

Article

O-Transformer-Mamba: An O-Shaped Transformer-Mamba Framework for Remote Sensing Image Haze Removal

Xin Guan ¹, Runxu He ¹, Le Wang ², Hao Zhou ², Yun Liu ³ and Hailing Xiong ^{4,*}

¹ College of Computer and Information Science, Southwest University, Chongqing 400715, China; guanxin2020@email.swu.edu.cn (X.G.); hrx221226@email.swu.edu.cn (R.H.)

² School of Computer Science and Technology, Anhui University of Technology, Ma'anshan 243032, China; wangle@ahut.edu.cn (L.W.); haozhou@ahut.edu.cn (H.Z.)

³ College of Artificial Intelligence, Southwest University, Chongqing 400715, China; yunliu@swu.edu.cn

⁴ College of Electronic and Information Engineering, Southwest University, Chongqing 400715, China

* Correspondence: xionghl@swu.edu.cn

Highlights

What are the main findings?

- We propose an enhanced visual state space model and an adaptive self-attention mechanism that jointly improve global-local feature modeling and selective feature enhancement.
- A novel framework for dehazing remote sensing images.

What are the implications of the main findings?

- Enhancing global-local feature interaction and selective feature refinement is essential for achieving robust remote sensing image restoration.
- On both real and synthetic remote sensing datasets, our method achieves state-of-the-art performance.

Abstract

Although Transformer-based and state-space models (e.g., Mamba) have demonstrated impressive performance in image restoration, they remain deficient in remote sensing image dehazing. Transformer-based models tend to distribute attention evenly, making them difficult to handle the uneven distribution of haze. While Mamba excels at modeling long-range dependencies, it lacks fine-grained spatial awareness of complex atmospheric scattering. To overcome these limitations, we present a new O-shaped dehazing architecture that combines a Sparse-Enhanced Self-Attention (SE-SA) module with a Mixed Visual State Space Model (Mix-VSSM), balancing haze-sensitive details in remote sensing images with long-range context modeling. The SE-SA module introduces a dynamic soft masking mechanism that adaptively adjusts attention weights based on the local haze distribution, enabling the network to more effectively focus on severely degraded regions while suppressing redundant responses. Furthermore, the Mix-VSSM enhances global context modeling by combining sequential processing of 2D perception with local residual information. This design mitigates the loss of spatial detail in the standard VSSM and improves the feature representation of haze-degraded remote sensing images. Thorough experiments verify that our O-shaped framework outperforms existing methods on several benchmark datasets.

Keywords: image dehazing; remote sensing; transformer; state space model; deep learning



Academic Editors: Salah Bourennane, Ashraf Saleem and Sidike Paheding

Received: 31 October 2025

Revised: 3 January 2026

Accepted: 4 January 2026

Published: 6 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Advances in remote sensing technology [1,2] have made satellite imagery an essential data source for applications such as agriculture, environmental monitoring, and urban planning. Nevertheless, atmospheric particles and gases near the Earth's surface scatter and absorb light during image acquisition, leading to haze-related degradations, including blurred textures, reduced contrast, detail loss, and color distortion. These degradations significantly undermine the reliability of downstream tasks such as object detection, surveillance, and land-cover mapping. Consequently, satellite image dehazing has become a crucial preprocessing step for restoring visual quality and ensuring robust performance in practical remote sensing applications.

Image dehazing [3–7] is essentially an underdetermined inverse problem that requires the simultaneous estimation of multiple parameters. Traditional physics-based approaches [8] can partially mitigate the effects of haze, but often yield unstable results when confronted with complex terrains and non-uniform atmospheric conditions. Earlier image dehazing techniques, which include methods based on filtering, Retinex theory, and histogram equalization, mainly aim to enhance contrast but often overlook the underlying physical causes of degradation, leading to over-enhancement. Model-based approaches rely on simplified atmospheric transmission models and focus on estimating the transmission map and atmospheric light to better align with the underlying degradation mechanisms. Wang et al. [9] proposed a haze removal strategy for individual images, integrating a physical degradation model with the MSRCR method to improve image clarity and color quality. He et al. [10] introduced a dehazing approach for individual images based on the Dark Channel Prior, effectively recovering haze-free scenes by estimating transmission and atmospheric light. Zhu et al. [11] presented an efficient single-image dehazing framework based on the color attenuation prior, enabling accurate depth estimation and clear image restoration.

Initially, convolutional neural networks (CNNs) [12] became the dominant choice. By stacking convolutional layers, CNNs progressively extract local features from images, effectively enhancing the perception of edges and textures. Cai et al. [13] developed DehazeNet, a CNN-based approach that learns to estimate the transmission map from hazy inputs, enabling image restoration through the atmospheric scattering principle. Another line of approaches employs end-to-end learning to generate haze-free results without estimating intermediate physical parameters. AOD-Net [14] simplifies the atmospheric scattering model by combining the transmission map and atmospheric light into a unified parameter, enabling direct optimization through reconstruction loss. GridDehazeNet [15] utilizes residual learning to enhance clean image prediction. Following this, FCTF-Net [16] builds upon residual structures by introducing cross-level fusion and enhanced feature transmission, further improving detail preservation and dehazing quality. FFA-Net [17], based on a feature fusion design, incorporates adaptive attention mechanisms to effectively integrate multi-scale features and enhance overall image clarity. However, due to the inherent constraints of convolutional operations, CNNs are challenged in capturing long-distance dependencies, even when employing dilated convolutions or U-Net architectures. This limitation becomes especially pronounced in remote sensing scenarios, where large-scale haze regions require effective global context modeling.

With the rise of self-attention mechanisms, Transformer-based models have been introduced into single-image dehazing to enhance global context modeling. For instance, DehazeFormer [18] leverages a multi-scale transformer backbone with specialized dehazing-oriented designs to achieve superior performance, especially in handling non-uniform haze. To overcome the locality constraints of CNNs, Transformers [19–21] have been introduced into image restoration tasks and quickly gained prominence.

Vision Transformers (ViT) [22], based on a global self-attention mechanism, can directly model dependencies between any two positions in an image, significantly improving long-range modeling capability. In dehazing tasks, this allows for better contextual integration. However, the standard self-attention mechanism treats all positions equally, making it difficult to distinguish the sparse and regionally concentrated nature of haze. As a result, Transformer-based methods may still suffer from missing details or blurred edges in the output.

In recent years, numerous advanced deep learning techniques [23–27] have been applied to image dehazing, primarily categorized into CNN-based methods and Transformer-based methods. Thanks to their powerful feature representation capabilities, learning-driven dehazing algorithms have demonstrated superior performance compared to traditional approaches. Continuing this trend, MixDehazeNet [28] introduces a Mix Structure Block that combines convolutional and attention-based pathways to balance local detail restoration and global context modeling, achieving improved dehazing performance. The Omni-Kernel Network [29] further explores adaptive receptive fields by integrating multiple convolution kernels of varying sizes, allowing the network to flexibly capture diverse degradation patterns in hazy images. AdaIR [30] proposes a unified image restoration framework that adaptively mines frequency information and modulates restoration strategies based on input content, enabling robust performance across different haze levels and scene types.

More recently, the state space models (SSMs) [31] have attracted attention as a sequence modeling framework with linear computational complexity. State Space Models (SSMs) are recurrent architectures that capture long-range dependencies via dynamic state transitions. Evolving from classical linear systems, recent developments like S4 introduce HiPPO-based mechanisms for efficient memory retention, while Mamba (S6) [32] enhances this by enabling dynamic parameterization, making SSMs effective for long-sequence modeling and context-aware reasoning. Mamba improves traditional SSMs by using a selective mechanism and a hardware-optimized cyclic scanning method, boosting efficiency while modeling long-range dependencies. This enables strong performance in recent low-level vision tasks. Vmamba [33] was the first to apply the Mamba architecture to visual tasks, demonstrating performance comparable to Transformers on recognition and scene understanding benchmarks. Building on this, MambaIR [34] adapted Mamba for tasks such as enhancing image resolution and removing noise, achieving notable results. Despite their strong sequential modeling capabilities, Mamba's design primarily emphasizes one-dimensional channel-wise processing and lacks fine-grained spatial modeling. This structural limitation reduces its adaptability in remote sensing image dehazing, where complex atmospheric degradation and uneven haze distribution demand both accurate local detail recovery and robust global estimation.

To address the limitations of CNNs, Transformers, and state space models (SSMs) in remote sensing image dehazing, we present an O-shaped network architecture capable of capturing global contextual information while maintaining local detail fidelity. Unlike U-shaped architectures [35] that emphasize hierarchical encoder–decoder feature reuse, or parallel designs [36] that fuse Transformer and Mamba features in a single forward pass, the proposed O-shaped topology enables two complementary branches to perform independent and complete feature reconstruction before forming a closed-loop interaction. This design allows the Transformer branch to focus on explicit global attention modeling, while the Mamba-based branch emphasizes efficient long-range state propagation with preserved local continuity. Motivated by the spatially sparse and regionally clustered nature of haze in remote sensing imagery, as well as the need for both long-range dependency modeling and precise local perception, we design two key components. First, a sparsely enhanced

self-attention (SE-SA) mechanism is introduced to dynamically emphasize haze-affected regions and enhance structural detail recovery. Unlike traditional attention, which treats all positions equally, this module selectively focuses on degraded areas, improving both the selectivity and accuracy of the dehazing process. Second, to address the spatial modeling limitations of existing SSMs like Mamba, we develop a mixed visual state space model (Mix-VSSM) that fuses the local receptive field of convolutions with the long-range modeling capability of state space representations. Integrated into an O-shaped topology, this module facilitates multi-scale and cross-level feature interactions, equipping the network with stronger robustness against complex, large-area, and non-uniform haze distributions.

