Interaction-Guided Two-Branch Image Dehazing Network

Huichun Liu[1], Xiaosong Li[1,⋆], and Tianshu Tan[2]

[1] School of Physics and Optoelectronic Engineering, Foshan University, Foshan 528225, China
[2] Hong Kong University of Science and Technology, Hong Kong, China
Feecuin@outlook.com, lixiaosong@buaa.edu.cn, ttanad@connect.ust.hk

Abstract. Image dehazing aims to restore clean images from hazy ones. Convolutional Neural Networks (CNNs) and Transformers have demonstrated exceptional performance in local and global feature extraction, respectively, and currently represent the two mainstream frameworks in image dehazing. **In this paper, we propose a novel dual-branch image dehazing framework that guides CNN and Transformer components interactively.** We reconsider the complementary characteristics of CNNs and Transformers by leveraging the differential relationships between global and local features for interactive guidance. This approach enables the capture of local feature positions through global attention maps, allowing the CNN to focus solely on feature information at effective positions. The single-branch Transformer design ensures the network's global information recovery capability. Extensive experiments demonstrate that our proposed method yields competitive qualitative and quantitative evaluation performance on both synthetic and real public datasets. Codes are available at https://github.com/Feecuin/Two-Branch-Dehazing

Keywords: Image dehazing · CNN · Transformer · Interaction-guided network

1 Introduction

Haze is caused by small particles in the atmosphere scattering light, which reduces the visibility of objects and leads to a decline in the performance of visual systems in practical tasks such as autonomous driving, object detection, and drone aerial photography. Image dehazing technology[25, 30, 21, 20] can eliminate the influence of haze, restore scene visibility, and provide high-quality images to visual systems. In early image dehazing techniques, the relationship between hazy and clean images was typically described by the following model[18]:

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

⋆ Corresponding Author

where $\hat{x}$ is the 2D spatial location, I is the captured hazy image, J is the clean image, A is the global atmospheric light, and $t(\hat{x})$ is the transmission map, which is expressed as

$$t(x) = e^{-\beta d(\hat{x})}, \tag{2}$$

the transmission map $t(\hat{x})$ depends on the depth of the scene $d(\hat{x})$ and the haze density coefficient β. According to this model formula, many priors[8, 32, 3] were proposed in the early stages to constrain the ill-posedness it brings. However, such prior-based approaches rely on empirical knowledge and are difficult to adapt to different scenarios, and may produce artifacts in areas where priors are not satisfied. With the rise of deep-learning-based dehazing methods, many CNN-based dehazing algorithms have emerged[6, 10, 4], which can achieve better performance than prior-based methods. However, the convolutional mechanisms of CNNs determines that they are limited by smaller receptive fields.

A pure convolutional model causes a network to focus excessively on the local features of an image (e.g., edges and texture information), which is not conducive to the overall restoration of the image. Recently developed Transformer models[17] have achieved superior global feature extraction capabilities compared to CNN on various computer vision tasks. Its attention mechanism ensures that it has good global feature extraction capabilities. However, Transformers often lead to unwanted blurring and rough details during image reconstruction. Existing method[7] do not consider the feature correlation between CNN and Transformer, resulting in feature redundancy. **To combine the advantages of CNN and Transformer for feature extraction**, we propose an interactive guidance method that utilizes the ability of a Transformer to extract global features to provide accurate global information and guide a CNN to focus on detailed information within an effective feature space.

CNN are able to identify most of the useful details of images, and considering that Transformer will intersect with CNN during feature extraction, adding features directly will lead to information redundancy. Therefore, the down-sampling operation is introduced into the Transformer branch to distinguish the features extracted by the two branches as much as possible, avoid redundancy caused by repeated extraction, and improve the performance of the model. The details lost by downsampling will be compensated in the CNN branch. The proposed method makes full use of the complementary advantages of CNN and transformer to provide high-quality dehazing results with limited computing resources. In summary, our contributions are as follows:

- We propose an interaction-guided dual-branch image dehazing framework that utilizes the global information provided by the Transformer to guide the CNN to focus on local details effectively, while the Transformer branch ensures the ability of the network to recover global information.

- Our method effectively reduces the redundant information generated by the repeated feature extraction of Transformer and CNN, thereby improving the performance of the two-branch model.

– Extensive experiments on existing synthetic and real datasets consistently confirm the superiority of our model, and the ablation experiments demonstrate the effectiveness of each module.

2 Related Works

Single Image Dehazing. Image dehazing has always been a challenging and important task in computer vision and image processing. It can restore hazy images to clean images and has important applications in many fields. In recent years, many image dehazing methods[8, 32, 3] have been proposed, and early methods generally considered the effects of particle scattering in the atmosphere. They attempted to derive the parameters of atmospheric scattering using mathematical formulas, but these manually derived priors such as dark channel priors[8], color attenuation priors[32], and non-local priors[3] were derived based on empirical knowledge and are often difficult to adapt to diverse scenarios. When the scene does not meet these priors, these prior-based dehazing algorithms often output some results that do not meet expectations. For example, the dark channel prior[8] cannot handle the sky region, which can lead to image distortion. The saturation line prior (SLP)[14] reveals the linear relationship between the inverse of the saturation component and the brightness component in the local pixels of the normalized hazy image, and proposes a novel image dehazing framework to exploit the linear distribution of local pixels, which helps to improve the transfer estimation for better detail restoration and color preservation. In recent years, many deep learning-based methods for image dehazing have emerged, and these methods have gradually become mainstream. Early deep learning methods were still related to Atmospheric Scattering Model(ASM), for example, the CNN model DehazeNet[4] aims to estimate the transmission map t, and then substitute the estimated transmission map into the ASM for calculation to obtain dehazed images. DehazeNet represents a pioneering effort in image dehazing.

