ITC 3/55
Information Technology and Control
Vol. 55 / No. 3/ 2026
pp. 1406-1423
DOI 10.5755/j01.itc.55.3.45302

A Lightweight U-Shaped Transformer Network for Single Image Dehazing

Received 2026/06/16 | Accepted after revision 2026/08/28

HOW TO CITE: Zhang, Y., Lian, Y. (2026) A Lightweight U-Shaped Transformer Network for Single Image Dehazing. *Information Technology and Control*, 55(3), 1406-1423. https://doi.org/10.5755/j01.itc.55.3.45302

A Lightweight U-Shaped Transformer Network for Single Image Dehazing

Yiping Zhang[#]
College of Civil Engineering, Tongji University, Shanghai, 200092, China; e-mail: 2351606@tongji.edu.cn

Youbo Lian[#*]
Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Guangdong, 519087, China; e-mail: t330026086@mail.bnbu.edu.cn

[#]Yiping Zhang and Youbo Lian contributed equally to this work, they are both first authors.

Corresponding author: t330026086@mail.bnbu.edu.cn

Single-image dehazing aims to restore a clear image from a single haze-degraded observation. Although recent deep learning methods have shown promising performance, they still struggle to balance long-range dependency modelling and computational efficiency. To address this issue, we propose a lightweight U-shaped Transformer network, termed UTNet. The proposed network adopts a U-shaped encoder-decoder architecture to progressively extract and reconstruct hierarchical features, enabling effective integration of shallow spatial details and deep contextual information. To reduce the computational burden of Transformer-based modelling, local-window spatial attention is introduced at each stage to efficiently model dependencies. Unlike Swin Transformer, UTNet leverages its hierarchical U-shaped architecture to enlarge the receptive field and progressively capture global contextual information. In addition, channel attention is incorporated in parallel with spatial attention to enhance feature representation further. Multi-scale auxiliary supervision is also introduced to improve optimization stability and promote consistent restoration from coarse structures to fine details. Moreover, gamma-based data augmentation is employed to improve the model generalization to real-world hazy images. Extensive experiments on the Haze4K and NH-Haze datasets demonstrate that the proposed UTNet achieves competitive dehazing performance while maintaining a compact architecture with fewer parameters and lower computational cost.

KEYWORDS: Single Image Dehazing, Lightweight U-Shaped Transformer, Spatial-Channel Attention, Multi-Scale Supervision, Gamma Augmentation

1. Introduction

Single-image dehazing aims to recover a clear image from a single haze-degraded observation. Haze reduces image contrast, distorts colour, obscures structural details, and degrades the reliability of subsequent vision tasks. It is therefore an important preprocessing step in autonomous driving, unmanned aerial vehicle perception, and intelligent surveillance. These applications often operate on resource-constrained devices, making restoration quality and computational efficiency equally important. The problem is severely ill-posed because scene radiance, transmission, atmospheric light, and depth are unknown from a single input, particularly under dense or spatially varying haze, complex illumination, and weak textures [11, 22].

Early dehazing methods estimated transmission and atmospheric light using handcrafted assumptions, including the dark channel prior, boundary constraints, and non-local colour modelling [2, 9, 11, 20, 28]. Although physically interpretable, these methods are sensitive to violations of their priors. Deep-learning approaches subsequently learned transmission estimation or direct image restoration from paired data [3, 15, 26, 32]. Multi-scale architectures, feature attention, dense fusion, and contrastive learning further improved contextual aggregation and restoration fidelity [4, 7, 16, 24, 30].

Transformer-based methods extend contextual modelling through self-attention. DehazeFormer adapts window-based Transformer components to low-level restoration and provides compact T and S variants [27]. C2PNet combines curricular contrastive regularisation with a physics-aware architecture [33], while DehazeDCT introduces deformable Transformer blocks for non-homogeneous haze removal [8]. Other efficiency-oriented approaches include MB-TaylorFormer, which employs Taylor-expanded linear attention [25], LDFormer for lightweight dehazing [23], and DehazeXL for global-context fusion in large images [5].

Despite these advances, broad contextual modelling remains difficult under a strict computational budget. Global self-attention is expensive at high spatial resolutions, whereas restricting attention to local regions can limit long-range interaction. This motivates an architecture that aggregates context across scales without substantially increasing model complexity.

To this end, we propose UTNet, a lightweight U-shaped Transformer for single-image dehazing. UTNet adopts a hierarchical encoder-bottleneck-decoder architecture with local-window spatial attention and parallel channel attention. Progressive downsampling expands contextual coverage across stages, while skip connections preserve spatial information during reconstruction. Multi-scale auxiliary supervision constrains intermediate decoder outputs, and gamma augmentation increases illumination diversity during training.

Experiments on the synthetic Haze4K dataset and the real-world NH-Haze dataset demonstrate that UTNet achieves competitive restoration quality with a compact computational footprint.

The main contributions of this work are summarised as follows:

1. We propose UTNet, a lightweight U-shaped Transformer that combines hierarchical feature processing with efficient contextual modelling for single-image dehazing.
2. We design a hybrid spatial-channel attention block that integrates local-window spatial attention with channel-wise feature recalibration.
3. We introduce multi-scale auxiliary supervision to constrain intermediate decoding stages and improve network optimisation.
4. We employ gamma augmentation to increase illumination diversity and improve robustness to real-world haze conditions.

2. Related Work

Single-image dehazing has progressed from physically motivated priors to CNN-based representation learning and Transformer-based contextual modelling. Representative developments are reviewed below.

2.1. Traditional Prior-based Dehazing Methods

Traditional methods were generally formulated under the atmospheric scattering model [19]. Because

transmission and atmospheric light cannot be directly recovered from a single observation, these approaches introduced assumptions based on local contrast, natural image statistics, or colour distributions.

Tan enhanced visibility by maximising local contrast [28], while Fattal estimated transmission using the statistical independence of surface shading and transmission [9]. He et al. proposed the dark channel prior [11]. Meng et al. combined boundary constraints with contextual regularisation [20], and Berman et al. modelled haze-free colours as non-local haze lines in RGB space [2].

These methods provide interpretable solutions, but their performance depends on the validity of the underlying priors and may degrade in bright regions, dense or non-homogeneous haze, and complex illumination.

2.2. CNN-based Dehazing Methods

CNN-based methods [14] replaced handcrafted optimisation with trainable mappings. DehazeNet estimated the transmission map [3], MSCNN combined coarse and fine-scale estimation [26], AOD-Net reformulated the atmospheric model into an end-to-end mapping [15], and DCPDN jointly estimated transmission, atmospheric light, and the restored image [32].

Later architectures placed greater emphasis on contextual aggregation and adaptive feature processing. GCANet used gated context aggregation [4], Grid-DehazeNet introduced cross-resolution feature exchange [16], FFA-Net combined channel and pixel attention [24], and MSBDN employed dense multi-scale feature fusion [7]. Contrastive learning further improved feature discrimination [30].

Although CNN-based methods substantially improved restoration quality, broad spatial dependencies often require deep or complex convolutional architectures.

2.3. Transformer-based Dehazing Methods

The Transformer was originally introduced for sequence modelling [29] and was later adapted to image dehazing. Guo et al. incorporated transmission-aware three-dimensional positional embedding [10]. DehazeFormer revised normalisation, activation, and window-based aggregation for low-level restoration [27]. C2PNet combined curricular contrastive regularisation with physics-aware modelling [33], and DehazeDCT employed deformable Transformer blocks for spatially varying haze [8].

Recent studies have also explored efficient Transformer designs. DehazeFormer-T/S provide compact window-attention variants [27], MB-TaylorFormer uses a multi-branch architecture with Taylor-expanded linear attention [25], and LDFormer focuses on lightweight dehazing [23]. DehazeXL instead targets global-context fusion for large and ultra-high-resolution images [5]. UTNet differs by using progressive resolution reduction to aggregate context across scales while retaining local-window attention.

Synthetic-to-real generalisation has been investigated through physical priors and unpaired learning. PSD uses physical guidance for domain transfer [6], while self-augmented unpaired dehazing decomposes haze density and scene depth [31]. These approaches address data-domain adaptation, whereas UTNet focuses on lightweight hierarchical feature modelling.

