughly; Dehamer and DehazeFormer-B are capable of outputting higher quality images; C^2PNet leads all the methods in terms of objective evaluation metrics, but our method produces competitive comparisons with fewer parameters than C^2PNet on the qualitative aspects while maintaining richer details and color information.

Therefore, combining the quantitative and qualitative results from the real-world and synthetic datasets, our WaveletFormerNet provides a more complete balance between dehazing performance and model complexity.

5.3. Parameters and runtime analysis

Table 5 demonstrates that our method has a significant advantage over the SOTA methods in parameters because decomposing the image by wavelet transform in high and low frequency before processing

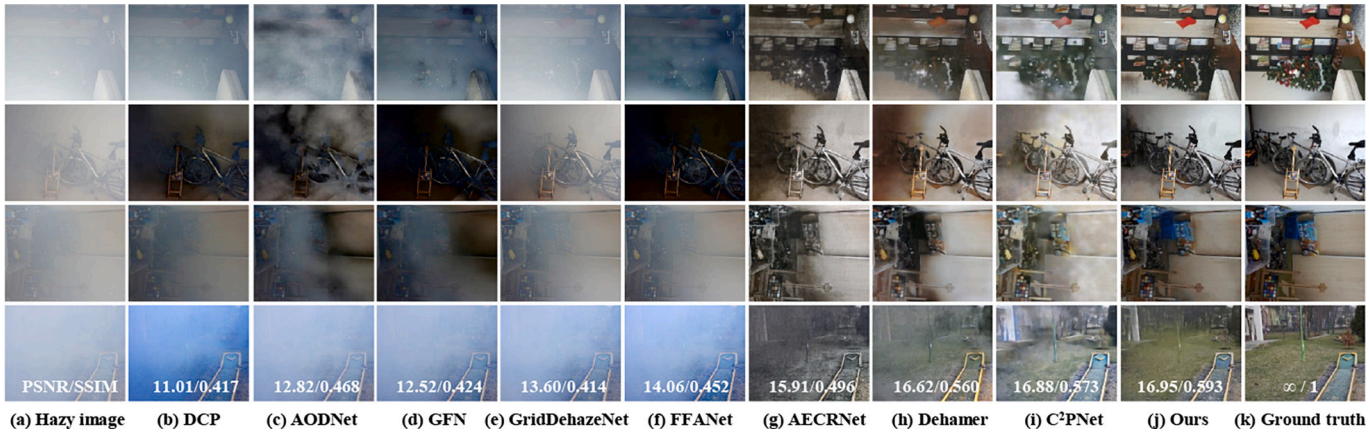

Fig. 13. Qualitative comparison results among WaveletFormerNet and SOTA methods on the real-world dense fog Dense-Haze dataset.

Table 1
Quantitative comparisons on referenced indicators between WaveletFormerNet and SOTA methods on the real-world datasets (I-Haze, O-Haze, Dense-Haze, and NH-Haze).

Methods	O-Haze		I-Haze		Dense-Haze		NH-Haze	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
TAPAMI'10 DCP [21]	14.68	0.520	14.43	0.642	11.01	0.417	12.72	0.519
TIP'16 DehazeNet [8]	14.65	0.510	14.59	0.620	11.84	0.430	14.62	0.523
ICCV'17 AODNet [25]	15.07	0.540	14.77	0.680	12.82	0.468	16.69	0.606
CVPR'18 GFN [43]	14.31	0.531	15.91	0.627	12.52	0.424	15.90	0.574
WACV'19 GCANet [9]	15.71	0.549	15.37	0.659	12.62	0.421	18.79	0.773
ICCV'19 GridDehazeNet [31]	16.53	0.550	15.75	0.676	13.60	0.414	16.12	0.595
CVPR'20 MSBDN [14]	16.69	0.596	15.68	0.711	14.18	0.411	17.54	0.581
AAAI'20 FFANet [39]	17.52	0.614	15.52	0.739	14.06	0.452	18.13	0.647
CVPR'21 AECRNet [53]	19.06	0.637	17.89	0.721	15.91	0.496	19.88	0.717
CVPR'22 DeHamer [19]	19.47	0.702	18.13	0.732	16.62	0.560	20.66	0.684
CVPR'23 C²PNet [58]	20.83	0.692	18.02	0.786	16.88	0.573	21.32	**0.825**
WaveletFormerNet(Ours)	**21.32**	**0.720**	**18.64**	**0.816**	**16.95**	**0.593**	**21.68**	0.822

Indicators marked with ↑ indicate higher and better data; ↓ indicate lower and better. We use **bold** and underline to mark the best and second-best methods.

Table 2
Quantitative comparisons on non-referenced indicators between WaveletFormerNet and SOTA methods on the real-world datasets (I-Haze, O-Haze, Dense-Haze, and NH-Haze).

