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.

Transformer Image Dehazing Method Based on Physical Prior Guidance and Adaptive Feature Modulation

Zhangshuo Hu
Southeast University https://orcid.org/0009-0007-2889-9504
Lin Chai
chailin1@seu.edu.cn
Southeast University https://orcid.org/0000-0002-3960-8828

Research Article

Keywords: image dehazing, Transformer, physical prior, dark channel prior, adaptive feature modulation, multi-scale features

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11186464/v2

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

Additional Declarations: The authors declare no competing interests.

Journal Name

Crossmark

RECEIVED
dd Month yyyy

REVISED
dd Month yyyy

PAPER

Transformer Image Dehazing Method Based on Physical Prior Guidance and Adaptive Feature Modulation

Hu Zhangshuo[1] and Chai Lin[1,*]

[1]School of Automation, Southeast University, Nanjing, China
[*]Author to whom any correspondence should be addressed.

E-mail: Hu Zhangshuo: 213232409@seu.edu.cn, +86 13645600960 , Chai Lin: chailin1@seu.edu.cn, +86 13851473410

Keywords: image dehazing, Transformer, physical prior, dark channel prior, adaptive feature modulation, multi-scale features

Abstract

Single-image dehazing remains challenging because haze is spatially non-uniform and its removal can easily compromise fine structures, especially in scenes containing strong illumination changes or dense atmospheric scattering. To address this problem, we propose a Transformer-based dehazing network that combines physical-prior guidance with adaptive feature modulation. The network follows a multi-scale encoder–decoder design. Its encoder couples Swin Transformer blocks with dilated convolutions to model long-range context while retaining haze-sensitive local structures, whereas the proposed adaptive feature modulation framework coordinates spatial attention, edge responses, and input-dependent convolutional kernels across feature scales. A transmission map estimated from the dark channel prior is embedded into the feature update process, allowing the atmospheric degradation pattern to influence feature reconstruction throughout the network rather than serving as a one-time auxiliary input. Experiments on RESIDE OTS/ITS and SOTS-Indoor/Outdoor show that the proposed method achieves 24.57 dB/0.907 on RESIDE OTS, 28.78 dB/0.928 on RESIDE ITS, 40.12 dB/0.9823 on SOTS-Indoor, and 34.28 dB/0.9746 on SOTS-Outdoor. Additional experiments evaluate the method on the held-out test pairs of I-HAZE, O-HAZE, and NH-HAZE, including spatially non-uniform haze. The highest structural similarity index (SSIM) values on the evaluated synthetic comparison sets indicate that the method is particularly effective at preserving structural consistency, while the NH-HAZE results also reveal room to improve structural recovery. Ablation and haze-level analyses further show that the individual modules address different failure modes, and that the complete model degrades more gradually as haze becomes denser. The proposed model's inspiration is illustrated in Fig. 1.

1 Introduction

Single-image dehazing is important for autonomous driving, video surveillance, remote sensing, and other vision systems that must operate under degraded visibility. Atmospheric scattering reduces contrast, weakens edges, and obscures small structures; these effects impair visual quality and can also propagate to downstream tasks such as detection and segmentation [1, 12]. The central difficulty is therefore not merely to increase image brightness, but to recover scene radiance without mistaking haze-induced attenuation for actual texture or color.

Classical methods formulate the degradation with the atmospheric scattering model and estimate the transmission map using assumptions such as the dark channel prior [2]. Their physical interpretation is attractive, yet the assumptions can become unreliable in bright sky regions, highly reflective surfaces, or scenes with non-uniform haze. Learning-based methods subsequently improved restoration quality by learning an end-to-end mapping from hazy images to clear targets [3]. Their flexibility comes at a cost: when the training distribution does not adequately cover the test scene, a network with no explicit degradation constraint may overfit appearance statistics and produce unstable color or structure reconstruction.

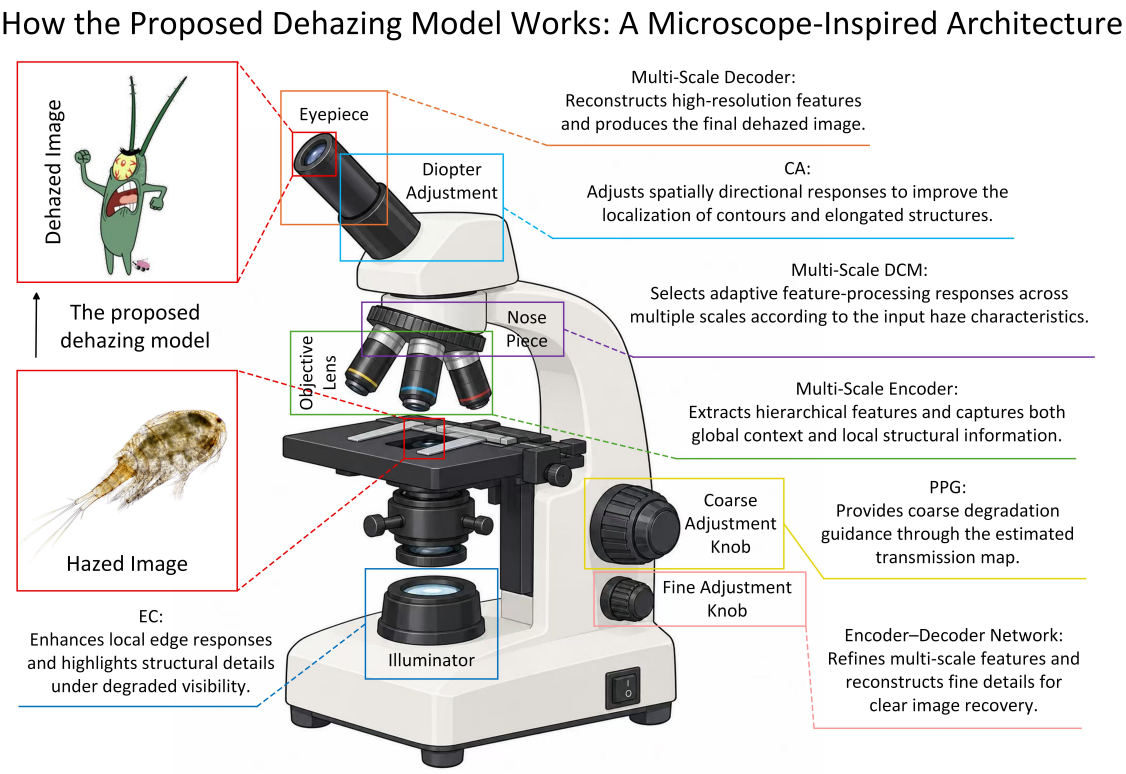


Figure 1. How the proposed dehazing model corresponds to microscope components.

Transformer architectures have recently been adopted to strengthen global dependency modeling in dehazing [4]. Window-based attention can relate distant regions while keeping the computational cost manageable, which is useful when haze affects large parts of a scene in a correlated manner. Nevertheless, most Transformer-based approaches still treat the problem primarily as feature transformation, while the imaging process itself remains weakly constrained [5]. This separation can lead to an unfavorable trade-off: global context may be reconstructed convincingly, but local boundaries and transmission changes can still be softened or misaligned under spatially varying haze.

Motivated by these observations, we develop a physical-prior-guided Transformer dehazing network with adaptive feature modulation. The multi-scale encoder combines Swin Transformer blocks and dilated convolutions so that global context and local high-frequency information can be represented within the same hierarchy. On top of this backbone, the Adaptive Feature Modulation (AFM) framework coordinates three complementary forms of evidence: coordinate attention assigns location-aware responses, Edge Convolution (EC) emphasizes local structural transitions, and Dynamic Convolution selects receptive-field mixtures according to the input content. The transmission estimate is coupled to this modulation process, giving the network an explicit indication of where atmospheric attenuation is likely to be strong.

From a feature-learning perspective, these components can be understood as successive refinements of feature allocation. The physical prior provides a degradation-aware reference, coordinate attention determines where information should be retained, EdgeConv identifies which local transitions are structurally meaningful, and Dynamic Conv adjusts the reconstruction response to the observed haze pattern. This interpretation also provides a common basis for analyzing the contributions of the individual modules.

The main contributions of this work are summarized as follows:

(1) We propose a physically guided adaptive feature modulation framework that carries transmission-related information from the input stage into multi-scale feature updates, thereby converting a physical prior from an auxiliary signal into a persistent modulation cue;

(2) We combine location-aware attention, edge-aware feature extraction, and input-conditioned convolution in a unified feature allocation scheme, allowing spatial, structural, and dynamic responses to be modeled jointly;

(3) We establish a multi-scale Transformer architecture in which physical guidance and adaptive

feature modulation interact throughout reconstruction, and evaluate its structural preservation and robustness across RESIDE, SOTS, and the real-haze datasets I-HAZE, O-HAZE, and NH-HAZE.

2 Related work

2.1 Methods based on physical a priori

Traditional dehazing methods generally estimate the transmission map and atmospheric light within the atmospheric scattering model. The dark channel prior exploits the observation that, in many haze-free natural images, at least one color channel tends to contain very low-intensity pixels within a local patch [2]. Subsequent work has refined transmission estimation and post-processing, for example by using guided filtering to reduce block artifacts or by incorporating color and luminance cues to correct the initial estimate.

The appeal of this family lies in its transparent physical interpretation. Its limitations become apparent when the scene violates the underlying statistics: sky and bright objects may not contain the expected dark pixels, while non-uniform haze can make a locally estimated transmission inconsistent with neighboring regions. In addition, the usual pipeline separates estimation from reconstruction, so errors in the prior can be amplified during later processing. These issues motivate a design in which physical information remains available during feature learning and can be corrected by image evidence.

2.2 Methods based on deep learning

End-to-end networks learn the mapping from hazy observations to clear images directly. Convolutional architectures extract hierarchical features [6, 7], while multi-scale fusion and attention mechanisms improve the representation of regions with different degradation patterns [8]. On public benchmarks, such models generally outperform hand-crafted pipelines because they can learn restoration cues that are difficult to encode with a fixed analytical rule.

Their data-driven nature also exposes two weaknesses. Performance can decline when the test scenes differ from the training distribution, and local convolution alone may struggle to represent a haze field whose effects extend over both broad regions and fine boundaries. In practice, this often appears as over-smoothed textures, incomplete edge recovery, or color compensation that is visually plausible but physically inconsistent. A useful architecture must therefore combine flexible appearance modeling with a representation of the degradation pattern itself.

Recent surveys organize single-image dehazing into enhancement-based, prior-based, and learning-based approaches, and discuss the associated architectures, datasets, evaluation metrics, and open challenges [46]. A complementary quantitative survey reviews deep-learning-based dehazing methods, benchmarks, evaluation metrics, and open challenges, including CNN-, GAN-, recurrent-, and Transformer-based approaches [47]. Recent learned methods have also diversified beyond conventional convolutional designs: QCNN-H uses quaternion neural networks, DHM combines hybrid restoration with detail refinement, T-Net adopts scale iteration, RIDCP uses codebook priors for real-image restoration, and RABNet introduces residual attention [34, 35, 36, 37, 39]. Other work targets the synthetic-to-real domain gap through unsupervised contrastive learning in UCL-Dehaze and unpaired diffusion-prior learning in Diff-Dehazer [42, 17]. These directions broaden the representational and training strategies available for dehazing, while generalization across haze distributions remains an important challenge.

2.3 Methods based on Transformer

Transformer models were introduced into low-level vision to improve the representation of long-range dependencies. In dehazing, window-based attention mechanisms [4] provide a practical compromise between global interaction and computational cost. They are well suited to scenes in which haze attenuation is correlated across distant regions, and they can complement convolutional operators that focus on local texture.