3. Methods

3.1. Atmospheric Scattering Model

The task of recovering a clean scene from haze-degraded input is closely related to the physical imaging process of atmospheric scattering. Unlike general image enhancement tasks, haze removal must account for the depth-dependent nature of atmospheric degradation. In outdoor scenes, image degradation is mainly caused by two coupled factors: attenuation of scene radiance during propagation and additive airlight scattered by atmospheric particles. As a result, the captured hazy image suffers from reduced contrast, faded colours, blurred wdetails, and reduced visibility, especially in scenes with dense haze, long viewing distances, or spatially varying atmospheric conditions.

The formation of a hazy image is commonly described by the atmospheric scattering model [19]. Given a hazy image at pixel location , the model can be written as

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

where $J(x)$ denotes the underlying haze-free image, $t(x)$ is the transmission map representing the portion of scene radiance that reaches the camera, and A

denotes the global atmospheric light. In this model, the first term $J(x)t(x)$ describes the attenuated scene radiance, while the second term $A(1–t(x))$ represents the scattered airlight introduced by the atmosphere.

The transmission map is usually related to scene depth and can be formulated as

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

where β is the atmospheric scattering coefficient and $d(x)$ is the scene depth. This expression indicates that the farther the scene point is from the camera, the smaller the transmission becomes, and the stronger the haze effect appears. Consequently, haze is naturally a depth-dependent and spatially non-uniform degradation. This also explains why distant objects often exhibit greater contrast attenuation and colour distortion than nearby regions.

Although the atmospheric scattering model provides a clear physical explanation for haze formation, recovering $J(x)$ from a single input image remains highly ill-posed, because the transmission map $t(x)$, the atmospheric light A, and the clean image $J(x)$ are all unknown. Meanwhile, only one degraded observation is available. Moreover, real-world haze often deviates from the ideal assumptions of the model due to a non-homogeneous atmosphere, varying illumination, and spatially inconsistent airlight. From the perspective of representation learning, effective dehazing requires the model to preserve fine local structures such as edges and textures, capture broader contextual information for estimating haze density and scene structure, and adaptively emphasize informative features while suppressing haze-dominant responses. Motivated by these observations, we design UTNet as a lightweight U-shaped Transformer framework that integrates hierarchical reconstruction, efficient spatial modelling, and adaptive channel-wise feature selection.

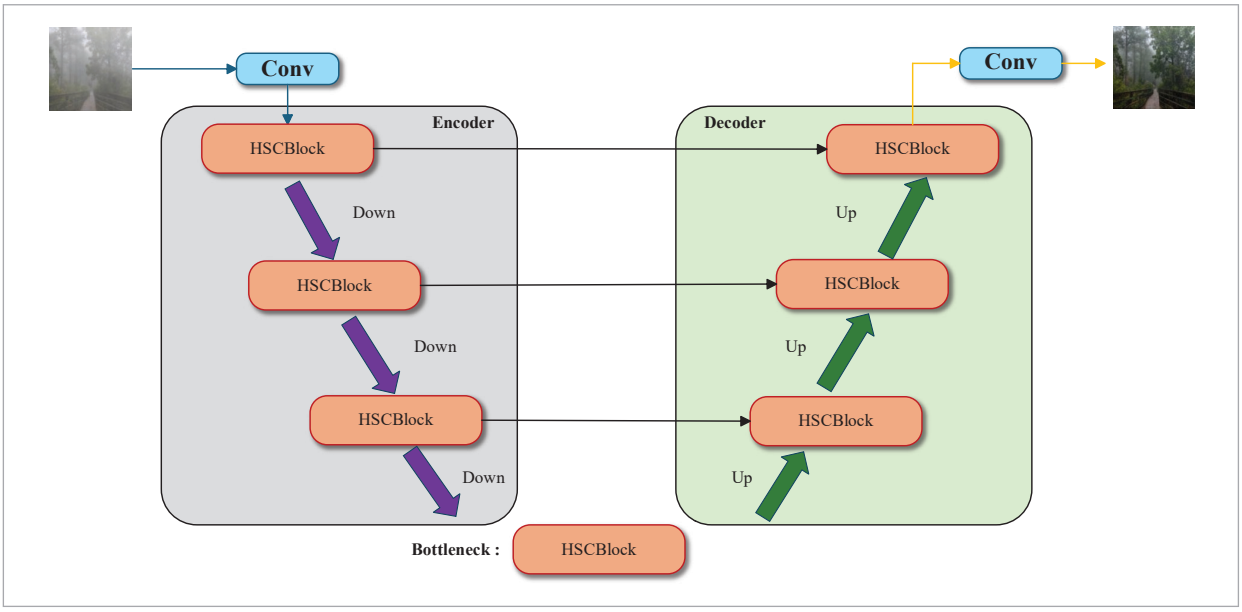
3.2. Overall Architecture of UTNet

UTNet adopts a U-shaped encoder–bottleneck–decoder architecture with HSCBlocks, skip connections, and multi-scale outputs, as shown in Figure 1. A convolutional embedding maps the hazy input to shallow features, which are progressively encoded and reconstructed across scales.

Encoder. Each encoder stage applies one HSCBlock followed by 2× downsampling. The reduced feature map merges neighbouring regions, enabling cross-window interaction, while the stage output is retained for the corresponding skip connection.

Bottleneck and receptive field. Let s denote the local-window side length. After i successive 2× downs-

Figure 1
Overall architecture of UTNet.



ampling operations, one bottleneck window covers the following side length in the input image:

$$RF(i) = s \cdot 2^{i}. \tag{3}$$

Accordingly, the number of downsampling stages required for a window to cover an L×L input is

$$n = log_2(L/s). \tag{4}$$

With L=64 and s=8, Equation (4) gives n=3. The bottleneck resolution is therefore 8×8, equal to the window size, so a single window covers the complete bottleneck feature map and RF(3)=64 covers the entire training patch.

Decoder. The decoder progressively upsamples the bottleneck features and fuses them with the corresponding encoder features through skip connections. Each stage applies an HSCBlock, and prediction heads at multiple resolutions provide auxiliary supervision.

3.3. Hybrid Spatial–Channel Attention Block (HSCBlock)

A central component of UTNet is the proposed Hybrid Spatial–Channel Attention Block (HSCBlock), which is designed to model spatial dependencies and channel-wise feature importance jointly. This design is motivated by the observation that haze degradation manifests in both spatial and channel domains. Spatially, haze is often non-uniform and coupled with scene depth, making it necessary to model structural dependencies beyond local neighbourhoods. Channel-wise, different feature channels respond differently to haze density, color attenuation, and airlight interference, implying that adaptive feature recalibration is also important for effective restoration. Based on these observations, HSCBlock integrates a spatial attention branch and a channel attention branch into a unified feature refinement module. The overall structure of HSCBlock is illustrated in Figure 2.

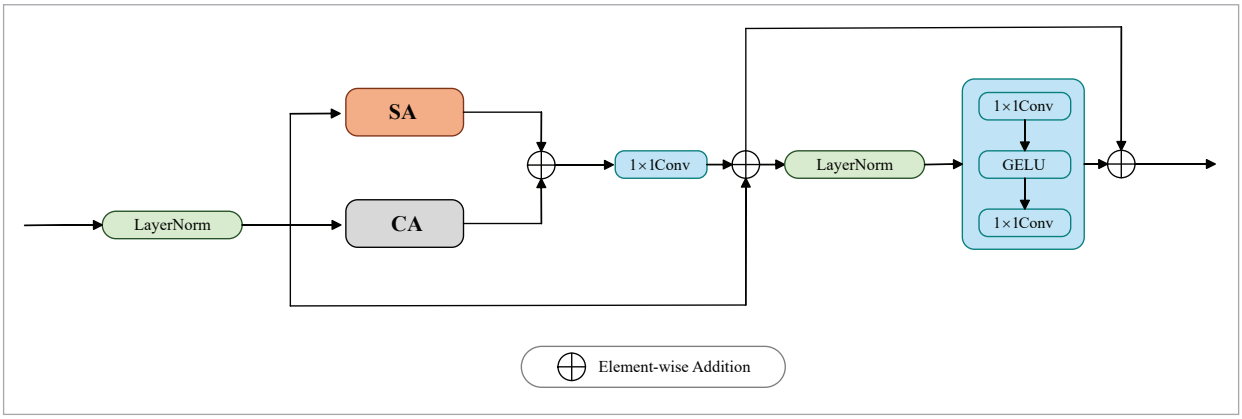
As illustrated in Figure 2, given an input feature map $X \in R^{C \times H \times W}$, the HSCBlock first normalizes and transforms the input features, and then sends them into two parallel branches: the spatial attention SA branch and the channel attention CA branch. The outputs of the two branches are then fused and added back to the input through a residual connection. This formulation allows the block to preserve the stability of feature propagation while enhancing the discriminative capability of the learned representations.