Methods	O-Haze			I-Haze			NH-Haze			Dense-Haze		
	Entropy↑	NIQE↓	FADE↓	Entropy↑	NIQE↓	FADE↓	Entropy↑	NIQE↓	FADE↓	Entropy↑	NIQE↓	FADE↓
TAPAMI'10 DCP [21]	7.091	4.847	0.913	6.979	4.613	0.917	6.698	4.962	0.932	6.652	5.042	0.949
TIP'16 DehazeNet [8]	7.173	4.773	0.846	7.107	4.494	0.885	6.741	4.544	0.911	6.709	4.831	0.923
ICCV'17 AODNet [25]	7.201	4.676	0.794	7.199	4.505	0.806	6.809	4.198	0.876	6.803	4.686	0.867
CVPR'18 GFN [43]	7.112	4.552	0.783	7.205	4.471	0.757	6.762	4.227	0.714	6.719	4.573	0.828
WACV'19 GCANet [9]	7.213	4.441	0.699	7.198	4.452	0.618	6.843	4.037	0.812	6.732	4.493	0.765
ICCV'19 GridDehazeNet [31]	7.332	4.277	0.711	7.292	4.323	0.685	6.971	3.977	0.649	6.862	4.347	0.721
CVPR'20 MSBDN [14]	7.695	3.321	0.513	7.206	3.464	0.539	7.239	3.808	0.535	7.211	3.931	0.617
AAAI'20 FFANet [39]	7.402	3.975	0.476	7.232	3.354	0.542	7.105	3.792	0.524	7.091	3.773	0.556
CVPR'21 AECRNet [53]	7.575	3.287	0.459	7.531	3.255	0.517	7.172	3.165	0.513	7.188	3.605	0.513
CVPR'22 DeHamer [19]	7.744	2.885	0.442	7.439	2.797	**0.431**	7.483	3.019	0.496	7.323	3.382	0.495
CVPR'23 C²PNet [58]	**7.811**	2.993	**0.429**	7.602	3.067	0.443	7.381	2.942	0.481	7.386	3.441	**0.483**
WaveletFormerNet(Ours)	7.792	**2.711**	0.437	**7.854**	**2.673**	0.445	**7.489**	**2.921**	**0.473**	**7.397**	**3.287**	0.491

Indicators marked with ↑ indicate higher and better data; ↓ indicate lower and better. We use **bold** and underline to mark the best and second-best methods.

reduces the network overhead. However, our approach is not superior in inference time because the wavelet transform takes up some time. However, WaveletFormerNet outperforms the SOTA methods on real-world datasets, which explains a comprehensive view of model complexity and dehazing performance.

5.4. Generality analysis for WaveletFormerNet

To our knowledge, the RB-Dust dataset [7] is the first publicly available agricultural landscape dusting dataset, consisting of 200 images with 1920 × 1080 pixels. Previous work by Peter et al. [7] demonstrated that dust properties are similar to haze properties and that some image dehazing algorithms are also suitable for image dedusting.

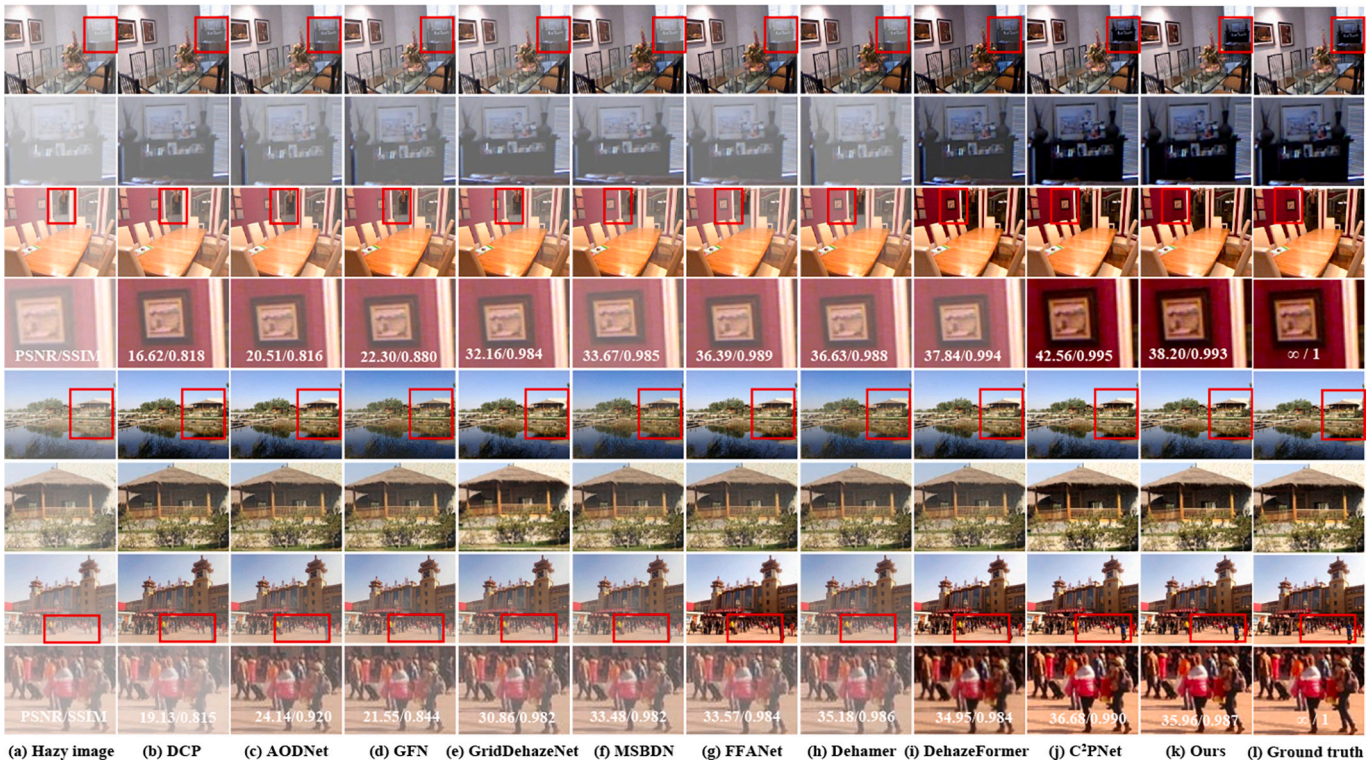

Fig. 14. Qualitative comparison results among WaveletFormerNet and SOTA methods on the synthetic hazy datasets (SOTS-indoor and SOTS-outdoor). The red frame lines represent enlarged details from the original images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

