Vision Transformer-Based Image Dehazing for Climate-Resilient Maritime Navigation

Xinqiang Chen, *Member, IEEE*, Zhengang Xin, Han Zhang, Yuzhen Wu, Chenxin Wei, and Octavian Postolache, *Senior Member, IEEE*

***Abstract*—As climate change intensifies, maritime transportation systems face critical challenges from frequent fog, high humidity, and adverse weather, which degrade visual perception and threaten navigation safety. To address these climate-induced disruptions, we propose TWRM-Net, a Transformer-based dehazing network designed for climate-resilient maritime vision. Firstly, the network employs a hierarchical encoder–decoder backbone that integrates convolutional priors with window-based self-attention, enabling joint modeling of local textures and global haze structures for maritime haze removal. Secondly, we introduce a Dual-Residual Attention Block (DRAB), which enhances structural awareness and haze localization by combining spatial and channel attention with residual learning, thus improving robustness in fog-dense environments. Lastly, the adaptive fusion module (AFM) adaptively balances low-frequency global context from the encoder and high-frequency details from the decoder, ensuring frequency-selective feature integration while suppressing aliasing artifacts. We construct a realistic maritime haze dataset using monocular depth estimation and the atmospheric scattering model, simulating climate-induced haze conditions. Extensive experiments on synthetic and real-world maritime datasets demonstrate that TWRM-Net achieves state-of-the-art performance, with PSNR of 35.07 dB, SSIM of 0.9735, LPIPS of 0.0880, and FID of 23.65, significantly outperforming existing methods. These results highlight the effectiveness of our approach for climate-adaptive intelligent transportation, providing reliable perception for safe and sustainable maritime navigation in the era of climate change.**

***Index Terms*—Computer vision, maritime transportation, maritime image dehazing, vision transformer.**

I. Introduction

With the rapid advancement of intelligent maritime transportation, vision-based perception has become indispensable for autonomous navigation, marine environment monitoring, and smart port management. However, climate-induced haze severely degrades image quality by blurring boundaries, reducing contrast, and distorting colors, thereby undermining downstream tasks such as ship detection, collision avoidance, and risk assessment [1]. Recent advances in maritime vision and adverse-weather perception highlight the need for models capable of handling low-texture sea surfaces, horizon ambiguity, and strong reflectance coupling between sky and water. Unlike terrestrial scenes, maritime imagery is dominated by large low-frequency gradients and non-uniform haze patterns, requiring restoration methods tailored to the ocean optical environment. Efficient image dehazing is thus critical for climate-resilient maritime transportation, as it addresses challenges from non-uniform haze distribution, multi-scale occlusion, and illumination artifacts. Beyond enhancing perception reliability, it also facilitates automation-driven emission reduction by improving decision accuracy and navigation efficiency [2]. In this study, the term 'climate-resilient' refers to the model's ability to adapt to climate-driven variations in maritime haze conditions. Its resilience refers to stable performance under different haze intensities, spatial patterns, and illumination variations associated with climate variability.

Early maritime dehazing methods primarily relied on classical prior-based approaches such as the Dark Channel Prior (DCP) color-line prior, and contextual regularization. While these methods demonstrated competence under controlled conditions, they frequently exhibited limitations in maritime scenarios characterized by dense or non-uniform haze distribution. Owing to their strong reliance on rigid physical assumptions, such techniques were often susceptible to color distortion and structural artifacts [3].

In recent years, advances in deep learning have shifted image dehazing research from experience-based models to AI-driven visual architectures with global perceptual capabilities. Among these, methods based on Convolutional Neural Networks, such as the AECR-Net which incorporates an attention-enhanced compression and multi-stage refinement mechanism, have achieved significant improvements in dehazing performance [4]. However, their limited receptive fields restrict global context modeling, which reduces effectiveness in dense maritime haze.

With the widespread adoption of the Transformer architecture in visual tasks, its exceptional global context modeling capabilities have provided new solutions for single image haze

Received 1 September 2025; revised 30 November 2025 and 20 January 2026; accepted 22 January 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 52472347, Grant 52331012, and Grant 52372316; and in part by the Opening Project of State Key Laboratory of Maritime Technology and Safety. The Associate Editor for this article was D. Sun. *(Corresponding author: Han Zhang.)*

Xinqiang Chen and Zhengang Xin are with the Institute of Logistics Science and Engineering, Shanghai Maritime University, Shanghai 201306, China (e-mail: chenxinqiang@stu.shmtu.edu.cn; xinzhengang@stu.shmtu.edu.cn).

Han Zhang is with the College of Ocean Science and Engineering, Shanghai Maritime University, Shanghai 201306, China (e-mail: zhanghan2024@stu.shmtu.edu.cn).

Yuzhen Wu is with SPG Qingdao Port Group Company Ltd., Qingdao 266011, China (e-mail: wuyz@qdport.com).

Chenxin Wei is with the School of Transportation, Southeast University, Nanjing 210096, China (e-mail: 230249241@seu.edu.cn).

Octavian Postolache is with the Iscte—Instituto Universitário de Lisboa, 1049-001 Lisbon, Portugal, and also with the Instituto de Telecomunicações, 1049-001 Lisbon, Portugal (e-mail: octavian.adrian.postolache@iscte-iul.pt).