Afterwards, AOD-Net[11] simultaneously estimates the transmission map t and atmospheric light A, and obtains the restored haze-free image through ASM. However, methods based on prior estimation often have some bias. Many recently proposed deep learning models do not require parameter estimation to be substituted into ASM for computation. Instead, these models directly restore blurry and hazy images to clean images. For example, GridDehazeNet[16] proposed that learning to recover images is better than directly estimating t, and designed attention-based multi-scale estimators to achieve dehazing. FFA-Net[19] introduced feature attention (FA) blocks, utilizing pixel and channel attention to improve the model's dehazing performance. MSBDN[6] skillfully combines the enhancement strategy and back-projection technology for image dehazing, and proposes a multi-scale enhanced dehazing network. SG-Net[9] proposes a novel end-to-end network to restore haze-free images, and the simple and efficient SG mechanism can be embedded into existing network families at will, with only a little extra time consumption while improving accuracy. In the early days, CNN dominated most computer vision tasks. In recent years, Transformer[26, 22] has

been widely used in computer vision, such as object detection, image segmentation, and other tasks.

Recently, many Transformer-based deep learning methods have emerged for image restoration[7, 28, 24, 13, 29]. For example, Restormer[28] designed multi-head attention and feedforward networks to enable models to capture remote pixel interactions, and DehazeFormer[21] adds reflection filling to the sliding window mechanism to reduce the loss of edge information in Swin Transformer[17] when processing hazy images. Fourmer[31] adds a Fourier-based general image degradation prior to the core structure of Fourier spatial modeling and Fourier channel evolution, which provides new insights into the design of image inpainting based on global modeling.

In recent years, some image dehazing algorithms with a two-branch structure have emerged, combining the advantages of CNN and Transformer to improve their dehazing performance. However, none of these two-branch dehazing algorithms considers the relationship between the features in the two-branch setup, and do not design the network in terms of feature differences and feature redundancy caused by repeated extraction. In contrast, we scrutinize the complementary properties between the two, integrate their strengths to enhance feature extraction, and also utilize the different relationships between features for interactive guidance to improve network performance. As shown in Fig.1, from left to right are the hazy image, and the dehazing results of SG-Net[9], Dehazeformer[21] and the proposed method, and the rightmost is the ground truth image. It can be seen that the image details recovered by SG-Net are clear, but the global consistency is poor, the overall tone and brightness are not natural enough, and there is a loss of details. The overall results appears fuzzy; Dehazeformer is superior to SG-Net in terms of global features and hue, and the overall image is harmonious, but its local details processing is weak, and some residual noise remains. Our method combines the advantages of CNN and Transformer model, and the recovered image not only maintains global consistency and naturalness but also handles local details carefully. This results in an overall high-definition image with significantly reduced noise.

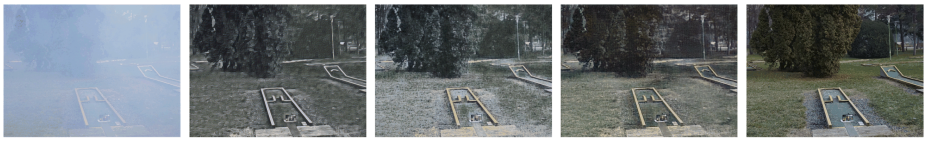

Fig. 1. (a) hazy image, (b) and (c) represent the dehazing results of CNN method and the Transformer method, respectively, (d) Dehazing result of our method, and (e) groundtruth image.

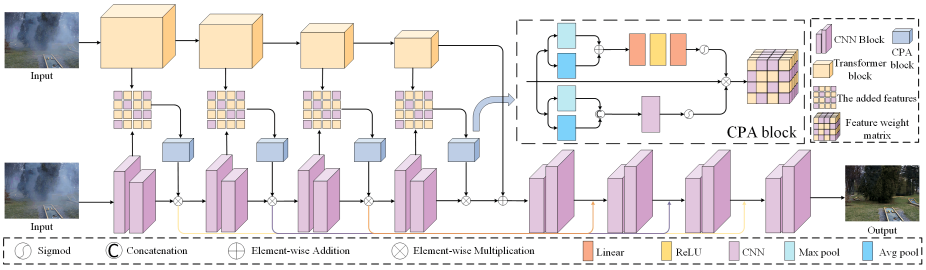

Fig. 2. Our method utilizes Transformer and CNN branches, where the output features in the middle of each layer are utilized in a CPA generated attention map to guide the CNN. The CNN results are combined with the results of the Transformer branch to perform CNN upsampling to recover image details.