Recent Transformer-based dehazing methods include DehazeFormer and MB-TaylorFormer, which explore hierarchical contextual modeling and computationally efficient multi-branch attention, respectively [5, 38]. For dense and spatially non-homogeneous haze, the NTIRE 2024 challenge report documents recent approaches evaluated on the DNH-HAZE benchmark [48]. Together, these studies show continued progress in global context modeling and real-haze restoration.

Most existing Transformer-based dehazing methods, however, still rely mainly on architectural feature modeling. Improvements in attention or backbone design do not by themselves guarantee that the reconstructed features follow the atmospheric imaging process. When haze is spatially

non-uniform or illumination changes sharply, global modeling alone may recover scene-level appearance while weakening boundaries or introducing local brightness errors. The remaining gap is the coupling between global representation and degradation-aware guidance.

2.4 Fusion methods and existing problems

Hybrid methods attempt to address this gap by supplying transmission estimates as additional inputs or by adding atmospheric-scattering-related losses to a deep network. Such strategies can improve robustness, but in many cases the physical cue remains an extra source of information rather than an active part of feature formation [8]. If the prior is injected only at the input, its influence may diminish after repeated down-sampling and nonlinear transformation.

This loose fusion also leaves two practical questions unresolved. How can prior information remain effective across multiple feature scales, particularly when haze boundaries do not coincide with semantic boundaries? How should spatial selection, edge enhancement, and receptive-field adaptation interact instead of operating as isolated add-on blocks? A unified modulation mechanism is needed to answer both questions while preserving the flexibility of data-driven reconstruction.

Accordingly, this work treats the transmission estimate as a continuous feature-modulation cue rather than a one-time auxiliary input. The resulting design couples physical guidance with spatial, structural, and dynamic feature selection, providing the basis for the network described below.

3 Method

3.1 Overall Framework

The architecture of the proposed physical-prior-guided multi-scale Transformer dehazing network is illustrated in Fig. 2. The network follows an encoder–decoder framework with three closely coupled components: a physical-prior guidance branch, a hierarchical Transformer backbone, and Dynamic Convolution modules. This organization combines the interpretability of atmospheric scattering with the flexibility of learned feature representations. The prior supplies degradation-aware information, while the backbone and modulation modules determine how that information should be used for reconstruction.

Throughout the equations, scalar variables, indices, and single-symbol functions are italic; vectors and feature tensors are bold italic; descriptive subscript labels, named operators, and metric abbreviations are upright.

Given a hazy image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, the dark channel prior (DCP) is used to estimate a transmission map $\boldsymbol{T}$. The physical-prior guidance (PPG) branch incorporates this estimate through a Prior Embedding module, which projects it into the feature space as a degradation-aware cue during feature updating. In parallel, an Edge Convolution (EC) module enhances boundary and local contrast responses. This is particularly useful in dense haze, where the attenuation of high-frequency components can make neighboring regions appear artificially similar.

The backbone contains four hierarchical Transformer stages. Patch embedding and spatial down-sampling produce multi-scale feature representations $\{\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3, \boldsymbol{F}_4\}$ with spatial resolutions of $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the input resolution, respectively. Each stage uses the proposed Dilated Swin Block (DSB), which combines the long-range dependency modeling of Swin Transformer [4] with the enlarged receptive field of dilated convolution. The two branches address complementary aspects of dehazing: the Transformer branch relates distant regions that share a degradation pattern, while the dilated branch retains local structures without discarding spatial resolution.

Let $\boldsymbol{F}_l^{\text{in}}$ denote the input feature of the l-th encoder stage. The corresponding DSB can be formulated as

$$\boldsymbol{F}_l^{\text{out}} = S\big(\boldsymbol{F}_l^{\text{in}}\big) + C_d\big(\boldsymbol{F}_l^{\text{in}}\big)\,, \tag{1}$$

where $S(\cdot)$ denotes the Swin Transformer branch and $C_d(\cdot)$ represents the dilated convolution branch with dilation rate $d = 2$. Shifted-window self-attention provides contextual interaction beyond a local convolutional neighborhood, whereas the dilated branch expands the receptive field while retaining the feature grid. Their sum therefore preserves both scene-level consistency and the local transitions that are easily weakened by haze.

At the bottleneck of the encoder, a Dynamic Convolution module (DCM) adapts feature extraction to the observed haze distribution. DCM receives the prior-guided representation and aggregates a set of candidate convolution kernels according to input-dependent attention weights. This allows the feature extractor to change its kernel composition according to the current feature statistics, while preserving the output-channel dimension of the corresponding static convolution.

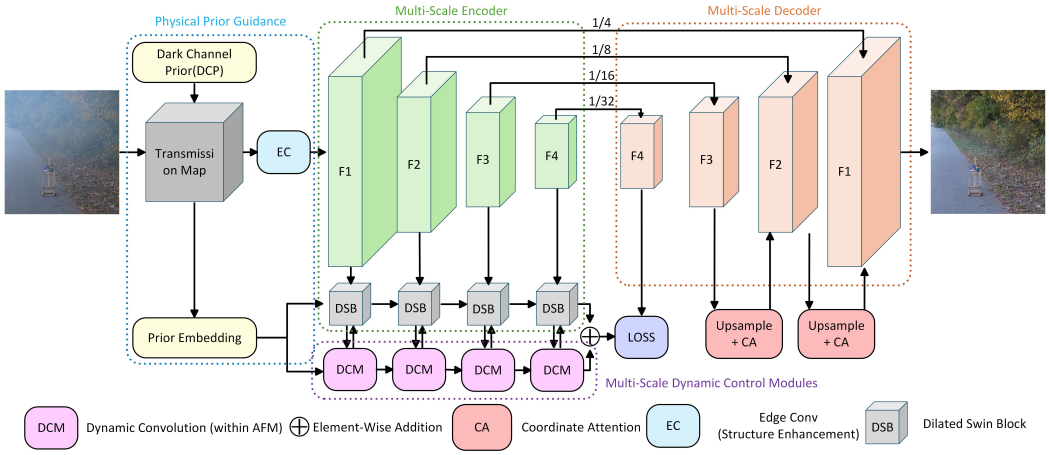


Figure 2. Model overall structure diagram

The decoder reconstructs high-resolution representations by combining hierarchical encoder features through skip connections. Coordinate Attention (CA) modules are inserted during this recovery process to encode horizontal and vertical positional dependencies, which improves the localization of elongated structures, contours, and other spatially organized details. Reconstruction proceeds from coarse context to fine detail and produces the final dehazed image $\boldsymbol{J}$.

The network is trained with a hybrid objective comprising physical consistency, perceptual, and pixel-wise reconstruction terms, following the mixed-loss design in [49]:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{phy}} + \lambda_3 \mathcal{L}_{\text{per}} + \mathcal{L}_{\text{rec}}, \tag{2}$$

The physical consistency term is defined as

$$\mathcal{L}_{\text{phy}} = \lambda_1 \left\| \nabla \widehat{\boldsymbol{T}} \right\|_1 + \lambda_2 \left\| \widehat{\boldsymbol{A}} - \boldsymbol{A}_{\text{gt}} \right\|_2^2, \tag{3}$$

where $\widehat{\boldsymbol{A}}$ is estimated by averaging the 0.1% brightest pixels in the hazy input $\boldsymbol{I}$, and $\widehat{\boldsymbol{T}}$ is estimated by inverting the atmospheric scattering model using $\boldsymbol{I}$, the network output $\boldsymbol{J}$, and $\widehat{\boldsymbol{A}}$:

$$\widehat{\boldsymbol{T}}(\boldsymbol{x}) = \frac{\boldsymbol{I}(\boldsymbol{x}) - \widehat{\boldsymbol{A}}}{\boldsymbol{J}(\boldsymbol{x}) - \widehat{\boldsymbol{A}} + \epsilon}, \qquad \epsilon = 10^{-3}. \tag{4}$$

Here, ∇ denotes the spatial gradient, and $\boldsymbol{A}_{\text{gt}}$ is the ground-truth atmospheric light available for the synthetic training data. The transmission-gradient term discourages spatially irregular transmission estimates, while the atmospheric-light term encourages consistency with the known physical degradation.

The perceptual and reconstruction terms are

$$\mathcal{L}_{\text{per}} = \sum_{k=1}^{2} \left\| \phi_k(\boldsymbol{J}) - \phi_k(\boldsymbol{J}_{\text{gt}}) \right\|_2^2, \qquad \mathcal{L}_{\text{rec}} = \left\| \boldsymbol{J} - \boldsymbol{J}_{\text{gt}} \right\|_1, \tag{5}$$

where $\boldsymbol{J}_{\text{gt}}$ is the haze-free ground-truth image. The feature mappings ϕ_1 and ϕ_2 are extracted from the `conv3_3` and `conv4_3` layers of a pretrained VGG-19 network [28], respectively. The perceptual term preserves feature-level appearance and texture, while the L_1 term constrains pixel-level fidelity and is less sensitive to outliers than an L_2 reconstruction loss. Together, these terms balance physical plausibility, perceptual quality, and accurate image reconstruction.

3.2 Physical Prior Guidance (PPG)

To make the feature representation sensitive to the imaging process, we introduce a physical prior derived from the atmospheric scattering model. The formation of a hazy observation can be written as

$$\boldsymbol{I}(\boldsymbol{x}) = \boldsymbol{J}(\boldsymbol{x}) \cdot \boldsymbol{T}(\boldsymbol{x}) + \boldsymbol{A} \cdot (\boldsymbol{1} - \boldsymbol{T}(\boldsymbol{x})) \tag{6}$$

which provides the physical basis for single-image dehazing.

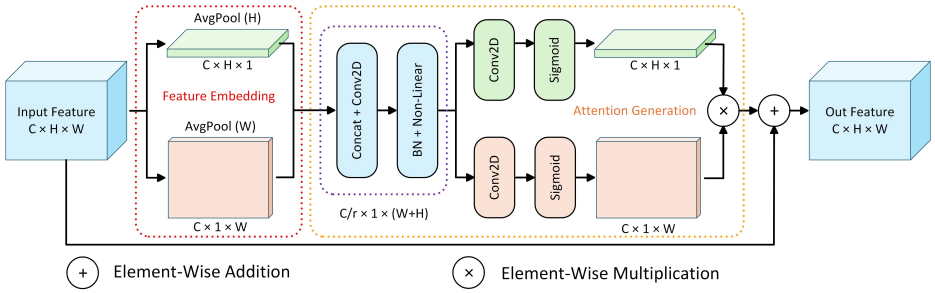


Figure 3. Coordinate Attention (CA) structure diagram

Here, $\boldsymbol{I}(\boldsymbol{x})$ is the observed hazy image, $\boldsymbol{J}(\boldsymbol{x})$ is the scene radiance in the absence of haze, $\boldsymbol{T}(\boldsymbol{x})$ is the transmission, and $\boldsymbol{A}$ is the global atmospheric light. Transmission is related to scene depth according to

$$\boldsymbol{T}(\boldsymbol{x}) = \mathrm{e}^{-\beta d(\boldsymbol{x})} \tag{7}$$

where β denotes the atmospheric scattering coefficient and $d(\boldsymbol{x})$ is the scene depth. Equation (7) gives the physical interpretation of transmission: lower values indicate stronger attenuation. Since $d(\boldsymbol{x})$ and β are not directly available from a single hazy image, we use the dark channel prior to estimate an approximate transmission map:

$$\boldsymbol{T}(\boldsymbol{x}) = \mathbf{1} - \omega \cdot \min_{\boldsymbol{y}\in\Omega(\boldsymbol{x})} \left(\min_{c\in\{\mathrm{r,g,b}\}} \frac{\boldsymbol{I}^c(\boldsymbol{y})}{\boldsymbol{A}^c} \right) \tag{8}$$

Here, $\Omega(\boldsymbol{x})$ denotes the local window centered at $\boldsymbol{x}$, and ω controls the strength of the prior. The estimated map is an approximate spatial cue rather than an exact physical solution. A convolution in the Prior Embedding module maps it to a feature $\boldsymbol{F}_{\mathrm{p}}$, which interacts with encoder features through Dynamic Convolution at multiple scales. Thus, the transmission prior guides feature modulation according to haze attenuation, while learned features can compensate for errors in the estimate:

$$\boldsymbol{F}_l' = \boldsymbol{F}_l + \psi(\boldsymbol{F}_l, \boldsymbol{F}_{\mathrm{p}}) \tag{9}$$

where $\boldsymbol{F}_l$ denotes the feature map at encoder level l, $\boldsymbol{F}_{\mathrm{p}}$ denotes the embedded prior feature, and $\psi(\cdot)$ denotes the Dynamic Conv-based modulation function. Here, l is an italic index for the encoder level, whereas the upright subscript p identifies the prior feature. This formulation allows the estimated attenuation pattern to influence the feature response at multiple scales. Regions with stronger haze cues can receive more restoration capacity, while regions with weak or ambiguous cues can retain information from the learned image representation instead of being forced into an overly rigid physical solution.