Digital Object Identifier 10.1109/TITS.2026.3661634

removal. Traditional dehazing methods based on CNN are often constrained by limited local receptive fields, resulting in insufficient global modeling capacity and restricted feature representation [5]. To address these limitations, improved Transformer-based models have recently been introduced to haze removal tasks. For instance, the Swin Transformer method achieves a balance between local and global information modeling through its hierarchical window and shifted window attention mechanisms [6]. The Pyramid Vision Transformer enhances multi-scale representation ability via pyramid feature extraction, while Restormer proposes an efficient architecture for modeling long-range dependencies tailored specifically to image restoration tasks [7], [8]. These approaches highlight Transformer's global modeling capability for haze removal.

Maritime visual perception is severely degraded by dense and spatially non-uniform haze, which reduces visibility and undermines the reliability of ship detection and autonomous navigation systems. To address this challenge, this study aims to develop TWRM-Net, a physically guided Transformer–convolution hybrid network designed to restore maritime images affected by atmospheric scattering and complex haze conditions. We employ a maritime-oriented Transformer architecture named TWRM-Net. The model follows a U-Net-style encoder–decoder structure. The input image is first transformed into feature representations via PatchEmbed. Through multi-stage stacking of the Basic Layer, which combines local convolution and window attention, the model progressively extracts multi-scale features from shallow to deep levels. During decoding, PatchUnEmbed is employed for gradual upsampling, while an adaptive fusion module integrates skip-connected features to restore spatial details. The final output is projected back to the image space through a PatchUnEmbed layer composed of convolution and PixelShuffle operations, producing a transmission map and a residual bias term. These outputs are constrained using a residual-form atmospheric scattering model to achieve physically consistent and high-quality image dehazing.

The network is designed modularly, incorporating multi-head self-attention, window/shifted-window attention mechanisms, and hierarchical multi-scale feature modeling [9]. This integration significantly enhances feature representation in hazy regions and global structural modeling capability, thereby achieving more robust dehazing performance in complex and variable maritime scenarios. The main contributions of this paper can be summarized as follows:

- We propose a hierarchical Transformer framework with a U-Net–style encoder–decoder architecture. In this design, convolution and window-based attention are jointly employed for adaptive multi-scale feature extraction, while skip connections combined with an Adaptive Fusion Module (AFM) effectively restore spatial details under physical constraints.
- We design a Dual-Residual Attention Block (DRAB) to replace standard Transformer attention, enhancing feature representation and cross-scale integration for improved structural recovery and perceptual consistency.
- We construct a high-fidelity maritime hazy image dataset using depth estimation and the atmospheric scattering model, and extensive experiments show that TWRM-Net outperforms state-of-the-art methods on multiple benchmarks.

II. Related Works

A. Physics-Guided Maritime Image Dehazing

Single image dehazing aims to restore a clear image from one degraded by atmospheric particles like haze. This technology is crucial for intelligent maritime systems, as dense haze and high humidity severely degrade image quality, directly impacting vessel detection, obstacle avoidance, and risk assessment [10]. Thus, efficient and robust dehazing has become a key enabler for building climate-resilient intelligent transportation systems.

Traditional model-based methods typically rely on the atmospheric scattering model, which formulates the formation of a hazy image as a combination of direct transmission and atmospheric light [11]. One of the most influential works in this line is the DCP proposed by He et al., based on the assumption that in haze-free outdoor images, most local regions have at least one color channel with very low pixel intensity. Although DCP performs well in many scenarios, it tends to fail in bright or sky regions and may introduce halo artifacts. Fattal subsequently introduced the Color-Line Prior [3], which models the linear distribution of colors in small patches and achieves better performance in sky and smooth regions. Meng et al. further incorporated boundary constraints and contextual regularization to refine the transmission map, thereby improving structural consistency [12]. Despite these improvements, prior-based methods often underperform in complex or non-uniform haze conditions, largely due to their over-reliance on hand-crafted assumptions.

B. CNN-Based Maritime Image Dehazing

To overcome the limitations of the aforementioned model-based approaches, CNN-based methods have become mainstream in single image dehazing. Among them, DehazeNet was the first to introduce CNN into image dehazing, enabling end-to-end transmission map estimation. AOD-Net reformulated the atmospheric scattering model into a unified framework that directly recovers a clear image [13]. Subsequently, DCPDN incorporated the dark channel prior into a trainable network architecture and further improved dehazing performance through multi-scale feature extraction and attention mechanisms, resulting in enhanced generalization capability [14]. To address complex and non-uniform haze distributions, FFA-Net introduced a feature fusion attention mechanism that dynamically assigns weights across spatial and channel dimensions, achieving higher restoration quality on synthetic datasets [15]. Zhou et al. proposed a method combining fast Fourier convolution with the ConvNeXt architecture for frequency-domain modeling [16]. In terms of feature fusion mechanisms, Chen et al. proposed DEA-Net, which integrates detail-enhanced convolution and content-guided attention to

effectively improve the restoration of local details and compensate for the shortcomings of previous models in recovering high-frequency information [17].

Although CNN-based methods have advanced image dehazing, their performance remains limited by local receptive fields, hindering the handling of non-uniform haze or complex scattering. Moreover, heavy reliance on synthetic training data often leads to performance degradation in real-world applications like maritime or aerial imagery.