3 Proposed Method

An overview of the proposed framework is presented in Fig.2. A blurred image is input into two branches, namely a CNN and a Transformer, to extract local and global features, respectively. The global information extracted by each layer of the Transformer is used to guide the CNN to focus on effective local information. Downsampling is performed on the blurred image before it enters the Transformer block to expand its receptive field and reduce computational complexity. The image details are then restored by the CNN decoder, and a final haze-free image is obtained. Various experiments have demonstrated that our network provides good performance on real datasets.

Network Structure

Global Perception Module. We utilize a DehazeFormer module to extract global features, and its improvements in normalization layer and spatial aggregation scheme make it more efficient in dehazing than the original Swin-Transformer. The normalization layer preserves the mean and standard deviation of the original image, ensuring that the restored image has the same contrast and brightness as those of image. The normalization layer is represented as follows:

$$y = F\left(\frac{x-\mu}{\sigma}\gamma + \beta\right) \cdot (\sigma W_\gamma + B_\gamma) + (\mu W_\beta + B_\beta), \tag{3}$$

the feature map $x \in \mathbb{R}^{b\times n\times c}$,where $n = h \times w$ (i.e. height and width), μ and σ represents the mean and standard deviation, respectively; γ and β denote scaling factors and biases, respectively; W_β and $W_\gamma$and B_β and B_γ are the weights and bias used for transformation μ and σ, respectively.

The sliding window mechanism of DehazeFormer utilizes reflection filling to ensure that the size of the edge window is the same as that of the set window to prevent missing edge information in an image, which can improve the

performance of the network. Additionally, DehazeFormer adds a layer of convolution after ordinary aggregation because the attention mechanism aggregates information within the edge window while ignoring information between windows. Therefore, convolutional operations can be used to aggregate information between neighboring windows, which is represented as follows:

$$\text{Aggregation}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right) + \text{Conv}(\hat{V}), \tag{4}$$

where $\hat{V} \in \mathbb{R}^{b \times h \times w \times c}$, is V prior to window division.

In the proposed method, we use four transformer layers to extract global features, with a downsampling layer preceding the transformer input. The receptive field of the Transformer branch is increased, and the redundant information generated by the repeated feature extraction of the Transformer and CNN branches is reduced, thereby improving the model's performance.

Local Perception Module. To achieve the function of extracting local features, we introduce CNN as another branch of the model. Due to the fact that CNN is mainly based on local perception, its convolutional mechanism also determines that it has a smaller receptive field, which can effectively extract local detail information. However, pure convolutional models can lead to the excessive extraction of image edges and texture details, which can lead to a decrease in model performance and an increase in computational costs. Therefore, we propose leveraging the advantages of the Transformer to extract global features to guide the CNN in local feature extraction. We use a four-layer CNN corresponding to the Transformer structure, where each layer has two outputs. One output is added to the output of the Transformer, and then an attention map is generated by the channel and pixel attention block (CPA). The attention map is then multiplied by the other output of the current CNN layer, and the results are input into the next layer of the CNN to provide global guidance for local extraction, thereby helping the CNN extract detail information more effectively. In the decoder, we utilize four convolution layers, corresponding to the number of layers in the Transformer and CNN branches.

Image upsampling is performed by the decoder to restore detail information. Skip connections are made with the CNN in the dual branches to help preserve the details of the original image. These connections also serve to avoid the loss of detail information caused by network training and improve overall network performance. Finally, this process outputs a clean and haze-free image.

Channel and pixel attention. We extracted the features between each layer of the dual branches and add them together to obtain effective information regarding the entire image space to guide the CNN. After these features pass through the CPA block, a weight matrix is obtained to guide the CNN. The CPA block comprises channel attention[23] and pixel attention[15], with x and y as input and output, respectively. The channel attention formula is defined as follows:

$$\text{channel}_{\text{att}} = \text{Sigmoid}\left(\text{FC}(\text{avgpool}(x), \text{maxpool}(x))\right), \tag{5}$$

$\text{channel}_{\text{att}}$ is the channel attention weight of the feature map. Channel attention models the global information of the entire channel, capturing the global impor-

tance of each channel in the input feature map, reducing redundant information, and improving the attention paid to key features. The pixel attention formula is defined as follows:

$$\text{pixel}_{\text{att}} = \text{Sigmoid}\left(\text{conv1}(\text{mean}(x), \max(x))\right), \tag{6}$$

$\text{pixel}_{\text{att}}$ is the pixel attention weight of the feature map. Pixel attention models the local information at each pixel location and can capture the importance of each location in the input feature map, enhancing the model's perception of the local features of the input image, with enhanced attention to detail information.

$$y = \text{channel}_{\text{att}} \times \text{pixel}_{\text{att}} \times x, \tag{7}$$

The combination of channel attention and pixel attention leverages their respective advantages, such that the model can pay attention to both channels and pixels at the same time, and has the ability to extract both local and global features, which can better guide the CNN to capture information in the effective feature space, thus improving the performance of the model.

4 Experiment

In this section, we conducted extensive experiments to demonstrate the effectiveness of our proposed method. Firstly, we introduce the experimental setup, and then compare it with advanced dehazing methods on both synthetic and real datasets. Additionally, ablation studies are presented to demonstrate the effectiveness of each component of the proposed model.