The contributions of this study can be summarized as follows:

- We propose a novel O-shaped topology that fosters comprehensive multi-scale and cross-level feature integration, thereby enhancing the network's capability to address large-scale and spatially uneven haze distributions in remote sensing imagery.
- We design a sparsely enhanced self-attention (SE-SA) mechanism that adaptively focuses on degraded regions, effectively recovering structural elements, including edges and surface patterns.
- A mixed visual state space model (Mix-VSSM) combining convolution and state space modeling to balance local spatial awareness and global sequence modeling.
- Our proposed network demonstrates superior effectiveness on both real-world and synthetic datasets, handling diverse scales and haze density levels with remarkable robustness.

2. Materials and Methods

2.1. Motivation

The haze distribution in remote sensing images is inherently non-uniform, making fixed sparse patterns a suboptimal choice. Unlike conventional dehazing scenarios with predominantly horizontal viewpoints, remote sensing images are captured from a near-vertical perspective, where atmospheric scattering effects exhibit large-scale spatial continuity and strong adjacency influence across wide regions. The dual-branch Transformer–Mamba architecture is motivated by its complementary strengths: the Transformer branch captures global spatial dependencies, while the Mamba branch efficiently models long-range contextual information. To further enhance the efficiency and adaptability of the attention mechanism, we design a Sparse-Enhanced Self-Attention (SE-SA) module. Unlike existing sparse attention mechanisms [18,20] that rely on fixed local windows or channel reduction, SE-SA adaptively adjusts the attention distribution to better handle spatially varying haze. Existing hybrid approaches [37,38] merely combine the two paradigms without establishing effective cross-hierarchical interaction. To overcome these limitations, we propose an O-shaped framework enhanced by the SE-SA and an enhanced Visual State Space Module (Mix-VSSM), where late fusion allows each branch to learn independently without early interference, and their integration at the final stage enables effective cross-level feature interaction, jointly improving texture restoration and haze adaptability.

2.2. Overview of Proposed O-Transformer-Mamba Architecture

Figure 1 depicts the O-Transformer-Mamba framework, which employs a dual-branch O-shaped configuration with three hierarchical levels in each branch, forming a balanced multi-stage encoder-decoder architecture.

The first branch is the Transformer branch, built upon stacked Transformer Blocks to model global contextual information. This branch is capable of capturing long-range dependencies, which helps to restore structural information severely degraded by widespread haze. Each Transformer Block consists of Layer Normalization, Sparse-Enhanced

Self-Attention (SE-SA), and residual connections. The encoded features are gradually upsampled to recover spatial resolution and are fused with corresponding decoder-stage features through skip connections to enhance detail representation.

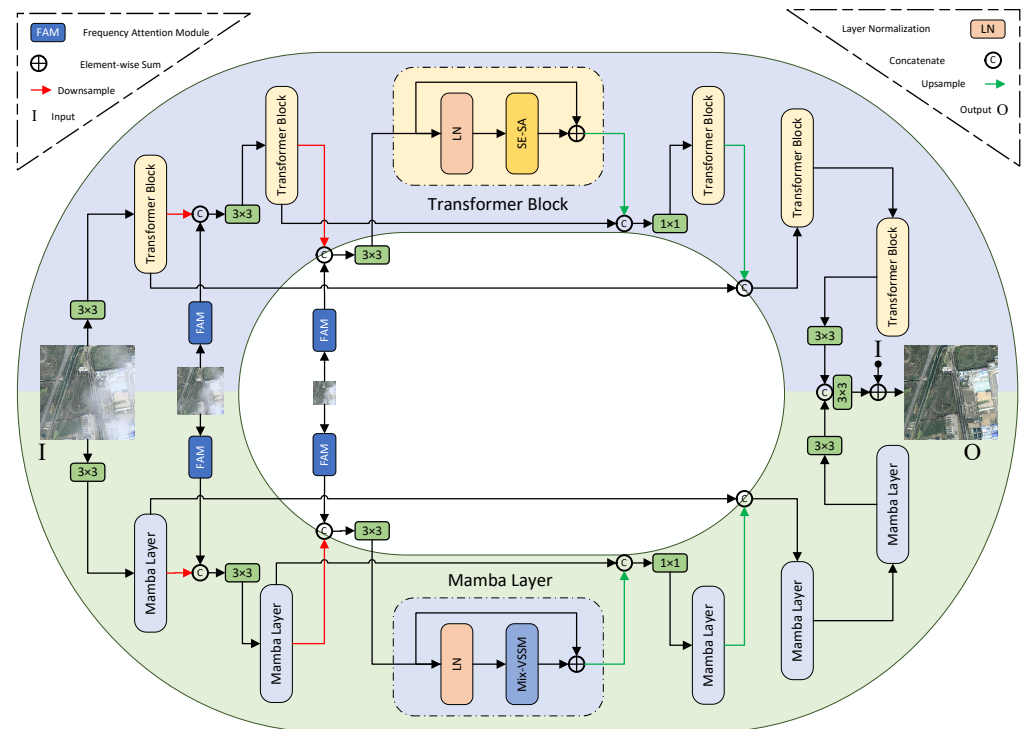


Figure 1. The architecture of our O-Transformer-Mamba. The overall architecture adopts an O-shaped design, consisting of two branches: one based on Transformer and the other on Mamba.

The second branch is the Mamba branch, centered on State Space Models (SSMs), and employs stacked Mamba Layers for sequential modeling. It serves to compensate for the missing information from the first branch. While the overall structure mirrors that of the Transformer branch, it differs by embedding a Mixed Visual State Space Model (Mix-VSSM) within the Mamba Layers.

In the encoder stages of both branches, Frequency Attention Modules (FAMs) of varying scales are embedded to guide the model's focus toward frequency-domain features, thereby enhancing sensitivity to edges and textures. During the decoding phase, both branches reconstruct features independently and form a circular information flow through the O-shaped topology, enabling cross-branch contextual interaction. The network integrates multi-scale processing, dual-path encoding, and frequency-aware mechanisms, collectively enhancing dehazing performance while maintaining efficient modeling capacity.

2.3. Sparse-Enhanced Self-Attention

The self-attention operation in the SE-SA module, acting on the channel dimension, is shown in Figure 2a. An input feature map $X \in \mathbb{R}^{H \times W \times C}$ is processed with a point-wise convolution and a subsequent 3×3 depth-wise convolution to encode inter-channel dependencies, resulting in the query (Q), key (K), and value (V) representations. These are then reshaped and used to compute a dense self-attention matrix $A \in \mathbb{R}^{C \times C}$ via dot-product operation.

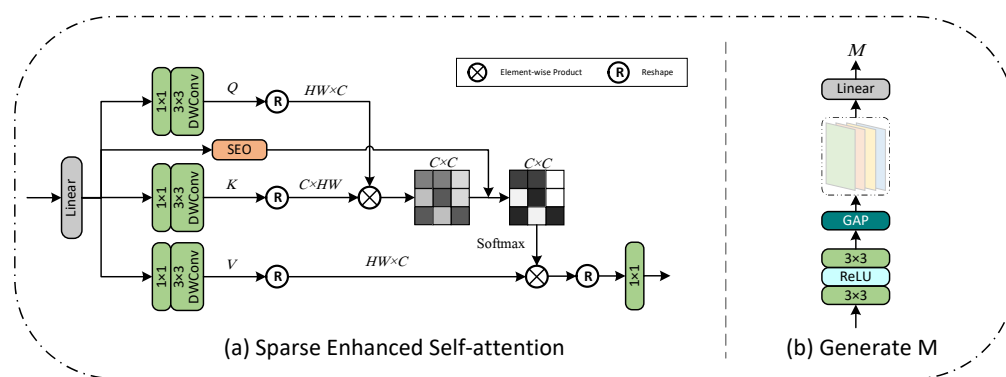


Figure 2. The architecture of our Sparse-Enhanced Self-Attention. In the figure, (a) represents the overall architecture, while (b) illustrates the generation of M within the SEO module.

However, we observe that not all tokens in the key matrix are strongly correlated with those in the query, and attending to irrelevant tokens may introduce noise and redundancy into the self-attention map, ultimately degrading the clarity of the reconstructed images. In order to address this challenge, we introduce a masking mechanism M that selectively filters the self-attention scores. Unlike standard self-attention, this mechanism amplifies important responses and suppresses uninformative ones, thereby improving self-attention focus during the dehazing process. Specifically, when the self-attention score exceeds a learned threshold, it is enhanced; otherwise, it is attenuated. As shown in Figure 3, the difference between the two types of attention is intuitively demonstrated.



Figure 3. The difference between traditional Self-Attention and Sparse-Enhanced Self-Attention. The Sparse-Enhanced Self-Attention module applies localized enhancement and suppression to the traditional Self-Attention mechanism.