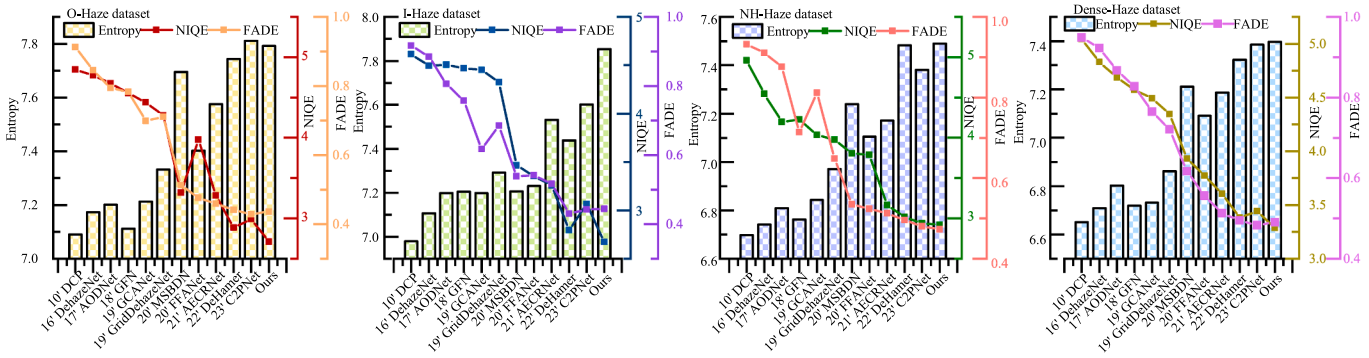

Fig. 15. Quantitative comparisons on non-referenced indicators (Entropy, NIQE, and FADE) results among WaveletFormerNet and SOTA methods on four real-world datasets (O-Haze, I-Haze, NH-Haze, and Dense-Haze).

Therefore, we selected the classical algorithms for image defogging in recent years to compare the generalizability among WaveletFormerNet and other SOTA methods. Fig. 17 shows more qualitative comparison results among WaveletFormerNet and SOTA methods, which illustrates that WaveletFormerNet removes more dense and non-homogeneous dust and retains more textural detail, demonstrating the promising robustness and better generalization ability of the proposed WaveletFormerNet.

5.5. Application test

The SIFT algorithm [33] is used to detect and describe the matching of feature points between different images by extracting the local features of the image. This approach has a wide range of applications in the fields of target recognition and target tracking. In this section, we perform a feature point matching test to evaluate the performance of WaveletFormerNet. We selected the last four years of SOTA methods for comparison. Fig. 18 shows more of the application test results among our method and SOTA methods, and we can observe that WaveletFormerNet has the highest number of matching points. The application test shows that WaveletFormerNet exhibits better performance in computer vision-related applications.

5.6. Ablation study

To verify the effectiveness of the proposed WaveletFormerNet, we performed a structure and loss function ablation study on the NH-Haze dataset in Table 6 and Fig. 19.

5.6.1. Effectiveness of the DWT and IDWT

Variant V1 represents WaveletFormerNet without the wavelet transform and inverse wavelet transform for upsampling and

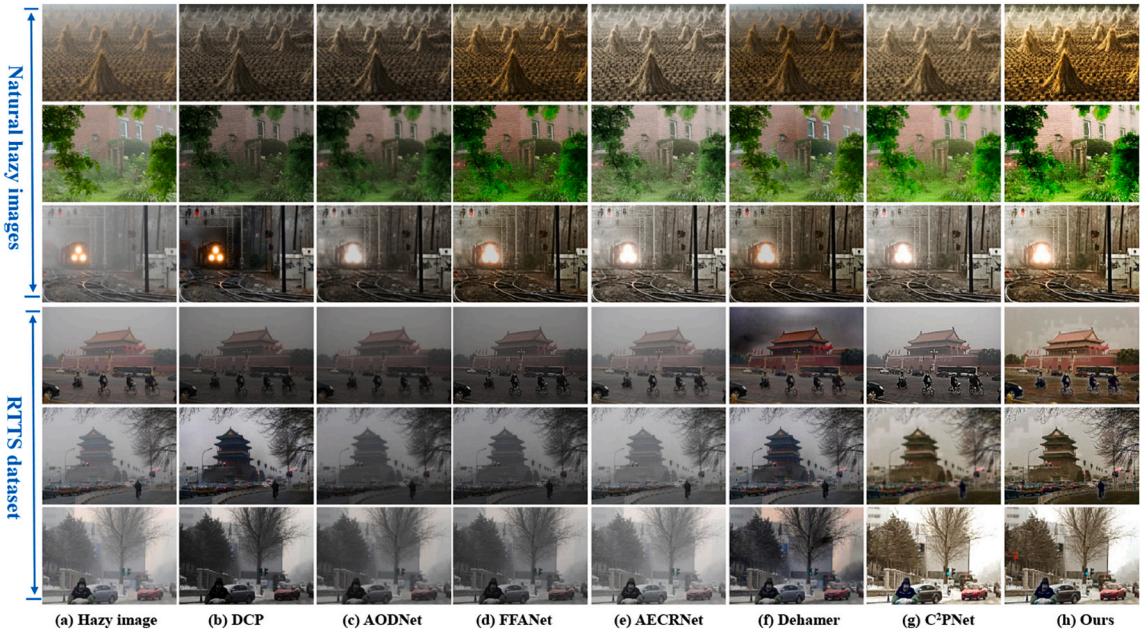

Fig. 16. Qualitative comparison among WaveletFormerNet and SOTA methods on the natural hazy images and RTTS dataset.