4.1 Experimental Setups

Training Detail. The proposed method was implemented in Python on an Nvidia GeForce RTX 3090 GPU. We used the Adam optimizer with default parameters to optimize our algorithm, setting the initial learning rate to 0.0001 and using only L1 loss. We use randomly cropped image blocks for training, gradually scaling up the size of the image blocks from 128× 128 to the full size during training.

Dataset. Experiments were conducted on the synthetic dataset RESIDE-6K[12], the real dataset NH-HAZE[2], and DENSE-HAZE[1]. RESIDE-6K is a mixed dataset of indoor and outdoor images on RESIDE, hence it is called SOTS mix. Its training set includes 3000 OTS image pairs and 3000 ITS image pairs, and its test set is also divided into mixed indoor and outdoor image pairs, with a total of 1000 image pairs combined. The DENSE-HAZE dataset and NH-HAZE dataset are both composed of 45 training images, five validation images, and five test images. The haze of DENSE-HAZE is dense and uniform, while the haze of NH-HAZE is dense and uneven.

Compared Methods and Metrics. To demonstrate the effectiveness of our method, we compare it with GridDehazeNet[16], FFA-Net[19], MSBDN[6], PSD[5],

SG-Net[9], D4[27], SLP[14], Dehazeformer[21] (Dehazeformer-T variant) and fourmer[31] on synthetic and real data. If no pretrained model is provided, we retrain the model using the authors' code. Otherwise, we evaluate them using their online code for a fair comparison. All of these representative methods are selected for visual comparison. For the quantitative evaluation of image quality assessment, we use the commonly used PSNR, SSIM, Entropy, and LPIPS to compare the performance of each method.

4.2 Experimental Results

Results on the Synthetic Datasets. We first test on the synthetic hazy image dataset RESIDE-6K. The images contained in RESIDE-6K can be divided into two types: indoor and outdoor, and the dehazing results of different methods are shown in Fig.3 and 4.

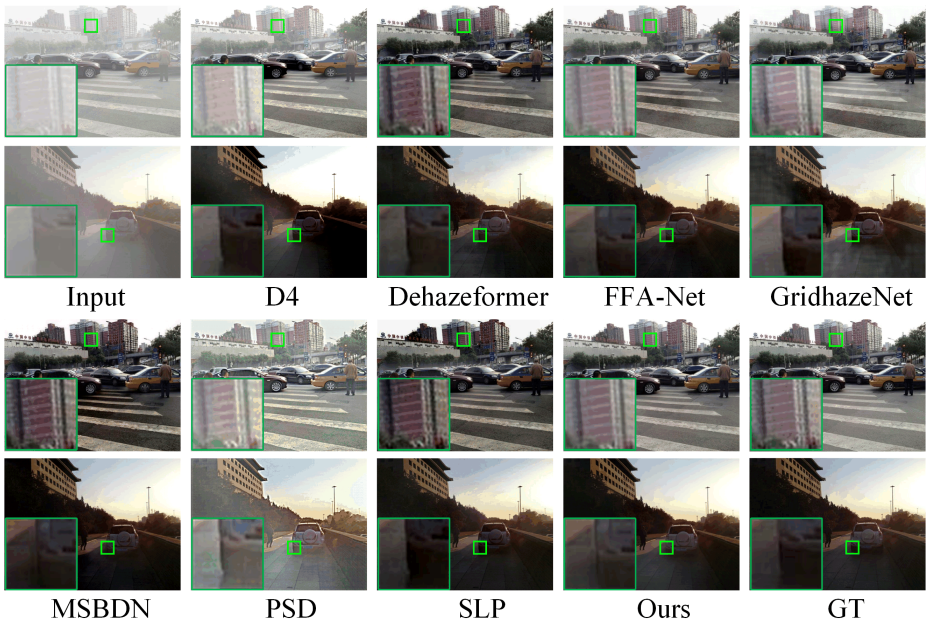

Fig. 3. Visual comparison of outdoor scenes of different dehazing methods on RESIDE-6K dataset. (Zooming in can obtain a clearer view)

As shown in Fig.3, there are significant differences between the restored images and the ground truth images for all comparison methods in the outdoor scene. There are residual haze and some undesired details (artifacts, blur, etc.), such as D4,GridDehazeNet and PSD, and most methods have color bias, such as MSBDN and SLP. As shown in Fig.4, for indoor scenes, almost all images

