Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062328

ARTICLE

FS-MSFormer: Image Dehazing Based on Frequency Selection and Multi-Branch Efficient Transformer

Chunming Tang* **and Yu Wang**

Tianjin Key Laboratory of Intelligent Control for Electrical Equipment, School of Artificial Intelligence, Tiangong University, Tianjin, 300387, China

*Corresponding Author: Chunming Tang. Email: tangchunminga@hotmail.com

Received: 16 December 2024; Accepted: 06 March 2025; Published: 19 May 2025

ABSTRACT: Image dehazing aims to generate clear images critical for subsequent visual tasks. CNNs have made significant progress in the field of image dehazing. However, due to the inherent limitations of convolution operations, it is challenging to effectively model global context and long-range spatial dependencies effectively. Although the Transformer can address this issue, it faces the challenge of excessive computational requirements. Therefore, we propose the FS-MSFormer network, an asymmetric encoder-decoder architecture that combines the advantages of CNNs and Transformers to improve dehazing performance. Specifically, the encoding process employs two branches for multi-scale feature extraction. One branch integrates an improved Transformer to enrich local and global contextual information while achieving linear complexity, and the other branch dynamically selects the most suitable frequency components in the frequency domain for enhancement. A single decoding branch is utilized to achieve feature recovery in the decoding process. After enhancing local and global features, they are fused with the encoded features, which reduces information loss and enhances the model's robustness. A perceptual consistency loss function is also designed to minimize image color distortion. We conducted experiments on synthetic datasets SOTS-Indoor, Foggy Cityscapes, and the real-world dataset Dense-Haze, showing improved dehazing results. Compared with FSNet, our method has shown improvements of 0.95 dB in PSNR and 0.007 in SSIM on SOTS-Indoor dataset, and enhancements of 1.89 dB in PSNR and 0.0579 in SSIM on the Dense-Haze dataset demonstrate the effectiveness of our method.

KEYWORDS: Asymmetric encoder-decoder architecture; perceived consistency loss; unified transformer

1 Introduction

Hazy weather can affect daily life and machine vision systems. In hazy conditions, the visibility of the scene is reduced, and the quality of images captured by the cameras deteriorates. In the field of autonomous driving, haze limits the performance of subsequent high-level visual tasks, such as vehicle reidentification and scene understanding. Therefore, image dehazing is a meaningful visual task.

Image dehazing is the process of predicting a clear image from an input hazy image. Currently, there are many research algorithms for image dehazing, which can be roughly divided into two categories from the perspective of image processing: prior-based methods and deep learning-based methods. Prior-based dehazing methods manually model the statistical discrepancy between the hazy and haze-free images as empirical priors, such as dark channel prior (DCP) [1], non-local prior (NLP) [2], and color attenuation prior (CAP) [3]. These methods rely on the physical properties of the image to infer the blurring effect caused by atmospheric scattering in the dehazing process. The development of deep learning networks has

made significant progress in image dehazing. Current dehazing methods are classified into two categories based on network architecture: (1) CNN-based methods and (2) Transformer-based methods. CNN-based algorithms [4–6] build deep learning models by increasing the depth or width of the network, utilizing the network learning capability to estimate the parameters of the atmospheric scattering model, or implementing enhancement technology [7–9] to enhance image contrast and sharpen the details. Transformer-based methods [10] use the ability to establish global dependencies and achieve large receptive fields [11–13] to restore images. Although both methods perform excellently, they still have their own drawbacks. CNNs are effective at extracting local features through small neighborhood convolution operations. Although dilated convolutions can extend the receptive field and capture more extensive features, due to fixed convolution kernel size, it is challenging to capture global dependencies. To address this limitation and effectively model long-range dependencies, Transformer architectures have become an excellent alternative. Transformer models, with their self-attention mechanism, excel at capturing global context and long-range dependencies across the entire image, complementing the strengths of CNNs in local feature extraction. On the other hand, Transformer-based networks require a large number of parameters and high-cost training due to the complex self-attention operations, which often lead to high redundancy and an inability to capture local feature details effectively.

To address these issues, we propose a new asymmetric encoder-decoder architecture for image dehazing called FS-MSFormer. This network combines CNNs and Transformers, benefiting from both to capture local and global dependencies, helping the network generate more natural and realistic images with less complexity.

- We propose an asymmetric encoder-decoder dehazing network FS-MSFormer, which innovatively combines the advantages of CNNs and Transformers. Specifically, the Encoder is devised as a dual-branch structure to enhance local features while capturing long-range dependencies dynamically; thereby, texture and structural details can be restored more effectively.
- To address color distortion challenges in real-world hazy images, we introduce a perceptual consistency loss, which mitigates color distortion to enhance the robustness and generalization ability of the network in different hazy conditions.
- Experimental results on public synthetic and real dehazing datasets show that our network outperforms some advanced methods in performance. Additionally, our ablation studies are conducted to illustrate the effectiveness of different modules in FS-MSFormer.