C. ViT-Based Maritime Image Dehazing

The emergence of Vision Transformer (ViT) has provided new insights to address these challenges. By introducing a global self-attention mechanism, the ViT approach breaks the limitations of convolutional inductive bias and enables effective modeling of long-range dependencies and global context. Various improved ViT-based models have demonstrated exceptional performance in image restoration tasks such as single image dehazing, denoising, and super-resolution [18].

To better meet the requirements of low-level vision tasks, researchers have proposed various structural improvements. For instance, Uformer incorporates a hierarchical locally-enhanced mechanism into a U-shaped Transformer backbone to improve the ability to capture detailed features [19]. To address the challenges posed by non-uniform haze in dehazing tasks, AECR-Net introduced attention-guided mechanisms to enhance haze-layer separation and structure preservation [20]. Restormer further improve structural consistency and restoration quality through efficient channel-spatial modeling and multi-scale attention mechanisms.

In summary, ViT-based methods show great potential in image dehazing, but current methods have two main limitations: First, they lack specific optimization for maritime challenges (e.g., dense haze, high humidity, glare), leading to insufficient generalization in practice; second, there is still room for improvement in haze boundary detail modeling and cross-scale feature fusion.

To address this, we propose TWRM-Net, a Transformer-based dehazing network tailored for maritime applications. It introduces a meteorology-aware dual-residual attention block and an adaptive fusion module to enhance feature representation in hazy areas, strengthen global dependency modeling, and improve the recovery of edges and structural details. Experiments show that TWRM-Net achieves superior dehazing quality in complex maritime environments, especially under thick haze, strong glare, and multi-scale scenes, demonstrating significant advantages over existing methods.

III. Methodology

A. Framework Overview

The proposed TWRM-Net tackles maritime image dehazing under haze and extreme climate for navigation. It employs a hierarchical encoder-decoder architecture that integrates convolutional strengths in local texture with Vision Transformer's capabilities in global dependency representation.

In the proposed framework, the input image is first mapped into a latent token space through a convolution-based Patch Embedding layer. The encoder consists of transformer dehazing block (TDB), which progressively downsample the spatial resolution while enhancing feature dimensions to extract hierarchical semantics. During decoding, spatial resolution is gradually restored via Transformer-based upsampling stages, while skip connections fused by the adaptive fusion module (AFM) provide frequency-aware integration between encoder and decoder features. Furthermore, a dual-residual attention block (DRAB) is incorporated into the convolutional stream to reinforce local structural awareness and haze-edge sensitivity. Finally, the decoded features are projected back into the image space using PatchUnEmbed, and a physics-inspired residual reconstruction formulation produces the haze-free output. This overall architecture equips TWRM-Net with the capability to simultaneously model local details, global haze structures, and cross-scale interactions.

B. Hierarchical Encoder for Multi-Scale Representation Learning

The encoder of TWRM-Net is designed to gradually extract hierarchical semantics while retaining fine-grained structural information, enabling the model to capture both local textures and global haze distributions [21]. Given an input hazy image $I \in R^{3\times H\times W}$, the encoder first employs a convolution-based PatchEmbed layer to transform the image into a latent token representation:

$$X_0 = PatchEmbed\,(I)\,, X_0 \in R^{C\times H\times W}, C = 2 \tag{1}$$

PatchEmbed maps the image into latent tokens for subsequent attention-based modeling.

After initialization, the encoder comprises three sequential stages, each implemented as a BasicLayer constructed from TDB [22]. At each stage s, spatial resolution is reduced by a PatchMerging operation, while the channel dimension is expanded:

$$X_s = PatchMerging_s\,(X_{s-1})\,, X_s \in R^{C_s\times \frac{H}{2^s}\times \frac{W}{2^s}}, C_s = 2^s C \tag{2}$$

This hierarchical downsampling ensures that shallow layers capture local details, whereas deeper layers encode broader semantic and contextual information.

Each stage consists of a BasicLayer, composed of TDB. Inside a TDB, the feature map is first normalized and passed through a window-based multi-head self-attention (W-MHSA) module to model long-range dependencies within local windows:

$$A^{(m)} = Softmax\left(\frac{Q^{(m)}K^{(m)^T}}{\sqrt{d}} + B^{(m)}\right) Y^{(m)} = A^{(m)}V^{(m)} \tag{3}$$

$$Y = Concat_m Y^{(m)}, \hat{X} = Proj(Y) \tag{4}$$

where $d = C/h$ is the head dimension and $B^{(m)}$ denotes the learnable relative position bias. Next, the self-attention output is fused with a depthwise separable convolution (DWConv) branch to strengthen local texture modeling. Finally, a lightweight feed-forward network (MLP) enhances feature expressiveness under residual learning:

$$X_{out} = X'' + MLP(Norm(X'')) \tag{5}$$

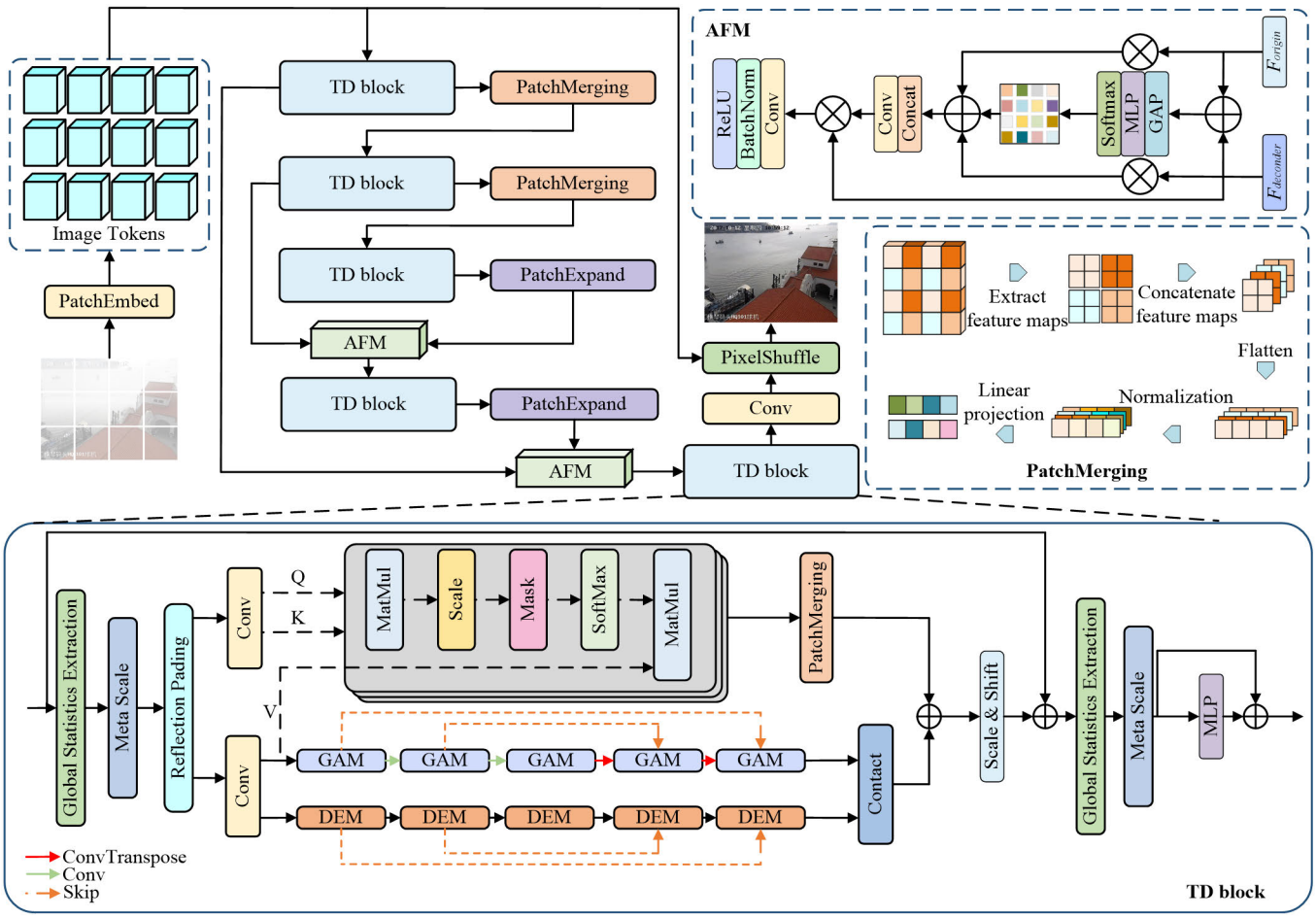


Fig. 1. The overall architecture of the proposed dehazing network combining TWRM-Net.

This hierarchical encoder thus balances local and global modeling. Shallow layers emphasize fine-grained cues such as edges, ship contours, and near-field textures, while deeper layers capture semantic structures including haze intensity variations and horizon lines.

DRAB enhances feature discrimination by combining spatial and channel attention. The DRAB contains two parallel submodules: the Global Attention Module (GAM) and the Detail Enhancement Module (DEM) (see Figure 2).

GAM stacks several residual blocks to capture local spatial structures and integrates a Spatial Attention Block (SAB) to emphasize haze-relevant regions. The SAB first compresses feature maps along the channel dimension using both global max-pooling and average-pooling:

$$x_{attn} = Concat\left(Max(x), Avg\left(x\right)\right), x_{attn} \in R^{2\times H\times W} \tag{6}$$

followed by convolution and sigmoid activation to generate the spatial attention mask:

$$M_s = \sigma\left(\phi\left(x_{attn}\right)\right), M_s \in [0,1]^{1\times H\times W} \tag{7}$$

The final output is given by

$$x' = x + x \cdot M_s \tag{8}$$

which strengthens haze edges and object boundaries.

DEM focuses on channel dependency modeling by cascading multiple depthwise convolutions and ReLU activations, followed by a Channel Attention Block(CAB). The CAB employs global average pooling and a two-layer residual MLP to generate adaptive channel weights:

$$w = \sigma(W_2\phi(W_1 \cdot Avg(x))) \in [0,1]^{C\times 1\times 1} \tag{9}$$

which reweights the feature maps across channels to emphasize content-aware information such as textures and color gradients.

Finally, the outputs of GAM and DEM are concatenated and compressed through convolution to form the final fused feature. By combining spatial and channel attention in a dual-residual manner, DRAB effectively enhances both position-aware and content-aware features. This design ensures that the network is simultaneously sensitive to haze density distributions and capable of restoring fine-grained details, which is critical for accurate dehazing in maritime navigation.