Spatial Attention. The purpose of the spatial attention branch is to model structural dependencies among different spatial locations. For a given input feature map X, we first project it into query, key, and value representations:

$$Q = W_q X, \quad K = W_k X, \quad V = W_v X, \tag{5}$$

where W_q, W_k, and W_v denote learnable linear projections. Instead of applying global self-attention, which is computationally expensive for image restoration tasks, we adopt a local-window attention

Figure 2
Overall architecture of the proposed HSCBlock.



strategy. Specifically, the feature map is partitioned into non-overlapping local windows of size 8×8 with a stride 8, and self-attention is computed within each window as

$$SA(X) = Softmax\left(\frac{QK^{\top}}{\sqrt{d}}\right)V, \tag{6}$$

where d denotes the feature dimension of each attention head.

This design has several advantages. First, it significantly reduces the computational burden compared with global self-attention, making it more suitable for lightweight dehazing networks. Second, local-window attention aligns well with the observation that many haze-related degradations, such as contrast attenuation and detail fading, exhibit localized spatial patterns. Therefore, computing attention within local regions allows the network to focus on meaningful structural relationships while maintaining efficiency. Third, because HSCBlocks are stacked across different scales in the U-shaped architecture, the effective receptive field of the overall network can still be progressively enlarged. As a result, UTNet achieves a favourable balance between local structural modelling and broader contextual interactions.

Channel Attention. While the spatial attention branch focuses on modelling spatial dependencies, it does not explicitly account for the fact that different feature channels may contribute unequally to dehazing. To complement the spatial branch, we introduce a channel attention mechanism [12] that performs adaptive feature recalibration across channels.

Given the input feature map X, we first aggregate spatial information by global average pooling:

$$z_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} X_c(i,j), \tag{7}$$

where z_c denotes the global descriptor of the c-th channel. Based on this compact channel representation, we apply a lightweight gating function to estimate channel-wise importance weights:

$$s = \sigma(W_2 \cdot \delta(W_1 z)), \tag{8}$$

where and are learnable parameters, denotes the ReLU activation function [21], and denotes the sigmoid function. The reweighted feature map is then obtained by

$$CA(X) = s \odot X, \tag{9}$$

where ⊙ denotes channel-wise multiplication.

This process enables the network to selectively emphasize channels that are more informative for haze removal while suppressing redundant or haze-dominant responses. In image dehazing, such adaptive feature selection is particularly important because haze does not affect all feature channels equally. By explicitly modelling channel importance, the CA branch improves feature discrimination and complements the structural modelling ability of the SA branch.

Fusion of Spatial and Channel Attention. After obtaining the outputs of the two branches, we fuse them as

$$Y = X + W_f(\mathrm{SA}(X) + \mathrm{CA}(X)), \tag{10}$$

where W_f denotes a learnable 1×1 convolution used for feature aggregation. This formulation allows the two branches to complement each other: the SA branch focuses on the locations of important structures, while the CA branch determines which feature channels to emphasize. Since haze degradation is both spatially variant and feature-dependent, effective restoration cannot rely solely on a single type of attention. By integrating spatial and channel attention into a unified block, HSCBlock simultaneously enhances local structural reasoning and adaptive feature selection, achieving a better balance between representation capacity and computational efficiency.

3.4. Loss Function

To effectively train UTNet and fully exploit its hierarchical reconstruction capability, we adopt a multi-scale supervision strategy. Instead of imposing supervision only on the final full-resolution prediction, UTNet produces multiple dehazed outputs at different decoding stages. These intermediate outputs correspond to different spatial resolutions and are jointly optimized during training. Such a design introduces additional gradient signals into intermediate layers, facilitates optimization, and encourag-

es the network to learn a consistent restoration behaviour from coarse structures to fine details.

Let O_i denote the predicted output at scale i, and let G_i denote the corresponding ground-truth image resized to the same scale. The multi-scale reconstruction loss is defined as

$$L_{ms} = \sum_{i=0}^{3} \lambda_i \|O_i - G_i\|_1, \quad (11)$$

where λ_i controls the contribution of the supervision at the i-th scale. In this work, we assign the loss weights as

$$[\lambda_0, \lambda_1, \lambda_2, \lambda_3] = [1, 0.5, 0.25, 0.125]$$

Under this setting, the final full-resolution output is assigned the largest weight, while lower-resolution outputs receive progressively smaller weights. This weighting strategy is motivated by the fact that the highest-resolution prediction is the ultimate restoration target, whereas intermediate outputs primarily serve as auxiliary supervision to stabilize optimization and guide hierarchical feature learning.

More specifically, coarse-scale supervision helps the network learn globally consistent dehazing behaviour in the early and middle stages of decoding, while stronger supervision at the finest scale encourages accurate reconstruction of high-frequency details. This combination improves gradient propagation through the U-shaped architecture and reduces the risk of unstable optimization. Although the formulation is simple, this multi-scale loss plays an important role in practice. It not only enhances convergence stability but also helps UTNet achieve a better balance between global structural restoration and local detail recovery.

3.5. Complexity Analysis

Let H×W×C denote a feature map and s the local-window side length. Following Swin Transformer [18], the computational costs of global multi-head self-attention (MSA) and window attention (W-MSA) are

$$\Omega(MSA) = 4HWC^2 + 2(HW)^2C \quad (12)$$

$$\Omega(W-MSA) = 4HWC^2 + 2s^2HWC. \quad (13)$$

At hierarchy level i, UTNet reduces the spatial resolution to $H/2^i \times W/2^i$ and increases the channel width to 2^iC. The corresponding attention cost is

$$\Omega_i = 4HWC^2 + 2s^2HWC/2^i. \quad (14)$$

Summing over n+1 hierarchy levels gives

$$\Omega hier = \Sigma_{i=0}^{n} \Omega_i \quad (15)$$

$$\Omega hier \leq 4(n+1)HWC^2 + 4s^2HWC. \quad (16)$$

The attention term decreases geometrically with depth, and the overall complexity remains linear in the number of spatial positions. The measured UTNet FLOPs increase from 0.315G at 64^2 to 5.04G at 256^2 and 20.2G at 512^2, consistent with this analysis. Table 1 provides an empirical comparison with representative Transformer models.For L=64 and s=8, shifted-window propagation at a fixed resolution requires approximately L/s=8 cross-window steps, whereas UTNet reaches full bottleneck coverage after (L/s)=3 downsampling stages.

Table 1

Empirical complexity comparison of representative Transformer models. Swin-T values are reported in [18]; the four dehazing models are evaluated at 64×64 under the same protocol.

Model	Attention	Input	Params (M)	GFLOPs
Swin-T [18]	W-MSA	224^2	29.00	4.50
DehazeFormer-T	Swin-style W-MSA	64^2	0.69	0.44
DehazeFormer-S	Swin-style W-MSA	64^2	1.29	0.83
MB-TaylorFormer-B	Taylor linear attention	64^2	2.68	1.99
UTNet (ours)	**Hierarchical W-MSA**	**64^2**	**1.34**	**0.32**

4. Experiments

4.1. Experimental Settings

Datasets. We conduct experiments on two representative dehazing benchmarks, namely Haze4K and NH-Haze. Haze4K [17] is a large-scale synthetic dehazing dataset containing 4,000 hazy-clean image pairs, which are divided into 3,000 pairs for training and 1,000 pairs for testing. The hazy images are synthesized using an atmospheric-scattering model with varying haze densities and scene depths. NH-Haze [1] is a real-world non-homogeneous haze dataset containing 55 pairs of hazy and corresponding haze-free images captured in outdoor scenes. In this work, the proposed UTNet is trained exclusively on the Haze4K training set. The trained model is then evaluated on both the Haze4K test set and the full NH-Haze dataset.

Model configuration. Unless otherwise specified, all experiments are conducted using the proposed UTNet model. According to the implementation, the network adopts a four-scale U-shaped encoder-bottleneck-decoder architecture with a base channel number of 32. After the initial 3×3 convolutional embedding, the encoder contains three hierarchical stages with channel dimensions of 32, 64, and 128, respectively, followed by a bottleneck stage with 256 channels. The decoder then reconstructs the dehazed image through three symmetric upsampling stages with channel dimensions of 128, 64, and 32. Each encoder stage, the bottleneck, and each decoder stage contains one HSCBlock, resulting in a total of seven HSCBlocks in the default UTNet configuration. In the spatial attention branch, the number of attention heads is set to 4 and the local window size to 8.