2 Related Work

2.1 Prior-Based Methods

Early image dehazing methods were mainly based on prior knowledge and grounded in the atmospheric scattering model. By statistically analyzing hazy and clear image pairs, these methods estimate atmospheric light and transmission rates and then recover the clear, haze-free image through inverse operations. He et al. proposed the Dark Channel Prior (DCP) image dehazing method [1], which accurately calculates the transmission rate using low-intensity local pixel values in clear images. Berman et al. [2] introduced the fog-line prior theory to estimate the transmission rate, using the non-local color prior (NCP) for image dehazing. Riaz et al. [14] proposed a simple yet effective multi-block image restoration technique that improved the limitations of DCP, enhancing its speed and efficiency when processing high-resolution images. The time complexity of prior-based methods is generally low, and they perform well when prior conditions are satisfied. However, when the prior conditions are not met, these methods may lead to color distortion and halo artifacts.

2.2 Enhancement-Based Methods

Image enhancement methods in the field of image dehazing mainly improve visual effects by adjusting the contrast, color, and details of the image, not directly relying on physical models. Li et al. [8] presented a low-light enhancement (LLE) solution that integrates Retinex theory with deep learning. Qu et al. proposed the EPDN [15], which restores the color and details of hazy images through a multi-resolution generator and enhancement module. Qin et al. introduced FFA [16], utilizing feature fusion architecture and attention mechanism to improve dehazing performance. Cui et al. [9] built FSNet, which generates clear images by decoupling frequency components and dynamically selecting the most informative frequency components for restoration. Zhang et al. presented CVANet [17], simulating the visual attention mechanism of the human eyes to generate high-resolution images. Although enhancement-based methods can effectively improve image contrast and sharpen details, they often struggle to fully restore the details and contrast that remain in heavily hazy images.

2.3 Learning-Based Methods

With the development of CNNs, many researchers have introduced deep neural networks for image dehazing. Methods such as DehazeNet [4], DCPDN [18], and MSCNN [5] achieve image dehazing by estimating the transmission rate and atmospheric light. GridDehazeNet [19] captures multi-scale features of an image and directly estimates the haze-free image, proposing that learning to restore images is more effective than directly estimating atmospheric scattering parameters. Methods like AECRNet [20] and AOD-Net [21], based on image transformation ideas, do not rely on the atmospheric scattering model but predict the haze-free image in an end-to-end manner. GFN [22] introduced a gating mechanism to dynamically adjust the weights of different features during the feature fusion process, improving dehazing performance on images with varying haze intensities. Lu et al. proposed MixDehazeNet [23], which uses large convolution kernels with multi-scale features to capture both global and local information and improve image quality. Shen et al. [24] introduced MITNet, which recovers haze-free images in both the time and frequency domains. Chen et al. [25] presented DEA-Net to address the less effective of vanilla convolution and haze non-uniformity issues in recovering high-quality haze-free images. Yin et al. [26] proposed a multi-model fusion framework that combines a visual attention network (VAN) and a ODE heuristic network (ODEN) to enhance dehazing performance by leveraging their advantages. Although CNNs perform well in image dehazing, particularly in image feature extraction and nonlinear mapping, it is challenging to learn long-range pixel dependencies.

2.4 Transformers in Image Dehazing

Inspired by the success of Vision Transformers (ViTs) [11] in various computer vision tasks, increasing research has focused on introducing attention mechanisms into dehazing architectures. With their ability to model global dependencies in images, Transformers have been transferred to image restoration tasks, demonstrating significant potential. Zhao et al. [12] were the first to introduce ViTs into dehazing networks, allowing the network to jointly learn image decomposition and dehazing, resulting in more natural and finer high-quality haze-free images. DehazeFormer [13], built on the Swin Transformer, designed an improved U-Net based on the Transformer specifically for the dehazing task, making more efficient use of contextual information to enhance network performance. Yang et al. [27] proposed a multi-scale Transformer fusion network MSTFDN for image dehazing. Zamir et al. [28] introduced Restormer, a network for multi-scale learning of high-resolution images by improving multi-head attention and the feed-forward network. Qiu et al. proposed the MB-TaylorFormer [29], which achieves linear complexity by approximating softmax-attention using a Taylor expansion. Guo et al. [30] leveraged the complementary strengths of CNNs and

Transformers in feature extraction, proposing a dehazing network that combines both. Dong et al. [31] introduced a dual-branch dehazing network combining Transformer and residual channel attention. Using a Transformer in image dehazing has significant advantages in modeling global context and capturing remote dependencies, but it also has shortcomings such as high computational resource consumption, requiring a large amount of supervised data to converge, and limited local context modeling.

3 Proposed Method

In this section, We first present the architecture of FS-MSFormer. Next, the unified asymmetric Encoder decoder dehazing network is described in detail. Finally, we introduce the perceived consistency loss.

3.1 Overall Architecture

In the real world, fog density varies with depth of field, making it challenging for FSNet to perform as well as on synthetic datasets. Additionally, although FSNet [9] expands the receptive field through global and window-based average pooling, the fixed convolution kernel size leads to inconsistent restoration between near and distant scenes. These limitations result in color distortion and a lack of detail in the final dehazed image. Therefore, we aim to address this issue by leveraging Transformer's strong global context modeling capability. However, the high computational resource consumption of Transformer poses a challenge. To overcome this challenge, we draw inspiration from MB-TaylorFormer [29], which approximates the softmax attention mechanism using a Taylor expansion, achieving linear complexity. It effectively reduces the model's computational cost and maintains high performance. It is the reason that we combine FSNet with MB-TaylorFormer to build FS-MSFormer, as shown in Fig. 1.