Table 3
Quantitative comparisons between WaveletFormerNet and SOTA methods on the synthetic dataset (RESIDE).

Methods	SOTS-Indoor		SOTS-Outdoor		SOTS-Indoor			SOTS-Outdoor		
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	Entropy↑	NIQE↓	FADE↓	Entropy↑	NIQE↓	FADE↓
TAPAMI'10 DCP [21]	16.62	0.818	19.13	0.815	6.407	4.825	0.419	6.312	4.925	0.428
TIP'16 DehazeNet [8]	19.82	0.821	24.75	0.927	6.687	4.761	0.395	6.499	4.892	0.420
ICCV'17 AODNet [25]	20.51	0.816	24.14	0.920	6.856	4.198	0.377	6.794	4.347	0.394
CVPR'18 GFN [43]	22.30	0.880	21.55	0.844	6.497	4.612	0.320	6.403	4.587	0.335
WACV'19 GCANet [9]	30.06	0.960	22.76	0.889	6.594	4.632	0.315	6.471	4.529	0.327
ICCV'19 GridDehazeNet [31]	32.16	0.984	30.86	0.982	7.227	4.327	0.346	6.854	4.433	0.359
CVPR'20 MSBDN [14]	33.67	0.985	33.48	0.982	6.687	4.031	0.309	6.57	4.124	0.328
AAAI'20 FFANet [39]	36.39	0.989	33.57	0.984	6.905	3.802	0.287	6.793	3.914	0.319
CVPR'21 AECRNet [53]	37.17	0.990	–	–	7.290	3.978	0.261	–	–	–
CVPR'22 DeHamer [19]	36.63	0.988	35.18	0.986	7.676	3.42	0.279	7.734	3.608	0.317
TIP'23 Dehazeformer-B [49]	37.84	0.994	34.95	0.989	7.503	**3.364**	0.289	7.455	3.494	**0.307**
CVPR'23 C²PNet [58]	**42.56**	**0.995**	**36.68**	**0.990**	7.893	3.375	**0.275**	7.692	3.506	0.323
WaveletFormerNet(Ours)	38.20	0.993	35.96	0.987	**7.923**	3.403	0.280	**7.766**	**3.488**	0.312

Indicators marked with ↑ indicate higher and better data; ↓ indicate lower and better. We use **bold** and underline to mark the best and second-best methods. Data marked with - is unavailable.

Table 4
Quantitative comparisons on non-referenced indicators between WaveletFormerNet and SOTA methods on the natural hazy images and RTTS dataset.

Methods	Natural hazy images			RTTS dataset		
	Entropy↑	NIQE↓	FADE↓	Entropy↑	NIQE↓	FADE↓
TAPAMI'10 DCP [21]	7.256	5.197	0.628	7.405	6.037	0.659
ICCV'17 AODNet [25]	7.302	5.032	0.617	7.592	5.654	0.643
AAAI'20 FFANet [39]	7.561	4.591	0.549	7.513	5.467	0.618
CVPR'21 AECRNet [53]	7.498	4.783	0.562	7.341	5.239	0.581
CVPR'22 DeHamer [19]	7.543	**4.552**	0.537	7.565	4.867	**0.505**
CVPR'23 C²PNet [58]	7.697	4.587	**0.503**	7.483	4.901	0.523
WaveletFormerNet(Ours)	**7.705**	4.591	0.518	**7.607**	**4.765**	0.537

Indicators marked with ↑ indicate higher and better data; ↓ indicate lower and better. We use **bold** and underline to mark the best and second-best methods.

downsampling. We only use normal convolution for upsampling and downsampling.

Variant V2 represents WaveletFormerNet with the DWT for downsampling.

Variant V3 represents WaveletFormerNet with the IDWT for upsampling.

We can observe from Fig. 19 (a)-(e) that using the wavelet transform instead of upsampling and downsampling can be more effective in extracting texture details and the overall structure of the image from the complex haze background. From Table 6, compared to Variant V1, the

Table 5
We conduct parameters (# Param), floating-point operations (# FLOPs), and inference time as the main metrics of computational efficiency on RGB image with a resolution of 256 × 256 between WaveletFormerNet and SOTA methods.

Methods	Overhead		
	#Param↓	#FLOPs	Runtime↓
ICCV'17 AODNet	**0.002 M**	0.115G	**0.316 ms**
ICCV'19 GridDehazeNet	0.956 M	21.49G	15.35 ms
CVPR'20 MSBDN	31.35 M	24.44G	9.826 ms
AAAI'20 FFANet	4.456 M	287.5G	52.76 ms
CVPR'21 AECRNet	2.611 M	52.20G	–
CVPR'22 DeHamer	132.45 M	48.93G	26.31 ms
TIP'23 DehazeFormer-B	2.514 M	25.79G	19.22 ms
CVPR'23 C^2PNet	7.17 M	429.52G	–
WaveletFormerNet (Ours)	2.26 M	4.08G	16.58 ms

Table 6
Structure and loss function ablation study of WaveletFormerNet on the NH-Haze Dataset.