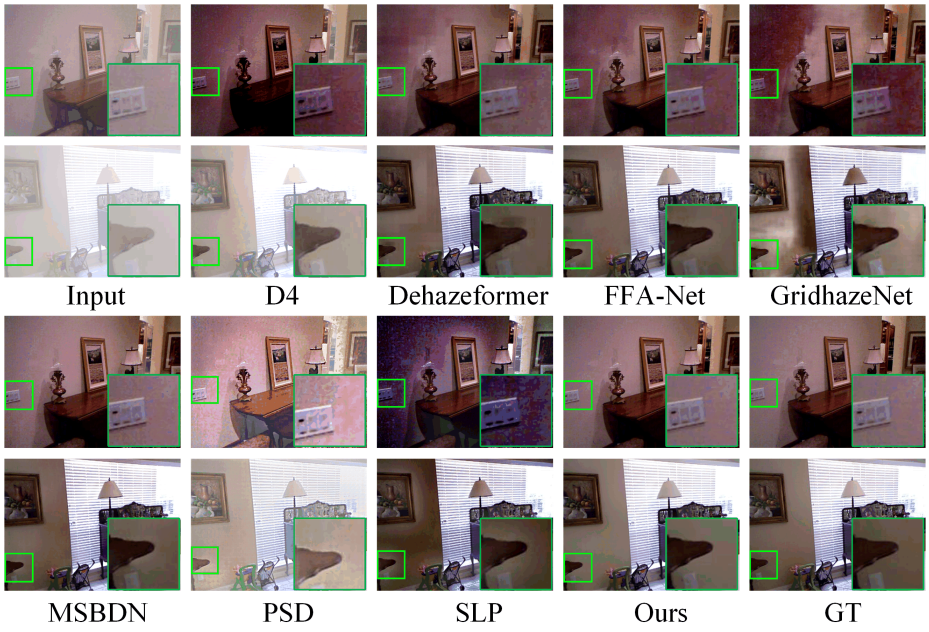

Fig. 4. Visual comparison of indoor scenes of different dehazing methods on RESIDE-6K dataset. (Zooming in can obtain a clearer view)

of FFA-Net, SLP, and Dehazeformer have problems such as color deviation (red or purple), blurred details, and low contrast. In contrast, for both outdoor and indoor scenes, the images generated by our method are closest to real haze-free images in terms of details as well as overall image tone.

Table 1 presents a comparison of the quantitative measures obtained on the synthetic datasets. Table 1 shows that our method is second only to Dehazeformer in PSNR and SSIM in the RESIDE-6K test set, and also shows good performance in Entropy and LPIPS. It can be seen that the experimental results show that the proposed method has excellent dehazing performance on the synthetic dataset.

Results on the Real Datasets. To further verify the dehazing ability of our method in real scenarios, we tested various methods on the uniform haze dataset DENSE-HAZE and the nonuniform haze NH-HAZE. The dehazing results are shown in Fig.5 and 6.

As shown in Fig.5, the NH-HAZE dataset test shows that D4, PSD and SLP still have obvious haze, and methods such as GridDehazeNet, MSBDN and Dehazeformer have some problems of details blur and distortion. As shown in Fig.6, in the DENSE-HAZE dataset test, D4, PSD and SLP still have a lot of residual haze, while other methods have serious color deviation and relatively obvious noise. Overall, in terms of subjective evaluation, the comparison methods produce some problems such as blur, distortion or noise in both NH-HAZE

Table 1. Quantitative comparison of the proposed algorithm and different comparison methods on the RESIDE-6K dataset. **Bold** is the best, Red is the second.

Methods	Venue& Year	RESIDE-6K			
		$PSNR\uparrow$	$SSIM\uparrow$	$Entropy\uparrow$	$LPIPS\downarrow$
GridDehazeNet [16]	*ICCV'19*	25.65	0.9371	7.4597	0.1915
FFA-Net [19]	*AAAI'20*	27.26	0.9567	7.4151	0.1757
MSBDN [6]	*CVPR'20*	27.44	0.9511	7.4309	0.1751
PSD [5]	*CVPR'21*	15.47	0.8149	**7.4673**	0.1697
SG-Net [9]	*ACCV'22*	-	-	-	-
D4 [27]	*CVPR 22*	18.97	0.8422	6.5999	0.1635
SLP [14]	*TIP'23*	21.35	0.9261	7.3925	0.0978
Dehazeformer-T [21]	*TIP'23*	**30.36**	**0.9730**	7.4419	**0.0303**
Fourmer[31]	*ICML'23*	-	-	-	-
Ours	-	30.20	0.9643	7.4325	0.1749

and DENSE-HAZE scenes. In contrast, the images recovered by our method are closest to the real haze-free images in terms of both dehazing and color restoration, so our method has the best dehazing performance in subjective evaluation.

Tables 2 and 3 present a comparison of the quantitative measures obtained on the real datasets NH-HAZE and DENSE-HAZE. As can be seen from Table 2, in the DENSE-HAZE dataset, our method has the best values in PSNR and SSIM, and Entropy and LPIPS are only 0.133 and 0.0237 lower than Dehazeformer-T. Table 3 shows that in the NH-HAZE dataset, our method has the best values in PSNR, Entropy and LPIPS, and SSIM is only 0.0498 lower than Fourmer. The experimental results show that our method also has the best performance in objective evaluation.

Table 2. Quantitative comparison of the proposed algorithm and different comparison methods on the DENSE-HAZE dataset. **Bold** is the best, Red is the second.