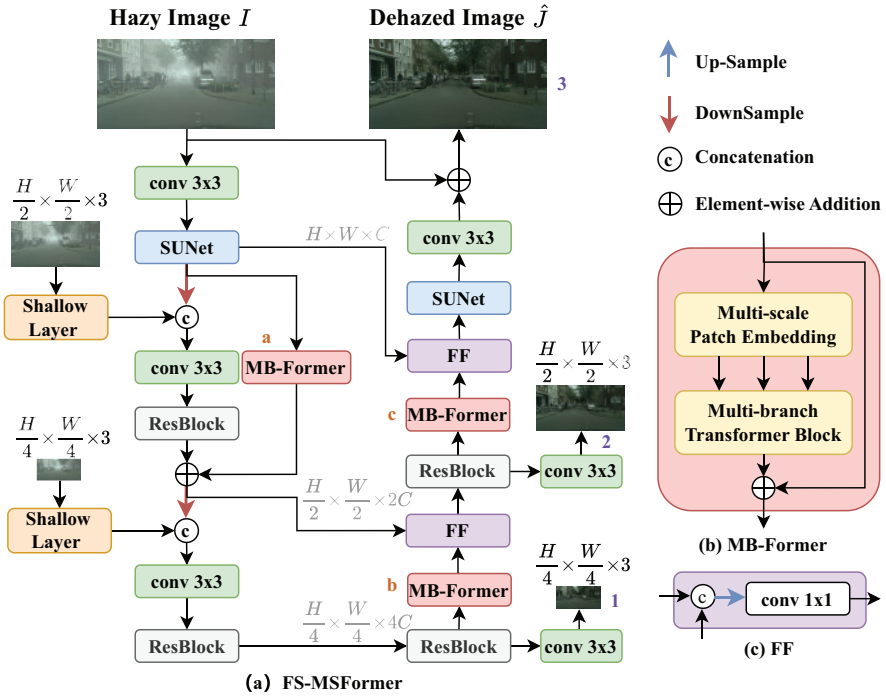

Figure 1: The overall architecture of FS-MSFormer. (a) Flowchart of FS-MSFormer. (b) MB-Former consists of a multi-branch hierarchical design based on multi-scale patch embedding and multi-branch transformer block. (c) Feature fusion module. Shallow Layer extracts shallow features for low-resolution input image I. Purple $1, 2, 3$ represent three scales of $\hat{J}$ and $1, 2$ only used in training (in Section 3.3). Orange a, b, c represent different positions of MB-Former (in Section 4.4.3)

Both the Encoder and decoder of our proposed FS-MSFormer consist of SUNet, ResBlock, and MB-Former. SUNet and ResBlock are shown in Fig. 2c and Fig. 3 in FSNet [9]. MB-Former is shown in Fig. 2a in MB-TaylorFormer [29]. The SUNet provides multi-scale learning and reduces complexity. The ResBlock contains several residual-type blocks for frequency selection and modulation. Both the SUNet and Resblock are integrated by MDSF and MCSF. MDSF mainly contains two elements: frequency decoupler (Fig. 2d) and modulator (Fig. 2e) in FSNet [9]. The Decoupler dynamically decomposes the feature map into different frequency patterns based on learned filters, and the modulator utilizes channel-wise attention to accentuate the useful frequency. The MCSF module has two branches with different receptive fields, the global branch, and the window-based branch, to efficiently enlarge the receptive field. MB-Former consists of multi-scale patch embedding and multi-branch transformer block [29], enabling more flexible embedding of coarse-to-fine features during the patch embedding stage and capturing long-distance pixel interactions with limited computational cost.

As shown in Fig. 1, given a hazy image of shape $H \times W \times 3$, where 3 is the number of channels and H, W denotes spatial coordinates. Initially, a 3×3 convolution is used to extract shallow features. Then, these shallow features pass into the SUNet module to obtain multi-scale features. Next, they are input into ResBlock and MB-Former, two branches separately, to capture deeper features and global features. Finally, these features are fused and passed into the next ResBlock to complete the encoding of the hazy image. In addition, during the encoding process, the low-resolution hazy images are concatenated with the extracted deep features to promote feature fusion and complementarity, and a 3×3 convolution is applied to adjust the number of channels. Next, the encoded features are fed to the decoder, where the features are gradually restored to the original size of the hazy image. In the decoding process, feature fusion is achieved through concatenation between the encoded and decoded features to aid in feature restoration. The encoded features go through two rounds of ResBlock, MB-Former, and FF modules, followed by the final SUNet. Finally, a 3×3 convolution and skip connections are applied to obtain the final clear image. During this process, two low-resolution images are also generated, which are used only during the training.