Training details. All experiments are implemented in PyTorch 2.9.0 with CUDA 12.8 and conducted on an NVIDIA GeForce RTX 5070 Ti Laptop GPU (12 GB). During training, hazy and haze-free image pairs are cropped into patches. The proposed model is optimized using the Adam optimizer [13] with an initial learning rate of , a batch size of 64, and a total of 500 epochs. Unless otherwise stated, the default multi-scale supervision weights are set to

$$[\lambda_0,\lambda_1,\lambda_2,\lambda_3]=[1,0.5,0.25,0.125].$$

Unless otherwise stated, all experiments follow the default training pipeline described above.

Evaluation metrics. We adopt both restoration-quality and efficiency-oriented metrics for comprehensive evaluation. For image restoration performance, PSNR and SSIM are used as the primary quantitative metrics on both Haze4K and NH-Haze. To evaluate model complexity and inference efficiency, we further report the number of parameters, FLOPs, and FPS, computed at an input size of 64×64 for fair comparison across methods. Through this evaluation protocol, we assess not only the dehazing capability of UTNet but also its computational efficiency and practicality as a lightweight restoration model.

4.2. Comparison with Existing Methods

To evaluate UTNet against representative CNN- and Transformer-based dehazing methods, we compare it with FFA-Net [24], C2PNet [33], DehazeFormer-T/S [27], and MB-TaylorFormer-B [25]. The quantitative results are summarized in Table 2.

Quantitative comparison. UTNet achieves 23.26 dB/0.9350 on Haze4K, comparable to DehazeFormer-T (23.24/0.9355) and DehazeFormer-S (23.21/0.9390), and higher than MB-TaylorFormer-B (22.38/0.9186). On NH-Haze, UTNet obtains 11.96 dB/0.5032 and remains competitive with the DehazeFormer variants. Although competing methods achieve higher values on individual metrics, UTNet provides a favourable balance between restoration quality and computational efficiency.

Visual comparison on the synthetic dataset. To further examine the restoration quality, we present visual comparisons on the Haze4K test set in Figure 3. Compared with FFA-Net [24] and C2PNet [33], the proposed UTNet produces dehazed results with clearer structural recovery, more natural colour restoration, and fewer residual haze artifacts. In particular, UTNet better restores scene contrast and preserves object boundaries in heavily degraded regions, while avoiding the over-smoothed appearance or colour inconsistency that may occur in competing methods. These observations are consistent with the quantitative results in Table 2 and suggest that the proposed hybrid spatial-channel design is beneficial for recovering both global scene structure and local details.

Visual comparison on the real-world dataset. We further provide qualitative results on NH-Haze in Figure 4 to evaluate the practical performance of

Table 2

"Pretrained" denotes results from officially released models. DehazeFormer-T, DehazeFormer-S, and MB-TaylorFormer-B were retrained for 300 epochs under the Haze4K tile-64 data-processing and evaluation protocol, whereas UTNet follows its 500-epoch schedule.

Method	Haze4K PSNR ↑	Haze4K SSIM ↑	NH-Haze PSNR ↑	NH-Haze SSIM ↑
FFA-Net (pretrained)	19.40	0.9040	11.41	0.4184
FFA-Net (retrained)	21.55	0.8878	11.82	0.4340
C2PNet (pretrained)	20.06	0.8792	10.21	0.3617
C2PNet (retrained)	**23.92**	**0.9396**	**12.29**	0.4861
DehazeFormer-T (retrained)	23.24	0.9355	12.16	**0.5042**
DehazeFormer-S (retrained)	23.21	0.9390	12.25	0.5031
MB-TaylorFormer-B (retrained)	22.38	0.9186	11.60	0.4968
UTNet (ours)	23.26	0.9350	11.96	0.5032

Figure 3

Visual comparisons with representative dehazing methods on the synthetic Haze4K dataset.

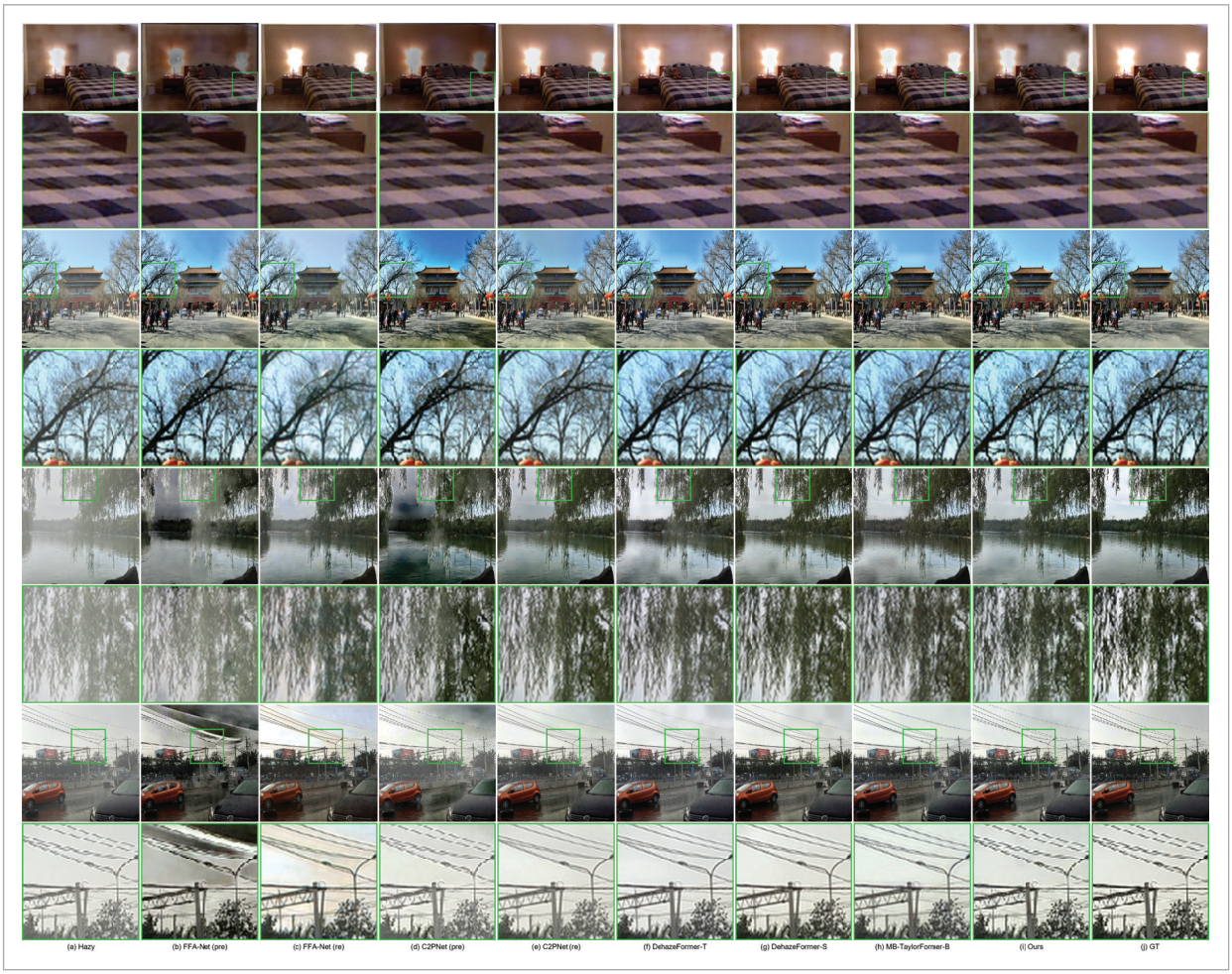


Figure 4

Visual comparisons with representative dehazing methods on the real-world NH-Haze dataset.

