Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

DLF-DehazeFormer: An Image Dehazing Method with Collaborative Feature Enhancement and Multi-Loss Supervision

Ruzhen Xie
Hebei University of Science and Technology
Ya Guo
guoya@hebust.edu.cn
Hebei University of Science and Technology

Research Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11811620/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

DLF-DehazeFormer: An Image Dehazing Method with Collaborative Feature Enhancement and Multi-Loss Supervision

Ruzhen Xie[1] and Ya Guo[1*]

[1]School of Information Science and Engineering,Hebei University of Science and Technology,No. 26 Yuxiang Street, Yuhua District,Shijiazhuang,050018,Hebei Province,China.

*Corresponding author(s). E-mail(s): guoya@hebust.edu.cn;

Abstract

Indoor images captured by robots are degraded by mist and dust, reducing contrast and obscuring details. Transformer-based dehazing models provide strong feature modeling but remain limited in local-detail representation and training supervision. We propose DLF-DehazeFormer, combining feature enhancement and multi-loss supervision. A Dual Frequency Adaptive Block (DFAB) strengthens structural and textural representations. For output supervision, perceptual and edge losses augment the original L1 loss to constrain reconstruction. Integrating DFAB with the best-performing tested composite loss improves indoor hazy image quality. Trained on the Indoor Training Set of the Realistic Single Image Dehazing (RESIDE) benchmark, the model is evaluated on its indoor Synthetic Objective Testing Set against published image dehazing methods. DLF-DehazeFormer achieves 35.97 dB peak signal-to-noise ratio and a structural similarity index of 0.9913, exceeding the original DehazeFormer by 0.93 dB and 0.0021, respectively. These results demonstrate the effectiveness of combining DFAB feature enhancement and composite-loss supervision for indoor image dehazing. The evaluation uses paired synthetic indoor hazy and clear images from the public RESIDE benchmark.

Keywords: Indoor image dehazing, Transformer, DFAB feature enhancement, Multi-loss joint supervision strategy

1 Introduction

With advances in computer vision, visual perception systems are used in remote sensing, intelligent transportation, UAV inspection, and indoor robot perception [1]. In relatively enclosed indoor spaces, water vapor from humidification and dust generated during operations can accumulate locally, causing nonuniform image degradation unlike widespread outdoor haze. Image quality directly affects feature extraction for object detection [2, 3], segmentation [4], recognition [5], and other high-level vision tasks. Scattering and attenuation by mist and dust reduce contrast, shift colors, blur details, and obscure textures [6, 7], potentially impairing accurate scene information. Effective indoor image dehazing is therefore important. Image dehazing has evolved from traditional prior-based modeling to deep learning. Research can be divided into image-restoration theory, traditional dehazing, and deep-learning-based dehazing [8]. Restoration theory

underpins haze-degradation models; traditional techniques use handcrafted rules, including non-physical and physical-model methods; and deep networks learn haze-removal mappings from large datasets, becoming the dominant approach. Over the past decade, the nonlinear modeling capacity of data-driven methods has markedly improved detail restoration and generalization. Research has advanced along three lines: Convolutional Neural Networks (CNNs) [9], Transformers, and diffusion models [10]. Early CNN dehazing progressed from partially model-guided methods to end-to-end prediction. DehazeNet [11] combined atmospheric-scattering priors with convolutional features, but its staged transmission estimation could accumulate errors. AOD-Net [12] reformulated the scattering model to directly output clear images with substantially lower computation. Later CNNs incorporated residual connections, multiscale features, and attention to improve dehazing accuracy. Nevertheless, their local receptive fields limit long-range pixel dependencies, especially under dense, spatially nonuniform haze, motivating Transformer and diffusion-based approaches. To overcome local CNN modeling limitations, Transformers use self-attention to model long-range image relationships and complex haze distributions. DehazeFormer [13] addresses limitations of the original Swin Transformer for dehazing by combining window attention, convolutional enhancement, and multiscale fusion. However, its local texture representation can be improved, and its training mainly relies on L1 pixel loss without direct perceptual or edge supervision. We therefore propose DLF-DehazeFormer to jointly enhance feature representation and optimize loss supervision. The main contributions are as follows:

- Dual-frequency adaptive feature enhancement: DFAB is adopted and adapted to combine high- and low-frequency feature extraction, channel attention, and adaptive weight fusion, improving structural and local-texture representation and detail recovery.
- Multi-loss joint supervision: VGG perceptual loss and Edge loss are added to the original L1 loss. The resulting L1+VGG+Edge objective jointly constrains pixel reconstruction, perceptual features, and edge gradients to reduce blurred textures and missing edge details and improve reconstruction quality.
- Comparative experiments: DFAB feature enhancement and multi-loss supervision are combined into DLF-DehazeFormer, whose indoor dehazing performance is evaluated against other methods.

2 Related Work

2.1 Development of Single-Image Dehazing Techniques

Image dehazing addresses contrast attenuation and detail degradation caused by scattering from suspended atmospheric particles. Its development encompasses image-restoration theory, traditional dehazing, and deeplearning-based techniques.

2.1.1 Research on Fundamental Image Restoration Theory

Image restoration reconstructs clear images from observations degraded by blur, noise, compression artifacts, or haze [14]. Unlike enhancement aimed mainly at visual appearance, restoration models image degradation and solves an inverse problem to recover scene radiance [15]. McCartney et al. [16] described atmospheric scattering, and Narasimhan et al. [17] introduced this optical theory into computer vision, deriving the classical scattering model in which observed haze combines attenuated scene radiance and atmospheric light. Since advances in single-image dehazing in 2009 [18], researchers have studied constraints on the solution space. Su et al. [19] analyzed regularization in dehazing objectives, while Li et al. [20] reviewed degradation mechanisms and models and emphasized regularization and loss optimization for the ill-posed reconstruction problem. This work supports physical-model improvement and deep-learning loss design [6]. Although physical restoration models are interpretable, their robustness and detail recovery remain limited in complex haze, prompting continued research.