C. Hierarchical Decoder With Multi-Frequency Fusion

The decoder of TWRM-Net is designed as a hierarchical mirror of the encoder, progressively restoring the spatial resolution while maintaining structural consistency and semantic fidelity. Starting from the bottleneck representation, the decoder performs a series of learnable upsampling operations

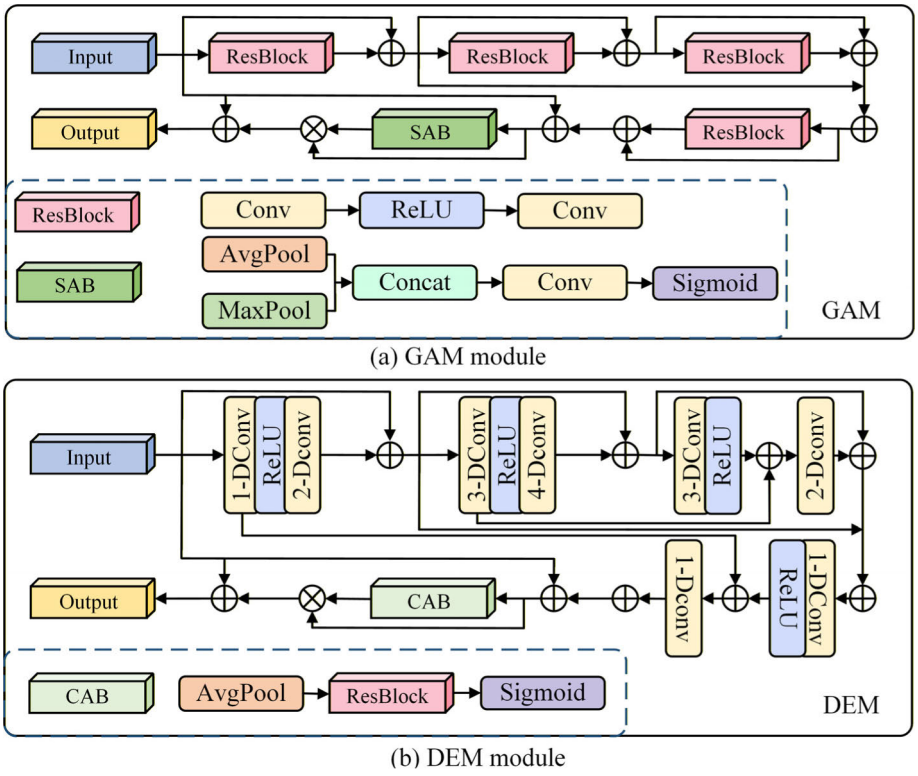


Fig. 2. Dual-Residual Attention Block: (a) GAM enhances spatial feature learning through residual spatial attention; (b) DEM introduces depth-aware channel attention via hierarchical depth-convolution branches.

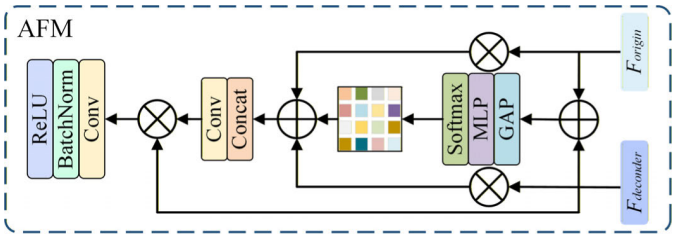


Fig. 3. Adaptive fusion module, which adaptively reweights fused features.

that gradually enlarge the feature maps back to the input resolution. Given the feature map from the previous stage X_{t-1}^{out}, the resolution is restored through a convolution-based upsampling layer:

$$\tilde{X}_t = upsampling(X_{t-1}^{out}), \tilde{X}_t \in R^{C_t \times H_t \times W_t} \tag{10}$$

The spatial resolution is restored through upsampling using a learnable upsampling strategy based on a convolutional projection followed by PixelShuffle. This ensures that the upsampled features align with the corresponding encoder stage in both spatial dimension and channel capacity.

To efficiently bridge the encoder and decoder stages while preserving hierarchical feature semantics, we propose a adaptive fusion module(AFM) for adaptive feature fusion (see Figure 3). Unlike direct skip connections or simple summation strategies, AFM performs dynamic frequency-aware selection across multi-resolution feature streams, allowing the decoder to emphasize informative structures and suppress redundant activations.

Given two input feature maps from the encoder and decoder, denoted as $F_1, F_2 \in R^{C \times H \times W}$, AFM first concatenates them along the channel dimension:

$$F_c = [F_1; F_2] \in R^{2C \times H \times W} \tag{11}$$

A global average pooling is applied to summarize the spatial content of each channel:

$$z = Avg(F_c), z \in R^{2C \times 1 \times 1} \tag{12}$$

Then, a lightweight channel attention network, composed of two fully connected layers and a non-linearity (typically ReLU), is applied to compute a modulation weight vector:

$$w = \sigma\left(W_2 \cdot \delta\left(W_1 \cdot z\right)\right), w \in [0,1]^{2C \times 1 \times 1} \tag{13}$$

where W_1, W_2 are the learnable parameters of the MLP, $\delta(\cdot)$ denotes ReLU, and $\sigma(\cdot)$ is the sigmoid activation.