Variants	NH-Haze dataset				
	$\mathscr{L}_{\ell_1}$	$\mathscr{L}_{\mathbf{MS-SSIM}}$	$\mathscr{L}_{\text{per}}$	PSNR↑	SSIM↑
Variant V1: Normal convolution and Transformer block	✓	✓	✓	18.06	0.687
Variant V2: Transformer block + DWT	✓	✓	✓	18.91	0.703
Variant V3: Transformer block + IDWT	✓	✓	✓	19.45	0.724
Variant V4: Transformer block + DWT + IDWT	✓	✓	✓	20.08	0.767
Variant V5: Variant V4 + parallel convolution	✓	✓	✓	20.22	0.788
Variant V6: Variant V5 + FAM	✓	✓	✓	21.17	0.809
Variant V7: Variant V6 + ASPP Module (Ours)	✓	✓	✓	**21.68**	**0.822**
Ours Variant V7: Variant V6 + ASPP Module	✓	✓	×	21.49	0.817
Ours Variant V7: Variant V6 + ASPP Module	✓	×	×	19.60	0.783

PSNR metric of Variant V4 is improved by 2.02 dB, which again proves the effectiveness of the DWT and IDWT method in our algorithm.

5.6.2. *Effectiveness of the parallel convolution*

We can observe the better visual changes from Fig. 19 (e)-(f), we perform an additional convolution of the features from the DWT and then achieve a dynamic aggregation style of information. The parallel convolution can better brighten the brightness and contrast of the dehazed image.

5.6.3. *Effectiveness of the FAM*

From Fig. 19 (f)-(g), We can intuitively feel that the color and texture of the picture is more natural and rich, and the PSNR indicator is also improved by 0.95 dB compared to Variant V5. FAM, as an effective method of feature aggregation in the encoding and decoding process, makes the network's dehazing performance more significantly improved, revealing the effectiveness of the proposed FAM.

5.6.4. *Effectiveness of the ASPP module*

We use the ASPP Module to obtain features in different receptive fields. From Fig. 19 (g)-(h), the details of the image are more prominent, in addition, as can be seen from Table 6, compared with Variant V6, the PSNR metric of our full model is improved by 0.41 dB, and the overall dehazing effect of the image is also enhanced.

5.6.5. *Effectiveness of loss function*

In addition, we further illustrate the importance of the loss function used in this work. As can be seen by observing in Table 6; each increase in the loss function is effective and crucial for improving PSNR and SSIM. The $\mathscr{L}_{\ell_1}$ loss provides a wider and more stable gradient. The $\mathscr{L}_{\mathbf{MS-SSIM}}$ loss integrates the variations of resolution and visualization conditions to consider structural differences. The $\mathscr{L}_{\text{per}}$ to promote the perceptual similarity of dimensional spatial features and perceive the image from a high-dimension. By integrating all the losses in the training stage, our model acquired a more promising performance.

6. Conclusion and discussion

The paper proposes WaveletFormerNet for real-world single-image dehazing, trading off the dehazing performance, generalization ability, and model complexity. We embed the wavelet transform into the ViT by presenting the WaveletFormer and IWaveletFormer blocks, alleviating structure and texture detail loss during the encoder and decoder. We devise FAM to capture the long-range dependencies among different levels of information and improve decoder efficiency. Extensive experiments demonstrate that our WaveletFormerNet outperforms SOTA methods on real-world fog datasets. Moreover, generality analysis and application tests show that WaveletFormerNet exhibits better generalization capability and superior performance in computer vision-related applications of our method.

Although the proposed WaveletFormerNet brings pleasant results in subjective visualization, we can still see from Fig. 20 that some artifacts are introduced when processing images of dense fog. We believe that is due to the limited real-world dense fog data pairs. On the other hand, the WaveletFormerNet needs to be improved and promoted in the runtime. Real-world data collection is complex, so training on limited data affects

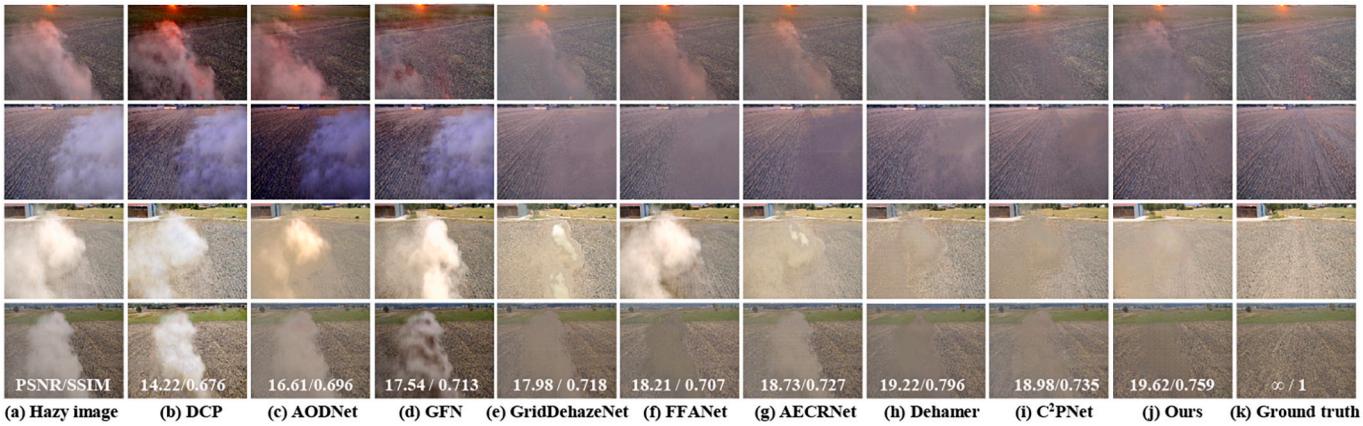

Fig. 17. Qualitative comparison results among WaveletFormerNet and SOTA methods on real-world dust dataset (RB-Dust dataset). The red frame lines represent enlarged details from the original images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