Compared with a purely model-based pipeline, the proposed strategy does not require the estimated prior to be exact. The prior supplies a physically meaningful direction, and the data-driven branches learn how that direction should be corrected using image content. This division of roles is important in complex scenes, where a useful transmission cue may be available even though direct inversion would amplify its errors.

Unlike methods that append the prior only at the network input, our formulation inserts physical information into feature updates through $\psi(\cdot)$. The prior can therefore continue to affect feature allocation across scales, which is the mechanism through which the network maintains degradation awareness during reconstruction.

3.3 CA (Coordinate Attention)

To strengthen spatially organized structure recovery, we introduce Coordinate Attention in the decoder. Rather than collapsing all spatial positions into a single channel descriptor, the module aggregates information along the horizontal and vertical directions separately. It can consequently preserve positional cues while assigning channel responses, which is useful for contours, boundaries, and elongated objects that may otherwise be weakened during up-sampling.

The structure of the CA module is shown in Fig. 3.

Let the input feature map be $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$. Following the Coordinate Attention design [9], the module performs one-dimensional global average pooling along the width and height directions,

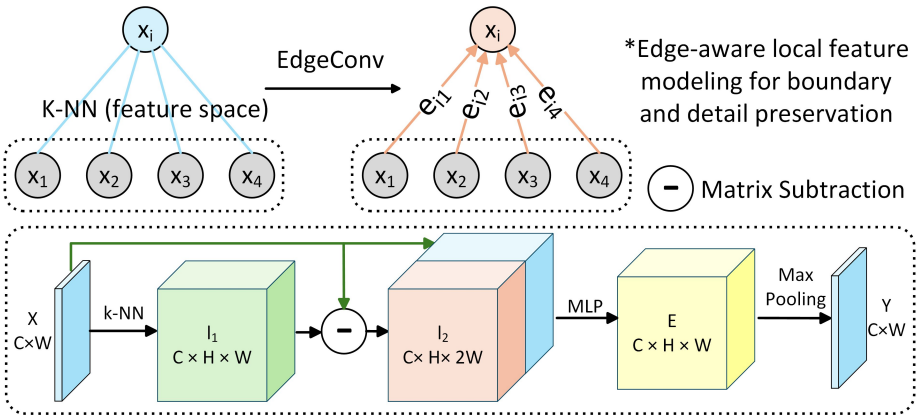


Figure 4. Edge Convolution (Edge Conv) structure diagram

respectively, to obtain two directional feature descriptions:

$$\boldsymbol{X}_{\mathrm{h}}(c,h)=\frac{1}{W}\sum_{w=1}^{W}\boldsymbol{X}(c,h,w),\quad \boldsymbol{X}_{\mathrm{w}}(c,w)=\frac{1}{H}\sum_{h=1}^{H}\boldsymbol{X}(c,h,w), \tag{10}$$

where $\boldsymbol{X}_{\mathrm{h}}\in\mathbb{R}^{C\times H\times 1}$ and $\boldsymbol{X}_{\mathrm{w}}\in\mathbb{R}^{C\times 1\times W}$. To concatenate these descriptors along their coordinate dimension, $\boldsymbol{X}_{\mathrm{w}}$ is transposed to shape $C\times W\times 1$.

The two descriptors are concatenated along the spatial dimension and passed through a shared 1×1 convolution that reduces the channel dimension to C/r, followed by a hard-swish activation ϕ:

$$\boldsymbol{F}=\phi\left(\boldsymbol{W}_{\mathrm{s}}*\left[\boldsymbol{X}_{\mathrm{h}},\boldsymbol{X}_{\mathrm{w}}^{\mathsf{T}}\right]\right), \tag{11}$$

where r is the channel reduction ratio, $\boldsymbol{W}_{\mathrm{s}}$ denotes the shared convolution weights, and $[\cdot]$ denotes concatenation. The resulting feature $\boldsymbol{F}\in\mathbb{R}^{(C/r)\times(H+W)\times 1}$ is split along its spatial dimension into $\boldsymbol{F}_{\mathrm{h}}\in\mathbb{R}^{(C/r)\times H\times 1}$ and $\boldsymbol{F}_{\mathrm{w}}\in\mathbb{R}^{(C/r)\times W\times 1}$.

The two split features are then mapped by independent 1×1 convolutions and a Sigmoid activation to produce directional attention weights:

$$\boldsymbol{A}_{\mathrm{h}}=\sigma(\boldsymbol{W}_{\mathrm{h}}*\boldsymbol{F}_{\mathrm{h}}),\quad \boldsymbol{A}_{\mathrm{w}}=\left[\sigma(\boldsymbol{W}_{\mathrm{w}}*\boldsymbol{F}_{\mathrm{w}})\right]^{\mathsf{T}}, \tag{12}$$

where the width-branch output is transposed back to shape $C\times 1\times W$; thus, $\boldsymbol{A}_{\mathrm{h}}\in\mathbb{R}^{C\times H\times 1}$, $\boldsymbol{A}_{\mathrm{w}}\in\mathbb{R}^{C\times 1\times W}$, and $\sigma(\cdot)$ represents the Sigmoid activation function.

The input features are recalibrated by the two broadcastable directional attention weights:

$$\boldsymbol{Y}=\boldsymbol{X}\odot\boldsymbol{A}_{\mathrm{h}}\odot\boldsymbol{A}_{\mathrm{w}}, \tag{13}$$

where $\odot$ denotes element-wise multiplication. This factorized reweighting encodes channel attention together with horizontal and vertical positional information.

The resulting attention map jointly models channel relevance and directional position information. In the decoder, this gives structurally important locations a more precise response and provides the subsequent reconstruction layers with features that distinguish true boundaries from haze-induced low contrast [9].

3.4 Edge Conv (Edge Convolution)

EdgeConv was originally introduced as a graph-based operator for point-cloud learning, where local geometric relations are represented through edge features between a point and its neighbors [10]. In this work, we adapt the same relational principle to image features and use it to connect the PPG representation with the subsequent multi-scale encoder. The purpose is not to impose a point-cloud graph on the image, but to make local feature differences explicit so that boundaries attenuated by haze remain available to the reconstruction path.

The adapted Edge Conv structure is shown in Fig. 4.

Given an image feature map $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}\subseteq\mathbb{R}^F$, each $\boldsymbol{x}_i$ denotes the feature at one location of the two-dimensional grid. We construct a directed local graph $G=(V,E)$ by connecting each feature to its k nearest neighbors in the current feature space, where $V=\{1,\ldots,n\}$ and each

edge connects a central feature $\boldsymbol{x}_i$ to a neighboring feature $\boldsymbol{x}_j \in \mathcal{N}(i)$. This feature-space k-NN construction allows the graph to represent feature similarity rather than relying only on fixed spatial adjacency.

The EdgeConv operation is formally defined as:

$$\boldsymbol{x}'_i = \bigsqcup_{j \in \mathcal{N}(i)} h_\Theta \left(\boldsymbol{x}_i, \boldsymbol{x}_j \right) \tag{14}$$

where $h_\Theta : \mathbb{R}^F \times \mathbb{R}^F \to \mathbb{R}^{F'}$ is a learnable nonlinear edge function and $\bigsqcup$ denotes a symmetric aggregation operator, such as max or sum. The operation describes local relations rather than isolated pixel responses.

Following the asymmetric edge function in the original EdgeConv formulation, we combine the central feature with its relative difference from a neighbor:

$$h_\Theta \left(\boldsymbol{x}_i, \boldsymbol{x}_j \right) = \bar{h}_\Theta \left(\boldsymbol{x}_i, \boldsymbol{x}_j - \boldsymbol{x}_i \right) \tag{15}$$

In our implementation, the edge function $\bar{h}_\Theta(\cdot)$ is realized by a shared multilayer perceptron with ReLU activation:

$$e'_{ijm} = \mathrm{ReLU} \left(\theta_m \cdot \left(\boldsymbol{x}_j - \boldsymbol{x}_i \right) + \phi_m \cdot \boldsymbol{x}_i \right) \tag{16}$$

where $\Theta = (\theta_1, \ldots, \theta_M, \phi_1, \ldots, \phi_M)$ contains the learnable weights and $M = F'$ is the output dimension. Max pooling is selected because it is invariant to neighbor ordering and retains the strongest local response:

$$x'_{im} = \max_{j \in \mathcal{N}(i)} e'_{ijm} \tag{17}$$

An important feature of our implementation is the dynamic graph update. The neighborhood is recomputed from the current feature space at each layer [10], rather than being permanently tied to the initial image grid. Consequently, the graph can connect locations that are structurally or semantically similar even when they are far apart in the original image grid. This is useful near haze boundaries, where the most informative comparison may be with a structurally similar location rather than with the nearest pixel alone.

The max aggregation makes EdgeConv invariant to the ordering of neighbors, while the relative term $\boldsymbol{x}_j - \boldsymbol{x}_i$ provides translation-invariant local contrast and the central term $\boldsymbol{x}_i$ preserves contextual information. These properties make the operator suitable for extracting edge-like changes from image features without discarding the appearance of the central region.

3.5 Dynamic Conv (Dynamic Convolution)

To increase the representation capability of the feature extractor without increasing the network depth or width, we employ Dynamic Convolution [11]. Unlike a static convolution, which uses one fixed kernel for all inputs, Dynamic Convolution maintains a set of K candidate convolution kernels and aggregates them according to input-dependent attention weights. The candidate kernels have the same kernel size and input/output channel dimensions, and the aggregated kernel is adapted to the current feature input.

As shown in Fig. 5, the input feature is processed by a set of parallel candidate convolution kernels. The attention branch first applies global average pooling to obtain a compact representation of the input. This representation is then transformed by two fully connected layers with a ReLU activation between them, and a Softmax function produces normalized attention weights for the candidate kernels.