In feature fusion, the AFM employs modulation weights w to reweight the concatenated feature map F_c, thereby achieving cross-scale adaptive feature selection. Specifically, the low-frequency components from the encoder emphasize scene structures and large-scale semantics, which are essential for restoring the global outline during dehazing, while the high-frequency components from the decoder highlight local textures and edge details, facilitating the recovery of fine structures in degraded regions. This dynamic selection mechanism effectively mitigates the aliasing artifacts commonly observed in conventional skip connections, ensuring a more balanced frequency distribution in the fused features. The final output is formulated as an adaptively weighted representation:

$$F_{out} = \sum_i \alpha_i \cdot F_i \tag{14}$$

where α_i is derived from the weight vector w, representing the network's attention allocation across different frequency branches. By design, the AFM enhances edge responses, leading to stronger activations in prominent structural regions such as near-field contours, which benefits precise spatial reconstruction. Moreover, it provides effective haze compensation by increasing the saliency of low-response areas in heavily degraded regions.

The Adaptive Fusion Module (AFM) integrates complementary features from parallel branches using a lightweight channel-attention mechanism. Unlike frequency-domain operations, AFM works entirely in the spatial domain by extracting global channel descriptors through pooling and generating attention weights via a compact gating network. This formulation enables AFM to emphasize informative channels and suppress degraded ones, providing an accurate balance between global context and local structures.

The AFM is an adaptive fusion module. It integrates low-frequency components to capture global illumination and color gradients that are significantly altered in maritime haze. These low-frequency cues enable AFM to correct large-scale tonal shifts while high-frequency components enhance local texture details, resulting in more balanced and visually consistent dehazing.

After AFM, the integrated features are processed by TDB, which refine the spatial distribution of haze and enhance contextual coherence. These blocks mirror the encoder's TDB design, but operate at progressively higher resolutions, ensuring that haze-relevant features learned in the encoder are fully propagated to the reconstruction stage. By combining long-range self-attention with local convolutional refinement, the decoder simultaneously restores global consistency and fine-grained visual details.

Overall, the hierarchical decoder ensures the gradual recovery of spatial resolution through structured upsampling, and it integrates encoder information through frequency-selective fusion. This combination enables TWRM-Net to restore haze-degraded ship images with natural color balance, sharp edges, and structurally coherent content, even in the presence of dense and spatially varying haze.

D. Physically Inspired Reconstruction and Adaptive Inference Strategy

After feature decoding and fusion, the final output of TWRM-Net is projected back to the image space using a convolution-based PatchUnEmbed operation. This layer converts the tokenized feature map into a four-channel output tensor $Y \in R^{4\times H\times W}$, where the first channel corresponds to a learned transmission coefficient map $K \in R^{1\times H\times W}$, and the remaining three channels form the residual bias term $B \in R^{3\times H\times W}$. Motivated by the classical atmospheric scattering model and previous data-driven enhancements, we adopt a residual-based refinement formulation for dehazing:

$$\hat{I} = K \cdot I - B + I \tag{15}$$

where I is the original hazy input image, and $\hat{I}$ is the predicted haze-free output. This formulation effectively compensates for the attenuation.

This formulation can be interpreted as a learnable residual approximation of the standard atmospheric model

$$I(x) = J(x)\,t(x) + A\,(1 - t(x)),\, t(x) = e^{-\beta d(x)} \tag{16}$$

where the learnable term K approximates the transmission map $t(x)$, and the bias term B compensates for atmospheric light $A(1-t(x))$ and modeling deviations. Unlike traditional explicit estimation of A and $t(x)$, this residual formulation allows the network to implicitly adapt to diverse maritime haze patterns. Eq. (16) acts at the reconstruction stage and constrains the network output to satisfy the residual atmospheric model. Through backpropagation, this physical constraint regularizes the earlier feature extraction modules, ensuring that the learned representations remain compatible with physically plausible dehazing.

To ensure compatibility with arbitrary input sizes during inference, the model incorporates a reflective padding strategy. Given an input image of dimensions $H \times W$, the height and width are adjusted to be divisible by the patch size via reflection padding. The image is then padded along the bottom and right edges with mirrored content, which guarantees that all patch partitioning and reconstruction operations maintain spatial alignment and avoid truncation. After the forward pass, the padding is removed to restore the original resolution.

The reconstructed haze-free prediction I' is supervised by a combination of pixel-level, perceptual, and physics-guided constraints [23]. Specifically, we employ an l_1-based reconstruction loss to encourage faithful recovery of clear images, expressed as

$$\mathcal{L}_{rec} = \| I' - J \|_1 \tag{17}$$

where J denotes the ground-truth haze-free image. To enhance structural and perceptual consistency, we incorporate a perceptual loss computed on VGG feature space, defined as

$$\mathcal{L}_{perc} = \sum\nolimits_l \| \phi_l(I') - \phi_l(J) \|_2^2 \tag{18}$$

where $\phi_l(\cdot)$ denotes the activation of the l-th VGG layer. In addition, the residual reconstruction formulation introduced in Eq.(15) naturally gives rise to a physics-guided consistency loss that enforces adherence to the atmospheric scattering model. This term is written as

$$\mathcal{L}_{phys} = \| I' - (K \cdot I - B + I) \|_1 \tag{19}$$

which regularizes the learning of the transmission coefficient K and the bias term B. The overall training objective integrates these components into a unified loss function:

$$\mathcal{L}_{total} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{phys}\mathcal{L}_{phys} \tag{20}$$

In our experiments, the weighting factors are set to $\lambda_{rec} = 1.0$, $\lambda_{perc} = 0.1$, and $\lambda_{phys} = 0.5$, which achieves a balanced trade-off between pixel accuracy, perceptual fidelity, and physical consistency [24]. This formulation yields stable convergence during training and effectively constrains the model to produce haze-free outputs that are both visually realistic and physically interpretable.