Methods	Venue& Year	DENSE-HAZE			
		$PSNR\uparrow$	$SSIM\uparrow$	$Entropy\uparrow$	$LPIPS\downarrow$
GridDehazeNet [16]	*ICCV'19*	14.49	0.4401	5.8736	0.7168
FFA-Net [19]	*AAAI'20*	15.17	0.3243	6.3659	0.7069
MSBDN [6]	*CVPR'20*	15.51	0.3478	6.9032	0.7197
PSD [5]	*CVPR'21*	9.73	0.4345	5.5030	0.8184
SG-Net [9]	*ACCV'22*	14.91	0.4641	-	-
D4 [27]	*CVPR'22*	11.49	0.4821	6.3123	0.7555
SLP [14]	*TIP'23*	13.81	0.4857	6.7357	0.8489
Dehazeformer-T [21]	*TIP'23*	15.52	0.4635	**7.093**	**0.6459**
Fourmer[31]	ICML 23	15.95	0.4917	-	-
Ours	-	**16.42**	**0.5235**	6.9600	0.66966

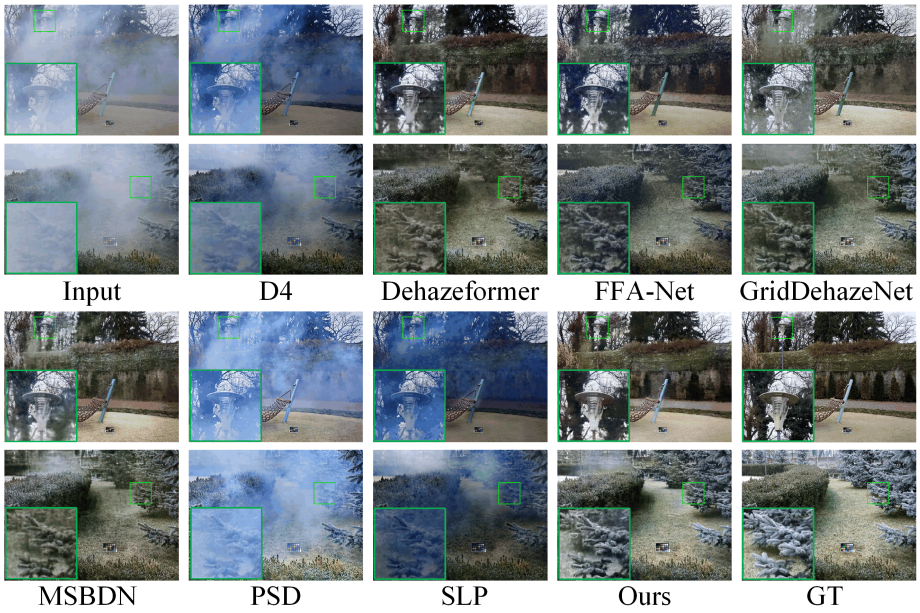

Fig. 5. Visual comparison of different dehazing methods on NH-HAZE dataset.(Zooming in can obtain a clearer view)

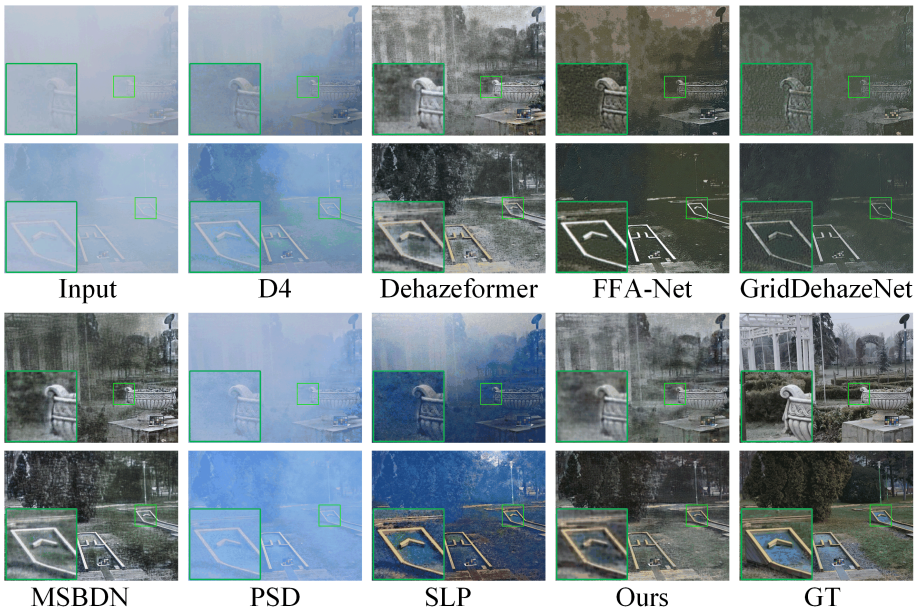

Fig. 6. Visual comparison of different dehazing methods on DENSE-HAZE dataset.(Zooming in can obtain a clearer view)

Table 3. Quantitative comparison of the proposed algorithm and different comparison methods on the NH-HAZE dataset. **Bold** is the best, Red is the second.