The core operation is defined as follows. Let the candidate kernels and biases be

$$\{(\widetilde{\boldsymbol{W}}_1, \widetilde{\boldsymbol{b}}_1), (\widetilde{\boldsymbol{W}}_2, \widetilde{\boldsymbol{b}}_2), \cdots, (\widetilde{\boldsymbol{W}}_K, \widetilde{\boldsymbol{b}}_K)\}. \tag{18}$$

The global average pooled feature is

$$z_c = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \boldsymbol{X}_c(i, j), \tag{19}$$

where $\boldsymbol{X}_c(i, j)$ denotes the value at spatial location (i, j) in channel c. The attention logits are computed as

$$\boldsymbol{a} = \boldsymbol{W}_2 \delta(\boldsymbol{W}_1 \boldsymbol{z}), \tag{20}$$

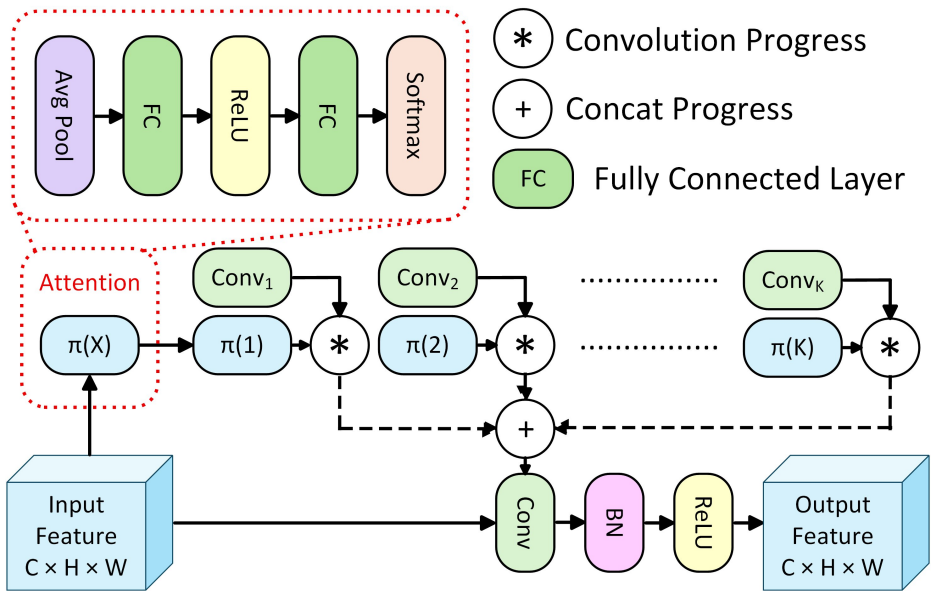


Figure 5. Dynamic Convolution (Dynamic Conv) structure diagram

where $\boldsymbol{z}$ is the pooled feature vector, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the weights of the two fully connected layers, and $\delta(\cdot)$ denotes the ReLU function. The attention coefficient for the k-th candidate kernel is obtained by Softmax:

$$\pi_k(\boldsymbol{X}) = \frac{\exp(a_k/\tau)}{\sum_{j=1}^{K}\exp(a_j/\tau)}, \qquad 0 \le \pi_k(\boldsymbol{X}) \le 1, \quad \sum_{k=1}^{K}\pi_k(\boldsymbol{X}) = 1, \tag{21}$$

where τ is the Softmax temperature. The input-dependent aggregated kernel and bias are

$$\widetilde{\boldsymbol{W}}(\boldsymbol{X}) = \sum_{k=1}^{K}\pi_k(\boldsymbol{X})\widetilde{\boldsymbol{W}}_k, \qquad \widetilde{\boldsymbol{b}}(\boldsymbol{X}) = \sum_{k=1}^{K}\pi_k(\boldsymbol{X})\widetilde{\boldsymbol{b}}_k. \tag{22}$$

After batch normalization (BN) and activation, the output of the Dynamic Convolution layer is written as

$$\widetilde{\boldsymbol{Y}} = \phi\Big(\mathrm{BN}\Big(\widetilde{\boldsymbol{W}}(\boldsymbol{X}) * \boldsymbol{X} + \widetilde{\boldsymbol{b}}(\boldsymbol{X})\Big)\Big), \tag{23}$$

where $*$ denotes convolution, $\mathrm{BN}(\cdot)$ applies this normalization, and $\phi(\cdot)$ is the subsequent activation function. Thus, the convolution kernel is not fixed during inference: its composition changes with the input feature through the attention vector $\{\pi_k(\boldsymbol{X})\}_{k=1}^{K}$. This input-dependent kernel aggregation increases the nonlinearity and representation capacity of the layer while retaining the same output-channel dimension as its static counterpart. In the proposed dehazing network, this mechanism enables the feature extractor to adapt its response to the degradation characteristics encoded in the input feature.

4 Experimental Design and Results

4.1 Dataset

The experiments cover synthetic and captured-haze data. RESIDE OTS and ITS provide outdoor and indoor synthetic hazy/clear pairs, respectively. Each is divided into training, validation, and test subsets in an 8:1:1 ratio. The OTS-trained and ITS-trained models are evaluated on SOTS-Outdoor and SOTS-Indoor, respectively; SOTS is kept entirely separate from training and validation and used only for testing [13]. The captured-haze datasets I-HAZE, O-HAZE, and NH-HAZE are each divided into training and held-out test pairs, with NH-HAZE specifically representing spatially non-uniform haze [43, 44, 45]. Their partitions are specified below.

4.1.1 RESIDE Dataset RESIDE (REalistic Single Image DEhazing Dataset) is a benchmark introduced by Li et al. for evaluating single-image dehazing methods [13]. Its synthetic subsets include the Outdoor Training Set (OTS), the Indoor Training Set (ITS), and the Synthetic Objective Testing Set (SOTS). OTS and ITS contain paired synthetic hazy/clear images for

outdoor and indoor scenes, respectively. Each of OTS and ITS is divided into training, validation, and test subsets at an 8:1:1 ratio. Separate models are trained using the OTS and ITS training subsets, and the corresponding validation subsets are used during model development. The OTS-trained model is evaluated on SOTS-Outdoor, and the ITS-trained model is evaluated on SOTS-Indoor; no SOTS images are used for training or validation.

(1) RESIDE OTS subset

Data composition: OTS is a large-scale collection of outdoor hazy/clear pairs synthesized from clear outdoor scenes using the atmospheric scattering model;

Scene coverage: It focuses on open outdoor environments, including urban buildings, rural roads, natural landscapes, and transportation-related scenes;

Fog concentration characteristics: Haze is grouped into light ($t(x) \in [0.7, 1.0]$), medium ($t(x) \in [0.4, 0.7]$), and heavy ($t(x) \in [0.1, 0.4]$) regimes, enabling evaluation across increasing degradation strength;

Selection basis: Its scale and scene diversity make it suitable for assessing generalization and robustness to haze concentration.

(2) RESIDE ITS Subset

Data composition: ITS contains paired synthetic hazy and clear indoor images generated with the same physical model as OTS;

Scene characteristics: Indoor images contain weak illumination, backlighting, and point-source interference, while haze can change sharply near doors, windows, and corners. These properties place stronger demands on local feature extraction and illumination-aware reconstruction;

Selection basis: It complements the outdoor evaluation and tests whether the model remains reliable when lighting and haze distribution are more localized.

(3) RESIDE SOTS subsets

SOTS is RESIDE's paired synthetic test benchmark. It contains 500 indoor and 500 outdoor hazy/clear image pairs. All SOTS images are reserved for testing and are excluded from the training and validation data. We use SOTS-Indoor and SOTS-Outdoor for quantitative comparisons, and SOTS-Indoor for the haze-level experiment [13].

4.1.2 Real-Haze Datasets I-HAZE contains 35 paired indoor hazy and haze-free images captured under controlled conditions using a professional haze generator [43]; 30 pairs are used for training and 5 are held out for testing. O-HAZE contains 45 paired outdoor scenes captured with and without haze under matched illumination [44]; 40 pairs are used for training and 5 are held out for testing. NH-HAZE contains 55 paired outdoor scenes and was designed to represent non-homogeneous haze distributions [45]; 50 pairs are used for training and 5 are held out for testing. For each dataset, the model is trained on its designated training pairs, and the held-out pairs are used for quantitative and visual evaluation.

4.2 Experimental Setup and Scheme

To support reproducibility, the experiments were conducted with a fixed hardware and software configuration. The graphics processing unit (GPU), central processing unit (CPU), random-access memory (RAM), and solid-state drive (SSD) used in the experiments are listed below.

(1) Hardware Configuration

Table 1. Hardware Configuration

Hardware Component	Specific Parameters
GPU	NVIDIA GeForce RTX 5060 (Blackwell architecture; CUDA Toolkit 12.8 or later for native support
CPU	Intel Core i9-12900K (16 cores, 24 threads, 3.2GHz base frequency, 5.2GHz max turbo)
RAM	16GB
Storage	1TB SSD

(2) Software Environment

Table 2. Software Environment

Software Component	Specific Version
Operating System	Ubuntu (64-bit)
Deep Learning Framework	PyTorch
Programming Language	Python
Acceleration Library	CUDA, CUDA Deep Neural Network (cuDNN)
Image Processing Library	OpenCV (Open Source Computer Vision Library), Scikit-image
Numerical Computing Library	NumPy, SciPy
Visualization Tool	Matplotlib, TensorBoard

(3) Experimental Parameter Settings

Training parameters: We use AdamW with a weight decay of 1×10^{-4}, an initial learning rate of 1×10^{-4}, cosine annealing with $T_{\max} = 100$, a batch size of 16, and 100 training epochs;

Loss function: The hybrid objective in Eq. 2 uses $\lambda_1 = 0.5$ and $\lambda_2 = 0.1$ for the transmission-gradient and atmospheric-light components of $\mathcal{L}_{\text{phy}}$, respectively, and $\lambda_3 = 1.0$ for $\mathcal{L}_{\text{per}}$. The perceptual features are extracted from the `conv3_3` and `conv4_3` layers of a pretrained VGG-19 network; the pixel-wise reconstruction term is L_1.

4.2.1 Experimental Scheme Design Two complementary experimental protocols are used. The ablation study isolates the effect of each module, while the comparison study evaluates the complete network against representative dehazing methods.

(1) Ablation Experiment

To examine the contribution of PPG, CA, Edge Conv, and Dynamic Conv, five progressively enriched configurations are trained. All configurations use the same data splits and optimization settings; only the module combination changes:

Group 1 (Baseline): Basic Transformer model without PPG, CA, Edge Conv, and Dynamic Conv;

Group 2: Baseline + PPG module;

Group 3: Baseline + PPG + CA module;

Group 4: Baseline + PPG + CA + Edge Conv module;

Group 5 (Ours): Baseline + PPG + CA + Edge Conv + Dynamic Conv module.

PSNR and SSIM are measured on RESIDE OTS for the ablation comparison. Because the configurations are nested, the change between adjacent rows reveals the role of the newly introduced module, while the difference between the baseline and the complete model reflects the cumulative effect of the full modulation strategy.

(2) Comparative Experiment

The comparison covers synthetic and real-haze benchmarks, with the methods selected for each dataset as reported in the corresponding results tables. On RESIDE OTS and ITS, we compare with DCP [2], AOD-Net [3], CycleGAN [16], UCL-Dehaze [42], DehazeFormer [5], and Diff-Dehazer [17]. On SOTS-Indoor, SOTS-Outdoor, and the SOTS haze-level evaluation, the comparison includes DCP [2], AOD-Net [3], MSCNN [18], DCPDN [19], DehazeNet [20], EPDN [21], DRN [22], MSBDN [23], and FFA-Net [24].

For I-HAZE, O-HAZE, and NH-HAZE, the comparisons include DCP [2], AOD-Net [3], DehazeNet [20], Fast-Net [31], and EDN-GTM [32]. The additional NH-HAZE comparison includes MSID [30], DeHamer [33], DehazeFormer [5], QCNN-H [34], DHM [35], RIDCP [36], T-Net [37], MB-TaylorFormer [38], and RABNet [39]. Our complete model is included in each applicable comparison. PSNR and SSIM are used for paired datasets.