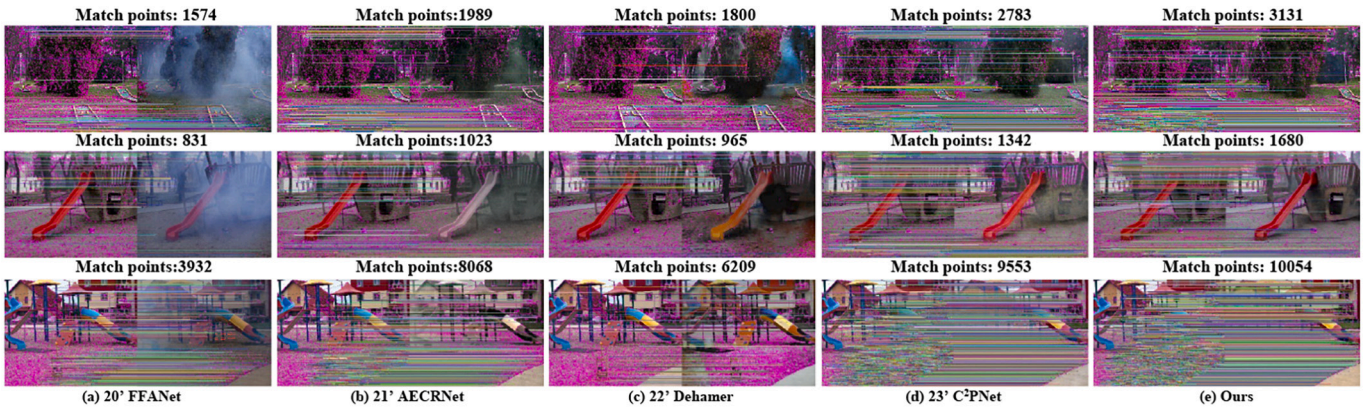

Fig. 18. Application test results of WaveletFormerNet and SOTA methods on NH-Haze dataset. The purple dots represent feature points, and the horizontal lines represent the matching of feature points between the dehazed result by different methods (right one) and a clear reference image (left one); the denser the matching lines are, the higher the degree of feature matching. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

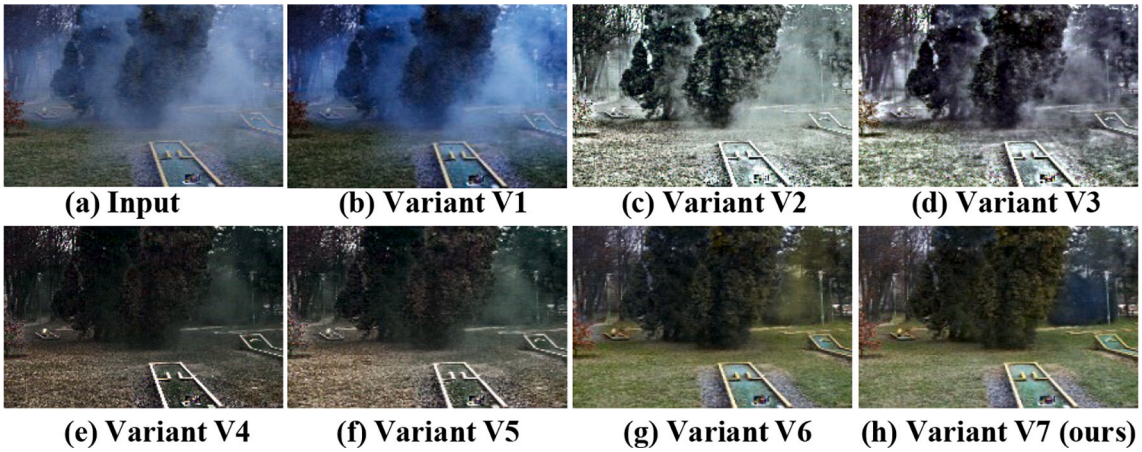

Fig. 19. Visualization comparison of different variants of WaveletFormerNet on the NH-Haze dataset.

Fig. 20. Some output results of WaveletFormerNet on the Dense-Haze dataset may result in artifacts or incomplete defogging. The red frame represents a zoomed-in detail. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the expression capacity of the data-driven models. Therefore, this is our subsequent work to improve our model and introduce more and better quality real-world datasets in the future. Our next phase of improvement also includes aspects such as adjusting the direct strategy, optimizing the structure of the model, and combining it with more advanced computer vision tasks, to improve the current applicability of the model.

CRediT authorship contribution statement

Shengli Zhang: Conceptualization of this study, Methodology, Software, Validation, Data Curation, Formal analysis, Writing - Original Draft, Writing - Revised Draft. **Zhiyong Tao:** Supervision, Writing - Review Editing. **Sen Lin:** Supervision, Writing - Review Editing.