3.2 Unified Asymmetric Encoder Decoder Dehazing Network

FS-MSFormer optimizes local detail restoration and global information modeling in image dehazing tasks using an asymmetric encoder-decoder structure. In the decoder, to distinguish the low-level local texture features from the high-level global features, we add another MB-Former, which decodes features of different scales to improve the quality of the restored image by fusion.

3.2.1 Encoder

After extracting multi-scale features with SUNet, we use two branches to extract deep features and global features of the image separately: one is the ResBlock branch, and the other is the MB-Former branch. The reason for adding the MB-Former branch at this stage is to effectively reduce computational resource consumption while alleviating the model's sensitivity to shallow noise and mitigating the risk of overfitting. The ResBlock branch dynamically selects the most suitable frequency components for enhancement within local windows, while the MB-Former branch extracts global features using a self-attention mechanism, establishing long-range dependencies. By fusing features from both branches, multi-scale feature extraction is achieved, enabling the Encoder to not only better capture the global context of the image but also preserve fine details, thereby enhancing its feature extraction capabilities.

3.2.2 Decoder

In the decoder, we design a branch that includes two iterations of ResBlock + MB-Former + FF modules to enhance the network's performance in feature recovery. Specifically, the encoded features are first input into the ResBlock, where its dynamic and multi-scale frequency selection mechanism enhances the representation of effective feature components within local windows, enabling initial feature recovery. Then, the MB-Former's self-attention mechanism is applied to deeply integrate local and global information, which not only strengthens the recovery of local details but also ensures the coherence of global content. The FF module subsequently fuses the decoded features with the encoded features to compensate for any potential information loss during the decoding process. Through these three stages of feature fusion, the network's ability to recover details and textures is improved while also enhancing the model's global understanding of the image, ultimately boosting dehazing performance.

3.3 Perceived Consistency Loss Function

Given an image pair I and J, where J is the clear image corresponding to I, the clear image generated by FS-MSFormer is denoted as $\hat{J}$. During training, in addition to calculating the absolute difference between $\hat{J}$ and I in both the time domain and frequency domain, we incorporate the Structural Similarity Index (SSIM) loss to reduce color distortion in the dehazed image and help the model converge. The total loss is shown in Eq. (1), where α is an empirical value set to 0.1. L_{TD} and L_{FD} represent the L_1 loss calculated in the time and frequency domains, as shown in Eqs. (2) and (3), respectively. L_{SSIM} is the SSIM loss, as shown in Eq. (4). By combining L_1 loss and SSIM loss, we reduce pixel-level errors while preserving image quality, ultimately enhancing the robustness of the model.

$$L_{total} = L_{TD} + \alpha L_{FD} + L_{SSIM}, \tag{1}$$

$$L_{TD} = \sum_{s=1}^{3} \frac{1}{E_s} \left\| \hat{J} - J \right\|_1, \tag{2}$$

$$L_{FD} = \sum_{s=1}^{3} \frac{1}{E_s} \left\| \mathscr{F}(\hat{J}) - \mathscr{F}(J) \right\|_1, \tag{3}$$

where E_s represents the total number of pixels in an image, and s represents the three scales of $\hat{J}$ as shown by purple *1*, *2*, *3* in Fig. 1. $\mathscr{F}$ represents fast Fourier Transform.

$$L_{SSIM} = 1 - \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}, \tag{4}$$

where x and y represent $\hat{J}$ and J. μ_x, μ_y and σ_x, σ_y represent the mean and standard deviations of x and y. σ_{xy} represents their covariance. C_1 and C_2 are two constants introduced to prevent division by zero. Here, $C_1 = 0.0001$, $C_2 = 0.0009$.

4 Experiments

4.1 Experiment Setup

4.1.1 Implementation Details

We apply random horizontal flipping for data augmentation with a probability of 0.5. During training, the batch size is set to 2, and the number of epochs is 30. The initial learning rate is set to $1e-4$ and is gradually reduced to $1e-6$ using cosine annealing. The optimizer used is Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$). We train our model using PyTorch on an NVIDIA RTX 3090 GPU.

4.1.2 Datasets

We evaluated the proposed FS-MSFormer on synthetic datasets SOTS Indoor (ITS) [32], Foggy Cityscapes (FC) [33], and real dataset Dense-Haze [34].

ITS: Contains 13,990 pairs of clear and hazy indoor images, along with a comprehensive objective test set.

Dense-Haze: Includes 55 paired images, with the last 5 images used as the test set and the rest for training.

FC: A fog dataset simulating real-world scenes, where each hazy image is rendered from a clear image in the Cityscapes dataset. It consists of 550 images, with 498 for training and 52 for testing, each with three levels of fog density: light, medium, and dense.

4.1.3 Evaluation Metrics

We employed two evaluation metrics for both synthetic and real image datasets: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). PSNR is a pixel-based quality metric commonly used to quantify the difference between dehazed and clear images. A higher PSNR value indicates more remarkable similarity between the two images, suggesting better dehazing quality. However, PSNR mainly focuses on pixel-wise absolute errors and does not account for factors such as image structure and texture, meaning it cannot fully capture the subjective visual quality of the image. In contrast, SSIM is more aligned with the human visual system's perception, providing a better assessment of structural and detail restoration in images. Therefore, we incorporate SSIM loss, which offers a more intuitive and comprehensive evaluation of image quality.