In addition, we introduce a lightweight learnable module called the Sparsity Enhancement Operator (SEO), which has FLOPs and parameters of 1.28 G and 0.02 M, respectively. Specifically, SEO generates dynamic guidance signals directly from the input features, which encode contrast attenuation, frequency suppression, and spatial inconsistency caused by haze. These signals modulate the self-attention weights by enhancing informative responses and suppressing noisy or redundant ones. As shown in Figure 2b, the generation process of M in SEO is as follows:

$$\hat{X} = \text{Conv}^{3 \times 3}(\text{ReLU}(\text{Conv}^{3 \times 3}(X))) \quad (1)$$

$$M = \text{Linear}(\text{GAP}(\hat{X})) \quad (2)$$

where $GAP(\cdot)$ stands for global average pooling, $Linear(\cdot)$ represents Linear layer, and $Conv^{3 \times 3}(\cdot)$ denotes 3×3 convolution. Formally, the SE-SA process is expressed as:

$$SE - SA = Softmax(SE(\frac{QK^T}{\sqrt{d}}))V \quad (3)$$

here, $SE(\cdot)$ denotes the sparsity enhancement operation guided by prompt modulation, and $\sqrt{d}$ represents the number of attention heads. Similar to most existing methods, we adopt a multi-head attention mechanism: the outputs of multiple attention heads are concatenated along the channel dimension and then passed through a linear projection layer to produce the final attention output.

SEO produces a dynamic modulation mask M in a data-driven manner. This design avoids rigid structural constraints and enables the attention distribution to adapt flexibly to local feature variations and spatial complexity. Through adaptive scaling, informative and discriminative features are emphasized, while redundant or less relevant responses are suppressed, leading to a more efficient and balanced attention allocation. The detailed procedure of the SEO-guided adjustment is summarized in Algorithm 1.

Algorithm 1: Sparse Enhancement Operation

Input: Input feature map $X \in \mathbb{R}^{H \times W \times C}$

Output: Attention-enhanced feature map $Y \in \mathbb{R}^{C \times C}$

Step 1: Reducing Channels:

$$\hat{X} \leftarrow Conv^{3 \times 3}(ReLU(Conv^{3 \times 3}(X)));$$

Step 2: Generate M :

$$M \leftarrow Linear(GAP(\hat{X}));$$

Step 3: Calculating the gap between the attention score and M :

$$A' \leftarrow A - M;$$

Step 4: Gaps Enlargement and Smoothing:

$$\hat{A} \leftarrow Sigmoid(A' * scale);$$

// scale is the adjustable scaling factor

Step 5: Compute attention output:

$$Y \leftarrow \hat{A} * A;$$

return Y

2.4. Mixed Visual State Space Model

The standard Mamba framework was first developed for modeling sequential data in a one-dimensional format, but its direct application to 2D images presents notable challenges due to the rich spatial structures and contextual information inherent in visual data. Flattening the 2D feature representation into a linear sequence can separate spatial neighbors, compromising local consistency and causing what is termed “local pixel forgetting.” This issue is particularly pronounced in remote sensing image dehazing, where the central pixel is highly influenced by the spatial arrangement and intensity values of its surrounding pixels, a phenomenon known as the adjacency effect [39]. As such, the strategy used to convert 2D features into sequential representations plays a critical role in determining dehazing performance.

To deal with the challenges outlined above, we integrate the Vision State Space Model (VSSM) into the dehazing framework. Specifically, Mix-VSSM converts two-dimensional feature maps into sequential representations through a structured scan that preserves spatial ordering, enabling effective long-range dependency modeling. To better preserve local continuity and spatial details, an additional convolutional residual branch is introduced to capture contextual information around each anchor pixel, after which the outputs are

reshaped back to the two-dimensional feature space. The architecture of the proposed Mix-VSSM is illustrated in Figure 4a.

Building upon prior work [33], Mix-VSSM processes the input feature $X \in \mathbb{R}^{H \times W \times C}$ through three parallel branches. In the first branch, a linear projection expands the channel dimension to λC , where λ is a predefined expansion factor. This is followed by depthwise convolution, a SiLU activation, and a 2D Selective Scanning Module (2D-SSM). In the second branch, the feature channels are similarly expanded to λC via a linear layer, followed by depthwise convolution, SiLU activation, and a Residual Convolutional Group (ResGroup). The outputs of these two branches are concatenated and fused via a linear layer, followed by Layer Normalization. Meanwhile, the third branch also expands the channels to λC using a linear layer, followed by SiLU activation. The resulting features are then combined with the fused features from the first two branches via element-wise (Hadamard) multiplication. Finally, a linear projection reduces the channel dimension back to C , yielding the output feature X_{out} with the same spatial size as the input. X_{out} can be written as:

$$x_1 = \text{SiLU}(\text{DWConv}(\text{Linear}(X))) \quad (4)$$

$$x_2 = \text{SiLU}(\text{DWConv}(\text{Linear}(X))) \quad (5)$$

$$X_1 = 2D-SSM(x_1) \quad (6)$$

$$X_2 = \text{ResGroup}(x_2) \quad (7)$$

$$X_c = \text{LN}(\text{Linear}(\text{Cat}(X_1, X_2))) \quad (8)$$

$$X_3 = \text{SiLU}(\text{Linear}(X)) \quad (9)$$

$$X_{out} = \text{Linear}(X_3 * X_c) \quad (10)$$

where $\text{DWConv}(\cdot)$ denotes depth-wise convolution, $\text{Linear}(\cdot)$ represents Linear layer, $\text{Cat}(\cdot)$ denotes Concatenate, and $*$ stands for the Hadamard product.

The proposed 2D Selective Scanning Module (2D-SSM) extends the capabilities of Mamba, which was originally optimized to process sequential streams causally, making it suitable for one-dimensional sequence applications such as NLP. However, the architecture shows limitations when processing visual data without inherent sequential order, including images. To enhance the modeling of spatial relationships in two dimensions, we employ the 2D-SSM.

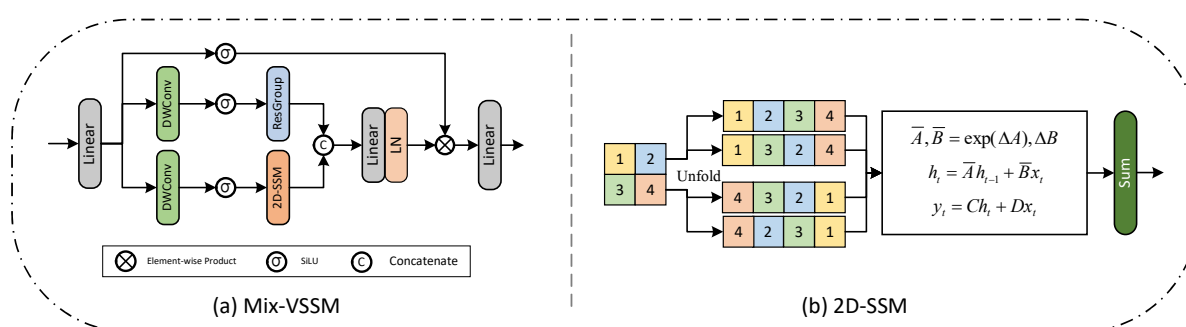


Figure 4. The architecture of our Mixed Visual State Space Model. In the figure, (a) represents the overall architecture, while (b) illustrates the internal mechanism of the 2D-SSM.

Figure 4b demonstrates how the 2D-SSM converts two-dimensional feature maps into a series of linear sequences by scanning along four diagonal orientations: top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right. Each directional sequence is then processed using a discrete state-space formulation to capture long-range dependencies. After processing, the outputs from all directions are aggregated and reshaped to reconstruct the original 2D spatial structure. This approach allows the

model to better utilize spatial context across multiple orientations, thereby improving its capacity for image-based restoration tasks.

2.5. Loss Function

The L_1 loss is adopted as the main optimization target because of its effectiveness in preserving structural fidelity and improving PSNR and SSIM. Additionally, to refine frequency representation, a frequency-domain constraint is incorporated by performing the Fourier Transform (FT) on both predicted and reference images and computing their L_1 distance in the complex domain.

$$Loss = \alpha \cdot \left[\|J_{real} - GT_{real}\|_1 + \|J_{imag} - GT_{imag}\|_1 \right] + \|J - GT\|_1 \quad (11)$$

here, J denotes the dehazed remote sensing image produced by O-Transformer-Mamba, while GT is its corresponding ground truth. J_{real} and GT_{real} represent the real components of the generated and ground truth images in the frequency domain, respectively, and J_{imag} and GT_{imag} refer to their imaginary components. α set at 0.1, it modulates the significance of the frequency loss component.

3. Results

3.1. Dataset

The effectiveness of the proposed O-Transformer-Mamba is evaluated on the synthetic remote sensing datasets SateHaze1k [40] and HRSD [41], along with the real-world RRSHID [42] and UAV-based HazyDet [43] datasets, to comprehensively evaluate its dehazing capability across diverse scenarios and complex degradation conditions.