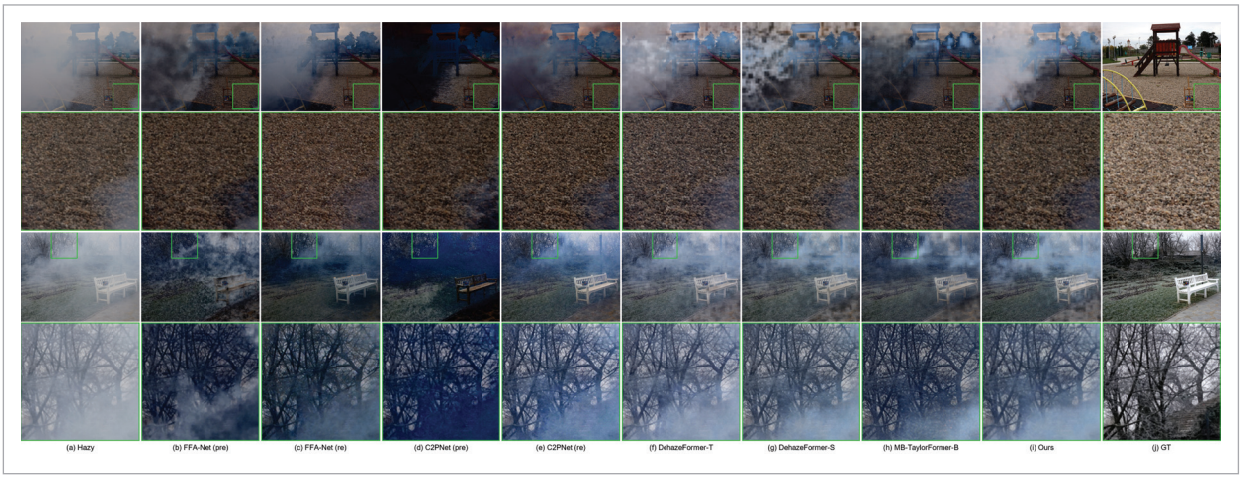
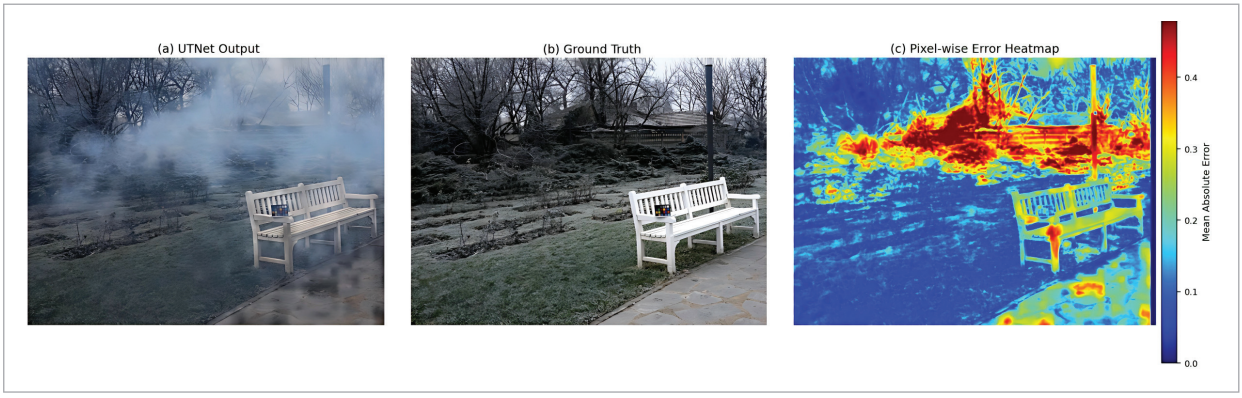


Figure 5

Pixel-wise error heatmap of UTNet on an NH-Haze test image.



UTNet on real-world haze. Since real hazy scenes usually involve more complicated haze distributions, illumination variations, and texture ambiguity, they are substantially more challenging than synthetic benchmarks. As shown in Figure 4, all compared methods exhibit noticeable performance degradation on real-world scenes. The proposed UTNet still suffers from incomplete haze removal in dense regions and slight color deviation in areas with strong illumination variation. However, compared with FFA-Net [24] and C2PNet [33], UTNet produces relatively cleaner results with better contrast recovery and more faithful color reproduction. The competing methods tend to leave more prominent haze residues and produce less natural color tones, particularly in distant regions and areas with complex scene structures. Overall, although real-world dehazing remains a challenging problem for all evaluated methods, UTNet achieves comparatively better visual quality while maintaining its lightweight architecture.

To further analyze the reconstruction quality of UTNet on NH-Haze, we compute the pixel-wise mean absolute error between the model output and the ground truth and visualize it as an error heatmap in Figure 5. As shown, large reconstruction errors are mainly concentrated in regions with dense non-homogeneous haze, complex sky–background boundaries, and fine vegetation textures. These regions exhibit strong spatially varying attenuation, weak structural cues, and pronounced illumination changes, which are insufficiently represented by the synthetic Haze4K training data. In UTNet, high-resolution features are primarily processed within local windows, while broader contextual information is progressively aggregated

through downsampling. This design may provide limited direct cross-window interaction at early stages when local structures are severely obscured. Moreover, the compact model capacity and the absence of explicit physical and colour-consistency constraints reduce robustness to extreme real-world haze, resulting in residual haze and slight colour deviation.

4.3. Failure Case Analysis

To further examine the influence of the synthetic-to-real distribution gap, we conducted a leakage-free held-out real-domain experiment. The 55 NH-Haze image pairs were split into 44 training pairs and 11 strictly held-out test pairs. UTNet was retrained for 300 epochs using the same configuration as the Haze4K-only control, with the 44 real-image pairs oversampled by 30 times. The held-out images were never used for training or model selection. On the 11 held-out images, the Haze4K-only model achieves 11.79 dB PSNR and 0.476 SSIM, whereas adding the 44 NH-Haze training pairs increases the performance to 17.05 dB and 0.634 SSIM. Representative difficult images Nos. 47, 46, and 53 improve by 7.47, 4.44, and 8.37 dB, respectively, as shown in Figure 6.

Since the model architecture, capacity, training epochs, and optimization settings remain unchanged, the substantial improvement on never-seen NH-Haze images indicates that the synthetic-to-real distribution gap is a major contributor to the residual haze and colour deviation. The jointly trained model, however, decreases from 22.48 to 21.48 dB on the Haze4K test set under the same 300-epoch setting, revealing a trade-off between synthetic-domain accuracy and real-domain adaptation. Therefore, the main comparison continues to report the original Haze4K-only 500-epoch model. The regional analysis further shows that dense-haze areas remain the most difficult, with an average PSNR of 7.57±3.25 dB, compared with 14.82±4.36 dB in sky regions.

4.4. Cross-Resolution Evaluation

To evaluate the resolution adaptability of UTNet, the same trained model is tested on the Haze4K test set at 256×256, the native 400×400 resolution, and 512×512. All images are processed using full-image inference without tiling or additional fine-tuning. The results are reported in Table 3.

Table 3
Cross-resolution evaluation of UTNet on the Haze4K test set.

Input resolution	PSNR (dB)	SSIM
256×256	22.81	0.9141
Native (400×400)	22.96	0.9182
512×512	23.05	0.9201

UTNet maintains stable restoration performance across the three input resolutions, with a maximum PSNR variation of 0.23 dB and an SSIM variation of 0.0060. The results show that the hierarchical encoder–decoder structure remains effective as the spatial dimensions of the intermediate features change. Although larger inputs produce multiple local windows at the bottleneck, progressive downsampling and multi-scale feature aggregation preserve consistent dehazing performance without resolution-specific retraining.

4.5. Efficiency and Memory Analysis

In addition to restoration quality, computational efficiency and inference memory are important for practical image dehazing. We compare UTNet with FFA-Net, C2PNet, DehazeFormer-T, DehazeFormer-S, and MB-TaylorFormer-B in terms of parameters, FLOPs, FPS, latency, peak GPU memory, and peak CPU memory. All models are evaluated with a 64×64 input and a batch size of 1 using PyTorch 2.9.0 on Windows 11. Since no dedicated edge device was available, an NVIDIA GeForce RTX 5070 Ti Laptop GPU with 12 GB of memory and the host CPU were used as a consistent proxy platform.

As shown in Table 4, UTNet achieves the lowest FLOPs and inference latency among the compared methods, reaching 108.93 FPS with a latency of 9.18 ms. Its peak GPU memory is 25.76 MB, close to FFA-Net and lower than C2PNet, DehazeFormer-S, and MB-TaylorFormer-B, although DehazeFormer-T requires less GPU memory. UTNet uses 109.7 MB peak CPU memory; FFA-Net records the lowest value overall at 69.1 MB, while UTNet has the lowest CPU memory among the Transformer-based baselines (109.7 versus 165.7–196.0 MB). Overall, the main efficiency advantage of UTNet lies in its low computational cost and fast inference, together with moderate GPU and CPU memory requirements.