4.2 Experiments on Synthetic Hazy Images

4.2.1 Comparison on ITS Dataset

Our FS-MSFormer has been compared with seven other algorithms, including MSBDN [35], AECR-Net [20], HDM [36], Dehamer [30], FSNet [9] CL2S [37], and DEA-Net [25]. The experimental results are summarized in Table 1. The results show that our algorithm outperforms the baseline FSNet, with a 0.95 improvement in PSNR and a 0.007 improvement in SSIM, which were also better than other methods, indicating the effectiveness of our approach. The subjective comparison results for dehazing are shown in Fig. 2. We evaluated the same dehazed image generated by different methods, and our proposed method can recover more precise edges.

Table 1: Comparison with the results of current advanced dehazing methods

Method	SOTS-indoor		Dense-Haze		FC	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
MSBDN [35] (2020)	33.67	0.985	15.13	0.555	–	–
AECRNet [20] (2021)	37.17	0.9901	15.80	0.466	–	–
HDM [36] (2022)	38.56	0.991	–	–	–	–
Dehamer [30] (2022)	36.63	0.988	**16.62**	0.560	–	–
FSNet* [9] (2023)	38.25	0.988	13.38	0.5391	16.51	0.8402
CL2S [37] (2024)	35.36	0.9808	–	–	–	–
DEA-Net [25] (2024)	**41.31**	0.9945	–	–	–	–
Ours	39.15	**0.995**	15.27	**0.5971**	**18.23**	**0.9109**

Note: *means using the same training parameters to train the model.

4.2.2 Comparison on FC Dataset

We compared FS-MSFormer with the baseline FSNet on the FC Dense datasets. As shown in Table 1, we observed improvements in both PSNR and SSIM. The subjective comparison results for dehazing are shown in Fig. 3, where it is clear that FS-MSFormer removes fog in the distance of images more cleanly.

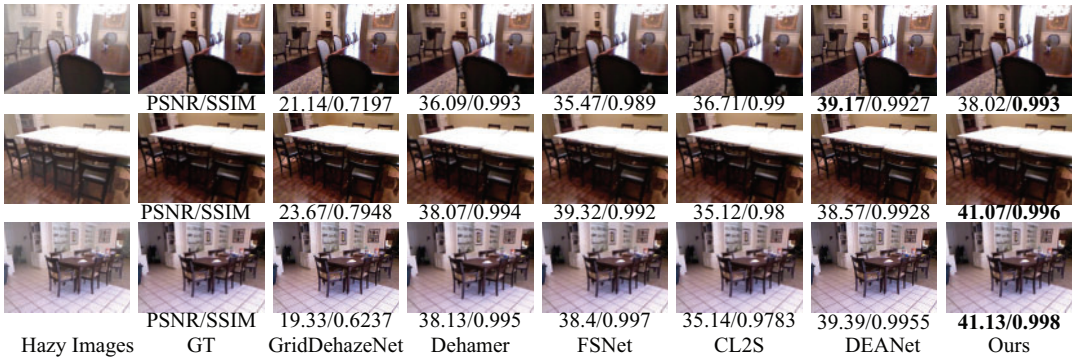

Figure 2: Comparison of dehazing visualization results with other methods on ITS

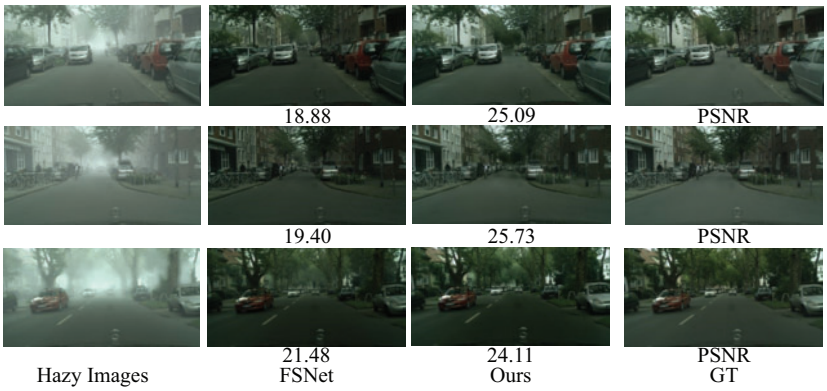

Figure 3: Comparison of the PSNR of some dehazing visualization results on FC Dense

4.3 Comparison on Real Hazy Images

We conduct PSNR and SSIM evaluations on the real dataset Dense-Haze, and the results are presented in Table 1. The results indicate that FS-MSFormer has improved SSIM. We also use LPIPS to evaluate the perceptual quality of dehazed images, as shown in Table 2. Subjective comparison is shown in Fig. 4, which shows that our method has less color distortion and produces more explicit images compared to FSNet.