The SateHaze1k dataset is intended for remote sensing image dehazing and includes images with varying haze intensity. It is divided into three subsets corresponding to light, medium, and dense haze levels. Each subset contains a training set of 320 images and a test set of 45 images. Light haze images are created based on masks obtained from authentic cloud formations, medium haze combines features typical of mist and moderate haze, and dense haze is synthesized via transmittance maps to emulate severe atmospheric conditions. This hierarchical arrangement enables a thorough evaluation of dehazing methods across diverse haze scenarios.

HRSD comprises two subsets: LHID and DHID. LHID contains 30,517 images allocated for training and 500 for evaluation, generated using an atmospheric scattering model to capture varying haze intensities and enhance model generalization. DHID, in contrast, consists of 14,990 images synthesized with real haze patterns for higher realism, with 14,490 assigned to training and 500 to testing. By combining both simulated and realistic haze distributions, HRSD provides a diverse benchmark for thoroughly assessing the performance of O-Transformer-Mamba across different dehazing scenarios.

RRSHID is a pioneering large-scale real-world dataset for remote sensing image dehazing, consisting of 3053 pairs of hazy and corresponding haze-free images. The dataset was collected from various regions in China, including urban areas, agricultural zones, and coastal landscapes, under diverse lighting and seasonal conditions, thus capturing rich geographic and atmospheric variations. Unlike datasets produced through oversimplified atmospheric simulation processes, RRSHID was constructed jointly with meteorological agencies, ensuring that haze distribution and sensor-induced spectral distortions are authentically represented. This makes RRSHID a reliable benchmark for evaluating dehazing algorithms and a valuable resource for studying model generalization in real-world remote sensing scenarios.

To evaluate the real-world applicability and robustness of our model, we utilize the recently released HazyDet dataset, which includes drone-captured images degraded by

haze. It provides 8000 training images, 1000 validation images, and 2000 test images, combining naturally hazy scenes with synthetic samples generated via the Atmospheric Scattering Model (ASM).

3.2. Implementation Details

Experiments were performed on a system equipped with two NVIDIA RTX 4090 GPUs. To expand the training set, images were randomly rotated by 90°, 180°, or 270° and subjected to horizontal flipping. During training, the initial learning rate was set to 0.0003, with a batch size of 8 and a weight decay factor of 5×10^{-4} . All input images were uniformly resized to 240×240 pixels. Optimization was performed using the Adam algorithm, with exponential decay rates configured as $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Additionally, a cosine annealing schedule was employed to dynamically adjust the learning rate throughout training.

3.3. Quantitative Results

Tables 1 and 2 present the quantitative evaluation results on the SateHaze1k and RRSID datasets, where the performance is evaluated using PSNR and SSIM metrics, with mean performance reported across different haze density levels. Table 3 summarizes the corresponding results for the HRSD dataset. In addition, Table 4 reports the performance on the HazyDet dataset, further demonstrating the strong generalization ability of our model for drone-based remote sensing image dehazing tasks. Table 5 shows the number of network parameters and floating-point operations (FLOPs) for the proposed method and other similar methods on a 512×512 image.

Table 1. For the SateHaze1k dataset, the leading method is marked in **bold**, while the second-best approach is shown with underline.

Methods	Thin		Moderate		Thick		Average	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DCP [10]	20.15	0.8645	20.51	0.8932	15.77	0.7117	18.81	0.8241
FCTF-Net [16]	19.13	0.8532	22.32	0.9107	17.78	0.7617	19.74	0.8419
AOD-Net [14]	15.97	0.8169	15.39	0.7442	14.44	0.7013	15.27	0.7541
GridDehaze-Net [15]	19.81	0.8556	22.75	0.9085	17.94	0.7551	20.17	0.8397
Dehazeformer [18]	24.90	0.9104	27.13	<u>0.9431</u>	<u>22.68</u>	<u>0.8497</u>	<u>24.90</u>	<u>0.9011</u>
MixDehaze-Net [28]	22.12	0.8822	23.92	0.9040	19.96	0.7950	22.00	0.8604
FFA-Net [17]	24.04	<u>0.9130</u>	25.62	0.9336	21.70	0.8422	23.79	0.8963
OK-Net [29]	20.68	0.8860	25.39	0.9406	20.21	0.8186	22.09	0.8817
AdaIR [30]	24.14	0.9092	25.26	0.9376	21.01	0.8209	23.47	0.8892
Ours	25.62 ^{+0.72}	0.9227 ^{+0.0097}	27.47 ^{+0.34}	0.9451 ^{+0.0020}	22.94 ^{+0.26}	0.8563 ^{+0.0066}	25.34 ^{+0.44}	0.9080 ^{+0.0069}

Table 2. For the RRSID dataset, the leading method is marked in **bold**, while the second-best approach is shown with underline.

Methods	Thick Fog		Moderate Fog		Thin Fog		Average	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DCP [10]	12.43	0.4493	13.32	0.4771	16.16	0.4762	13.97	0.4675
FCTF-Net [16]	22.43	0.6686	21.72	0.6313	21.20	0.5579	21.78	0.6193
AOD-Net [14]	15.10	0.3847	15.26	0.3923	18.22	0.4850	16.19	0.4207
GridDehaze-Net [15]	23.84	0.7125	22.54	0.6475	22.54	0.6203	22.97	0.6601
Dehazeformer [18]	24.71	0.7309	20.93	0.6183	23.09	0.6452	22.92	0.6648
MixDehaze-Net [28]	23.64	0.6728	22.43	0.6233	22.02	0.5666	22.70	0.6209
FFA-Net [17]	<u>25.09</u>	0.7397	23.24	0.6713	23.09	0.6409	23.81	0.6840
OK-Net [29]	24.76	<u>0.7415</u>	<u>24.18</u>	<u>0.7025</u>	<u>23.55</u>	<u>0.6599</u>	<u>24.16</u>	<u>0.7013</u>
AdaIR [30]	23.98	0.7146	23.12	0.6671	22.54	0.6007	23.21	0.6608
Ours	25.52 ^{+0.43}	0.7541 ^{+0.0126}	24.48 ^{+0.30}	0.7096 ^{+0.0071}	23.92 ^{+0.37}	0.6798 ^{+0.0199}	24.64 ^{+0.48}	0.7145 ^{+0.0132}

Table 3. For the HRSD dataset, the leading method is marked in **bold**, while the second-best approach is shown with underline.

Methods	LHID		DHID		Average	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DCP [10]	21.34	0.7976	19.15	0.8195	20.25	0.8086
FCTF-Net [16]	28.55	0.8727	22.43	0.8482	25.49	0.8605
AOD-Net [14]	21.91	0.8144	16.03	0.7291	18.97	0.7718
GridDehaze-Net [15]	25.80	0.8584	26.77	0.8851	26.29	0.8718
MixDehaze-Net [28]	<u>29.47</u>	0.8631	27.36	0.8864	<u>28.42</u>	0.8748
FFA-Net [17]	29.33	0.8755	24.62	0.8657	26.98	0.8706
OK-Net [29]	29.03	0.8766	<u>27.80</u>	<u>0.8973</u>	<u>28.42</u>	<u>0.8870</u>
AdaIR [30]	28.08	<u>0.8776</u>	24.01	0.8676	26.05	0.8726
Ours	29.81 ^{+0.34}	0.8819 ^{+0.0043}	28.64 ^{+0.84}	0.8993 ^{+0.0020}	29.23 ^{+0.81}	0.8906 ^{+0.0036}

Table 4. For the HazyDet dataset, the leading method is marked in **bold**, while the second-best approach is shown with underline.

Methods	HazyDet	
	PSNR ↑	SSIM ↑
DCP [10]	17.03	0.8024
FCTF-Net [16]	24.89	0.8552
AOD-Net [14]	18.99	0.7808
GridDehaze-Net [15]	26.66	0.8801
MixDehaze-Net [28]	<u>28.75</u>	<u>0.9068</u>
FFA-Net [17]	27.12	0.8782
OK-Net [29]	27.76	0.8875
AdaIR [30]	27.60	0.8857
Ours	29.08 ^{+0.33}	0.9080 ^{+0.0012}

Table 5. Comparison of FLOPs and parameters across models.

Method	FLOPs	Parameters
DCP	-	-
FCTF-Net	40.19 (G)	163.48 (K)
AOD-Net	457.70 (M)	1.76 (K)
GridDehaze-Net	85.72 (G)	955.75 (K)
MixDehaze-Net	114.30 (G)	3.17 (M)
FFA-Net	624.20 (G)	4.68 (M)
Dehazeformer	358.89 (G)	9.68 (M)
OK-Net	158.20 (G)	4.43 (M)
AdaIR	588.73 (G)	28.74 (M)
O-Transformer-Mamba	395.98 (G)	6.20 (M)

According to Table 1, the proposed O-Transformer-Mamba model demonstrates notable improvements in PSNR and SSIM for varying haze densities, indicating strong dehazing capabilities. Earlier dehazing techniques like DCP and AOD-Net, which are based on atmospheric scattering models, exhibit poor overall performance. Similarly, FCTF-Net and GridDehaze-Net also yield suboptimal results. MixDehaze-Net and OK-Net achieve moderate performance, while FFA-Net, Dehazeformer, and AdaIR perform comparatively well. Nonetheless, O-Transformer-Mamba consistently surpasses all other methods in key metrics. Notably, it achieves an average PSNR and SSIM gain of 0.44 dB and 0.0069, respectively, over the second-best method, Dehazeformer. These results validate the effectiveness and robustness of our approach in addressing the challenges of remote sensing image dehazing.