Declaration of competing interest

None.

Data availability

Data will be made available on request.

Acknowledgements

This work was partly supported by the Applied Basic Research Project of Department of Science & Technology of Liaoning province under Grant 2022JH2/101300274 and partly by the Educational Department of Liaoning Province under Grant No. LJKMZ20220679.

References

[1] Usman Ali, Jeongdan Choi, KyoungWook Min, Young-Kyu Choi, Muhammad Tariq Mahmood, Boundary-constrained robust regularization for single image dehazing, Pattern Recogn. 140 (2023) 109522.

[2] Codruta O. Ancuti, Cosmin Ancuti, Mateu Sbert, Radu Timofte, Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 1014–1018.

[3] Codruta O. Ancuti, Cosmin Ancuti, Radu Timofte, Christophe De Vleeschouwer, O-haze: a dehazing benchmark with real hazy and haze-free outdoor images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 754–762.

[4] Codruta O. Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, Ntire 2020 challenge on nonhomogeneous dehazing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 490–491.

[5] Cosmin Ancuti, Codruta O. Ancuti, Radu Timofte, Christophe De Vleeschouwer, I-haze: a dehazing benchmark with real hazy and haze-free indoor images, in: Advanced Concepts for Intelligent Vision Systems: 19th International Conference, ACIVS 2018, Poitiers, France, September 24–27, 2018, Proceedings 19, Springer, 2018, pp. 620–631.

[6] Balasubramanyam Appina, A 'complete blind' no-reference stereoscopic image quality assessment algorithm, in: 2020 International Conference on Signal Processing and Communications (SPCOM), IEEE, 2020, pp. 1–5.

[7] Peter Buckel, Timo Oksanen, Thomas Dietmueller, Rb-dust-a reference-based dataset for vision-based dust removal, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1140–1149.

[8] Bolun Cai, Xu Xiangmin, Kui Jia, Chunmei Qing, Dacheng Tao, Dehazenet: An end-to-end system for single image haze removal, IEEE Trans. Image Process. 25 (11) (2016) 5187–5198.

[9] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Lu Dongdong Hou, Yuan, and Gang Hua., Gated context aggregation network for image dehazing and deraining, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019, pp. 1375–1383.

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, Hartwig Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.

[11] Lark Kwon Choi, Jaehee You, Alan Conrad Bovik, Referenceless prediction of perceptual fog density and perceptual image defogging, IEEE Trans. Image Process. 24 (11) (2015) 3888–3901.

[12] R.L. Claypoole, R.G. Baraniuk, R.D. Nowak, Adaptive wavelet transforms via lifting, in: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181) vol. 3, 1998, pp. 1513–1516.

[13] Sourya Dipta Das and Saikat Dutta. Fast deep multi-patch hierarchical network for nonhomogeneous image dehazing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 482–483.

[14] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, Ming-Hsuan Yang, Multi-scale boosted dehazing network with dense feature fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2157–2167.

[15] Yu Dong, Yihao Liu, He Zhang, Shifeng Chen, Yu Qiao, Fd-gan: Generative adversarial networks with fusion-discriminator for single image dehazing, in: Proceedings of the AAAI Conference on Artificial Intelligence 34, 2020, pp. 10729–10736.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al.. An image is worth 16x16 words: Transformers for image recognition at scale, ICLR, 2021.

[17] Raanan Fattal, Single image dehazing, ACM Trans. Graph. (TOG) 27 (3) (2008) 1–9.

[18] Minghan Fu, Huan Liu, Yankun Yu, Jun Chen, Keyan Wang, Dw-gan: A discrete wavelet transform gan for nonhomogeneous dehazing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 203–212.

[19] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, Chongyi Li, Image dehazing transformer with transmission-aware 3d position embedding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5812–5820.

[20] Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, Vishal Monga, Deep wavelet prediction for image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 104–113.

[21] Kaiming He, Jian Sun, Xiaoou Tang, Single image haze removal using dark channel prior, IEEE Trans. Pattern Anal. Mach. Intell. 33 (12) (2010) 2341–2353.

[22] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, Mu Li, Bag of tricks for image classification with convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 558–567.

[23] Kui Jiang, Zhongyuan Wang, Peng Yi, Junjun Jiang, Jing Xiao, Yuan Yao, Deep distillation recursive network for remote sensing imagery super-resolution, Remote Sens. 10 (11) (2018) 1700.

[24] Apurva Kumari, Subhendu Kumar Sahoo, A new fast and efficient dehazing and defogging algorithm for single remote sensing images, Signal Process. 215 (2024) 109289.

[25] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, Dan Feng, Aod-net: All-in-one dehazing network, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4770–4778.

[26] Boyi Li, Wenqi Ren, Fu Dengpan, Dacheng Tao, Dan Feng, Wenjun Zeng, Zhangyang Wang, Benchmarking single-image dehazing and beyond, IEEE Trans. Image Process. 28 (1) (2018) 492–505.

[27] Zhengguo Li, Chaobing Zheng, Haiyan Shu, Wu. Shiqian, Dual-scale single image dehazing via neural augmentation, IEEE Trans. Image Process. 31 (2022) 6213–6223.

[28] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, Wangmeng Zuo, Multi-level wavelet-cnn for image restoration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 773–782.

[29] Shiben Liu, Huijie Fan, Sen Lin, Qiang Wang, Naida Ding, Yandong Tang, Adaptive learning attention network for underwater image enhancement, IEEE Robot. Autom. Lett. 7 (2) (2022) 5326–5333.