2.1.2 Research on Conventional Image Dehazing Techniques

Before deep learning became widespread, traditional dehazing used image characteristics or

atmospheric physics. It can be divided into nonphysical and physical-model-based approaches. Nonphysical methods include image enhancement and fusion. Without explicitly modeling haze formation, they improve appearance through pixel transformations, frequency decomposition, or image reconstruction. Early histogram equalization and gamma correction enhanced contrast but could overenhance clear areas, amplify noise, and distort colors [21]. Later Retinex theory [22], wavelet transforms, and homomorphic filtering separated high- and low-frequency information to retain some textures. Image fusion methods [23] combined complementary information from multiple frames to reduce haze. These approaches are simple and efficient without requiring degradation models, but their lack of physical constraints can lead to insufficient detail recovery and color distortion under nonuniform haze and complex lighting. Physical-model methods estimate degradation parameters. The Dark Channel Prior (DCP) introduced by He et al. [24] markedly improved single-image dehazing and became a milestone. Meng et al. [25] used boundary constraints and contextual regularization to improve transmission estimation and restoration in complex scenes. Zhu et al. [26] proposed the Color Attenuation Prior (CAP), estimating depth from brightness and saturation to reduce computational complexity. Such priors can fail in sky regions, large white areas, and complex lighting. Recent studies therefore seek adaptive compensation and combined constraints. Liao et al. [27] combined haze lines and color attenuation priors to reduce oversaturation, detail loss, and artifacts. The inverted haze density correction prior (IHDCP) [28] further improved restoration efficiency through pixelwise gamma correction and transmission-estimation optimization. In summary, image enhancement, physical models, and priors advanced traditional dehazing, but reliance on handcrafted features and assumptions limits adaptation to complex scenes. Deep learning offers a data-driven alternative.

2.1.3 Research on Deep-Learning-Based Image Dehazing

With advances in deep learning and large paired datasets, data-driven dehazing has become a mainstream research direction [29]. Neural networks learn nonlinear mappings from hazy to clear images, improving detail recovery, color preservation, and adaptability to complex scenes. Early deep-learning dehazing relied mainly on CNNs. DehazeNet [11] used convolutional features to estimate transmission and the scattering model to restore clear images, moving from handcrafted priors to learned parameters. AOD-Net [12] then enabled end-to-end dehazing by combining physical parameters. Dynamic convolution was later introduced into AOD-Net [30] to balance compactness and feature representation. CNNs incorporating UNet, multiscale features, and attention further improved texture recovery, but convolution still primarily models local regions and has limited long-range dependency modeling. Transformer architectures [31] were introduced to address limited long-range CNN modeling, although standard self-attention is costly for high-resolution images. Swin Transformer uses window-based and shifted-window attention to reduce computation while enabling information exchange, supporting low-level vision tasks. DehazeFormer [13] combines window attention, local convolution, and SK Fusion to merge multiscale encoder–decoder features. It achieves strong PSNR and SSIM results on public datasets such as RESIDE compared with traditional CNNs, making it a representative Transformer dehazing method. Diffusion models have recently been applied to image restoration, using iterative denoising to reconstruct degraded images and improve texture and perceptual quality. Latent Diffusion Models (LDMs) [32] lower computation by operating in latent rather than pixel space. Cui et al. [33] proposed conditional diffusion with parallel multiconvolution attention to address blurred details and incomplete real-scene dehazing. Wang et al. [34] used frequency compensation to enhance mid- and high-frequency details in complex real-world haze. However, multistep generation increases inference time and computation, limiting real-time applications. In summary, CNNs, Transformers, and diffusion models have advanced dehazing through local feature extraction, dependency modeling, and generative restoration, respectively. Given the balance between feature modeling and computational efficiency, this study uses DehazeFormer to investigate indoor dehazing.

2.2 Research on Feature Enhancement Methods for Image Dehazing

Attention-based feature enhancement assigns response weights according to dependencies among features, improving the use of informative representations. Restormer, proposed by Zamir et al. [35], uses efficient attention and gated transformations to balance modeling capacity and computational cost in high-resolution restoration. These studies show that attention can emphasize useful information during multiscale feature transfer. PromptIR, proposed by Potlapalli et al. [36], uses task-related prompts to adjust features for different degradations. DehazeFormer [13] combines window attention and SK Fusion for hazy-image feature extraction and cross-scale fusion. Frequency-based enhancement analyzes different frequency components to extract and fuse structure and texture information in degraded images. Low frequencies mainly reflect overall brightness and structure, whereas high frequencies capture edges and fine details. FFTformer, proposed by Kong et al. [37], introduced frequency processing into image deblurring to enhance restoration. Adaptive fusion exploiting complementary high- and low-frequency information can improve structure and detail recovery in dehazing. In summary, we introduce and adapt DFAB in DehazeFormer, drawing on an existing dual-frequency enhancement method [38] to strengthen both structure and detail representations.

3 Proposed DLF-DehazeFormer Method

3.1 Overall Framework

DLF-DehazeFormer uses DehazeFormer as its backbone, combining DFAB-based feature enhancement and multi-loss supervision to improve intermediate features and output supervision, reduce texture blurring and edge-detail loss, and enhance indoor dehazing quality. Its architecture is shown in Fig. 1. As shown in Fig. 1, DLF-DehazeFormer includes a feature-enhanced dehazing network, multi-loss supervision, and parameter optimization. DFAB enhances structure and texture representation in the decoder, while a composite L1, VGG, and Edge loss guides parameter updates through backpropagation. Together, these strategies improve feature representation and supervision.