As shown in Table 2, our proposed O-Transformer-Mamba demonstrates superior dehazing performance across various haze levels tested with the RRSHID dataset comprising challenging real-world remote sensing scenes. The complexity of haze distribution, diverse degradation patterns, and absence of reliable priors in real-world images make accurate modeling difficult, resulting in generally lower performance across existing methods. Methods like DCP, AOD-Net, and FCTF-Net achieve relatively low performance, whereas GridDehaze-Net, MixDehaze-Net, and Dehazeformer yield intermediate results. More competitive performance is observed with FFA-Net, OK-Net, and AdaIR. In contrast, O-Transformer-Mamba achieves notable gains, surpassing the second-best OK-Net by 0.48 dB in PSNR and 0.0132 in SSIM on average. These outcomes highlight the strong adaptability and effectiveness of our model in tackling the complexities of real remote sensing dehazing tasks.

Table 3 presents the dehazing performance of O-Transformer-Mamba on two synthetic benchmark datasets, LHID and DHID, both generated using an atmospheric scattering-based approach to generate a range of haze intensities. On the LHID dataset, our method outperforms the second-best MixDehaze-Net and AdaIR, achieving a PSNR improvement of 0.34 dB and a SSIM gain of 0.0043. For the DHID dataset, while OK-Net benefits from large-kernel convolutions and delivers competitive results, O-Transformer-Mamba still surpasses it with a PSNR increase of 0.5 dB and a SSIM improvement of 0.0017. The findings as a whole highlight the robust performance and dehazing effectiveness of O-Transformer-Mamba in synthetic remote sensing environments.

To further show that our model performs well beyond specific datasets, extending to diverse and complex real-world conditions, we evaluated it on a UAV-based hazy image dataset to assess its adaptability, robustness, and generalization capabilities. As shown in Table 4, our O-Transformer-Mamba consistently outperforms all competing methods across all evaluation metrics. Notably, it achieves a PSNR gain of 0.33 dB and a SSIM improvement of 0.0012 compared to the second-best approach, indicating substantial enhancements in dehazing quality under challenging aerial imaging conditions.

3.4. Qualitative Results

Visual results of multiple dehazing methods on the light-haze test images are presented in Figure 5. As observed, DCP exhibits limited dehazing capability, leaving a significant amount of residual haze and producing a noticeable green tint. GridDehaze-Net and OK-Net can remove most haze but still suffer from minor haze retention. While FFA-Net, MixDehaze-Net, Dehazeformer, and AdaIR achieve relatively effective dehazing results, they fall short in terms of detail and color fidelity. For instance, in the regions marked by red boxes, the color reproduction deviates from that of the ground truth. In contrast, our O-Transformer-Mamba demonstrates superior haze removal and better aligns with the reference images. Specifically, in the first image, the rooftop within the red box shows color recovery much closer to the ground truth. In the second image, the vegetation area exhibits more accurate green tones, while other methods tend to produce overly light hues. Overall, O-Transformer-Mamba outperforms other methods in haze removal, color accuracy, and frequency detail restoration, yielding results that are visually more consistent with real-world scenes.

Figure 6 evaluates the performance of several dehazing methods on moderately hazy remote sensing images. In the first image, DCP fails to effectively remove the haze, while in the second image, it achieves partial dehazing but still leaves residual haze. In comparison, methods such as GridDehaze-Net, Dehazeformer, MixDehaze-Net, FFA-Net, OK-Net, and AdaIR deliver visually convincing results, with dehazed outputs generally close to the ground truth. However, closer inspection of the highlighted regions reveals noticeable

differences. In the first image, within the red box highlighting a land area, most methods show varying degrees of distortion, whereas our O-Transformer-Mamba produces results that more closely resemble the reference image. In the second image, the rooftop with white stripes in the red box shows darkening or loss of high-frequency details in most methods, while our approach maintains more accurate color and structural fidelity. These results demonstrate that O-Transformer-Mamba offers enhanced restoration of high-frequency content and more faithful color reproduction, largely due to its integration of multi-scale frequency-domain features.

Figure 7 presents a comparison of various methods on heavily degraded images from the thick haze dataset. DCP fails to effectively suppress atmospheric interference, resulting in hazy outputs with strong color distortions. GridDehaze-Net also suffers from noticeable color shifts across the scene. Although FFA-Net, MixDehaze-Net, OK-Net, and AdaIR produce relatively cleaner images, their reconstructions of the rooftop area (highlighted by the red box) remain insufficient in terms of color fidelity. Both Dehazeformer and our O-Transformer-Mamba generate visually similar results; however, our method provides slightly more accurate reconstruction in the rooftop region. Overall, O-Transformer-Mamba demonstrates strong robustness and superior capability in recovering fine details under severe haze conditions.



Figure 5. Visual evaluation of two thin hazy samples from the Haze1k-thin dataset.



Figure 6. Visual evaluation of two moderate hazy samples from the Haze1k-moderate dataset.

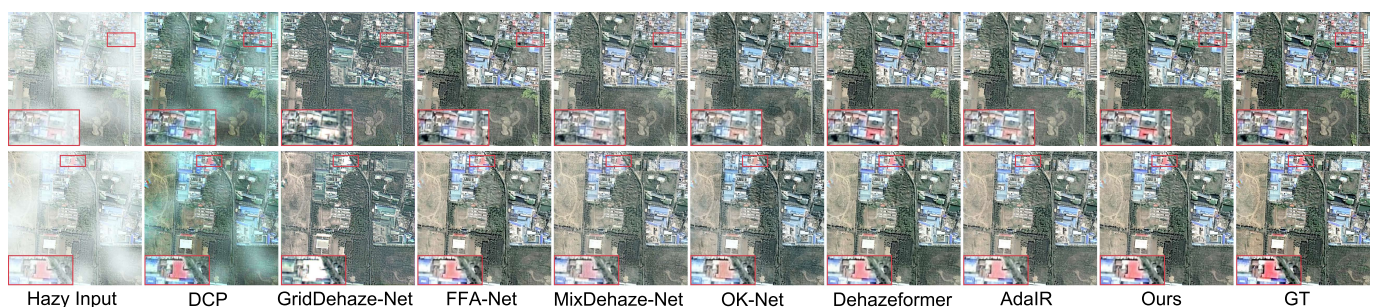


Figure 7. Visual evaluation of two thick hazy samples from the Haze1k-thick dataset.

Figure 8 presents a visual comparison of various dehazing methods on degraded images from a real-world remote sensing dataset captured under heavy haze conditions. The DCP method exhibits severe color distortion, causing the entire image to appear reddish. Within the red rectangle of the first image, the rooftop's high-frequency stripe patterns are blurry in the outputs of GridDehaze-Net, Dehazeformer, MixDehaze-Net, FFA-Net, OK-Net, and AdaIR, whereas our method preserves these patterns with greater clarity. In the second image, GridDehaze-Net, MixDehaze-Net, and AdaIR fail to recover the red rooftop structure. Compared to FFA-Net, OK-Net, and Dehazeformer, our method demonstrates superior restoration performance, particularly in recovering fine details.

Figure 9 illustrates the performance of various methods on real-world remote sensing images affected by moderate haze. The DCP method continues to exhibit a strong reddish color cast. GridDehaze-Net, MixDehaze-Net, and Dehazeformer fail to effectively remove the haze, resulting in suboptimal restoration. In the red-marked regions of both images, the outputs from FFA-Net, OK-Net, and AdaIR show significant blurring of rooftops and roads, while our method achieves clearer structural details. Overall, dehazing for real remote sensing imagery remains a highly challenging task, with all existing methods demonstrating limited restoration capability.

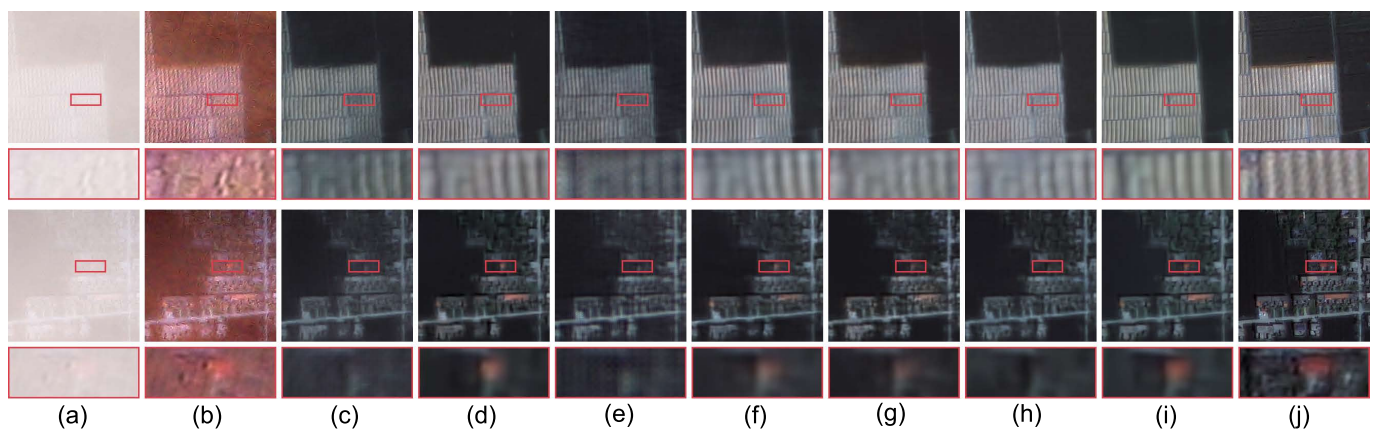


Figure 8. Visual evaluation of two thick fog samples from the RRSID dataset. (a) Hazy Input; (b) DCP; (c) GridDehaze-Net; (d) FFA-Net; (e) MixDehaze-Net; (f) OK-Net; (g) Dehazeformer; (h) AdaIR; (i) O-Transformer-Mamba; (j) GT.