This design ensures that TWRM-Net can operate on high-resolution inputs of arbitrary sizes without architectural modifications, making it practical for real-world deployment

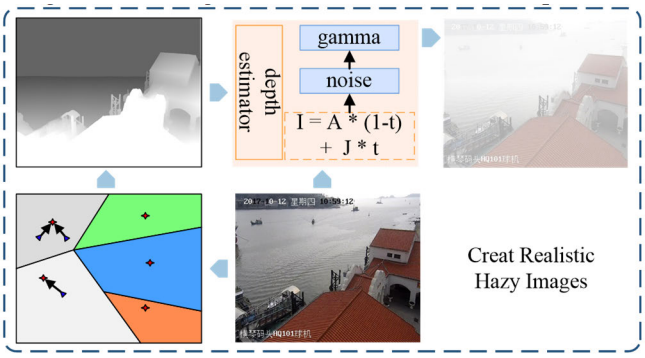


Fig. 4. Realistic hazy image synthesis using depth estimation, atmospheric scattering, and gamma–noise augmentation.

on maritime imagery. The combination of physically inspired reconstruction and flexible input handling makes the network robust and generalizable across diverse haze conditions and aspect ratios.

IV. Experimental Design

A. Data Description

In this study, we selected representative image scenes from the publicly available Singapore dataset and the Seaship dataset as input for our experiments [25]. These selected scenes include typical maritime conditions such as ship occlusions and nighttime maritime environments. The original image resolutions are 1920×1080 and 3840×2160; however, considering the constraints of the network architecture and memory efficiency, all video frames were uniformly resized to 640×480.

The dataset consists of 3,000 paired maritime images and is categorized into five representative scene types: (1) riverside video frames, (2) nighttime riverside video frames, (3) open-sea scenes, (4) bay and port scenes, and (5) nearshore maritime scenes. Riverside scenes include buildings and multiple vessel occlusions, which are used to evaluate the model's ability to recover fine texture details under cluttered conditions, while nighttime riverside images further assess robustness under low-light environments. Open-sea scenes are characterized by ambiguous sea–sky boundaries, posing challenges for large-scale structural restoration. Bay, port, and nearshore scenes typically involve dynamic wave patterns, complex maritime objects, and non-uniform haze distributions, providing a realistic testbed for evaluating multi-layer haze removal and spatial reconstruction performance. The dataset is randomly divided at the image-pair level into training, validation, and testing sets following a standard 80% / 10% / 10% split.

To construct a realistic hazy dataset for supervised training, we adopt a synthesis pipeline based on monocular depth estimation and the atmospheric scattering model (see Figure 4). Specifically, given a clear image, a depth estimation network first predicts the depth map, which provides pixel-wise distance information. To enhance realism, additional degradations such as random gamma correction and Gaussian noise are applied, simulating variations in illumination and sensor imperfections. This process generates diverse hazy scenes with non-uniform distributions that closely resemble real-world maritime environments, providing a reliable benchmark for network training and evaluation.

All experiments were conducted on a workstation equipped with an Intel Core i7-10700K CPU @ 3.80 GHz with 16 physical cores and two NVIDIA RTX A4000 graphics cards. The models were implemented in PyTorch 2.6 and trained under CUDA 11.8 on Ubuntu 20.04. All experiments were performed under the same hardware and software environment for consistency.

B. Evaluation Indicators

In this experiment, we use four metrics to evaluate dehazing performance, namely Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and Frechet Inception Distance (FID). Among these, PSNR is computed based on the Mean Squared Error (MSE) between the predicted and reference images, providing a pixel-level quantitative measurement of reconstruction fidelity [26].

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}(I(i,j) - K(i,j))^2 \tag{21}$$

$$PSNR_{\text{dB}} = 10 \times lg\left(\frac{MAX_I^2}{MSE}\right) \tag{22}$$

Here, m and n denote the number of rows (image height) and columns (image width), respectively. $I(i, j)$ represents the pixel intensity of the original image at position (i, j),and $K(i, j)$ denotes the pixel value of the processed image at position (i, j), while MAX represents the maximum grayscale value among the RGB channels, typically set to 255 for 8-bit images.

The Structural Similarity Index evaluates image similarity based on three components: luminance, contrast, and structure, thereby offering a more perceptually aligned assessment compared to traditional pixel-wise metrics.

$$SSIM(x,y) = \left[l(x,y)^{\alpha} \cdot c(x,y)^{\beta} \cdot s(x,y)^{\gamma}\right] \tag{23}$$

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{24}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{25}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{26}$$

Let x and y denote the pixel values of two image patches to be compared. Specifically, μ_x and μ_y represent the mean intensity values of images x and y, while σ_x and σ_y are their standard deviations, and σ_{xy} is the covariance between x and y, The SSIM index is computed using three constants:

$$c_1 = (k_1 L)^2, c_2 = (k_2 L)^2, c_3 = c_2/2 \tag{27}$$

where L is the dynamic range of the pixel values (typically 255 for 8-bit images), and k_1, k_2 are small positive constants used to stabilize the division in the similarity formulas.