3.2 DFAB-Based Feature Enhancement

3.2.1 DFAB Module

The Dual Frequency Adaptive Block (DFAB) in Fig. 2 comprises dual-frequency dynamic filtering, Channel Attention, and Adaptive Weight Generation. An initial 1×1 convolution maps the input channels to a suitable dimension. Parallel low-frequency (LP) and high-frequency (HP) branches then model information at different scales. Channel attention adaptively selects useful features.

As shown in Fig. 3, the LP branch uses adaptive average pooling at 1×1, 2×2, 3×3, and 6×6 to extract low-frequency structure, while the HP branch uses adaptive max pooling at 1×1, 4×4, 8×8, and 16×16 to enhance local textures and details. Both branches upsample features to the original size, concatenate them along channels, and apply ReLU, yielding complementary low- and high-frequency features for subsequent fusion.

As shown in Fig. 2(b), LP and HP outputs are concatenated along the channel dimension. Global average pooling produces a spatial descriptor, followed by two 1×1 convolutions, ReLU, and Softmax to obtain two normalized branch weights: the low-frequency weight (w_l) and the high-frequency weight (w_h). The low- and high-frequency features are multiplied by w_l and w_h, respectively, and added elementwise to yield a dynamically fused representation. These weights indicate the relative importance of each frequency branch in the current input features.

Channel attention further adjusts feature-channel responses after dual-frequency fusion, as shown in Fig. 2(c). As shown in Fig. 4, Global Average Pooling extracts channelwise information; two convolutions and ReLU provide nonlinear mapping, and Sigmoid generates channel attention weights as follows:

$$w = \sigma(\mathrm{Conv}_2(\mathrm{ReLU}(\mathrm{Conv}_1(\mathrm{GAP}(X))))) \quad (1)$$

where X denotes the input feature map, and C, H, and W denote its number of channels,

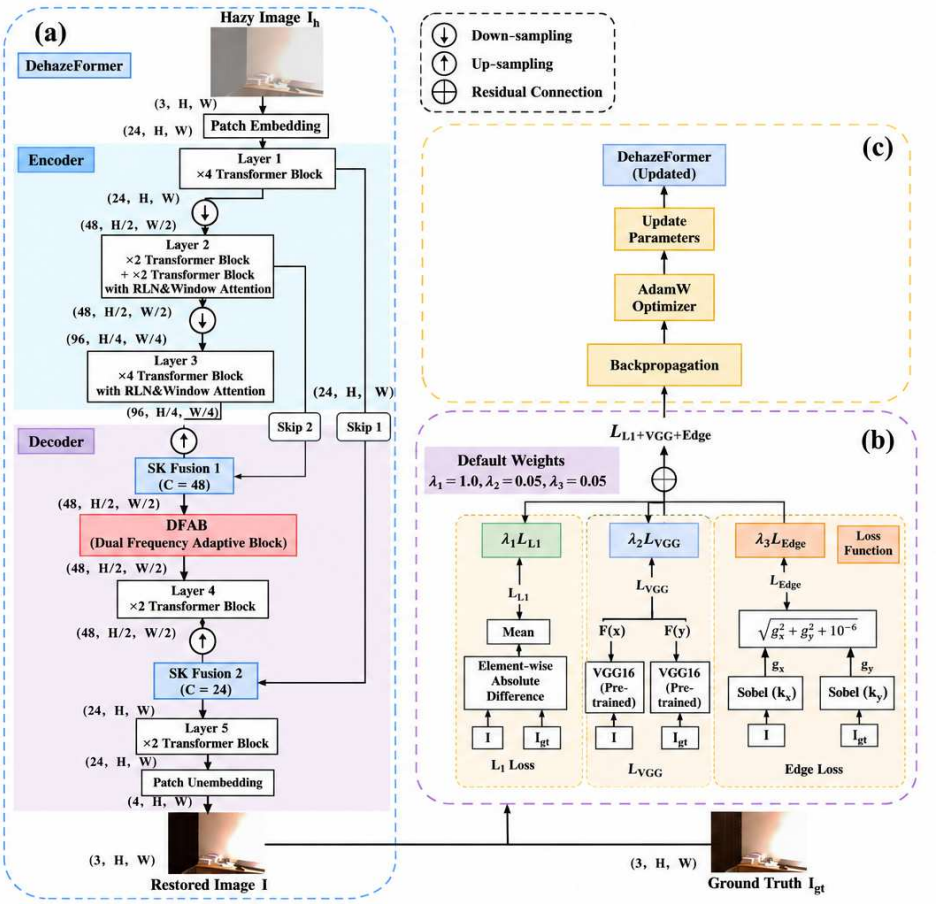

Fig. 1 Overall architecture of the DLF-DehazeFormer dual-level collaborative optimization framework. (a) Feature-enhanced dehazing network: a hazy input image passes through shallow feature embedding, multistage Transformer encoding and decoding, and cross-stage SK Fusion; DFAB is introduced after the first fusion to strengthen high- and low-frequency representations before reconstructing a clear image. (b) Multi-loss joint supervision strategy: L1 pixel loss, VGG perceptual loss, and Edge loss jointly constrain the recovered image, with weighted summation producing the total loss. (c) Model parameter optimization: backpropagation of the composite loss and AdamW optimization update the model parameters to achieve collaborative optimization of feature enhancement and loss supervision.

height, and width, respectively. GAP denotes global average pooling. Conv_1 and Conv_2 are 1×1 convolutions with channel transformation $C \to C/8 \to C$. ReLU and σ are activation functions, and w denotes the channel attention weights. The generated weight w is multiplied channel-wise with the input features, adaptively adjusting channel responses without changing the feature dimensions.