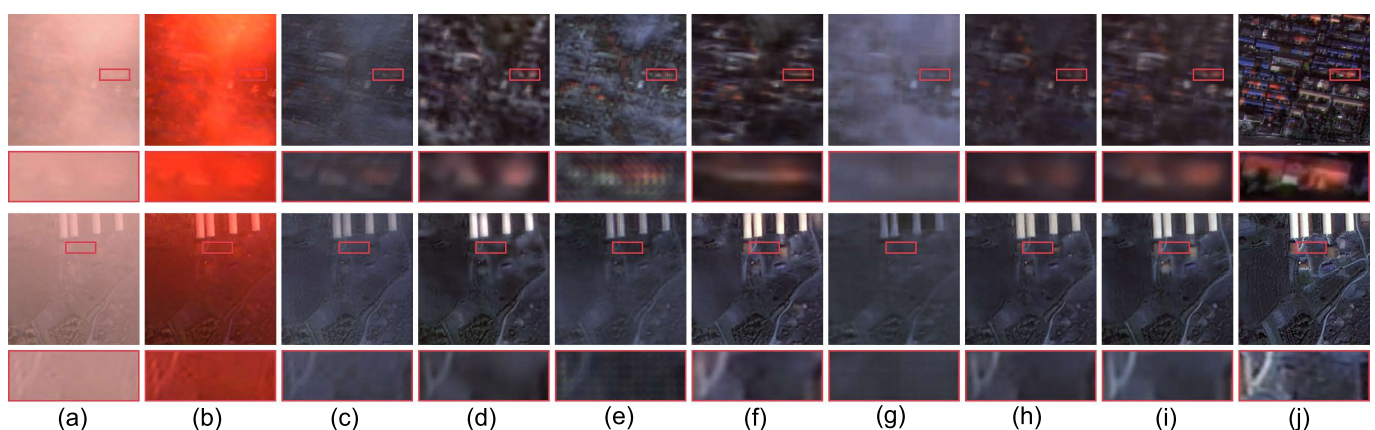


Figure 9. Visual evaluation of two moderate fog samples from the RRSID dataset. (a) Hazy Input; (b) DCP; (c) GridDehaze-Net; (d) FFA-Net; (e) MixDehaze-Net; (f) OK-Net; (g) Dehazeformer; (h) AdaIR; (i) O-Transformer-Mamba; (j) GT.

Figure 10 presents a visual comparison of a test set of real-world remote sensing images affected by light haze. Apart from DCP, most methods successfully eliminate

the majority of haze, with the main differences lying in their ability to reconstruct fine details. Specifically, in the red-marked area of the first image, O-Transformer-Mamba can recover the general structure, whereas the outputs from other methods remain highly blurred. In the second image, within the red box, O-Transformer-Mamba reconstructs the outline of the road to a noticeable extent, while the other approaches show very limited restoration performance.

Figure 11 showcases the performance comparison of several dehazing algorithms on the LHID dataset. The DCP method produces overly saturated results, which leads to the loss of important structural information. While all approaches deliver visually reasonable outputs, there are notable discrepancies in their PSNR and SSIM scores. In particular, our method surpasses the second-best in the first image by 0.32 dB in PSNR and 0.0013 in SSIM. For the second image, the improvements are 0.1 dB and 0.0002, respectively. In the third case, our approach achieves a gain of 0.12 dB in PSNR and 0.0015 in SSIM over the next best method.

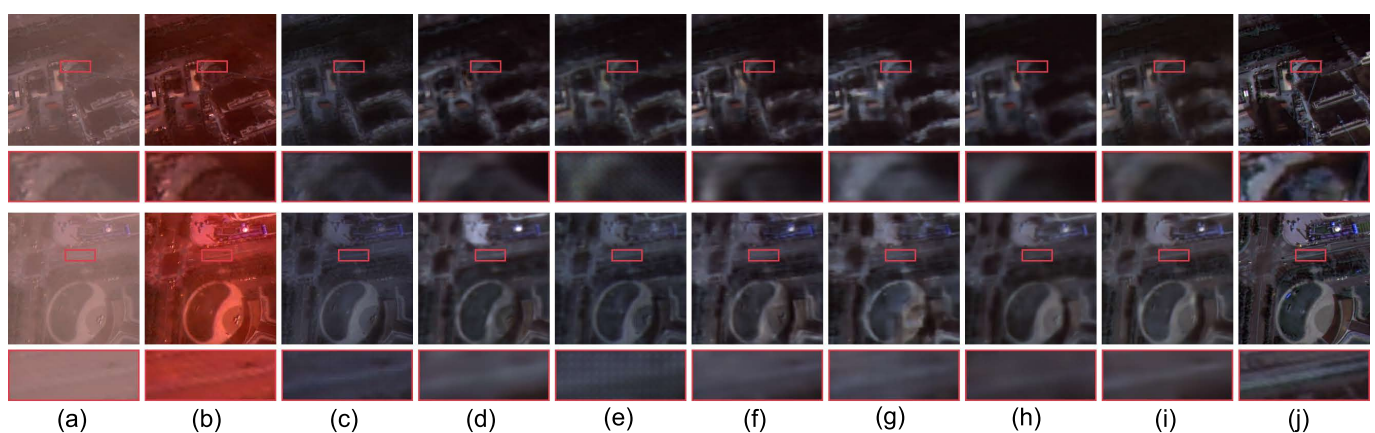


Figure 10. Visual evaluation of two thin fog samples from the RRSHID dataset. (a) Hazy Input; (b) DCP; (c) GridDehaze-Net; (d) FFA-Net; (e) MixDehaze-Net; (f) OK-Net; (g) Dehazeformer; (h) AdaIR; (i) O-Transformer-Mamba; (j) GT.



Figure 11. Visual evaluation of three examples from the LHID dataset.

Figure 12 presents a benchmark evaluation of dehazing performance on the DHID dataset. The output from DCP exhibits severe brightness reduction and loss of details, failing to restore essential high-frequency features. AOD-Net introduces an obvious dark overlay across its outputs. In the third image, FFA-Net and AdaIR struggle to effectively remove haze. Although GridDehaze-Net, MixDehaze-Net, and OK-Net manage to eliminate

haze to a large extent and restore the general scene structure, their PSNR and SSIM scores fall noticeably short compared to those achieved by our proposed O-Transformer-Mamba.

Figure 13 provides a benchmark comparison of dehazing performance on UAV-captured images under varying atmospheric conditions. DCP performs poorly across the board—failing to clear heavy haze in the first image, introducing sky distortion in the second, and leaving substantial haze residues in both the third and fourth images. AOD-Net also falls short, showing limited dehazing capability in all four cases. GridDehaze-Net, MixDehaze-Net, FFA-Net, OK-Net, and AdaIR successfully restore clarity to the third and fourth images by removing haze; however, our method offers superior restoration of ground structures and buildings. The first and second images are heavily obscured by dense haze, concealing key visual information. Compared to other approaches, our method can remove more haze and restore finer details throughout the scene.

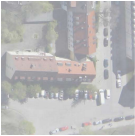
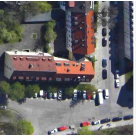
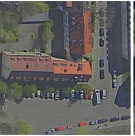
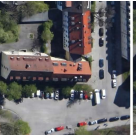
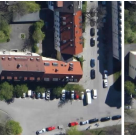
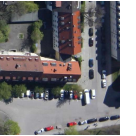
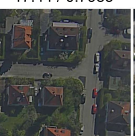
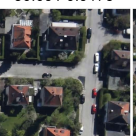
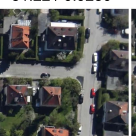
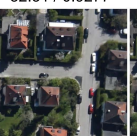
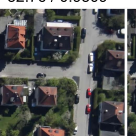
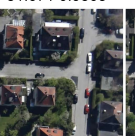
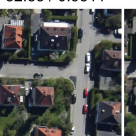
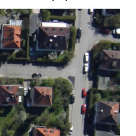
									
PSNR / SSIM	27.14 / 0.8856	17.44 / 0.7985	30.09 / 0.9175	31.22 / 0.9230	32.31 / 0.9277	32.76 / 0.9309	31.07 / 0.9303	32.83 / 0.9311	∞ / 1
									
PSNR / SSIM	19.30 / 0.8626	16.85 / 0.7647	30.58 / 0.9190	30.23 / 0.9194	32.48 / 0.9267	32.47 / 0.9289	29.26 / 0.9244	32.91 / 0.9294	∞ / 1
									
PSNR / SSIM Hazy Image	15.23 / 0.7984 DCP	14.82 / 0.7780 AOD-Net	27.06 / 0.9114 GridDehaze-Net	19.85 / 0.8701 FFA-Net	26.66 / 0.9175 MixDehaze-Net	28.57 / 0.9246 OK-Net	20.04 / 0.8763 AdaIR	28.89 / 0.9253 Ours	∞ / 1 GT

Figure 12. Visual evaluation of three examples from the DHID dataset.