Figure 6

Failure-case diagnostic on held-out NH-Haze images Nos. 47, 46, and 53. From left to right: hazy input, output of the Haze4K-only model, output of the model retrained on Haze4K plus 44 NH-Haze training pairs, and ground truth. The displayed test images were never used for training or model selection. Green boxes indicate representative regions with large reconstruction errors.

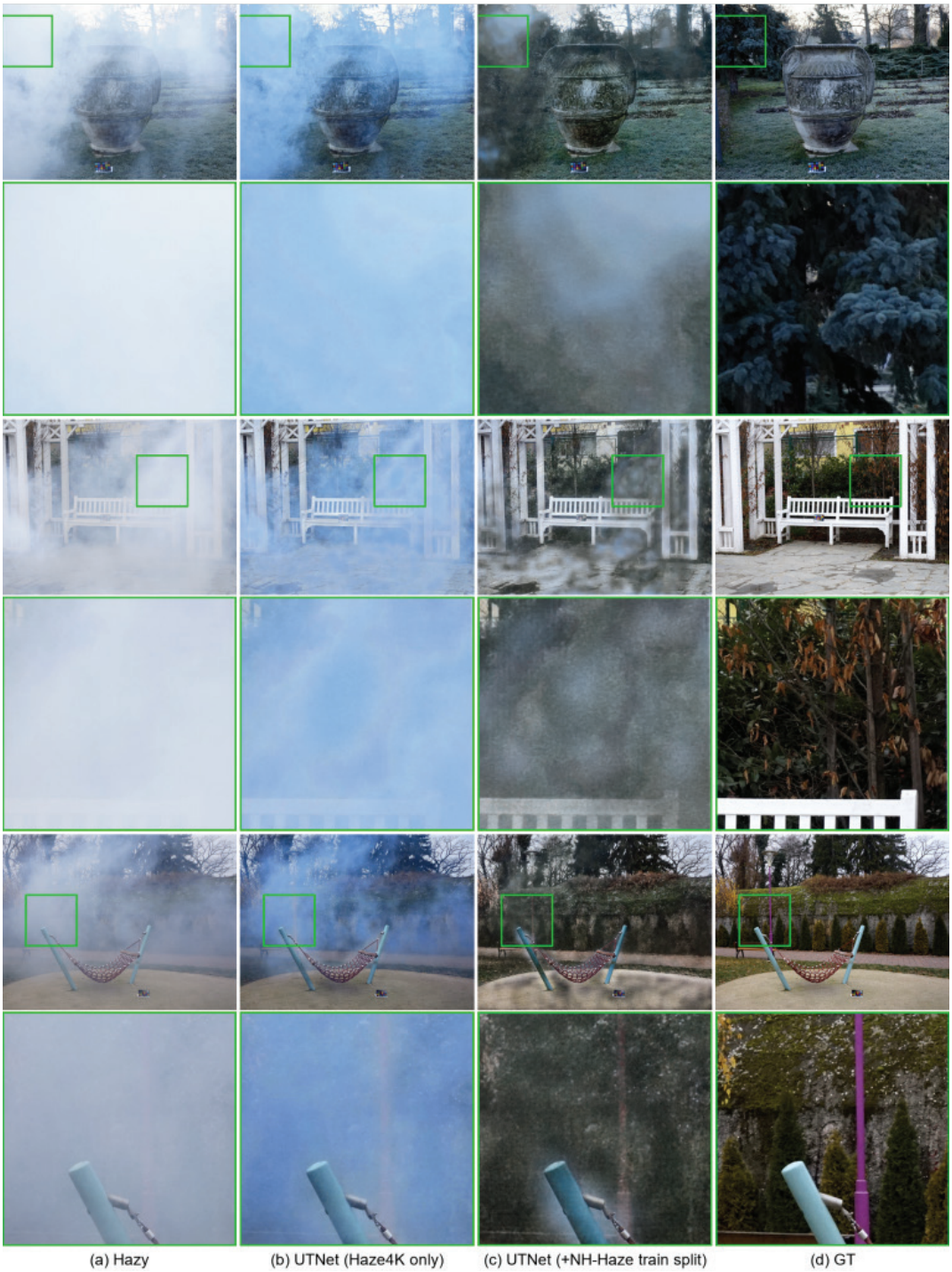


Table 4

Quantitative comparison of model complexity, inference efficiency, and memory consumption.Peak CPU memory is measured as the per-process increment over the post-import baseline at 64×64 resolution with a batch size of 1, excluding common framework overhead.

Method	Params	FLOPs	FPS	Latency	Peak GPU Memory	Peak CPU Memory
FFA-Net	4.456M	17.971G	26.84	37.26 ms	25.34 MB	**69.1 MB**
C2PNet	7.169M	28.810G	16.4	60.96 ms	36.56 MB	97.1 MB
DehazeFormer-T	**0.687M**	0.435G	76.77	13.03 ms	**21.81 MB**	165.7 MB
DehazeFormer-S	1.285M	0.83 4G	40.5	24.69 ms	27.84 MB	173.6 MB
MB-TaylorFormer-B	2.677M	1.989G	8.05	124.18 ms	42.94 MB	196.0 MB
UTNet (ours)	1.342M	**0.315G**	**108.93**	**9.18 ms**	25.76 MB	109.7 MB

4.6. Gamma Augmentation

Gamma augmentation models exposure and luminance variation by applying the same exponent to each paired hazy and ground-truth image:

$$I' = I^{\gamma}, \tag{17}$$

where γ is applied identically to both images in a pair, preserving their structural correspondence. Figure 7 illustrates the exposure changes. We first determine the sampling interval and then evaluate the contribution of gamma augmentation.

Table 5 shows that [0.8,1.2] gives the highest Haze4K PSNR and NH-Haze SSIM, whereas [0.6,1.4] gives the highest Haze4K SSIM and NH-Haze PSNR. The wider [0.4,1.6] range is not optimal on either dataset. We therefore select [0.6,1.4] as the default interval, balancing synthetic-domain accuracy and real-domain generalisation. These results are used only for interval selection and are not directly compared with the 500-epoch main results.

Figure 7

Gamma-based augmentation at different exposure levels. The display gamma values are g_{new}=2.0, 2.2, and 2.4 from left to right, where g_{new}=2.2 is the identity case. The training transformation follows Equation (17).

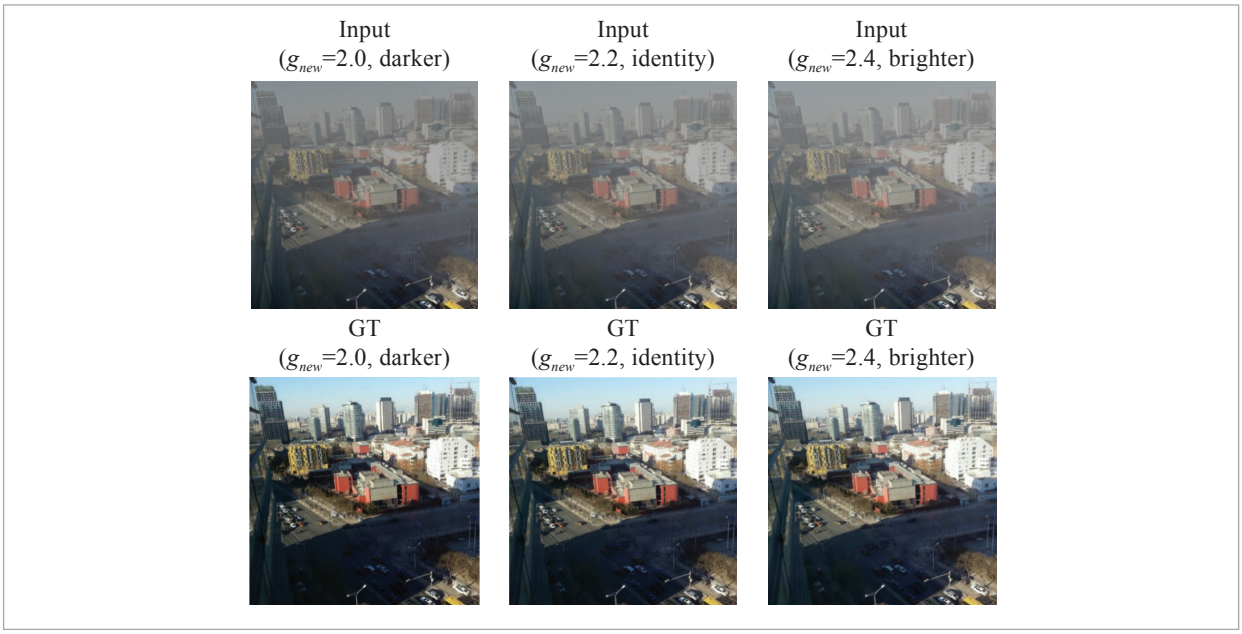


Table 5
Gamma interval selection under a common 300-epoch setting. Results are compared only within this table; the best result is bold and the second-best result is red.