3.3 Multi-Loss Joint Supervision Strategy

The original DehazeFormer trains with L1 loss by minimizing absolute pixel errors between restored and ground-truth images. This constrains pixel reconstruction but not perceptual features or edge details, potentially yielding blurred textures and insufficient local detail. Therefore, this study also examines VGG perceptual loss, SSIM loss, and Edge loss to assess multidimensional supervision in dehazing.

L1 loss measures the absolute pixelwise error between the restored image and the ground-truth image:

$$L_1 = \frac{1}{N}\sum_{i=1}^{N} |I_i - J_i| \tag{2}$$

where N is the number of compared pixel values, and I_i and J_i denote values from the restored and ground-truth images, respectively.

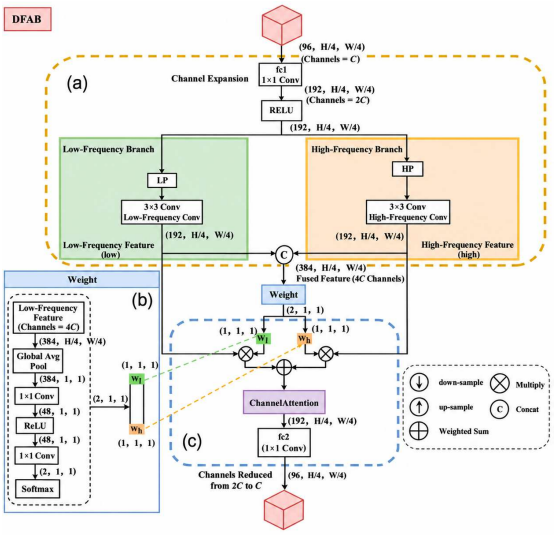

Fig. 2 Architecture of DFAB. (a) Dual-frequency feature extraction: the input feature channels are expanded and processed by low-frequency (LP) and high-frequency (HP) branches, followed by feature concatenation. (b) Adaptive weight generation: global average pooling, convolutions, and Softmax yield fusion weights for the two branches. (c) Dual-frequency fusion and channel enhancement: the high- and low-frequency features are adaptively weighted and fused, followed by channel attention and convolutional mapping to obtain enhanced output features.

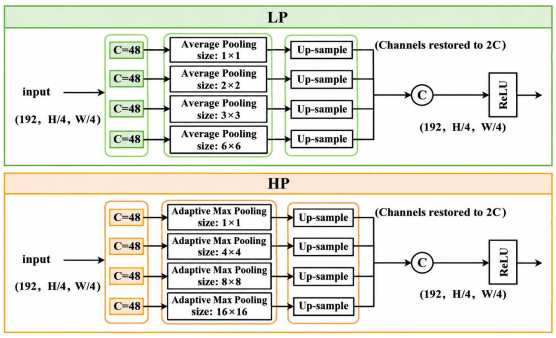

Fig. 3 Dual-frequency dynamic filtering module, in which low- and high-frequency branches extract structural information and texture details at multiple scales.

VGG perceptual loss measures differences between features extracted by a pretrained VGG16 network:

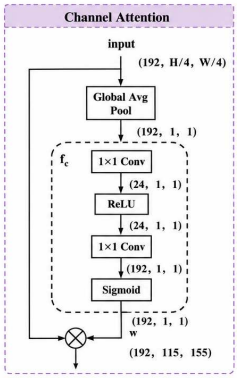

Fig. 4 Channel attention module, which recalibrates input features through adaptive channel weighting to strengthen informative feature responses.

$$L_{\text{VGG}} = \frac{\sum_{b,c,h,w} |\phi(I)_{b,c,h,w} - \phi(J)_{b,c,h,w}|}{BC_\phi H_\phi W_\phi} \quad (3)$$

where $\phi(\cdot)$ denotes the feature map output by the first 16 feature extraction modules of a pretrained VGG16 network, with the final module being the $\text{ReLU}_{3,3}$ activation; B is the batch size, and C_ϕ, H_ϕ, and W_ϕ denote the number of channels, height, and width of the extracted feature map, respectively.

Edge loss compares image edge information using Sobel gradients:

$$L_{\text{Edge}} = \frac{1}{N} \sum_{i=1}^{N} |G(I)_i - G(J)_i| \quad (4)$$

where $G(\cdot)$ denotes the gradient extraction operator, and $G(I)_i$ and $G(J)_i$ are the gradient responses of the restored and ground-truth images at location i.

SSIM loss uses luminance, contrast, and structural similarity as supervision criteria. It is evaluated only as a candidate in the loss-function ablation study and is not included in the final loss combination.

$$L_{\text{SSIM}} = 1 - \text{SSIM}(I, J) \quad (5)$$

where $\text{SSIM}(I, J)$ denotes the structural similarity index between the restored and ground-truth images. SSIM loss is examined as a candidate in the ablation study but is excluded from the final composite loss. To assess individual loss contributions, multiple combinations are constructed

using L1 loss as the baseline. According to the experiments, the final composite loss combines L1, VGG, and Edge losses as follows:

$$L_{\text{total}} = \alpha L_1 + \beta L_{\text{VGG}} + \gamma L_{\text{Edge}} \tag{6}$$

The loss weights are $\alpha = 1.0$ for L_1, $\beta = 0.05$ for L_{VGG}, and $\gamma = 0.05$ for L_{Edge}.