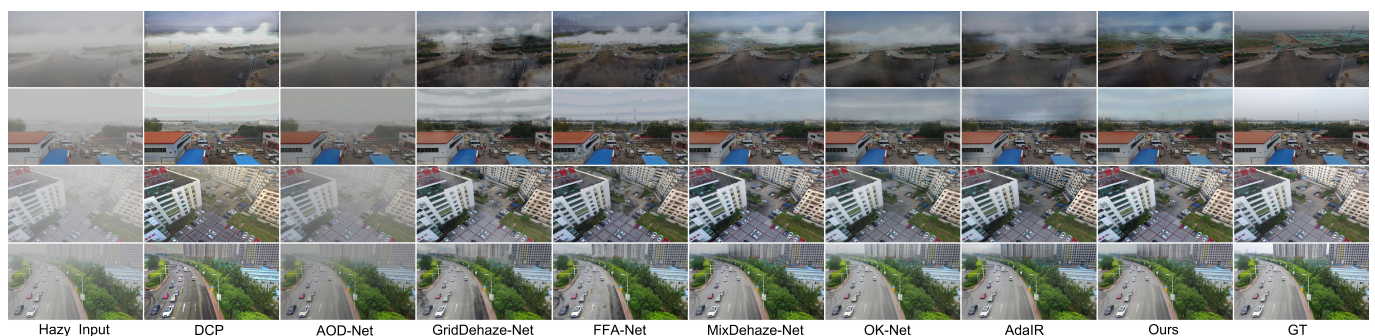


Figure 13. Visual evaluation of four examples from the HazyDet dataset.

3.5. Ablation Study

To comprehensively evaluate the effectiveness of our proposed design, we performed a series of ablation experiments on a light haze dataset, focusing on key modules within the framework. These include the Sparse-Enhanced Self-Attention (SE-SA), the Hybrid Visual State Space Module (Mix-VSSM), and the O-shaped architecture. Two baseline structures were constructed for comparison: one based on standard self-attention (SA), and the other on a conventional visual state space model (VSSM). The SE module was used as a control to assess SE-SA, while VSSM served as the reference for Mix-VSSM.

For computational efficiency during training, 120×120 patches were extracted from the original RGB images, while keeping the same hyperparameters and training settings as the full-resolution model. Table 6 presents the numerical evaluation, showing that each key

module contributes a noticeable performance gain when applied individually, confirming the effectiveness of their design. Additionally, the visual comparisons in Figure 14 further illustrate the individual contributions of these components, highlighting their roles in enhancing the overall dehazing capability of the network.

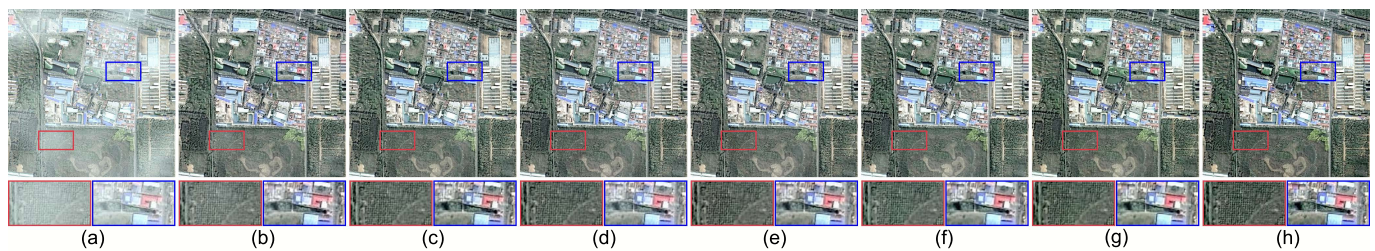


Figure 14. Visual evaluation of ablation experiments on the Thin dataset. (a) Hazy image; (b) SA; (c) SE-SA; (d) VSSM; (e) Mix-VSSM; (f) O(SE-SA + Mix-VSSM); (g) SE-SA + Mix-VSSM + FAM; (h) ground truth.

Table 6. Ablation studies on the thin-haze subset compare method performance, with the best outcomes shown in bold.

Methods	Thin Haze	
	PSNR $\uparrow$	SSIM $\uparrow$
SA	22.62	0.8986
SE-SA	23.36	0.9046
VSSM	23.01	0.8991
Mix-VSSM	24.07	0.9060
O(SE-SA + Mix-VSSM)	24.53	0.9121
SE-SA + Mix-VSSM + FAM	24.80	0.9152

Effectiveness of Sparse-Enhanced Self-Attention (SE-SA): To evaluate the impact of the SE-SA module, an ablation experiment was performed against a conventional self-attention (SA) baseline without the Sparse Enhancement Operation (SEO). As reported in Table 6, SE-SA demonstrates clear performance gains over SA, with improvements of 0.74 dB in PSNR and 0.006 in SSIM. A qualitative comparison in Figure 14 further supports this finding: in the red-boxed region containing dense haze over green vegetation, the input image exhibits significantly heavier haze compared to the rest of the scene. The baseline SA performs uniform dehazing but fails to sufficiently clear this highly degraded area. In contrast, SE-SA effectively removes the residual haze, highlighting its capacity to adaptively focus attention on regions with varying haze density. This confirms that our method enhances the model's robustness by dynamically adjusting attention distribution in response to spatial haze variation.

Effectiveness of Mixed Visual State Space Module (Mix-VSSM): An ablation study was conducted to assess the effectiveness of the proposed Mix-VSSM by comparing it with the baseline VSSM module. As shown in Table 6, Mix-VSSM achieves superior performance, with improvements of 1.06 dB in PSNR and 0.0069 in SSIM. While both methods effectively remove haze across the entire image, as observed in Figure 14, Mix-VSSM demonstrates noticeably better performance in restoring local details, particularly in the region highlighted by the blue box. This confirms that our approach significantly enhances local feature recovery, thereby improving the overall perceptual quality of the dehazed image.

Effectiveness of O-shaped architecture: To validate the effectiveness of the proposed O-shaped architecture, we integrated both the SE-SA and Mix-VSSM modules into its structure. This integration led to significant improvements in both quantitative metrics and overall visual quality. The O-shaped architecture successfully combines the comple-

mentary strengths of the Transformer and Mamba paradigms, resulting in high-quality image restoration and enhanced dehazing performance. For the O-shaped topology, we conducted additional ablation experiments, comparing the proposed O-shaped architecture with a U-shaped architecture and a simple parallel design. Table 7 shows that the O-shaped structure achieves more efficient feature fusion and consistently maintains superior performance.

Table 7. Ablation experiments on the mist subset compare the performance of different network structures; the best results are shown in bold.

Structures	Thin Haze	
	PSNR $\uparrow$	SSIM $\uparrow$
parallel	23.68	0.9051
U-shaped	24.11	0.9078
O-shaped	24.80	0.9152

4. Discussion

This paper proposes a remote sensing image dehazing framework called O-Transformer-Mamba. This framework combines Transformer-based global modeling with Mamba-based state-space representation through an O-shaped topology. This is further validated by specific ablation experiments targeting this topology. In addition, the SE-SA and Mix-VSSM modules enhance the model's adaptability to complex haze distributions and large-scale spatial structures in remote sensing images.

Nevertheless, several limitations remain. Like other strongly supervised methods, the proposed model relies on paired training data, which may constrain cross-dataset generalization under severe domain shifts. In extreme haze scenarios with near-opaque regions, restoration performance is fundamentally bounded by the amount of recoverable information in the input. Future work will explore extending the O-Transformer-Mamba framework to weakly supervised or semi-supervised learning paradigms, incorporating unpaired data and domain-invariant constraints to balance high-fidelity restoration with stronger generalization.

5. Conclusions

In this paper, we propose an O-shaped remote sensing image dehazing network based on Transformer and Mamba, named O-Transformer-Mamba, which is designed to address both the uneven haze distribution and long-range dependency modeling challenges in remote sensing imagery. The network incorporates two key components: First, the Sparse-Enhanced Self-Attention (SE-SA) module introduces a dynamic prompting mechanism to sparsely modulate the attention matrix. This allows the model to selectively focus on heavily degraded regions while suppressing redundant responses in irrelevant areas, thereby enhancing detail recovery. Second, the Mixed Visual State Space Module (Mix-VSSM) combines 2D state space modeling with a convolutional residual pathway to strengthen local spatial perception while preserving long-range context representation. This design is particularly effective in handling large-scale and complex atmospheric degradations common in remote sensing images. Experimental results demonstrate that O-Transformer-Mamba achieves superior performance across various synthetic and real-world remote sensing dehazing datasets, validating its robustness and generalization capability under multi-scale and non-uniform haze conditions.