4.3 Explanation of Evaluation Indicators (PSNR and SSIM)

We evaluate restoration quality with Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), two complementary measures widely used in image restoration. Both are reported with the convention that higher values indicate better agreement with the clear target (↑). PSNR emphasizes pixel-level fidelity, while SSIM is more sensitive to luminance, contrast, and structural organization.

4.3.1 Peak Signal-to-Noise Ratio (PSNR) (1) Definition and Calculation Formula

PSNR measures image distortion through the mean squared error (MSE) between the dehazed output and the haze-free ground truth [14]:

$$\text{PSNR} = 10 \times \log_{10}\left(\frac{\text{MAX}_{\text{I}}^2}{\text{MSE}}\right) \tag{24}$$

Where: - $\mathrm{MAX_I}$: The maximum pixel value for the representation used to calculate PSNR. It is 255 when images are represented in the 8-bit range $[0, 255]$ and 1 when normalized to $[0, 1]$; the output and reference images must use the same range; - MSE: Mean squared error, representing the average of the squared differences between pixel values of dehazed image $\boldsymbol{J}$ and haze-free ground-truth image $\boldsymbol{I}$:

$$\mathrm{MSE} = \frac{1}{H \times W \times C} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{C} |\boldsymbol{J}(i,j,k) - \boldsymbol{I}(i,j,k)|^2 \tag{25}$$

Where H, W, C are the height, width and number of channels of the image respectively ($C = 3$ for color images).

(2) Physical Meaning and Evaluation Logic

PSNR is expressed in decibels. A higher value indicates a smaller average pixel error, which generally means that the method has removed attenuation and color distortion more accurately on the evaluated target distribution.

In dehazing, a PSNR increase can indicate more faithful recovery of attenuated intensities, but it should not be interpreted in isolation: a model may improve average pixel agreement while still smoothing boundaries. For this reason, PSNR is analyzed together with SSIM.

4.3.2 Structural Similarity Index (1) Definition and Calculation Formula

SSIM was proposed by Wang et al. in 2004 [15]. Unlike PSNR, it compares local luminance, contrast, and structural relationships. In this formulation, μ_J and μ_I are local means, σ_J^2 and σ_I^2 are local variances, and σ_JI is the local covariance, all computed over the same image window. The image-level SSIM score is obtained by averaging the local SSIM values:

$$\mathrm{SSIM}(\boldsymbol{J}, \boldsymbol{I}) = \frac{(2\mu_\mathrm{J}\mu_\mathrm{I} + C_1)(2\sigma_\mathrm{JI} + C_2)}{(\mu_\mathrm{J}^2 + \mu_\mathrm{I}^2 + C_1)(\sigma_\mathrm{J}^2 + \sigma_\mathrm{I}^2 + C_2)} \tag{26}$$

An equivalent decomposition is commonly written as

$$\mathrm{SSIM}(\boldsymbol{J}, \boldsymbol{I}) = [l(\boldsymbol{J}, \boldsymbol{I})]^\alpha \times [c(\boldsymbol{J}, \boldsymbol{I})]^\beta \times [s(\boldsymbol{J}, \boldsymbol{I})]^\gamma \tag{27}$$

Where:

$l(\boldsymbol{J}, \boldsymbol{I})$: Luminance comparison function, $l(\boldsymbol{J}, \boldsymbol{I}) = \frac{2\mu_\mathrm{J}\mu_\mathrm{I} + C_1}{\mu_\mathrm{J}^2 + \mu_\mathrm{I}^2 + C_1}$, where μ_J and μ_I are local means computed over the same window;

$c(\boldsymbol{J}, \boldsymbol{I})$: Contrast comparison function, $c(\boldsymbol{J}, \boldsymbol{I}) = \frac{2\sigma_\mathrm{J}\sigma_\mathrm{I} + C_2}{\sigma_\mathrm{J}^2 + \sigma_\mathrm{I}^2 + C_2}$, where σ_J and σ_I are local standard deviations;

$s(\boldsymbol{J}, \boldsymbol{I})$: Structure comparison function, $s(\boldsymbol{J}, \boldsymbol{I}) = \frac{\sigma_\mathrm{JI} + C_3}{\sigma_\mathrm{J}\sigma_\mathrm{I} + C_3}$, where σ_JI is their local covariance;

$C_1 = (K_1 \times \mathrm{MAX_I})^2$, $C_2 = (K_2 \times \mathrm{MAX_I})^2$, $C_3 = C_2/2$ are constants to avoid division by zero, usually $K_1 = 0.01$, $K_2 = 0.03$;

$\alpha = \beta = \gamma = 1$, meaning equal weight contributions from luminance, contrast, and structure dimensions.

(2) Physical Meaning and Evaluation Logic

SSIM theoretically ranges from -1 to 1, with 1 indicating identical local structure; for most natural-image comparisons, reported values are between 0 and 1. A higher SSIM indicates stronger local agreement with the clear target and generally less distortion of edges, contours, and texture arrangement.

In the reported experiments, SSIM is therefore used to interpret whether a PSNR gain is accompanied by genuine structure preservation. A method that raises PSNR but lowers SSIM may be fitting intensities while losing boundaries; a simultaneous increase in both metrics provides stronger evidence that haze removal and detail retention are being improved together.

4.3.3 Basis for Indicator Selection The two metrics are selected for three related reasons:

1. Complementarity: PSNR measures pixel fidelity, whereas SSIM measures structural consistency, so their combination reflects both restoration accuracy and detail retention;

2. Universality: Both are established image-restoration measures and enable direct comparison with prior work;

3. Task relevance: dehazing requires attenuation removal without destroying scene structure, which corresponds naturally to the complementary behavior of PSNR and SSIM.

4.4 Experimental Results and Analysis

To evaluate the proposed network, we conduct comparative experiments on RESIDE OTS/ITS, SOTS-Indoor/Outdoor, and the real-haze datasets I-HAZE, O-HAZE, and NH-HAZE, including an evaluation under different haze levels on SOTS-Indoor. The ablation study is conducted on RESIDE OTS. PSNR assesses pixel-level restoration, while SSIM indicates whether intensity recovery preserves scene structure.

4.4.1 Comparative Experiment We compare the proposed network with representative methods from several design families: the convolutional AOD-Net, the generative CycleGAN, UCL-Dehaze, the Transformer-based DehazeFormer, and the diffusion-based Diff-Dehazer. This selection makes it possible to distinguish the effect of the proposed feature modulation from both conventional local convolution and newer global or generative restoration strategies. All methods are evaluated with the same PSNR and SSIM criteria.

Table 3 shows the quantitative comparison results of different methods on the RESIDE OTS and RESIDE ITS datasets.

Table 3. Quantitative comparison on RESIDE OTS and RESIDE ITS.

Method	RESIDE OTS		RESIDE ITS	
	PSNR	SSIM	PSNR	SSIM
DCP [2]	17.98	0.843	19.10	0.866
AOD-Net [3]	20.18	0.847	22.16	0.874
CycleGAN [16]	23.13	0.836	24.49	0.895
UCL-Dehaze [42]	23.11	0.851	25.57	0.914
DehazeFormer-M [5]	23.81	0.874	26.10	0.903
Diff-Dehazer [17]	24.61	0.896	29.11	0.925
Ours	24.57	0.907	28.78	0.928

Representative visual comparisons on the RESIDE datasets are shown in Fig. 6.

Representative visual comparisons on the SOTS subsets are shown in Fig. 7.

Table 3 reveals a consistent tension between pixel recovery and structure preservation across the comparison methods. The two metrics are therefore more informative when read together rather than as a single ranking.

AOD-Net obtains relatively low PSNR and SSIM on both subsets. Its compact convolutional design is efficient, but the limited receptive field and feature capacity make it difficult to distinguish haze attenuation from scene texture. CycleGAN raises PSNR on OTS to 23.13 dB, yet its SSIM remains 0.836. This gap suggests that the adversarial mapping can correct broad appearance differences while still altering local structures, which is a recognizable failure mode when texture fidelity is not explicitly constrained. UCL-Dehaze improves the indoor SSIM to 0.914, indicating more reliable structure recovery under localized illumination and haze changes.

DehazeFormer benefits from global dependency modeling and reaches 26.10 dB PSNR on ITS. Its SSIM of 0.903, however, remains below the strongest structure-preserving methods, suggesting that global context alone does not fully resolve local boundary ambiguity. Diff-Dehazer records the highest ITS PSNR, 29.11 dB, with an SSIM of 0.925. The result indicates strong pixel-level recovery, but its OTS SSIM of 0.896 is below that of the proposed method. The contrast between the two subsets is meaningful: indoor haze is often more localized and benefits from broad appearance reconstruction, whereas outdoor scenes contain larger spatially varying haze fields in which stable boundary alignment becomes more important.

The proposed method provides a more balanced response across the two subsets. It achieves 24.57 dB/0.907 on OTS and 28.78 dB/0.928 on ITS, with the highest SSIM in both cases. The advantage is most visible in structure preservation: the transmission prior supplies a spatial cue for attenuation, EdgeConv emphasizes local transitions, and the adaptive kernels can alter their receptive-field preference when the haze pattern changes. These mechanisms address complementary sources of error rather than relying on global modeling alone.

Relative to DehazeFormer, our SSIM improves by 0.033 on OTS and 0.025 on ITS. This consistent advantage in structural similarity is in line with the role of prior-guided modulation: the network retains long-range context while incorporating transmission changes and local boundaries. Compared with Diff-Dehazer, our method differs by only 0.04 dB in PSNR on OTS and 0.33 dB on ITS, while obtaining higher SSIM on both subsets. This trade-off suggests that the proposed model better preserves structural detail alongside pixel fidelity.

Taken together, the results support the view that physically guided feature modulation improves the stability of structural reconstruction while retaining competitive pixel fidelity. The