4 Experimental Results and Analysis

4.1 Experimental Setup

All experiments use one NVIDIA RTX 4090 GPU with 24 GB of VRAM, an AMD EPYC 7K62 48-Core Processor, and 58 GB of RAM. The software environment uses Python 3.8, PyTorch 1.11.0, and CUDA 11.3.

The public RESIDE (REalistic Single Image DEhazing) dataset is used for training and evaluation. Its Indoor Training Set (ITS) provides 13,990 image pairs for training, and SOTS-indoor provides 500 test pairs to evaluate restoration under synthetic indoor haze.

During training, images are randomly cropped into 256×256 patches and augmented through horizontal flips and random rotations to diversify samples and improve generalization. At test time, inference uses full-size images without random cropping or augmentation to assess complete-image restoration.

DehazeFormer-T is selected as the baseline, and all experiments use the same training configuration, as detailed in Table 1.

Image restoration quality and computational complexity are evaluated using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and Multiply–Accumulate Operations (MACs).

PSNR measures pixel-level reconstruction quality using the mean squared error between the restored and ground-truth images, and is calculated as:

Table 1 Training parameter settings

Parameter	Setting
Optimizer	AdamW
Initial learning rate	4×10^{-4}
Epoch	300
Batch Size	32
Input size	256 × 256
Learning rate strategy	CosineAnnealing

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right) \tag{7}$$

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (I_i - J_i)^2 \tag{8}$$

where MAX_I is the maximum permissible pixel value, N is the number of compared values, and I_i and J_i denote restored and ground-truth image values at pixel location i, respectively. SSIM quantifies similarity between the restored and ground-truth images in terms of luminance, contrast, and structure, and is calculated as:

$$\text{SSIM}(I, J) = \frac{(2\mu_I\mu_J + C_1)(2\sigma_{IJ} + C_2)}{(\mu_I^2 + \mu_J^2 + C_1)(\sigma_I^2 + \sigma_J^2 + C_2)} \tag{9}$$

where μ_I and μ_J are the local means of the two images; σ_I^2 and σ_J^2 are their local variances; σ_{IJ} is their local covariance; and C_1 and C_2 are constants for numerical stability. LPIPS extracts image features using a pretrained deep network and calculates weighted distances between corresponding feature layers, as follows:

$$\text{LPIPS}(I, J) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| \omega_l \odot (\hat{f}_l(I)_{h,w} - \hat{f}_l(J)_{h,w}) \right\|_2^2 \tag{10}$$

where $\hat{f}_l(\cdot)$ denotes the channel-normalized deep feature vector at spatial location (h, w) in layer l; H_l and W_l are the height and width of that layer's feature map; and ω_l denotes its learned channelwise weighting vector. The symbol $\odot$ denotes elementwise multiplication. MACs

measure computational complexity during inference; lower values indicate lower computational-resource use and higher efficiency.

4.2 Experimental Results and Discussion

Under the settings in Section 4.1, we report comparisons among dehazing methods, DFAB insertion and component ablations, composite-loss ablations, and visual results, together with parameter-count and computational-cost analysis.

4.2.1 Performance Comparison of Dehazing Methods

To verify the effectiveness of the proposed method, DLF-DehazeFormer is compared with the original DehazeFormer (DehazeFormer-T).

Table 2 shows clear differences in restoration accuracy and computation. AOD-Net is simple and computationally light, but its limited feature representation restricts texture and structure recovery. ASTA requires more computation yet achieves low restoration accuracy, indicating that additional computation does not necessarily improve performance. Among Transformer methods, UHDformer achieves relatively good PSNR and SSIM but requires many parameters and substantial computation. UHD Pormer has lower MACs but more parameters and weak restoration on the indoor test set. DehazeFormer combines window attention and multiscale fusion to achieve good restoration at relatively low cost, although its limited local-detail representation and single supervision objective leave room for improvement. DLF-DehazeFormer achieves the best PSNR and SSIM among the compared methods, improving over DehazeFormer by 0.93 dB and 0.0021. DFAB adds 0.18 M parameters and 2.88 G MACs, but the model remains smaller than ASTA, UHDformer, and UHD Pormer and uses fewer MACs than ASTA and UHDformer. Thus, DFAB-based feature enhancement and multi-loss supervision improve restoration accuracy at a moderate cost in complexity.

4.2.2 Ablation Experiments on the DFAB Module

To evaluate DFAB independently of multi-loss gains, all DFAB ablations use only the baseline L1 loss. We examine its insertion position and internal components.

To assess where DFAB works best in DehazeFormer, we test Encoder DFAB, Decoder DFAB, and Encoder+Decoder DFAB, since encoder and decoder stages contain different semantic and spatial features. Table 3 presents the results.

Table 3 shows that Decoder DFAB performs best: PSNR improves by 0.50 dB and SSIM by 0.0008 over the baseline, with a low LPIPS value. Encoder DFAB also helps, but using DFAB in both stages gives no further improvement and slightly reduces PSNR below the baseline, possibly because of redundant enhancement. We therefore use Decoder DFAB after Fusion1 and before Layer4.

To evaluate the contribution of DFAB components, we ablate dynamic-frequency fusion and channel attention at the selected insertion position. Table 4 reports the results.

Table 4 shows that dynamic-frequency processing is the main contributor: removing it reduces PSNR from 35.54 to 35.10 dB and SSIM from 0.9900 to 0.9895, while increasing LPIPS from 0.0079 to 0.0089. Removing channel attention slightly increases SSIM (0.9903 versus 0.9900) and leaves LPIPS unchanged (0.0079), but reduces PSNR to 35.47 dB. Thus, channel attention provides a modest PSNR benefit without improving every metric. We retain the complete DFAB for subsequent joint optimization and assess its overall effectiveness alongside the final model results (35.97 dB PSNR and 0.9913 SSIM; Table 2).