Methods	Venue& Year	NH-HAZE			
		$PSNR\uparrow$	$SSIM\uparrow$	$Entropy\uparrow$	$LPIPS\downarrow$
GridDehazeNet [16]	*ICCV'19*	17.23	0.5042	7.2881	0.3787
FFA-Net [19]	*AAAI'20*	18.09	0.5173	7.2727	0.3635
MSBDN [6]	*CVPR'20*	17.12	0.4539	7.3045	0.3989
PSD [5]	*CVPR'21*	10.32	0.5274	7.0658	0.5247
SG-Net [9]	*ACCV'22*	18.68	0.6609	-	-
D4 [27]	*CVPR'22*	12.66	0.5072	7.1318	0.5259
SLP [14]	*TIP'23*	15.84	0.5956	6.9696	0.4687
Dehazeformer-T [21]	*TIP'23*	18.73	0.5326	7.3274	0.3649
Fourmer[31]	*ICML'23*	19.91	**0.7214**	-	-
Ours	-	**20.10**	0.6716	**7.5319**	**0.3210**

The subjective and objective experimental results described above demonstrate the superiority of our method and the effectiveness of the proposed interaction-guided global and local feature extraction models. We also attach in Table 4 the number of parameters for each deep learning model along with MACs.

Table 4. Performance comparison of the comparative methods on the quantity of model parameters and MACs.

Methods	GridDehazeNet	FFA-Net	MSBDN	PSD	SG-Net	D4	SLP	Dehazeformer-T	Fourmer	Our
Parameters	0.956M	4.456M	31.35M	6.21M	3.33M	10.7M	-	0.686M	1.29M	22.87M
MACs	21.49G	287.8G	41.54G	143.91G	3.34G	2.25G	-	6.658G	20.6G	17.52G

4.3 Ablation Study

In this section, we conducted ablation analysis on each component of the proposed method and verified the impact of each component on the performance of dehazing. Firstly, we build the basic network framework as the Base of the dehazing network, which consists of two branches of Transformer and CNN as Encoder then the features are summed up and pass through the Decoder composed of CNN. Then we add different modules to base, including:
(1) **Base+DownS**: downsample the image once before feeding it into the Transformer.
(2) **Base+DownS+FA**: downsample an input image before it is fed into the Transformer and outputs the features of the Transformer and CNN between each layer to be summed.
(3) **Base+FA+CPA**: between each layer, the Transformer and CNN feature outputs are summed and fed into the CPA to obtain a weight matrix, which is

multiplied by the CNN outputs.
(4) **Ours**: Our model includes all the above blocks.

Table 5. Ablation studies with different modules on the NH-HAZE dataset.

Methods	**DownS**	**FA**	**CPA**	PSNR	SSIM
Base	-	-	-	17.70	0.5324
Base+DownS	✔	-	-	19.14	0.5985
Base+DownS+FA	✔	✔	-	19.48	0.6530
Base+FA+CPA	-	✔	✔	18.96	0.5608
Ours	✔	✔	✔	**20.10**	**0.6716**

For all models, we used L1 loss for image reconstruction and used the NH-HAZE dataset for training and testing in our ablation experiments. The quantitative evaluation results of the models described above are presented in Table 5. All modules improved model performance compared to the Base model, demonstrating the overall effectiveness of our design.
(1) Base+DownS: Although Transformer and CNN tend to extract inconsistent feature information, there will still be many redundant features extracted by dual-branch feature extraction under the same dimension, which affects the performance of the model. Therefore, we add downsampling before the Transformer branch to increase the global receptive field of the Transformer branch and reduce the feature redundancy caused by the repeated extraction of double-branch features. Therefore, as shown in Table 5, adding downsampling increased the PSNR and SSIM of the Base model from 17.70 and 0.5324 to 19.14 and 0.5985, respectively.
(2) Base+DownS+FA: On the basis of downsampling, the features of each layer of the two branches are extracted and added together, and then used to guide the CNN, so that the CNN branch has the ability to extract local features and global context, so that the model can better take into account global features and local features. Incorporating downsampling and feature addition is observed to increase the PSNR and SSIM from 19.14 and 0.5985 to 19.48 and 0.6530, respectively, compared to the model above.
(3) Base+CPA+FA: The model adds feature summation to the base model, and uses CPA to generate a weight matrix to guide the CNN, so that the model can focus on channels and pixels at the same time, which can better guide the CNN to capture information in the effective feature space. Without downsampling, these two modules improve the PSNR and SSIM of the Base model from 17.70 and 0.5324 to 18.96 and 0.5608.
(4) Ours: The full proposed model includes includes all the above blocks, and yielded the highest performance with PSNR and SSIM values of 20.10 and 0.6716, respectively, representing a PSNR increase of 2.40 dB compared with the Base model. These results clearly demonstrate the effectiveness of each module in the proposed model.

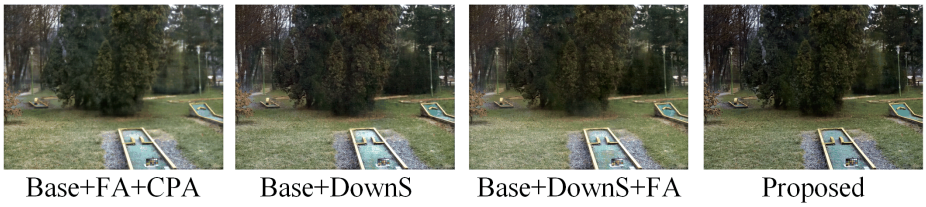

Fig. 7. Visual comparison of different dehazing methods on NH-HAZE dataset.(Zooming in can obtain a clearer view)

The visual comparison of the ablation model is shown in Figure.7. It can be seen from the figure that our complete model has better dehazing performance, the overall tone and brightness of the image are natural, the overall picture is full, and there is no edge blur or distortion. The visual comparison further verifies the effectiveness of the proposed module.