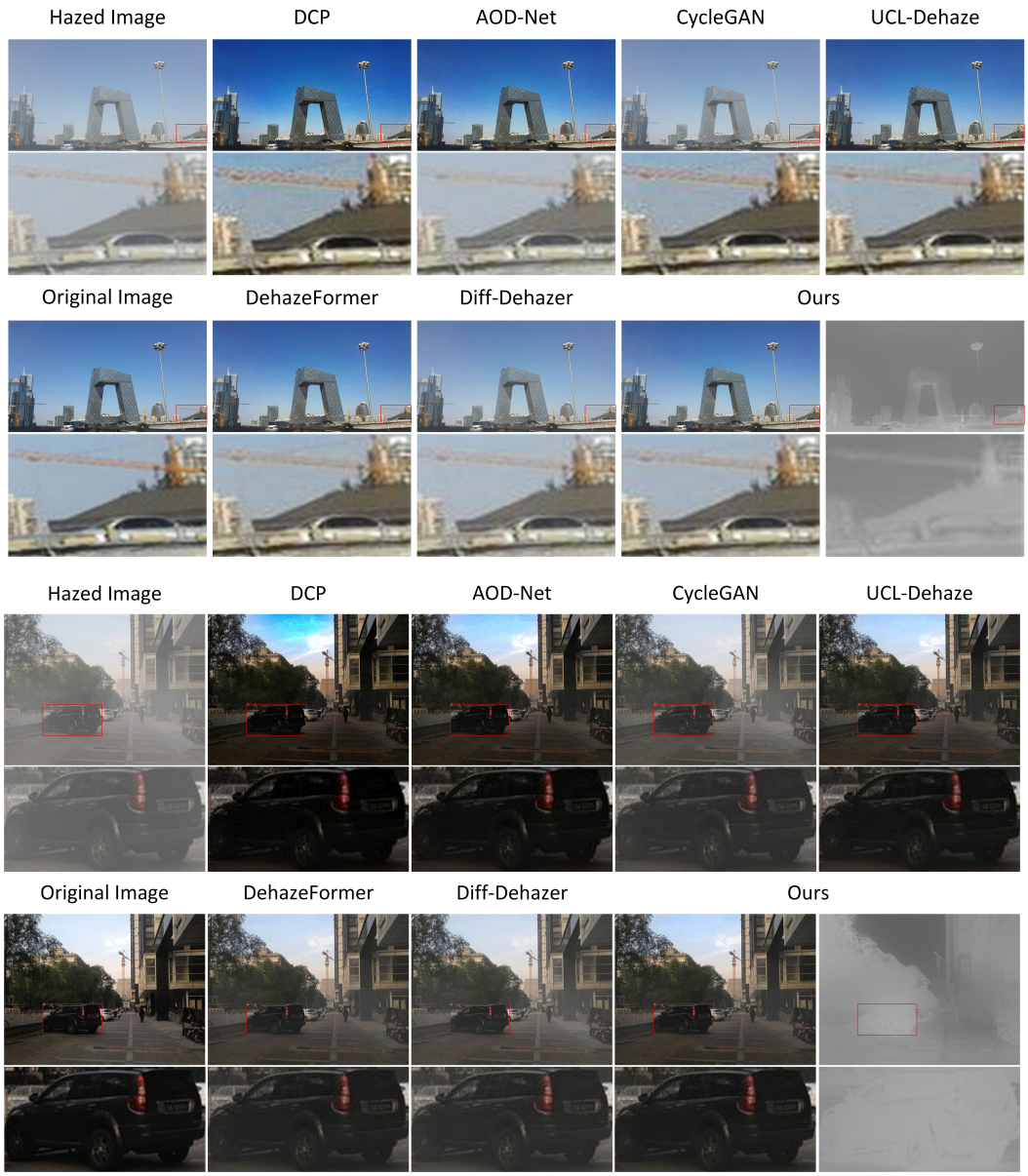


Figure 6. Visual comparison results on RESIDE datasets.

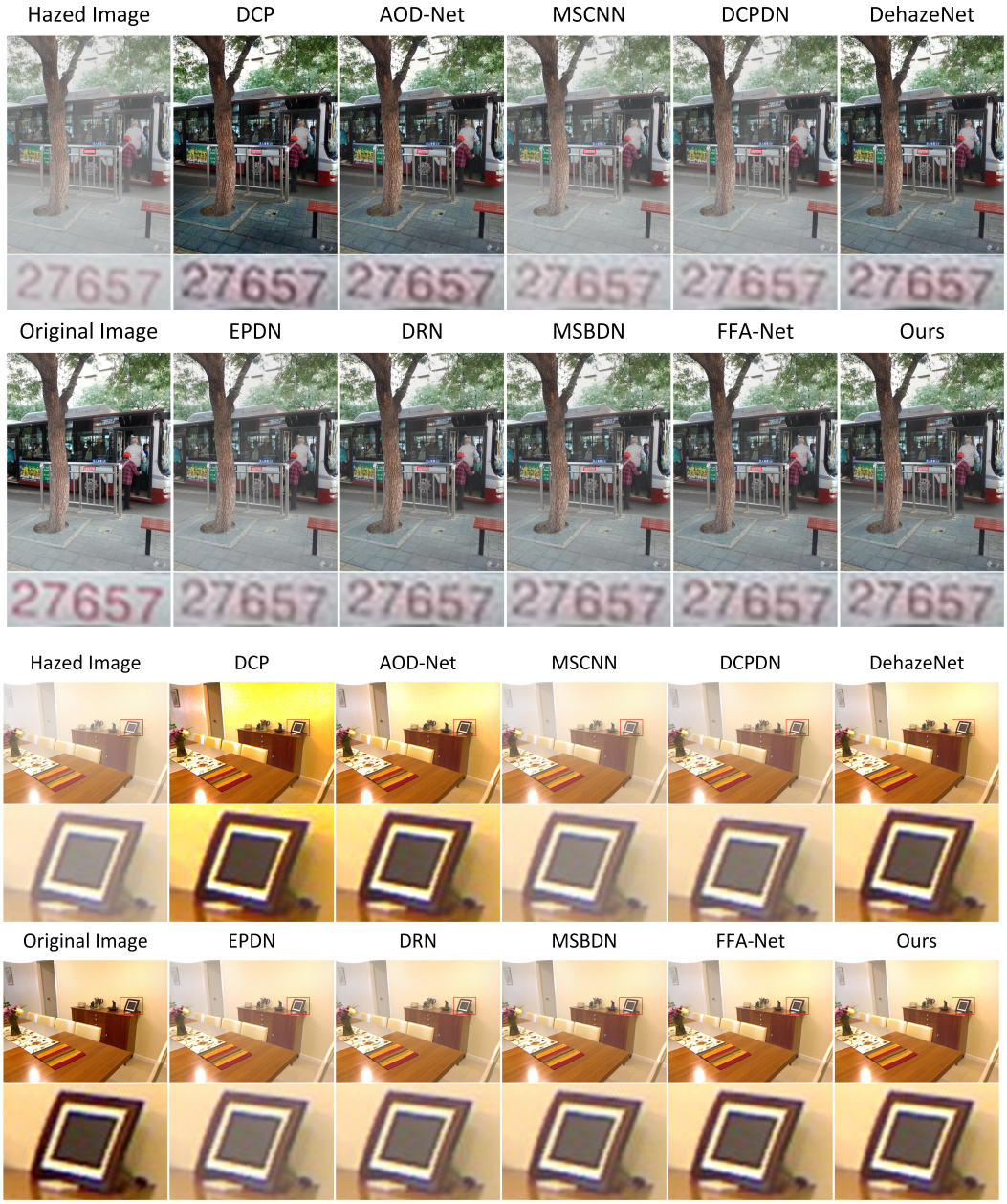


Figure 7. Visual comparison results on RESIDE SOTS-Outdoor (top) and RESIDE SOTS-Indoor (bottom).

benefit is not limited to a single comparison family; it appears against local convolution, adversarial restoration, Transformer modeling, and diffusion-based reconstruction.

These results should nevertheless be interpreted within the scope of the benchmarks. RESIDE and SOTS provide paired synthetic observations for controlled analysis but do not reproduce every artifact of real atmospheric imaging. The additional evaluations on I-HAZE, O-HAZE, and NH-HAZE assess performance on captured haze, although broader testing is still needed before drawing conclusions about unconstrained environments.

To examine generalization beyond the OTS/ITS training-oriented subsets, we further evaluate the models on the standard RESIDE SOTS-Indoor and SOTS-Outdoor test sets. The results are reported in Table 4.

These test sets provide a stricter view of how the learned reconstruction rules behave on unseen haze-image pairs.

Table 4. Quantitative comparison on RESIDE SOTS-Indoor and RESIDE SOTS-Outdoor.

Method	RESIDE SOTS-Indoor		RESIDE SOTS-Outdoor	
	PSNR	SSIM	PSNR	SSIM
DCP [2]	17.08	0.8016	18.57	0.8021
AOD-Net [3]	18.42	0.8317	21.06	0.8914
MSCNN [18]	18.94	0.8265	21.41	0.8039
DCPDN [19]	19.76	0.8564	18.93	0.8147
DehazeNet [20]	20.37	0.8728	23.18	0.8296
EPDN [21]	24.21	0.9476	23.44	0.8842
DRN [22]	31.58	0.9698	32.26	0.9730
MSBDN [23]	33.27	0.9724	29.84	0.9691
FFA-Net [24]	35.11	0.9752	33.54	0.9617
Ours	40.12	0.9823	34.28	0.9746

The SOTS results show a clear progression from fixed physical estimation to learned multi-scale reconstruction. DCP performs poorly on both subsets, which is consistent with the fact that a fixed prior cannot fully account for bright surfaces, local illumination, and texture loss. AOD-Net, MSCNN, and DCPDN improve upon DCP in selected cases, but their moderate SSIM values reveal that local convolutional processing still struggles to preserve fine structures when attenuation is strong or spatially irregular.

EPDN, DRN, MSBDN, and FFA-Net obtain substantially stronger results through multi-scale fusion and attention. FFA-Net reaches 35.11 dB PSNR on SOTS-Indoor, whereas DRN attains 0.9730 SSIM on SOTS-Outdoor. The difference between these two examples illustrates the same issue seen above: strong intensity recovery and strong structural retention are related but not interchangeable, and the preferred design depends on how the model handles high-frequency information.

Our method achieves the best PSNR and SSIM on both SOTS subsets. On SOTS-Indoor, it reaches 40.12 dB and 0.9823, improving over FFA-Net by 5.01 dB and 0.0071. On SOTS-Outdoor, it reaches 34.28 dB and 0.9746, exceeding FFA-Net by 0.74 dB and 0.0129. The larger indoor PSNR gain is consistent with effective recovery of attenuated intensities, while the SSIM gains on both subsets suggest that prior-guided multi-scale features also support structural preservation.

Across RESIDE OTS/ITS and SOTS, the results show a consistent structural advantage: the proposed method obtains the highest SSIM on OTS and ITS and the highest PSNR and SSIM on both SOTS subsets. This pattern is consistent with the interaction of prior-guided attenuation cues, edge-sensitive features, and multi-scale context, which helps preserve local boundaries while adapting to broader haze patterns [13, 25].

These findings remain limited by the use of synthetic haze for the main quantitative evaluation. The paired construction is valuable for isolating algorithmic effects, but real images can contain sensor noise, multiple scattering, and atmospheric light variations that are not fully represented by the synthesis process. The additional real-haze experiments below therefore provide a complementary evaluation of cross-domain generalization.

4.4.2 Comparative experiment under different haze levels To examine how performance changes as the degradation becomes more severe, we conduct a controlled evaluation on SOTS-Indoor at haze levels 2, 5, and 8, which are treated as light, medium, and heavy haze, respectively [13]. The same methods and evaluation protocol are used at all three levels, so the absolute scores and the rate of performance decay can be analyzed together.

Table 5. Comparison results on SOTS Indoor under different haze levels.

Method	Light Haze		Medium Haze		Heavy Haze	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DCP [2]	18.02	0.8108	17.15	0.8042	15.63	0.7924
AOD-Net [3]	18.54	0.8376	18.31	0.8289	17.01	0.8235
MSCNN [18]	20.11	0.8139	19.02	0.8234	17.48	0.8171
DCPDN [19]	20.87	0.8647	19.68	0.8591	18.22	0.8489
DehazeNet [20]	21.96	0.8801	20.44	0.8753	19.11	0.8643
EPDN [21]	25.73	0.9524	24.08	0.9348	22.67	0.9395
DRN [22]	33.02	0.9756	31.74	0.9712	30.12	0.9642
MSBDN [23]	34.56	0.9688	33.11	0.9705	31.45	0.9668
FFA-Net [24]	36.84	0.9814	35.29	0.9771	35.27	0.9699
Ours	41.27	0.9882	39.93	0.9815	37.54	0.9786

As haze becomes denser, every method loses some performance because transmission decreases and the visible evidence for texture and boundaries becomes weaker [26]. The amount of degradation, however, is method-dependent. DCP falls from 18.02 dB/0.8108 under light haze to 15.63 dB/0.7924 under heavy haze. The simultaneous decline in PSNR and SSIM indicates that the fixed prior is not only less accurate at estimating intensities, but also less reliable at locating structures once the image departs from its local statistical assumptions. AOD-Net, MSCNN, and DCPDN start from higher scores in some settings, yet their continued SSIM loss under heavy haze suggests that local convolutional features cannot fully recover the global organization of a strongly attenuated scene.