4.2.3 Ablation Experiments on Composite Loss Functions

To assess loss combinations, L1 alone is used as the baseline, with VGG perceptual, Edge, and SSIM losses added in different combinations. Table 5 reports the results.

The quantitative comparisons yield the following observations:

When optimized using only L1 loss, the network achieves a PSNR of 35.04 dB and an SSIM of 0.9892. With the L1 baseline: Adding VGG perceptual loss (L1+VGG) or Edge loss (L1+Edge) improves performance, whereas L1+SSIM produces an anomalous result. One possible explanation is that RESIDE synthetic indoor haze is

Table 2 Performance evaluation of the dual-level collaborative optimization method combining feature enhancement and multi-loss supervision

Methods	ITS / SOTS-indoor		Overhead	
	PSNR ↑	SSIM ↑	#Param (M) ↓	MACs (G) ↓
AOD-Net	17.59	0.7886	0.0018	0.50G
ASTA	16.70	0.7618	3.32	39.03G
UHDformer	34.24	0.9841	336.72	16.01G
UHD Pormer	16.07	0.7710	740.89	3.02G
DehazeFormer	35.04	0.9892	0.69	6.25G
DLF-DehazeFormer	**35.97**	**0.9913**	0.87	9.13G

Note: ↑ denotes larger-is-better metrics, and ↓ denotes smaller-is-better metrics. Bold values indicate the best results for the corresponding evaluation metric. #Param(M) denotes the number of parameters (in millions), and MACs denotes multiply–accumulate operations.

Table 3 Quantitative evaluation of different DFAB insertion strategies

Variant	DFAB insertion position	RESIDE-IN		
		PSNR ↑	SSIM ↑	LPIPS ↓
Baseline	—	35.04	0.9892	—
Encoder DFAB	After Layer3, before PatchSplit1	35.12	0.9898	0.0082
Decoder DFAB	After Fusion1, before Layer4	**35.54**	**0.9900**	**0.0079**
Encoder+Decoder DFAB	Both encoder and decoder	34.97	**0.9900**	0.0081

Note: Bold values indicate the best results for the corresponding evaluation metrics.

relatively uniform and strong edge-gradient constraints may amplify inherent dataset noise. In the current network, SSIM gradients may also conflict with L1 gradients during backpropagation, leading to a local optimum and reduced structural coherence.

Building on pairwise loss combinations, the experiments further evaluate combinations of three and four loss terms: The L1+VGG+Edge combination achieves the best overall performance with a clear combined gain. Compared with L1+VGG, PSNR rises to 35.34 dB, SSIM to 0.9904, and LPIPS falls to the group minimum of 0.0077. Combinations containing SSIM generally perform poorly. L1+VGG+SSIM drops to 29.36 dB PSNR and 0.9561 SSIM, while L1+VGG+Edge+SSIM reaches only 31.32 dB and 0.9674, far below combinations without SSIM. These results strongly indicate that, under the tested settings, SSIM loss has an optimization direction inconsistent with the other high-frequency or perceptual losses; combining them causes serious gradient interference and disrupts the learned feature mappings.

In summary, VGG and Edge losses substantially improve DehazeFormer, while SSIM loss causes optimization conflicts and harms restoration under the current training configuration. Considering PSNR, SSIM, and LPIPS, L1+VGG+Edge achieves the best overall performance.

4.2.4 Visual Comparison of Results

5 Conclusions and Future Work

This paper proposes DLF-DehazeFormer to address the limitations of local-detail feature representation and single-loss supervision in DehazeFormer. DFAB strengthens feature representation through high- and low-frequency extraction and adaptive fusion, while a composite-loss supervision strategy improves training. On SOTS-indoor, the model achieves 35.97 dB PSNR and 0.9913 SSIM, demonstrating improved image restoration and reconstruction quality. Future work will evaluate complex datasets from real outdoor scenes

Table 4 Ablation study of key components in DFAB

Variant	Dynamic Frequency	Channel Attention	PSNR	SSIM	LPIPS
DFAB w/o Frequency	×	✓	35.10	0.9895	0.0089
DFAB w/o Attention	✓	×	35.47	**0.9903**	**0.0079**
DFAB	✓	✓	**35.54**	0.9900	**0.0079**

Note: ✓ indicates that the corresponding component is retained, whereas × indicates that it is removed.

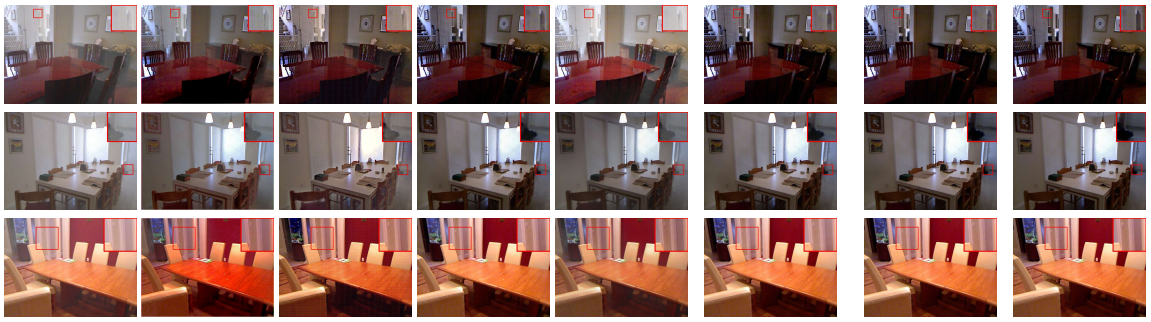

Fig. 5 Qualitative comparison of image dehazing results on three indoor scenes from the SOTS-indoor test set. Each row represents a different test scene. From left to right, the columns show the hazy input image (Input), the dehazing results obtained using AOD-Net, ASTA, UHDformer, UHD Pormer, DehazeFormer-T, and the proposed DLF-DehazeFormer (Ours), followed by the corresponding ground-truth clear image (GT).