Table 2: Comparison with the results of different methods on Dense-Haze

Method	Dehamer	FSNet	Ours
LPIPS↓	**0.5416**	0.5647	0.5491

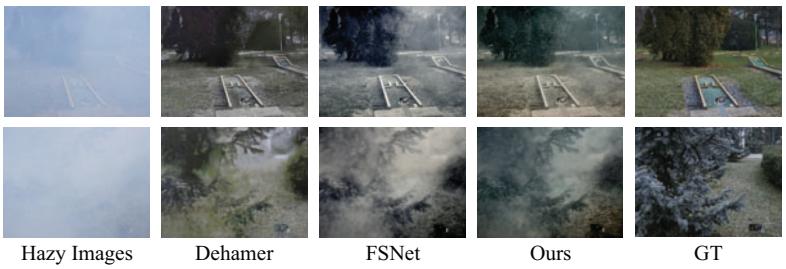

Figure 4: Comparison of dehazing visualization results on Dense-Haze

4.4 Ablation Studies

To further validate the effectiveness of our proposed method, we conducted ablation studies from the aspects of encoder-decoder structure, module parameters, and loss function.

4.4.1 Different Network Composition Modules

To achieve an expanded receptive field and capture longer distance dependencies, we adopted two approaches on the Dense-Haze dataset: adding MB-Former modules and using dilated convolutions. We used different modules at positions *a*, *b* and *c* in Fig. 1 in orange color. The experimental structure is shown in Table 3. We compared three methods: FSNet, adding MB-Former, and using dilated convolutions with three dilation rates. The results indicate that using MB-Former yields better results in both PSNR and SSIM.

Table 3: Comparison with the results of different modules

Method	Dilation rate	PSNR↑	SSIM↑
FSNet	–	13.38	0.53919
MB-Former	–	**15.27**	**0.5971**
Dilated convolution 1	[2, 2, 2]	14.00	0.5601
Dilated convolution 2	[2, 4, 4]	14.40	0.5569
Dilated convolution 3	[4, 4, 4]	14.09	0.5638

4.4.2 Different Feature Fusion Methods

We have designed two methods for feature fusion on Dense-Haze dataset. $\oplus$ means element-wise addition, and convolutional layer means we first use concat operation and then use 1×1 convolution to adjust channels. The results are shown in Table 4, indicating that using element-wise addition is better.

Table 4: Comparison with the results of different feature fusion methods

FF method	PSNR↑	SSIM↑
$\oplus$	**15.27**	**0.5971**
Convolutional layers	13.56	0.5694

4.4.3 MB-Former in FS-MSFormer

We have added MB-Former at different positions in the Encoder and decoder, as shown in orange a, b, c in Fig. 1. Table 5 shows the dehazing results we tested on the Dense-Haze dataset after adding MB-Former at different locations, where 0 represents the FSNet network, and a, b, c represent adding MB-Former at different locations in the network. The results in Table 5 demonstrate the effectiveness of our MB-Former approach.

Table 5: Comparison with the results of MB-Former at different locations

Position	PSNR↑	SSIM↑
0	13.38	0.5391
b + c	13.97	0.5327
a + c	14.38	0.5809
a + b	14.08	0.5765
a + b + c	**15.27**	**0.5971**

4.4.4 Different Encoder and Decoder

We designed three network architectures for the encoder-decoder on the Dense-Haze dataset, as shown in Fig. 5: Fig. 5 (1), which uses dual-branch ResBlock and MB-Former for feature extraction and recovery in both the Encoder and decoder. Fig. 5 (2) is an improved version of Fig. 5 (1), where SUNet is enhanced. Fig. 5 (3) represents our FS-MSFormer structure. The dehazing results are shown in Table 6. The results indicate that simply adding more Transformer modules does not necessarily improve performance. The key is to introduce the MB-Former module at appropriate locations based on the network's characteristics to achieve optimal improvement.

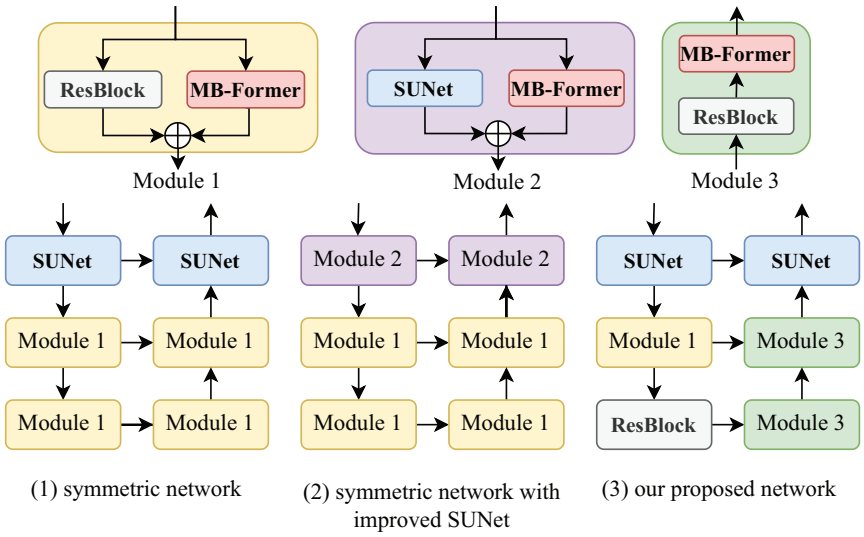

Figure 5: Different encoder-decoder network structures

Table 6: Comparison with the results of different encoder-decoder network structures