Gamma Range	Haze4K PSNR ↑	Haze4K SSIM ↑	NH-Haze PSNR ↑	NH-Haze SSIM ↑
Narrow [0.8,1.2]	**23.0281**	0.9229	11.9639	**0.5015**
Default [0.6,1.4]	22.4798	**0.9262**	**12.3757**	0.4900
Wide [0.4,1.6]	22.4805	0.9257	11.9289	0.4932

Table 6
Augmentation comparison using the selected gamma range [0.6,1.4]. The best result is bold and the second-best result is red.

Augmentation	Haze4K PSNR ↑	Haze4K SSIM ↑	NH-Haze PSNR ↑	NH-Haze SSIM ↑
None (baseline)	21.42	0.9081	**12.30**	0.4988
Flip	**23.50**	**0.9371**	11.82	0.4864
Flip + Gamma (ours)	23.26	0.9350	11.96	**0.5032**

Using the selected interval, Table 6 shows that Flip+Gamma improves NH-Haze SSIM from 0.4864 to 0.5032 compared with Flip, indicating greater robustness to illumination variation. However, its NH-Haze PSNR remains below the unaugmented baseline (11.96 versus 12.30 dB). Gamma transformation changes global luminance but does not reproduce the spatial distribution of real non-homogeneous haze. This is consistent with Section 4.3, where dense-haze regions reach only 7.57±3.25 dB compared with 14.82±4.36 dB in sky regions. Thus, gamma augmentation improves structural similarity under brightness variation but cannot eliminate the main domain gap caused by localized dense haze.

4.7. Multi-scale Loss Weights

To investigate how the supervision weights at different output scales affect the performance of UTNet, we conduct a comparative study on the multi-scale loss design. To isolate the effect of loss weighting, all models in this experiment are trained without data augmentation. Specifically, we evaluate four weight settings for each of the four output losses. L_0, L_1, L_2, and L_3, as listed in Table 7. Among them, [1,0,0,0] corresponds to the single-scale supervision setting, where only the final output is directly constrained, while the other three settings introduce progressively decaying auxiliary supervision at intermediate scales.

The quantitative results are reported in Table 7. It can be observed that different weight distributions yield notably different restoration behaviours across the two datasets. The single-scale supervision setting achieves relatively balanced performance, but lacks additional guidance from intermediate outputs. Introducing multi-scale supervision generally improves the structural learning ability of the network, especially on the more challenging NH-Haze dataset. In particular, the setting

Table 7
Quantitative evaluations of different multi-scale loss weight settings on the synthetic and real-world datasets.

Group	Loss Weight [L_0, L_1, L_2, L_3]	Haze4K PSNR	Haze4K SSIM	NH-Haze PSNR	NH-Haze SSIM
A (baseline)	[1,0,0,0]	21.71	0.9017	11.76	0.4979
B (ours)	**[1,0.5,0.25,0.125]**	21.42	**0.9081**	**12.30**	**0.4988**
C	[1,0.2,0.04,0.008]	21.35	0.9011	11.48	0.4915
D	[1,0.1,0.001,0.0001]	**21.73**	0.9043	11.72	0.4923

achieves the best performance on NH-Haze, yielding 12.30 dB PSNR and 0.4988 SSIM, indicating that a moderately decaying supervision scheme is beneficial for improving restoration quality in real-world hazy scenes.

Compared with the single-scale setting, introducing auxiliary supervision provides richer gradient signals at intermediate decoding stages, facilitating optimization and encouraging the network to learn coarse-to-fine restoration behaviour. However, the results also suggest that the relative weighting among scales can be carefully controlled. When the auxiliary losses decay too slowly or too strongly, the resulting supervision may either over-constrain intermediate outputs or become too weak to provide meaningful guidance. For example, the setting [1,0.2,0.04,0.008] shows weaker performance on NH-Haze, suggesting that insufficient intermediate supervision limits its contribution to real-world performance. On the other hand, the setting [1,0.1,0.001,0.0001] achieves the best Haze4K PSNR, but does not perform as well as [1,0.5,0.25,0.125] on NH-Haze, indicating that overly small auxiliary weights may bias the optimization toward the final high-resolution output alone.

Overall, these results show that multi-scale supervision is beneficial, but its effectiveness depends strongly on the weighting strategy. Considering the better overall balance between synthetic and real-world performance, we adopt $[\lambda_0, \lambda_1, \lambda_2, \lambda_3]=[1,0.5,0.25,0.125]$ as the default loss weighting scheme in the subsequent experiments.

4.8. Ablation on the Hybrid Attention Design

To evaluate the effectiveness of the proposed hybrid spatial–channel attention design, we conduct an ablation study by removing either the spatial attention *SA* or the channel attention *CA* branch from the HSCBlock. All models are trained under the same setting without data augmentation, and the loss weights are fixed to [1, 0, 0, 0] to eliminate the influence of multi-scale supervision. The comparison includes three variants: *SA*+*CA* full model, *SA* only, and *CA* only.

The quantitative results are reported in Table 8. As can be observed, the full model SA+CA achieves the best overall performance on the synthetic Haze4K dataset, with 21.71 dB PSNR and 0.9017 SSIM, outperforming both the SA-only (21.45 dB / 0.8861) and CA-only (20.22 dB / 0.8708) variants by clear margins. On the real-world NH-Haze dataset, SA+CA also achieves a competitive 11.76 dB PSNR and the highest SSIM of 0.4979, indicating that the combination of spatial and channel attention provides a more balanced feature representation for real-world hazy scenes.

When using only the spatial attention branch, the model achieves 21.45 dB PSNR on Haze4K, which is close to the full model, suggesting that spatial attention is highly effective at capturing local structural dependencies. However, its SSIM drops noticeably to 0.8861, and its NH-Haze performance (11.68 dB / 0.4971) also falls behind the full model. The CA-only variant achieves the highest NH-Haze PSNR of 11.78 dB, but its Haze4K performance degrades significantly (20.22 dB / 0.8708), indicating that channel attention alone lacks sufficient spatial modelling capability for synthetic haze removal. It is because channel attention primarily focuses on recalibrating features across channels, without explicitly modelling spatial relationships, which are essential for handling spatially varying haze.

Overall, these results demonstrate that spatial attention and channel attention play complementary roles in image dehazing. Spatial attention focuses on modelling structural dependencies across spatial locations, while channel attention enhances feature discrimination across channels. By integrating both mechanisms, the proposed HSCBlock enables UTNet to achieve a better balance between local detail restoration and global feature selection, thereby improving performance on both synthetic and real-world datasets.

Table 8

Quantitative evaluations of different attention configurations on the synthetic and real-world datasets.

Model	Haze4K PSNR	Haze4K SSIM	NH-Haze PSNR	NH-Haze SSIM
SA + CA (ours)	**21.71**	**0.9017**	11.76	**0.4979**
SA only	21.45	0.8861	11.68	0.4971
CA only	20.22	0.8708	**11.78**	0.4954

5. Conclusion

In this paper, we proposed UTNet, a lightweight U-shaped Transformer for single-image dehazing. UTNet combines hierarchical encoder–decoder reconstruction with local-window spatial attention and channel-wise feature recalibration. The hierarchical pathway progressively expands contextual coverage, while skip connections and multi-scale supervision support detail reconstruction and stable optimization. Gamma augmentation further improves robustness to illumination variation. Experiments on Haze4K and NH-Haze show that UTNet provides competitive restoration quality with a compact computational footprint. The model contains 1.34M parameters, requires 0.32G FLOPs for a 64×64 input, and achieves 108.93 FPS with a latency of 9.18 ms on the laptop GPU used in our evaluation. Cross-resolution experiments also demonstrate stable performance from 256×256 to 512×512 inputs.