[30] Wei Liu, Qiong Yan, Yuzhi Zhao, Densely self-guided wavelet network for image denoising, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 432–433.

[31] Xiaohong Liu, Yongrui Ma, Zhihao Shi, Jun Chen, Griddehazenet: Attention-based multi-scale network for image dehazing, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7314–7323.

[32] Ze Liu, Yutong Lin, Yue Cao, Hu Han, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[33] David G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.

[34] Earl J. McCartney, Optics of the Atmosphere: Scattering by Molecules and Particles, New York, 1976.

[35] Aditya Mehta, Harsh Sinha, Murari Mandal, Pratik Narang, Domain-aware unsupervised hyperspectral reconstruction for aerial image dehazing, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 413–422.

[36] Anish Mittal, Rajiv Soundararajan, Alan C. Bovik, Making a "completely blind" image quality analyzer, IEEE Sign. Proc. Lett. 20 (3) (2012) 209–212.

[37] Srinivasa G. Narasimhan, Shree K. Nayar, Vision and the atmosphere, Int. J. Comput. Vis. 48 (3) (2002) 233.

[38] Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, Luc Soler, U-net transformer: Self and cross attention for medical image segmentation, in: Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12, Springer, 2021, pp. 267–276.

[39] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, Huizhu Jia, Ffa-net: Feature fusion attention network for single image dehazing, in: Proceedings of the AAAI Conference on Artificial Intelligence 34, 2020, pp. 11908–11915.

[40] Pejman Rasti, Tonis Uiboupin, Sergio Escalera, Gholamreza Anbarjafari, Convolutional neural network super resolution for face recognition in surveillance monitoring, in: Articulated Motion and Deformable Objects: 9th International Conference, AMDO 2016, Palma de Mallorca, Spain, July 13–15, 2016, Proceedings 9, Springer, 2016, pp. 175–184.

[41] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, Ming-Hsuan Yang, Single image dehazing via multi-scale convolutional neural networks, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14, Springer, 2016, pp. 154–169.

[42] Wenqi Ren, Siwei Liu, Huaizu Zhang, Jian Pan, Xuelong Cao, Ming-Hsuan Yang, Learning to see through fog, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1–9.

[43] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, Ming-Hsuan Yang, Gated fusion network for single image dehazing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3253–3261.

[44] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[45] Olga Russakovsky, Jia Deng, Su Hao, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.

[46] Huanfeng Shen, Chi Zhang, Huifang Li, Quan Yuan, Liangpei Zhang, A spatial–spectral adaptive haze removal method for visible remote sensing images, IEEE Trans. Geosci. Remote Sens. 58 (9) (2020) 6168–6180.

[47] Karen Simonyan, Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, ICLR, 2015.

[48] Ayush Singh, Ajay Bhave, Dilip K. Prasad, Single image dehazing for a variety of haze scenarios using back projected pyramid network, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, Springer, 2020, pp. 166–181.

[49] Yuda Song, Zhuqing He, Hui Qian, Xin Du, Vision transformers for single image dehazing, IEEE Trans. Image Process. 32 (2023) 1927–1941.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30, 2017.

[51] Pengyu Wang, Hongqing Zhu, Hui Huang, Han Zhang, Nan Wang, Tms-gan: A twofold multi-scale generative adversarial network for single image dehazing, IEEE Trans. Circuits Syst. Video Technol. 32 (5) (2022) 2760–2772.
[52] Zhongyuan Wang, Peng Yi, Kui Jiang, Junjun Jiang, Zhen Han, Lu Tao, Jiayi Ma, Multi-memory convolutional neural network for video super-resolution, IEEE Trans. Image Process. 28 (5) (2018) 2530–2544.
[53] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, Lizhuang Ma, Contrastive learning for compact single image dehazing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10551–10560.
[54] Hao-Hsiang Yang, Fu. Yanwei, Wavelet u-net and the chromatic adaptation transform for single image dehazing, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 2736–2740.
[55] Hao-Hsiang Yang, Chao-Han Huck Yang, Yi-Chang James Tsai, Y-net: Multi-scale feature aggregation network with wavelet structure similarity loss function for single image dehazing, in: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 2628–2632.
[56] Guoqing Zhang, Wenxuan Fang, Yuhui Zheng, Ruili Wang, Sdbad-net: A spatial dual-branch attention dehazing network based on meta-former paradigm, in: IEEE Transactions on Circuits and Systems for Video Technology, 2023.
[57] Hang Zhao, Orazio Gallo, Iuri Frosio, Jan Kautz, Loss functions for image restoration with neural networks, IEEE Trans. Comp. Imag. 3 (1) (2016) 47–57.
[58] Yu Zheng, Jiahui Zhan, Shengfeng He, Junyu Dong, Yong Du, Curricular contrastive regularization for physics-aware single image dehazing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5785–5794.
[59] Chu Zhou, Minggui Teng, Yufei Han, Xu Chao, Boxin Shi, Learning to dehaze with polarization, in: Advances in Neural Information Processing Systems 34, 2021, pp. 11487–11500.
[60] Qingsong Zhu, Jiaming Mai, Ling Shao, A fast single image haze removal algorithm using color attenuation prior, IEEE Trans. Image Process. 24 (11) (2015) 3522–3533.
[61] Wenbin Zou, Mingchao Jiang, Yunchen Zhang, Liang Chen, Zhiyong Lu, Yi Wu, Sdwnet: A straight dilated network with wavelet transformation for image deblurring, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1895–1904.