Method	PSNR↑	SSIM↑
Structure(1)	14.87	0.5909
Structure(2)	14.71	0.5585
Structure(3)	**15.27**	**0.5971**

4.4.5 Different MB-Former Positions

MB-Former, as shown in Fig. 1, consists of multi-scale patch embedding and multi-branch transformer block. The size of the multi-scale patch embedding is adjusted to enable the model to capture multi-scale features. To evaluate its effectiveness, we tested different parameter settings for the multi-scale patch embedding on the Dense-Haze dataset, as shown in Table 7. We ultimately selected method 1 as the optimal parameter setting for the model.

Table 7: Comparison with the results of different MB-Former positions

Method	Embedding parameters	PSNR↑	SSIM↑
1	[32, 64, 128]	**15.27**	**0.5971**
2	[48, 96, 48]	14.99	0.5809
3	[24, 48, 24]	14.87	0.5789

4.4.6 Composition of Total Loss Function

During training, we investigated the total loss L_{total}, as shown in Eq. (5), composed of L_1 loss, perceptual loss, and SSIM loss, on the Dense-Haze dataset. The perceptual loss L_{perc} was calculated using VGG16. By adjusting different parameters, the experimental results of L_{total} are presented in Table 8. The results indicate that the system performs best without adding L_{perc}. Therefore, we ultimately selected the combination of L_1 loss and SSIM loss, where the hyperparameters λ_1 and λ_2 are both set to 1, as shown in Eq. (1), to converge the system.

$$L_{total} = \lambda_1(L_{TD} + \alpha L_{FD}) + \lambda_2 L_{SSIM} + \lambda_3 L_{perc} \tag{5}$$

Table 8: Dehazing results under different weight settings

λ_1	λ_2	λ_3	PSNR↑	SSIM↑
1	1	1	13.76	0.5447
1	0.1	0	14.52	0.5755
1	0.5	0	14.45	0.5872
1	1	0	**15.27**	**0.5971**
1	0	0	14.17	0.5575

5 Limitations

In developing our FS-MSFormer, we notice there are still some difficulties that need to be addressed. We will conduct further research in future.

- Domain shift: While our method has shown progress on both synthetic and real datasets, the generalization from synthetic to real datasets remains suboptimal. Bridging the gap between these datasets is an ongoing challenge that still requires further attention.
- Model efficiency: Our method combines the MB-Former module with FSNet, which may reduce computational complexity. However, compared to CNN-based methods, it still faces challenges such as high computational complexity and significant hardware requirements. We will explore the lightweight model.

6 Conclusion

This paper introduces a novel asymmetric encoder-decoder network, FS-MSFormer, which cleverly combines the advantages of CNNs and Transformer to improve image dehazing quality. In the encoding process of FS-MSFormer, a dual-branch feature extraction and fusion strategy is employed, effectively enriching both global and local features. In the decoding process, a single-branch structure is used to extract and fuse local and global features, which not only improves the quality of image recovery but also enhances the model's expressive capability. Furthermore, the perceptual consistency loss function effectively mitigates image color distortion and boosts the model's generalization ability. Overall, experimental results on both synthetic and real-world datasets demonstrate that FS-MSFormer achieves strong performance in image dehazing tasks.

Acknowledgement: The authors would like to thank the anonymous reviewers for their careful reading and valuable comments.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Data collection: Yu Wang; analysis and interpretation of results: Chunming Tang, Yu Wang; draft manuscript preparation: Chunming Tang, Yu Wang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Datasets Foggy Cityscapes, Dense-Haze, and ITS are publicly available.

Ethics Approval: Studies not involving humans or animals.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. He K, Sun J, Tang X. Single image haze removal using dark channel prior. IEEE Trans Pattern Anal Mach Intell. 2010;33(12):2341–53. doi:10.1109/TPAMI.2010.168.
2. Berman D, Treibitz T, Avidan S. Non-local image dehazing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 1674–82.
3. Zhu Q, Mai J, Shao L. A fast single image haze removal algorithm using color attenuation prior. IEEE Trans Image Process. 2015;24(11):3522–33. doi:10.1109/TIP.2015.2446191.
4. Cai B, Xu X, Jia K, Qing C, Tao D. DehazeNet: an end-to-end system for single image haze removal. IEEE Trans Image Process. 2016;25(11):5187–98. doi:10.1109/TIP.2016.2598681.