The results further show that performance on real non-homogeneous haze is strongly affected by the synthetic-to-real distribution gap. Introducing real NH-Haze training samples substantially improves performance on strictly held-out NH-Haze images, although a trade-off with synthetic-domain accuracy is observed. Difficult cases remain concentrated in regions with dense localized haze, weak structural cues, and complex illumination. Future work will therefore investigate broader real-domain training data, lightweight cross-window interaction, and physical or colour-consistency constraints. Validation on dedicated edge hardware will also be conducted to further assess practical deployment.

Acknowledgement

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Ancuti, C. O., Ancuti, C., Timofte, R. NH-HAZE: An Image Dehazing Benchmark with Non-Homogeneous Hazy and Haze-Free Images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, 1798-1805. Available: https://openaccess.thecvf.com/content_CVPRW_2020/html/w31/Ancuti_NH-HAZE_An_Image_Dehazing_Benchmark_With_Non-Homogeneous_Hazy_and_Haze-Free_CVPRW_2020_paper.html. https://doi.org/10.1109/CVPRW50498.2020.00230
2. Berman, D., Treibitz, T., Avidan, S. Non-Local Image Dehazing. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 1674-1682. https://doi.org/10.1109/CVPR.2016.179
3. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D. DehazeNet: An End-to-End System for Single Image Haze Removal. IEEE Transactions on Image Processing, 2016, 25(11), 5187-5198. https://doi.org/10.1109/TIP.2016.2598681
4. Chen, D., He, M., Fan, Q., Liao, J., Zhang, L., Hou, D., Yuan, L., Hua, G. Gated Context Aggregation Network for Image Dehazing and Deraining. Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2019, 1375-1383. https://doi.org/10.1109/WACV.2019.00151
5. Chen, J., Yan, X., Xu, Q., Li, K. Tokenize Image Patches: Global Context Fusion for Effective Haze Removal in Large Images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, 2258-2268. Available: https://openaccess.thecvf.com/content/CVPR2025/html/Chen_Tokenize_Image_Patches_Global_Context_Fusion_for_Effective_Haze_Removal_CVPR_2025_paper.html. https://doi.org/10.1109/CVPR52734.2025.00216
6. Chen, Z., Wang, Y., Yang, Y., Liu, D. PSD: Principled Synthetic-to-Real Dehazing Guided by Physical Priors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021,

7180-7189. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Chen_PSD_Principled_Synthetic-to-Real_Dehazing_Guided_by_Physical_Priors_CVPR_2021_paper.html. https://doi.org/10.1109/CVPR46437.2021.00710

7. Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., Yang, M.-H. Multi-Scale Boosted Dehazing Network with Dense Feature Fusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 2154-2164. https://doi.org/10.1109/CVPR42600.2020.00223

8. Dong, W., Zhou, H., Wang, R., Liu, X., Zhai, G., Chen, J. DehazeDCT: Towards Effective Non-Homogeneous Dehazing via Deformable Convolutional Transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2024, 6405-6414. https://doi.org/10.1109/CVPRW63382.2024.00642

9. Fattal, R. Single Image Dehazing. ACM Transactions on Graphics, 2008, 27(3), 72:1-72:9. https://doi.org/10.1145/1399504.1360671

10. Guo, C.-L., Yan, Q., Anwar, S., Cong, R., Ren, W., Li, C. Image Dehazing Transformer with Transmission-Aware 3D Position Embedding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 5802-5810. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Guo_Image_Dehazing_Transformer_With_Transmission-Aware_3D_Position_Embedding_CVPR_2022_paper.html. https://doi.org/10.1109/CVPR52688.2022.00572

11. He, K., Sun, J., Tang, X. Single Image Haze Removal Using Dark Channel Prior. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, 1956-1963. https://doi.org/10.1109/CVPR.2009.5206515

12. Hu, J., Shen, L., Sun, G. Squeeze-and-Excitation Networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

13. Kingma, D. P., Ba, J. Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations, 2015. DOI: 10.48550/arXiv.1412.6980.

14. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, 1998, 86(11), 2278-2324. https://doi.org/10.1109/5.726791

15. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D. AOD-Net: All-In-One Dehazing Network. Proceedings of the IEEE International Conference on Computer Vision, 2017, 4770-4778. https://doi.org/10.1109/ICCV.2017.511

16. Liu, X., Ma, Y., Shi, Z., Chen, J. GridDehazeNet: Attention-Based Multi-Scale Network for Image Dehazing. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, 7314-7323. https://doi.org/10.1109/ICCV.2019.00741

17. Liu, Y., Zhu, L., Pei, S., Fu, H., Qin, J., Zhang, Q., Wan, L., Feng, W. From Synthetic to Real: Image Dehazing Collaborating with Unlabeled Real Data. Proceedings of the 29th ACM International Conference on Multimedia, 2021, 50-58. https://doi.org/10.1145/3474085.3475331

18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

19. McCartney, E. J. Optics of the Atmosphere: Scattering by Molecules and Particles. Wiley, New York, 1976. Available: https://openlibrary.org/books/OL4882198M/Optics_of_the_atmosphere.

20. Meng, G., Wang, Y., Duan, J., Xiang, S., Pan, C. Efficient Image Dehazing with Boundary Constraint and Contextual Regularization. Proceedings of the IEEE International Conference on Computer Vision, 2013, 617-624. https://doi.org/10.1109/ICCV.2013.82

21. Nair, V., Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. Proceedings of the 27th International Conference on Machine Learning, 2010, 807-814. DOI: 10.5555/3104322.3104425.

22. Narasimhan, S. G., Nayar, S. K. Vision and the Atmosphere. International Journal of Com-

puter Vision, 2002, 48(3), 233-254. https://doi.org/10.1023/A:1016328200723

23. Pushpalatha, D., Prithvi, P. LDFormer: Lightweight Dehazing Transformer. Neural Computing and Applications, 2025, 37(24), 20049-20068. https://doi.org/10.1007/s00521-025-11440-x
24. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H. FFA-Net: Feature Fusion Attention Network for Single Image Dehazing. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7), 11908-11915. https://doi.org/10.1609/aaai.v34i07.6865
25. Qiu, Y., Zhang, K., Wang, C., Luo, W., Li, H., Jin, Z. MB-TaylorFormer: Multi-Branch Efficient Transformer Expanded by Taylor Formula for Image Dehazing. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, 12756-12767. Available: https://arxiv.org/abs/2308.14036. https://doi.org/10.1109/ICCV51070.2023.01176
26. Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., Yang, M.-H. Single Image Dehazing via Multi-Scale Convolutional Neural Networks. Computer Vision - ECCV 2016, Lecture Notes in Computer Science, 2016, 9906, 154-169. https://doi.org/10.1007/978-3-319-46475-6_10
27. Song, Y., He, Z., Qian, H., Du, X. Vision Transformers for Single Image Dehazing. IEEE Transactions on Image Processing, 2023, 32, 1927-1941. https://doi.org/10.1109/TIP.2023.3256763
28. Tan, R. T. Visibility in Bad Weather from a Single Image. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008, 1-8. https://doi.org/10.1109/CVPR.2008.4587643
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017, 30, 5998-6008. Available: https://proceedings.neurips.cc/paper/7181-attention-is-all-you-need.
30. Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., Ma, L. Contrastive Learning for Compact Single Image Dehazing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 10551-10560. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Wu_Contrastive_Learning_for_Compact_Single_Image_Dehazing_CVPR_2021_paper.html. https://doi.org/10.1109/CVPR46437.2021.01041
31. Yang, Y., Wang, C., Liu, R., Zhang, L., Guo, X., Tao, D. Self-Augmented Unpaired Image Dehazing via Density and Depth Decomposition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 2027-2036. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Yang_Self-Augmented_Unpaired_Image_Dehazing_via_Density_and_Depth_Decomposition_CVPR_2022_paper.html. https://doi.org/10.1109/CVPR52688.2022.00208
32. Zhang, H., Patel, V. M. Densely Connected Pyramid Dehazing Network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 3194-3203. https://doi.org/10.1109/CVPR.2018.00337
33. Zheng, Y., Zhan, J., He, S., Dong, J., Du, Y. Curricular Contrastive Regularization for Physics-Aware Single Image Dehazing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 5785-5794. https://doi.org/10.1109/CVPR52729.2023.00560

This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).