EPDN preserves structure comparatively well, while DRN and MSBDN maintain high SSIM through multi-scale feature fusion. FFA-Net reaches 36.84 dB and 0.9814 under light haze, showing strong detail enhancement when the input still contains sufficient local evidence. Its heavy-haze SSIM decreases to 0.9699, which suggests that aggressive feature enhancement cannot completely compensate for the loss of reliable boundary cues when attenuation becomes severe.

The proposed method remains the strongest performer across all three levels. It achieves 41.27 dB/0.9882 under light haze, 39.93 dB/0.9815 under medium haze, and 37.54 dB/0.9786 under heavy haze. The scores show that increased haze affects both pixel accuracy and structural consistency, but the output remains substantially closer to the clear target than the comparison methods at the same degradation level.

From light to heavy haze, the proposed model loses 3.73 dB PSNR and 0.0096 SSIM, as obtained from $41.27 - 37.54$. Even so, the SSIM decrease is smaller than that of most competing methods, indicating that the model preserves structural organization more consistently as the visible signal weakens.

The stability across haze levels is consistent with the intended mechanism of the architecture. The transmission prior supplies a coarse estimate of where attenuation is strongest, while the multi-scale modulation path allocates contextual support to those regions. Edge-aware features protect local transitions, and Dynamic Conv adjusts the receptive-field response as the image becomes less informative. Although performance still declines under dense haze, the model retains the best PSNR and SSIM among the compared methods at all three tested levels.

4.4.3 Comparative experiment on real non-uniform haze datasets The proposed method was additionally evaluated on the held-out test pairs from the real-haze datasets I-HAZE, O-HAZE, and NH-HAZE. I-HAZE and O-HAZE represent indoor and outdoor real-haze conditions, respectively, whereas NH-HAZE focuses on spatially non-uniform haze. PSNR and SSIM were used for quantitative evaluation, and higher values indicate better restoration quality.

Table 6. Quantitative comparison on I-HAZE, O-HAZE, and NH-HAZE.

Method	I-HAZE		O-HAZE		NH-HAZE	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DCP [2]	13.78	0.746	15.28	0.673	12.38	0.523
AOD-Net [3]	14.02	0.622	15.08	0.709	15.74	0.558
DehazeNet [20]	14.19	0.747	18.65	0.790	17.42	0.539
Fast-Net [31]	17.97	0.811	21.47	0.802	20.77	0.701
EDN-GTM [32]	21.89	0.826	23.67	0.821	20.17	0.708
Ours	22.41	0.832	24.56	0.834	21.82	0.733

Representative images from I-HAZE, O-HAZE, and NH-HAZE are shown in Fig. 8.



Figure 8. Qualitative comparison on I-HAZE, O-HAZE, and NH-HAZE.

The results in Table 6 show that the proposed method achieves the highest PSNR and SSIM among the six methods on both I-HAZE and O-HAZE. The method reaches 22.41 dB/0.832 on I-HAZE and 24.56 dB/0.834 on O-HAZE, exceeding EDN-GTM by 0.52 dB/0.006 and 0.89 dB/0.013, respectively. These gains are consistent with the benefit of combining a spatial attenuation cue with feature modulation for real indoor and outdoor haze.

On NH-HAZE, the proposed method achieves 21.82 dB PSNR and 0.733 SSIM. It obtains the highest PSNR in Table 6, improving over EDN-GTM by 1.65 dB; its SSIM is also 0.025 higher than EDN-GTM's. This indicates that the transmission prior provides useful degradation cues under locally varying haze, while EdgeConv helps preserve weakened transitions between haze regions and scene structures.

The NH-HAZE results were also compared with additional dehazing methods:

Table 7. Additional comparison on NH-HAZE.

Method	PSNR	SSIM
MSID [30]	19.62	0.726
DeHamer [33]	20.66	0.784
DehazeFormer [5]	20.14	0.748
QCNN-H [34]	16.86	0.613
DHM [35]	18.76	0.818
RIDCP [36]	12.34	0.508
T-Net [37]	16.78	0.721
MB-TaylorFormer [38]	19.94	0.811
RABNet [39]	20.16	0.700
Ours	21.82	0.733

As shown in Table 7, the proposed method achieves the highest PSNR among the listed methods, exceeding DeHamer by 1.16 dB. DHM, however, obtains a higher SSIM of 0.818, compared with 0.733 for the proposed method. The difference indicates that NH-HAZE remains challenging for structural recovery even when pixel-level restoration is strong. These results support the effectiveness of the proposed method on real and non-uniform haze, while also indicating room for improvement in fine-structure preservation.

4.4.4 Performance Parameter Comparison Table 8 shows that parameter count, floating-point operations (FLOPs), and runtime capture different aspects of deployment cost [27]. DCP has no learned parameters, and its runtime is 1.82 ms. AOD-Net is the fastest method at 0.68 ms, with only 0.002M parameters and 0.12G FLOPs, although its compact design limits its ability to represent complex haze and fine structures.

GCANet and GFN keep the parameter count below 1M but use 18.41G and 14.94G FLOPs, respectively, with runtimes of 31.32 and 28.95 ms. DehazeFormer-M uses 4.630M parameters and

48.64G FLOPs, with a runtime of 40.81 ms. FFA-Net has 4.456M parameters and 287.80G FLOPs, with a runtime of 178.18 ms. These results show why parameter count alone is insufficient for judging practical efficiency.

Table 8. Comparison of model complexity and inference efficiency

Method	Param(M)	FLOPs(G)	Runtime(ms)
DCP [2]	—	—	1.82
AOD-Net [3]	0.002	0.12	0.68
GCANet [40]	0.702	18.41	31.32
GFN [41]	0.499	14.94	28.95
DehazeFormer-M [5]	4.630	48.64	40.81
FFA-Net [24]	4.456	287.80	178.18
Ours	7.305	20.10	34.17

The proposed model contains 7.305M parameters and requires 20.10G FLOPs, with a runtime of 34.17 ms. It has the largest parameter count among the listed methods, while its FLOPs are lower than those of DehazeFormer-M and FFA-Net. Its runtime is higher than those of DCP, AOD-Net, GFN, and GCANet, but lower than those of DehazeFormer-M and FFA-Net. This profile is consistent with the design: the physical prior provides an early indication of the degradation direction, while adaptive modulation concentrates the learned capacity on correction and detail recovery instead of repeatedly fitting the entire haze process from appearance alone [18].

The efficiency result is therefore a trade-off rather than a claim that the proposed network is the smallest or fastest. With 20.10G FLOPs and a 34.17 ms runtime, it uses less computation and runs faster than DehazeFormer-M and FFA-Net, while its additional parameters support physical guidance and adaptive kernels. Under the tested hardware and input setting, this provides a balanced accuracy–complexity profile [8].

4.4.5 Ablation Experiment To isolate the contribution of each component, we conduct a progressive ablation study on RESIDE OTS [13]. Starting from a Transformer baseline without the proposed enhancement modules, PPG, CA, Edge Conv, and Dynamic Conv are added in sequence. Since each row changes one mechanism relative to the preceding row, the results reveal not only the cumulative gain but also the type of error each module helps to reduce.

Table 9. Ablation experiment results

Model	PPG	CA	Edge Conv	Dynamic Conv	PSNR	SSIM
Baseline	×	×	×	×	20.89	0.823
1	√	×	×	×	22.34	0.852
2	√	√	×	×	22.78	0.868
3	√	√	√	×	23.13	0.889
4	√	√	√	√	24.57	0.907

The baseline obtains 20.89 dB PSNR and 0.823 SSIM. Without physical or adaptive modulation, the network must infer both the haze distribution and the structure of the clear scene from generic feature responses. The relatively low values indicate that this unconstrained representation is insufficient for separating attenuation from texture and for maintaining stable boundaries.

Adding PPG raises PSNR to 22.34 dB and SSIM to 0.852, corresponding to gains of 1.45 dB and 0.029. The simultaneous improvement is important: the transmission cue does not merely correct average intensity, but also gives the encoder a spatial reference for deciding where haze-induced attenuation is likely to be present. The network can therefore devote more capacity to degraded regions instead of treating all low-contrast areas as texture loss.

CA further increases PSNR to 22.78 dB and SSIM to 0.868, gains of 0.44 dB and 0.016 over the PPG configuration. The larger SSIM improvement suggests that coordinate information contributes particularly to spatial alignment and structural localization. In practical terms, the module helps the decoder place edges and elongated structures more precisely.

Introducing Edge Conv raises PSNR to 23.13 dB and SSIM to 0.889, further gains of 0.35 dB and 0.021 over the preceding configuration. The SSIM increase highlights the value of explicit local differences for preserving contours and texture transitions. By comparing neighboring features, EdgeConv represents weakened boundaries as relational changes rather than isolated low-amplitude activations.

Adding Dynamic Conv produces the largest incremental gain: PSNR reaches 24.57 dB and SSIM reaches 0.907, which is 1.44 dB and 0.018 above the preceding configuration, or 3.68 dB and 0.084 above the baseline. The result indicates that the remaining error is not caused only by missing information, but also by a mismatch between a fixed feature-extraction rule and the variety of haze patterns. Input-conditioned kernels can favor local detail recovery in texture-rich regions and broader context in strongly attenuated regions, thereby reducing both residual pixel error and structural inconsistency.

The progression in Table 9 links each added component to a distinct feature-processing role: PPG provides attenuation guidance, CA improves spatial localization, Edge Conv strengthens boundary evidence, and Dynamic Conv adapts the response to the input. The cumulative gains suggest that these components complement one another, improving both pixel fidelity and structural similarity [24].

The ablation study consequently supports the central hypothesis of this work: performance gains arise from progressively refining feature allocation, not from simply increasing the number of blocks.

Although the main study uses synthetic haze, RESIDE provides paired targets and a common evaluation protocol, making it suitable for isolating the effect of the proposed modules and comparing them with prior methods. The additional experiments on I-HAZE, O-HAZE, and NH-HAZE complement this controlled study by evaluating performance on captured and spatially non-uniform haze.

Together, the results provide complementary evidence for the proposed design. On SOTS, the model achieves the highest PSNR and SSIM on both indoor and outdoor subsets, and it leads at all three tested haze levels. The ablation shows that adding PPG improves the baseline by 1.45 dB in PSNR and 0.029 in SSIM; adding CA, Edge Conv, and Dynamic Conv in sequence brings the total gains to 3.68 dB and 0.084. On real-haze data, the model leads in both metrics on I-HAZE and O-HAZE and in PSNR on NH-HAZE, while DHM retains higher SSIM on NH-HAZE. This combination of results highlights the benefits of prior-guided and adaptive feature processing, while identifying structural recovery under non-homogeneous haze as a remaining challenge.

5 Conclusion and Future Work

5.1 Summary

This work presents a physical-prior-guided multi-scale Transformer network for single-image dehazing. The encoder combines global dependency modeling with dilated local processing, while the decoder uses coordinate attention to recover spatially organized structures. EdgeConv supplies relational boundary cues, and Dynamic Conv adapts the effective receptive field to the observed haze pattern. Together with the transmission-guided modulation path, these components allow the network to use physical information throughout feature reconstruction rather than only at the input.

The resulting AFM framework provides a unified way to coordinate degradation cues, spatial location, local structure, and dynamic response. Each source of information addresses a different failure mode, which is why the complete model is more stable than the baseline under both scene changes and increasing haze density.