5 Conclusion

In this paper, we proposed an interaction-guided two-branch image dehazing network. The proposed model leverages the global and local feature extraction capabilities of a Transformer and CNN, respectively. It outputs the features between each layer for synthesis. The CPA block then considers these global and local features simultaneously to generate a weight matrix, which is multiplied by the outputs of the CNN to guide local feature extraction using global features. Additionally, the introduction of downsampling before the Transformer branch can effectively reduce computational complexity and increase the receptive field to improve overall model performance. The results of extensive experiments demonstrate that our method performs competitively in terms of both subjective and objective evaluations on synthetic and real datasets. Additionally, ablation analyses demonstrates the effectiveness of each module of the proposed method.

Acknowledgment. This work was supported by the Basic and Applied Basic Research of Guangdong Province under Grant 2023A1515140077, the Natural Science Foundation of Guangdong Province under Grant 2024A1515011880, the National Natural Science Foundation of China under Grant 62201149, the Guangdong Higher Education Innovation and Strengthening of Universities Project under Grant 2023KTSCX127.

References

1. Ancuti, C.O., Ancuti, C., Sbert, M., Timofte, R.: Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In: 2019 IEEE international conference on image processing (ICIP). pp. 1014–1018. IEEE (2019)

2. Ancuti, C.O., Ancuti, C., Timofte, R.: Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 444–445 (2020)
3. Berman, D., Avidan, S., et al.: Non-local image dehazing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1674–1682 (2016)
4. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D.: Dehazenet: An end-to-end system for single image haze removal. IEEE transactions on image processing **25**(11), 5187–5198 (2016)
5. Chen, Z., Wang, Y., Yang, Y., Liu, D.: Psd: Principled synthetic-to-real dehazing guided by physical priors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7180–7189 (2021)
6. Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., Yang, M.H.: 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 (2020)
7. Guo, C.L., Yan, Q., Anwar, S., Cong, R., Ren, W., Li, C.: Image dehazing transformer with transmission-aware 3d position embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5812–5820 (2022)
8. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE transactions on pattern analysis and machine intelligence **33**(12), 2341–2353 (2010)
9. Hong, T., Guo, X., Zhang, Z., Ma, J.: Sg-net: Semantic guided network for image dehazing. In: Asian Conference on Computer Vision. pp. 274–289. Springer (2022)
10. Laha, S., Foroosh, H.: Haar wavelet-based attention network for image dehazing. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 3948–3952. IEEE (2022)
11. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: Aod-net: All-in-one dehazing network. In: Proceedings of the IEEE international conference on computer vision. pp. 4770–4778 (2017)
12. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing **28**(1), 492–505 (2018)
13. Li, X., Liu, W., Li, X., Tan, H.: Physical perception network and an all-weather multi-modality benchmark for adverse weather image fusion. arXiv preprint arXiv:2402.02090 (2024)
14. Ling, P., Chen, H., Tan, X., Jin, Y., Chen, E.: Single image dehazing using saturation line prior. IEEE Transactions on Image Processing **32**, 3238–3253 (2023)
15. Liu, N., Han, J., Yang, M.H.: Picanet: Learning pixel-wise contextual attention for saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3089–3098 (2018)
16. Liu, X., Ma, Y., Shi, Z., Chen, J.: Griddehazenet: Attention-based multi-scale network for image dehazing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7314–7323 (2019)
17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
18. McCartney, E.J.: Optics of the atmosphere: scattering by molecules and particles. New York (1976)

19. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: Ffa-net: Feature fusion attention network for single image dehazing. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11908–11915 (2020)
20. Qiu, Y., Zhang, K., Wang, C., Luo, W., Li, H., Jin, Z.: Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12802–12813 (2023)
21. Song, Y., He, Z., Qian, H., Du, X.: Vision transformers for single image dehazing. IEEE Transactions on Image Processing **32**, 1927–1941 (2023)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
23. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11534–11542 (2020)
24. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17683–17693 (2022)
25. Wu, R.Q., Duan, Z.P., Guo, C.L., Chai, Z., Li, C.: Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22282–22291 (2023)
26. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
27. Yang, Y., Wang, C., Liu, R., Zhang, L., Guo, X., Tao, D.: Self-augmented unpaired image dehazing via density and depth decomposition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2037–2046 (2022)
28. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)
29. Zhang, Y., Zhou, S., Li, H.: Depth information assisted collaborative mutual promotion network for single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2846–2855 (2024)
30. Zheng, Y., Zhan, J., He, S., Dong, J., Du, Y.: Curricular contrastive regularization for physics-aware single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5785–5794 (2023)
31. Zhou, M., Huang, J., Guo, C.L., Li, C.: Fourmer: An efficient global modeling paradigm for image restoration. In: International conference on machine learning. pp. 42589–42601. PMLR (2023)
32. Zhu, Q., Mai, J., Shao, L.: A fast single image haze removal algorithm using color attenuation prior. IEEE transactions on image processing **24**(11), 3522–3533 (2015)