Author Contributions: Methodology, X.G. and R.H.; Software, L.W.; Supervision, H.X.; Validation, Y.L.; Writing—original draft, X.G.; Writing—review and editing, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (62401012), and the National Natural Science Foundation of China (62576290).

Data Availability Statement: The SateHaze1k [40], HRSD [41], RRSID [42] and HazyDet [43] datasets are for academic and research use only. Please refer to the paper for details.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, N.; Yang, A.; Cui, Z.; Ding, Y.; Xue, Y.; Su, Y. Capsule attention network for hyperspectral image classification. *Remote Sens.* **2024**, *16*, 4001. [\[CrossRef\]](#)
2. Wang, N.; Cui, Z.; Lan, Y.; Zhang, C.; Xue, Y.; Su, Y.; Li, A. Large-Scale Hyperspectral Image-Projected Clustering via Doubly Stochastic Graph Learning. *Remote Sens.* **2025**, *17*, 1526. [\[CrossRef\]](#)
3. Liu, Y.; Yan, Z.; Tan, J.; Li, Y. Multi-purpose oriented single nighttime image haze removal based on unified variational retinex model. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 1643–1657. [\[CrossRef\]](#)
4. Liu, Y.; Yan, Z.; Chen, S.; Ye, T.; Ren, W.; Chen, E. Nighthazeformer: Single nighttime haze removal using prior query transformer. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 4119–4128.
5. Zhou, H.; Chen, Z.; Liu, Y.; Sheng, Y.; Ren, W.; Xiong, H. Physical-priors-guided DehazeFormer. *Knowl.-Based Syst.* **2023**, *266*, 110410. [\[CrossRef\]](#)
6. Liu, Y.; Wang, X.; Hu, E.; Wang, A.; Shiri, B.; Lin, W. VNDHR: Variational single nighttime image Dehazing for enhancing visibility in intelligent transportation systems via hybrid regularization. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 10189–10203. [\[CrossRef\]](#)
7. Zhou, H.; Wang, Y.; Zhang, Q.; Tao, T.; Ren, W. A Dual-Stage Residual Diffusion Model with Perceptual Decoding for Remote Sensing Image Dehazing. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4109312. [\[CrossRef\]](#)
8. Rong, Z.; Jun, W.L. Improved wavelet transform algorithm for single image dehazing. *Optik* **2014**, *125*, 3064–3066. [\[CrossRef\]](#)
9. Wang, J.; Lu, K.; Xue, J.; He, N.; Shao, L. Single image dehazing based on the physical model and MSRCR algorithm. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 2190–2199. [\[CrossRef\]](#)
10. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 2341–2353. [\[CrossRef\]](#)
11. Zhu, Q.; Mai, J.; Shao, L. A fast single image haze removal algorithm using color attenuation prior. *IEEE Trans. Image Process.* **2015**, *24*, 3522–3533. [\[CrossRef\]](#)
12. Dong, H.; Pan, J.; Xiang, L.; Hu, Z.; Zhang, X.; Wang, F.; Yang, M.H. Multi-scale boosted dehazing network with dense feature fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Washington, WA, USA, 16–18 June 2020; pp. 2157–2167.
13. Cai, B.; Xu, X.; Jia, K.; Qing, C.; Tao, D. Dehazenet: An end-to-end system for single image haze removal. *IEEE Trans. Image Process.* **2016**, *25*, 5187–5198. [\[CrossRef\]](#)
14. Li, B.; Peng, X.; Wang, Z.; Xu, J.; Feng, D. Aod-net: All-in-one dehazing network. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4770–4778.
15. Liu, X.; Ma, Y.; Shi, Z.; Chen, J. Griddehazenet: Attention-based multi-scale network for image dehazing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7314–7323.
16. Li, Y.; Chen, X. A coarse-to-fine two-stage attentive network for haze removal of remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 1751–1755. [\[CrossRef\]](#)
17. Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; Jia, H. FFA-Net: Feature fusion attention network for single image dehazing. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11908–11915.
18. Song, Y.; He, Z.; Qian, H.; Du, X. Vision transformers for single image dehazing. *IEEE Trans. Image Process.* **2023**, *32*, 1927–1941. [\[CrossRef\]](#)
19. Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; Li, H. Uformer: A general u-shaped transformer for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 17683–17693.
20. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 5728–5739.

21. Nie, J.; Xie, J.; Sun, H. Remote sensing image dehazing via a local context-enriched transformer. *Remote Sens.* **2024**, *16*, 1422. [\[CrossRef\]](#)
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
23. Guan, X.; He, R.; Wang, L.; Zhou, H.; Liu, Y.; Xiong, H. DWTMA-Net: Discrete Wavelet Transform and Multi-Dimensional Attention Network for Remote Sensing Image Dehazing. *Remote Sens.* **2025**, *17*, 2033. [\[CrossRef\]](#)
24. Zhou, H.; Wang, L.; Li, Q.; Guan, X.; Tao, T. Multi-Dimensional and Multi-Scale Physical Dehazing Network for Remote Sensing Images. *Remote Sens.* **2024**, *16*, 4780. [\[CrossRef\]](#)
25. Wu, J.; Ai, H.; Zhou, P.; Wang, H.; Zhang, H.; Zhang, G.; Chen, W. Low-Light Image Dehazing and Enhancement via Multi-Feature Domain Fusion. *Remote Sens.* **2025**, *17*, 2944. [\[CrossRef\]](#)
26. Zhou, H.; Wang, Y.; Peng, W.; Guan, X.; Tao, T. ScaleViM-PDD: Multi-Scale EfficientViM with Physical Decoupling and Dual-Domain Fusion for Remote Sensing Image Dehazing. *Remote Sens.* **2025**, *17*, 2664. [\[CrossRef\]](#)
27. Wang, H.; Ding, Y.; Zhou, X.; Yuan, G.; Sun, C. Dehazing of Panchromatic Remote Sensing Images Based on Histogram Features. *Remote Sens.* **2025**, *17*, 3479. [\[CrossRef\]](#)
28. Lu, L.; Xiong, Q.; Xu, B.; Chu, D. Mixdehazenet: Mix structure block for image dehazing network. In Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 30 June–5 July 2024; pp. 1–10.
29. Cui, Y.; Ren, W.; Knoll, A. Omni-Kernel Network for Image Restoration. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; pp. 1426–1434.
30. Cui, Y.; Zamir, S.W.; Khan, S.; Knoll, A.; Shah, M.; Khan, F.S. AdaIR: Adaptive All-in-One Image Restoration via Frequency Mining and Modulation. In Proceedings of the the Thirteenth International Conference on Learning Representations, Singapore, 24–28 April 2025.
31. Hamilton, J.D. State-space models. *Handb. Econom.* **1994**, *4*, 3039–3080.
32. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752. [\[CrossRef\]](#)
33. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063.
34. Guo, H.; Li, J.; Dai, T.; Ouyang, Z.; Ren, X.; Xia, S.T. Mambair: A simple baseline for image restoration with state-space model. In *Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 222–241.
35. Sun, S.; Ren, W.; Zhou, J.; Gan, J.; Wang, R.; Cao, X. A hybrid transformer-mamba network for single image deraining. *arXiv* **2024**, arXiv:2409.00410. [\[CrossRef\]](#)
36. Hatamizadeh, A.; Kautz, J. Mambavision: A hybrid mamba-transformer vision backbone. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 25261–25270.
37. Su, X.; Li, S.; Cui, Y.; Cao, M.; Zhang, Y.; Chen, Z.; Wu, Z.; Wang, Z.; Zhang, Y.; Yuan, X. Prior-guided hierarchical harmonization network for efficient image dehazing. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; pp. 7042–7050.
38. Cui, Y.; Wang, Q.; Li, C.; Ren, W.; Knoll, A. EENet: An effective and efficient network for single image dehazing. *Pattern Recognit.* **2025**, *158*, 111074. [\[CrossRef\]](#)
39. Wang, T.; Du, L.; Yi, W.; Hong, J.; Zhang, L.; Zheng, J.; Li, C.; Ma, X.; Zhang, D.; Fang, W.; et al. An adaptive atmospheric correction algorithm for the effective adjacency effect correction of submeter-scale spatial resolution optical satellite images: Application to a WorldView-3 panchromatic image. *Remote Sens. Environ.* **2021**, *259*, 112412. [\[CrossRef\]](#)
40. Huang, B.; Zhi, L.; Yang, C.; Sun, F.; Song, Y. Single satellite optical imagery dehazing using SAR image prior based on conditional generative adversarial networks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1806–1813.
41. Zhang, L.; Wang, S. Dense haze removal based on dynamic collaborative inference learning for remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5631016. [\[CrossRef\]](#)
42. Zhu, Z.H.; Lu, W.; Chen, S.B.; Ding, C.H.Q.; Tang, J.; Luo, B. Real-World Remote Sensing Image Dehazing: Benchmark and Baseline. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4705014. [\[CrossRef\]](#)
43. Feng, C.; Chen, Z.; Kou, R.; Gao, G.; Wang, C.; Li, X.; Shu, X.; Dai, Y.; Fu, Q.; Yang, J. HazyDet: Open-source Benchmark for Drone-view Object Detection with Depth-cues in Hazy Scenes. *arXiv* **2024**, arXiv:2409.19833.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.