Table 5 Ablation results of different loss-function combinations

Methods	PSNR	SSIM	LPIPS
L1	35.04	0.9892	—
L1+VGG	35.14	0.9896	0.0089
L1+Edge	35.31	0.9897	0.0087
L1+SSIM	31.22	0.9689	0.0377
L1+VGG+Edge	**35.34**	**0.9904**	**0.0077**
L1+VGG+SSIM	29.36	0.9561	0.0385
L1+Edge+SSIM	31.66	0.9718	0.0253
L1+VGG+Edge+SSIM	31.32	0.9674	0.0224

Note: Bold values indicate the best results for the corresponding evaluation metrics.

and high-altitude remote sensing to improve generalization and robustness under adverse weather conditions.

Acknowledgements This work was supported by Science Research Project of Hebei Education Department (QN2026771) and the International Science and Technology Cooperation Project of Hebei Provincial Department of Science and Technology(Grant No. 25290102D).

Statements and Declarations

Competing Interests The authors declare that they have no competing interests.

Data Availability The data used in this study are publicly available from the RESIDE (REalistic Single Image DEhazing) benchmark. The Indoor Training Set (ITS) was used for training, and the indoor subset of the Synthetic Objective Testing Set (SOTS-indoor) was used for evaluation. Both subsets are available from the official RESIDE dataset page at https://sites.google.com/view/reside-dehaze-datasets/reside-v0. No custom-developed source code is publicly released for this work.

Author Contributions Ruzhen Xie: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. Ya Guo: Software, Validation, Writing – review & editing, Supervision, Project administration, Funding acquisition.

References

[1] Han, X., Chen, S., Fu, Z., Feng, Z., Fan, L., An, D., Wang, C., Guo, L., Meng, W., Zhang, X., Xu, R., Xu, S.: Multimodal fusion and vision–language models: a survey for robot vision. Inf. Fusion 126, 103652 (2026). https://doi.org/10.1016/j.inffus.2025.103652

[2] Yang, J., Chen, J., Xu, H.P., Wang, X.L., You, Y.N., Feng, X.: A review of joint target detection and recognition techniques for microwave and optical remote sensing imagery. J. Image Graph. 31, 1671–1688 (2026). https://doi.org/10.11834/jig.250648 (in Chinese)

[3] Deng, W., Gao, Y.: Remote sensing image target detection model based on improved YOLOX. In: Proceedings of the 42nd Chinese Control Conference (CCC), pp. 153–158. IEEE, Tianjin (2023)

[4] Shi, J., Wang, T.T., Zhu, Z.Q., Zhao, M.F., Wang, B.X., An, H.: A review of deep learning-based medical image segmentation methods. J. Image Graph. 30, 2161–2186 (2025). https://doi.org/10.11834/jig.240467 (in Chinese)

[5] Wang, N., Zhi, M.: A review of single-stage generic object detection algorithms based on deep learning. J. Front. Comput. Sci. Technol. 19, 1115–1140 (2025). https://doi.org/10.3778/j.issn.1673-9418.2411032 (in Chinese)

[6] Valanarasu, J.M.J., Yasarla, R., Patel, V.M.: TransWeather: transformer-based restoration of images degraded by adverse weather conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2343–2353. IEEE, New Orleans, LA (2022). https://doi.org/10.1109/CVPR52688.2022.00239

[7] Liu, B., Chen, S.B., Wang, J.X., Tang, J., Luo, B.: An oriented object detector for hazy remote sensing images. IEEE Trans. Geosci. Remote Sens. 62, 1–11 (2024). https://doi.org/10.1109/TGRS.2024.3451492

[8] Jia, T.Y., Zhuo, L., Li, J.F., Zhang, J.: Research progress in deep learning-based single-image dehazing. Acta Electron. Sin. 51, 231–245 (2023). https://doi.org/10.12263/DZXB.20220838 (in Chinese)

[9] Guo, F., Yang, J., Liu, Z., Tang, J.: Haze removal for single image: a comprehensive review. Neurocomputing 537, 85–109 (2023). https://doi.org/10.1016/j.neucom.2023.03.061

[10] Wei, Y.L., Xu, H.W.: A review of deep learning-based image dehazing methods. Inf. Technol. Inform. (4), 214–216, 220 (2023) (in Chinese)

[11] 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. 25, 5187–5198 (2016). https://doi.org/10.1109/TIP.2016.2598681

[12] 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 (ICCV), pp. 4780–4788. IEEE, Venice (2017). https://doi.org/10.1109/ICCV.2017.511

[13] Song, Y., He, Z., Qian, H., Du, X.: Vision transformers for single image dehazing. IEEE Trans. Image Process. 32, 1927–1941 (2023). https://doi.org/10.1109/TIP.2023.3256763

[14] Jin, W.S., Li, J., Gao, X.B.: A review of Mamba-based image restoration methods. J. Image Graph. 31, 1875–1896 (2026). https://doi.org/10.11834/jig.260111 (in Chinese)