On RESIDE and SOTS, the method obtains competitive PSNR and the highest SSIM among the reported comparisons, with particularly consistent structural retention on the SOTS haze-level evaluation. Additional experiments on I-HAZE, O-HAZE, and NH-HAZE evaluate the method on captured haze, including spatially non-uniform conditions. The method achieves the highest PSNR and SSIM among the compared methods on I-HAZE and O-HAZE, and the highest PSNR on NH-HAZE, although DHM obtains a higher SSIM on NH-HAZE. The ablation results further show that PPG, CA, EdgeConv, and Dynamic Conv contribute in distinct ways: the prior improves degradation awareness, CA improves spatial localization, EdgeConv protects local transitions, and Dynamic Conv handles input-dependent variation.

5.2 Limitations and Challenges

Several limitations remain. The model is more costly than compact convolutional neural network (CNN) baselines, and the dynamic and multi-scale paths increase computation relative to lightweight real-time designs [11]. Although the experiments include I-HAZE, O-HAZE, and NH-HAZE, they cover only a limited range of real-world scenes. Generalization to broader conditions may still be affected by the gap between the training distribution and real observations [29]. On NH-HAZE, the higher SSIM obtained by DHM also indicates room to improve structural recovery. Finally, very dense haze can remove information that is not

recoverable from a single image, so some fine details may remain uncertain even when the global structure is preserved [18].

5.3 Future Research Direction

Future work will focus on three directions. Lightweight attention and kernel-generation strategies could reduce runtime while retaining the current feature allocation behavior. Expanding the real-haze evaluation beyond I-HAZE, O-HAZE, and NH-HAZE, together with unsupervised or weakly supervised training, could further improve generalization to uncontrolled scenes. Finally, joint training with detection or segmentation, together with more detailed physical or multimodal cues, may help the model preserve task-relevant structures when the visible image is severely degraded.

References

References

[1] S.G. Narasimhan, S.K. Nayar, Vision and the atmosphere, *International Journal of Computer Vision* 48 (3) (2002) 233–254. https://doi.org/10.1023/A:1016328200723

[2] K. He, J. Sun, X. Tang, Single image haze removal using dark channel prior, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (12) (2011) 2341–2353. https://doi.org/10.1109/TPAMI.2010.168

[3] B. Li, X. Peng, Z. Wang, J. Xu, D. Feng, AOD-Net: All-in-one dehazing network, in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4780–4788. https://doi.org/10.1109/ICCV.2017.511

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical vision transformer using shifted windows, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

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

[6] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: *Advances in Neural Information Processing Systems* 25, 2012, pp. 1097–1105.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: *Advances in Neural Information Processing Systems* 30, 2017, pp. 5998–6008.

[8] H. Zhang, V.A. Sindagi, V.M. Patel, Joint transmission map estimation and dehazing using deep networks, *IEEE Transactions on Circuits and Systems for Video Technology* 30 (7) (2020) 1975–1986. https://doi.org/10.1109/TCSVT.2019.2912145

[9] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13708–13717. https://doi.org/10.1109/CVPR46437.2021.01350

[10] Y. Wang, Y. Sun, Z. Liu, S. Sarma, M. Bronstein, J. Solomon, Dynamic graph CNN for learning on point clouds, *ACM Transactions on Graphics* 38 (5) (2019) 1–12. https://doi.org/10.1145/3326362

[11] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, Z. Liu, Dynamic convolution: Attention over convolution kernels, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11027–11036. https://doi.org/10.1109/CVPR42600.2020.01104

[12] Z. Hu, L. Chai, YOLO-Spectra-Net: Reliability-Aware RGB–Infrared Fusion for Robust Object Detection Under Adverse Conditions, *Sensors* 26 (19) (2026) 6028. https://doi.org/10.3390/s26196028

[13] B. Li, W. Ren, D. Fu, D. Tao, X. Feng, W. Zeng, Z. Wang, Benchmarking single-image dehazing and beyond, *IEEE Transactions on Image Processing* 28 (1) (2019) 492–505. https://doi.org/10.1109/TIP.2018.2867951

[14] A.M. Eskicioglu, P.S. Fisher, Image quality measures and their performance, *IEEE Transactions on Communications* 43 (12) (1995) 2959–2965. https://doi.org/10.1109/26.477498

[15] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, *IEEE Transactions on Image Processing* 13 (4) (2004) 600–612. https://doi.org/10.1109/TIP.2003.819861

[16] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251. https://doi.org/10.1109/ICCV.2017.244

[17] Y. Lan, Z. Cui, C. Liu, J. Peng, N. Wang, X. Luo, D. Liu, Exploiting diffusion prior for real-world image dehazing with unpaired training, in: *Proceedings of the AAAI Conference on Artificial Intelligence* 39 (4) (2025) 4455–4463. https://doi.org/10.1609/aaai.v39i4.32469

[18] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, M.H. Yang, Single image dehazing via multi-scale convolutional neural networks, in: *European Conference on Computer Vision (ECCV)*, 2016, pp. 154–169. https://doi.org/10.1007/978-3-319-46475-6_10

[19] H. Zhang, V.M. Patel, Densely connected pyramid dehazing network, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3194–3203. https://doi.org/10.1109/CVPR.2018.00337

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

[21] Y. Qu, Y. Chen, J. Huang, Y. Xie, Enhanced Pix2pix Dehazing Network, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8160–8168. https://doi.org/10.1109/CVPR.2019.00835

[22] Y. Li, D. Cheng, D. Zhang, N. Wang, X. Gao, J. Sun, Single image dehazing with an independent detail-recovery network, *Knowledge-Based Systems* 254 (2022) 109579. https://doi.org/10.1016/j.knosys.2022.109579

[23] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, M.-H. Yang, Multi-scale boosted dehazing network with dense feature fusion, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2157–2167. https://doi.org/10.1109/CVPR42600.2020.00223

[24] X. Qin, Z. Wang, Y. Bai, X. Xie, H. Jia, FFA-Net: Feature fusion attention network for single image dehazing, in: *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (07) (2020) 11908–11915. https://doi.org/10.1609/aaai.v34i07.6865

[25] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

[26] R.T. Tan, Visibility in bad weather from a single image, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8. https://doi.org/10.1109/CVPR.2008.4587643

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: *International Conference on Learning Representations (ICLR)*, 2015.

[29] D. Engin, A. Genc, H.K. Ekenel, Cycle-dehaze: Enhanced CycleGAN for single image dehazing, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 825–833. https://doi.org/10.1109/CVPRW.2018.00127

[30] H. Liu, Z. Wu, L. Li, S. Salehkalaibar, J. Chen, K. Wang, Towards multi-domain single image dehazing via test-time training, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5831–5840. https://doi.org/10.1109/CVPR52688.2022.00574

[31] P. Morales, T. Klinghoffer, S.J. Lee, Feature forwarding for efficient single image dehazing, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 2078–2085. https://doi.org/10.1109/CVPRW.2019.00260

[32] L.-A. Tran, S. Moon, D.-C. Park, A novel encoder-decoder network with guided transmission map for single image dehazing, *Procedia Computer Science* 204 (2022) 682–689. https://doi.org/10.1016/j.procs.2022.08.082

[33] C.-L. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, C. Li, Image dehazing transformer with transmission-aware 3D position embedding, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5812–5820. https://doi.org/10.1109/CVPR52688.2022.00572

[34] V. Frants, S. Agaian, K. Panetta, QCNN-H: Single-image dehazing using quaternion neural networks, *IEEE Transactions on Cybernetics* 53 (9) (2023) 5448–5458. https://doi.org/10.1109/TCYB.2023.3238640

[35] N. Jiang, K. Hu, T. Zhang, W. Chen, Y. Xu, T. Zhao, Deep hybrid model for single image dehazing and detail refinement, *Pattern Recognition* 136 (2023) 109227. https://doi.org/10.1016/j.patcog.2022.109227

[36] R.-Q. Wu, Z.-P. Duan, C.-L. Guo, Z. Chai, C. Li, RIDCP: Revitalizing real image dehazing via high-quality codebook priors, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22282–22291. https://doi.org/10.1109/CVPR52729.2023.02134

[37] L. Zheng, Y. Li, K. Zhang, W. Luo, T-Net: Deep stacked scale-iteration network for image dehazing, *IEEE Transactions on Multimedia* 25 (2023) 6794–6807. https://doi.org/10.1109/TMM.2022.3214780

[38] Y. Qiu, K. Zhang, C. Wang, W. Luo, H. Li, Z. Jin, MB-TaylorFormer: Multi-branch efficient transformer expanded by Taylor formula for image dehazing, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 12802–12813. https://doi.org/10.1109/ICCV51070.2023.01176

[39] S.R. Zahra, F. Hussain, A. Hassan, T. Mazhar, A. Al-Rasheed, M.A. Khan, W. Ahmad, H. Hamam, RABNet: A residual attention-based network for visibility restoration in hazy images, *IET Computers & Digital Techniques* (2026) 8366822. https://doi.org/10.1049/cdt2/8366822

[40] D. Chen, M. He, Q. Fan, J. Liao, L. Zhang, D. Hou, L. Yuan, G. Hua, Gated context aggregation network for image dehazing and deraining, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 1375–1383. https://doi.org/10.1109/WACV.2019.00151

[41] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, M.-H. Yang, Gated fusion network for single image dehazing, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3253–3261. https://doi.org/10.1109/CVPR.2018.00343

[42] Y. Wang, X. Yan, F.L. Wang, H. Xie, W. Yang, X.-P. Zhang, J. Qin, M. Wei, UCL-Dehaze: Toward real-world image dehazing via unsupervised contrastive learning, *IEEE Transactions on Image Processing* 33 (2024) 1361–1374. https://doi.org/10.1109/TIP.2024.3362153

[43] C. Ancuti, C.O. Ancuti, R. Timofte, C. De Vleeschouwer, I-HAZE: A dehazing benchmark with real hazy and haze-free indoor images, in: *Advanced Concepts for Intelligent Vision Systems (ACIVS)*, 2018, pp. 620–631. https://doi.org/10.1007/978-3-030-01449-0_52

[44] C.O. Ancuti, C. Ancuti, R. Timofte, C. De Vleeschouwer, O-HAZE: A dehazing benchmark with real hazy and haze-free outdoor images, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 754–762. https://doi.org/10.1109/CVPRW.2018.00119

[45] C.O. Ancuti, C. Ancuti, R. Timofte, 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 (CVPRW)*, 2020, pp. 1798–1805. https://doi.org/10.1109/CVPRW50498.2020.00230

[46] P. Pandey, R. Gupta, N. Goel, Comprehensive review of single image defogging techniques: enhancement, prior, and learning based approaches, *Artificial Intelligence Review* 58 (4) (2025) 116. https://doi.org/10.1007/s10462-024-11034-4

[47] P. Dwivedi, S. Chakraborty, A comprehensive qualitative and quantitative survey on image dehazing based on deep neural networks, *Neurocomputing* 610 (2024) 128582. https://doi.org/10.1016/j.neucom.2024.128582 Institutional publication record

[48] C.O. Ancuti, C. Ancuti, F.-A. Vasluianu, R. Timofte, et al., NTIRE 2024 Dense and Non-Homogeneous Dehazing Challenge Report, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024, pp. 6453–6468. https://doi.org/10.1109/CVPRW63382.2024.00646

[49] D. Zhang, L. Ma, W. Du, X. Lin, Multi-scale adaptive Transformer-based image dehazing, *Journal of Computer-Aided Design & Computer Graphics* (2025), online first. DOI: 10.3724/SP.J.1089.2025-00200. Publisher article page