5. Ren W, Liu S, Zhang H, Pan J, Cao X, Yang MH. Single image dehazing via multi-scale convolutional neural networks. In: Computer Vision-ECCV 2016: 14th European Conference; 2016 Oct 11–14; Amsterdam, The Netherlands. p. 154–69.
6. Alenezi F. Image dehazing based on pixel guided CNN with PAM via graph cut. Comput Mater Contin. 2022;71(2):3425–43. doi:10.32604/cmc.2022.023339.
7. Wang S, Huang B, Wong TH, Huang J, Deng H. CLGA Net: cross layer gated attention network for image dehazing. Comput Mater Contin. 2023;74(3):4667–84. doi:10.32604/cmc.2023.031444.
8. Li Y, Wei X, Liao X, Zhao Y, Jia F, Zhuang X, et al. A deep retinex-based low-light enhancement network fusing rich intrinsic prior information. ACM Trans Multim Comput Commun Appl. 2024;20(11):1–23. doi:10.1145/3689642.
9. Cui Y, Ren W, Cao X, Knoll A. Image restoration via frequency selection. IEEE Trans Pattern Anal Mach Intell. 2024;46(2):1093–108. doi:10.1109/TPAMI.2023.3330416.
10. Vaswani A. Attention is all you need. In: Advances in neural information processing systems; 2017.
11. Dosovitskiy A. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
12. Zhao D, Li J, Li H, Xu L. Complementary feature enhanced network with vision transformer for image dehazing. arXiv:2109.07100. 2021.
13. Song Y, He Z, Qian H, Du X. Vision transformers for single image dehazing. IEEE Trans Image Process. 2023;32:1927–41. doi:10.1109/TIP.2023.3256763.
14. Riaz S, Anwar MW, Riaz I, Kim HW, Nam Y, Khan MA. Multiscale image dehazing and restoration: an application for visual surveillance. Comput Mater Contin. 2022;70(1):1–17. doi:10.32604/cmc.2022.018268.
15. Qu Y, Chen Y, Huang J, Xie Y. Enhanced pix2pix dehazing network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 8160–8.
16. Qin X, Wang Z, Bai Y, Xie X, Jia H. FFA-Net: feature fusion attention network for single image dehazing. Proc AAAI Conf Artif Intell. 2020;34:11908–15. doi:10.1609/aaai.v34i07.6865.
17. Zhang W, Zhao W, Li J, Zhuang P, Sun H, Xu Y, et al. CVANet: cascaded visual attention network for single image super-resolution. Neural Netw. 2024;170(2):622–34. doi:10.1016/j.neunet.2023.11.049.
18. Zhang H, Patel VM. Densely connected pyramid dehazing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 3194–203.
19. 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; 2019. p. 7314–23.
20. Wu H, Qu Y, Lin S, Zhou J, Qiao R, Zhang Z, et al. Contrastive learning for compact single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 10551–60.
21. 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; 2017. p. 4770–8.
22. Ren W, Ma L, Zhang J, Pan J, Cao X, Liu W, et al. Gated fusion network for single image dehazing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 3253–61.
23. Lu L, Xiong Q, Xu B, Chu D. MixDehazeNet: mix structure block for image dehazing network. In: 2024 International Joint Conference on Neural Networks (IJCNN); 2024; IEEE. p. 1–10.
24. Shen H, Zhao ZQ, Zhang Y, Zhang Z. Mutual information-driven triple interaction network for efficient image dehazing. In: Proceedings of the 31st ACM International Conference on Multimedia; 2023. p. 7–16.
25. Chen Z, He Z, Lu ZM. DEA-Net: single image dehazing based on detail-enhanced convolution and content-guided attention. IEEE Trans Image Process. 2024;33:1002–15. doi:10.1109/TIP.2024.3354108.
26. Yin S, Yang X, Lu R, Deng Z, Yang YH. Visual attention and ODE-inspired Fusion Network for image dehazing. Eng Appl Artif Intell. 2024;130(2):107692. doi:10.1016/j.engappai.2023.107692.
27. Yang Y, Zhang H, Wu X, Liang X. MSTFDN: multi-scale transformer fusion dehazing network. Appl Intell. 2023;53(5):5951–62.
28. Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang MH. Restormer: efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 5728–39.

29. 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; 2023. p. 12802–13.
30. Guo CL, 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; 2022. p. 5812–20.
31. Dong P, Wang B. TransRA: transformer and residual attention fusion for single remote sensing image dehazing. Multidimens Syst Signal Process. 2022;33(4):1119–38. doi:10.1007/s11045-022-00835-x.
32. Li B, Ren W, Fu D, Tao D, Feng D, Zeng W, et al. Benchmarking single-image dehazing and beyond. IEEE Trans Image Process. 2018;28(1):492–505. doi:10.1109/TIP.2018.2867951.
33. Sakaridis C, Dai D, Van Gool L. Semantic foggy scene understanding with synthetic data. Int J Comput Vis. 2018;126(9):973–92. doi:10.1007/s11263-018-1072-8.
34. Ancuti CO, 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); 2019; IEEE. p. 1014–8.
35. Dong H, Pan J, Xiang L, Hu Z, Zhang X, Wang F, et al. Multi-scale boosted dehazing network with dense feature fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 2157–67.
36. Liang Y, Wang B, Ren W, Liu J, Wang W, Zuo W. Learning hierarchical dynamics with spatial adjacency for image enhancement. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022. p. 2767–76.
37. Rohn Y. Rethinking the elementary function fusion for single-image dehazing. arXiv:2405.15817. 2024.