[15] Raju, N., Kankanala, S., Thatikonda, R., Chintakindi, B.M.: A review of real-time single image dehazing architectures for embedded vision under adverse weather conditions. IEEE Access 14, 13264–13304 (2026). https://doi.org/10.1109/ACCESS.2026.3655979

[16] McCartney, E.J.: Optics of the Atmosphere: Scattering by Molecules and Particles. Wiley, New York (1976)

[17] Narasimhan, S.G., Nayar, S.K.: Vision and the atmosphere. Int. J. Comput. Vis. 48, 233–254 (2002). https://doi.org/10.1023/A:1016328200723

[18] He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1956–1963. IEEE, Miami Beach, FL (2009)

[19] Su, Y.Z., He, C., Cui, Z.G., Jiang, K., Cai, Y.P., Li, A.H.: Semi-supervised image dehazing based on multiple prior constraints and consistency regularization. J. Electron. Inf. Technol. 44, 3409–3418 (2022). https://doi.org/10.11999/JEIT220381 (in Chinese)

[20] Li, Q.H., Cui, Z.G., Chen, Y.Q., Lan, Y.W., Cai, Y.P., Su, Y.Z.: A review of dehazing methods for degraded images in complex scenes. J. Rocket Force Univ. Eng. 40, 56–67 (2026) (in Chinese)

[21] Wang, W., Yin, B., Li, L., Li, L., Liu, H.: A low light image enhancement method based on dehazing physical model. Comput. Model. Eng. Sci. 143, 1595–1616 (2025). https://doi.org/10.32604/cmes.2025.063595

[22] Liu, S.T., Sun, C.Z., Lou, H.Y., Long, Y.Q., Xu, C.: Image dehazing based on histogram equalization and Retinex. China Comput. Commun. 35, 172–175 (2023) (in Chinese)

[23] Tang, L., Li, C., Ma, J.: Mask-DiFuser: a masked diffusion model for unified unsupervised image fusion. IEEE Trans. Pattern Anal. Mach. Intell. 48, 591–608 (2026). https://doi.org/10.1109/TPAMI.2025.3609323

[24] He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE Trans. Pattern Anal. Mach. Intell. 33, 2341–2353 (2011). https://doi.org/10.1109/TPAMI.2010.168

[25] Meng, G., Wang, Y., Duan, J., Xiang, S., Pan, C.: Efficient image dehazing with boundary constraint and contextual regularization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 617–624. IEEE, Sydney (2013). https://doi.org/10.1109/ICCV.2013.82

[26] Zhu, Q., Mai, J., Shao, L.: A fast single image haze removal algorithm using color attenuation prior. IEEE Trans. Image Process. 24, 3522–3533 (2015). https://doi.org/10.1109/TIP.2015.2446191

[27] Liao, M., Lu, Y., Zhang, J., Zhao, Y.Q., Di, S.H.: An image dehazing method based on haze lines and color attenuation prior. J. Commun. 44, 211–222 (2023). https://doi.org/10.11959/j.issn.1000-436x.2023013 (in Chinese)

[28] Liu, Y., Li, T., Tan, C., Ren, W., Ancuti, C., Lin, W.: IHDCP: single image dehazing using inverted haze density correction prior. IEEE Trans. Image Process. 35, 1448–1461 (2026). https://doi.org/10.1109/TIP.2026.3657636

[29] Cui, N.S., Tang, P., Zheng, X.D., Wei, M.Z., Guo, W.J.: Recent advances in image dehazing technology. Mech. Electr. Eng. Technol. 54, 1–8 (2025) (in Chinese)

[30] Wang, Y., He, B.: CasDyF-Net: image dehazing via cascaded dynamic filters. In: Proceedings of the IEEE 8th International Conference on Vision, Image and Signal Processing (ICVISP), pp. 1–8. IEEE, Kunming (2024). https://doi.org/10.1109/ICVISP64524.2024.10959384

[31] 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 (CVPR), pp. 17662–17672. IEEE, New Orleans, LA (2022). https://doi.org/10.1109/CVPR52688.2022.01716

[32] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR), pp. 10674–10685. IEEE, New Orleans, LA (2022). https://doi.org/10.1109/CVPR52688.2022.01042

[33] Cui, X.T., Wang, Y., Deng, Z.N., Wang, X.X., Liang, J., Deng, H.X.: Diffusion model image dehazing incorporating parallel multi-convolution attention. Comput. Eng. Des. 46, 2081–2088 (2025). https://doi.org/10.16208/j.issn1000-7024.2025.07.031 (in Chinese)

[34] Wang, J., Wu, S., Yuan, Z., Tong, Q., Xu, K.: Frequency compensated diffusion model for real-scene dehazing. Neural Netw. 175, 106281 (2024). https://doi.org/10.1016/j.neunet.2024.106281

[35] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.: Restormer: efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5718–5729. IEEE, New Orleans, LA (2022). https://doi.org/10.1109/CVPR52688.2022.00564

[36] Potlapalli, V., Zamir, S.W., Khan, S.H., Khan, F.S.: PromptIR: prompting for all-in-one image restoration. Adv. Neural Inf. Process. Syst. 36, 71275–71293 (2023). https://doi.org/10.52202/075280-3121

[37] Kong, L., Dong, J., Ge, J., Li, M., Pan, J.: Efficient frequency domain-based transformers for high-quality image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5886–5895. IEEE, Vancouver (2023). https://doi.org/10.1109/CVPR52729.2023.00570

[38] Song, T., Fan, S., Li, P., Jin, J., Jin, G., Fan, L.: Learning an effective transformer for remote sensing satellite image dehazing. IEEE Geosci. Remote Sens. Lett. 20, 8002305 (2023). https://doi.org/10.1109/LGRS.2023.3319832