LPIPS is a deep learning-based metric for perceptual image similarity assessment [27]. It operates by extracting deep feature representations from a pair of images using a pre-trained convolutional neural network (such as AlexNet

TABLE I
QUANTITATIVE COMPARISON OF VARIOUS DEBLURRING METHODS TRAINED ON MARITIME DATASETS

Model	Evaluation indicators			
	PSNR↑	SSIM↑	LPIPS↓	FID↓
TWRM-NET	**35.07**	**0.9735**	**0.0880**	**23.65**
ConvIR	31.95	0.9624	0.1060	26.32
DSANet	32.85	0.9653	0.1088	26.45
MixDehazeNet	30.46	0.9519	0.1083	34.84
FFA-Net	26.51	0.9408	0.1853	88.70

or VGG). The similarity is then computed by measuring the distance between corresponding feature maps across several intermediate layers. These distances are typically aggregated via a weighted sum, reflecting the human visual system's sensitivity to content, texture, and structure. A lower LPIPS score indicates that the two images are perceptually more similar. Compared to traditional pixel-wise metrics such as PSNR or SSIM, LPIPS better aligns with human subjective judgment and has been widely adopted in evaluating tasks such as image generation, inpainting, and style transfer.

FID is a distribution-based evaluation metric that quantifies the quality and diversity of generated images [28]. It utilizes a pre-trained Inception-v3 model to extract 2048-dimensional feature vectors from both generated and real images. The FID score is computed based on the Fréchet distance (also known as Wasserstein-2 distance) between the two multivariate Gaussian distributions fitted on these features. Formally, FID considers both the mean difference (reflecting image quality) and the covariance difference (reflecting sample diversity) between the two distributions. A lower FID score indicates that the generated images are statistically closer to real images in terms of both appearance and variation, and is thus preferred in image synthesis evaluations.

$$FID = \| \mu_r - \mu_g \|^2 + Tr(\Sigma_r + \Sigma_g 2 \left(\Sigma_r \Sigma_g\right)^{1/2}) \tag{28}$$

The term $\| \mu_r - \mu_g \|^2$ represents the squared distance between the mean feature vectors of the real and generated image distributions, reflecting the overall quality difference between them. The trace term $Tr(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{1/2})$ corresponds to the covariance-based component of the FID, where Σ_r and Σ_g denote the covariance matrices of real and generated features, respectively. This term captures the structural similarity and dispersion of the distributions, reflecting both the diversity and mode coverage of generated images. The expression $\left(\Sigma_r \Sigma_g\right)^{1/2}$ denotes the matrix square root, typically approximated using numerical methods such as singular value decomposition (SVD), to ensure a mathematically valid and stable estimation of the shared covariance structure.

PSNR and SSIM quantify low-level structural fidelity essential for restoring horizon lines and vessel contours, while LPIPS and FID capture high-level perceptual realism and distributional consistency that are critical in complex maritime haze conditions.

V. EXPERIMENTAL RESULTS

A. Experimental Results About Ocean Dehazing

To comprehensively evaluate the performance of our proposed method, we compare it with four state-of-the-art image dehazing algorithms: FFA-Net, MixDehazeNet, DSANet, and ConvIR. These methods represent diverse architectural paradigms. FFA-Net employs a combination of channel and pixel attention modules alongside local residual learning and multi-level feature fusion, excelling at modeling spatially non-uniform haze in a purely convolutional framework [15]. MixDehazeNet integrates large-kernel multi-scale convolutions with parallel attention mechanisms to effectively balance receptive field size and detail preservation, offering strong performance on haze removal [26]. DSANet introduces a dual-domain strip attention mechanism that jointly captures spatial and frequency domain dependencies, using directional aggregation and frequency modulation to enhance both contextual understanding and computational efficiency [29]. ConvIR utilizes a lightweight four-branch convolutional architecture with progressive resolution reduction to process multi-scale features while maintaining low complexity. These comparison models span attention-based CNNs, multi-scale hybrid designs, dual-domain mechanisms, and efficient convolutional schemes. By benchmarking against them, we demonstrate not only the superior performance of our model but also its generalization across varied haze modeling assumptions. Quantitative results are reported in Table I.

By introducing a dual residual attention mechanism and a modulation fusion module, the proposed TWRM-NET model exhibits remarkable advantages in the image dehazing task. Compared to existing state-of-the-art methods—including ConvIR, DSANet, MixDehazeNet, and FFA-Net—TWRM-NET achieves superior performance across all four key evaluation metrics: PSNR, SSIM, LPIPS, and FID. Specifically, TWRM-NET attains a PSNR of 35.07 dB and an SSIM of 0.9735, significantly outperforming other methods such as DSANet, which yields approximately 32.85 dB in PSNR and 0.9653 in SSIM. These improvements highlight TWRM-NET's enhanced capability in recovering fine textures and preserving structural details, resulting in more faithful and perceptually consistent haze-free reconstructions.

In addition, through the comparison of LPIPS and FID indicators, TWRM-NET also shows significant advantages in perceptual consistency and visual authenticity.

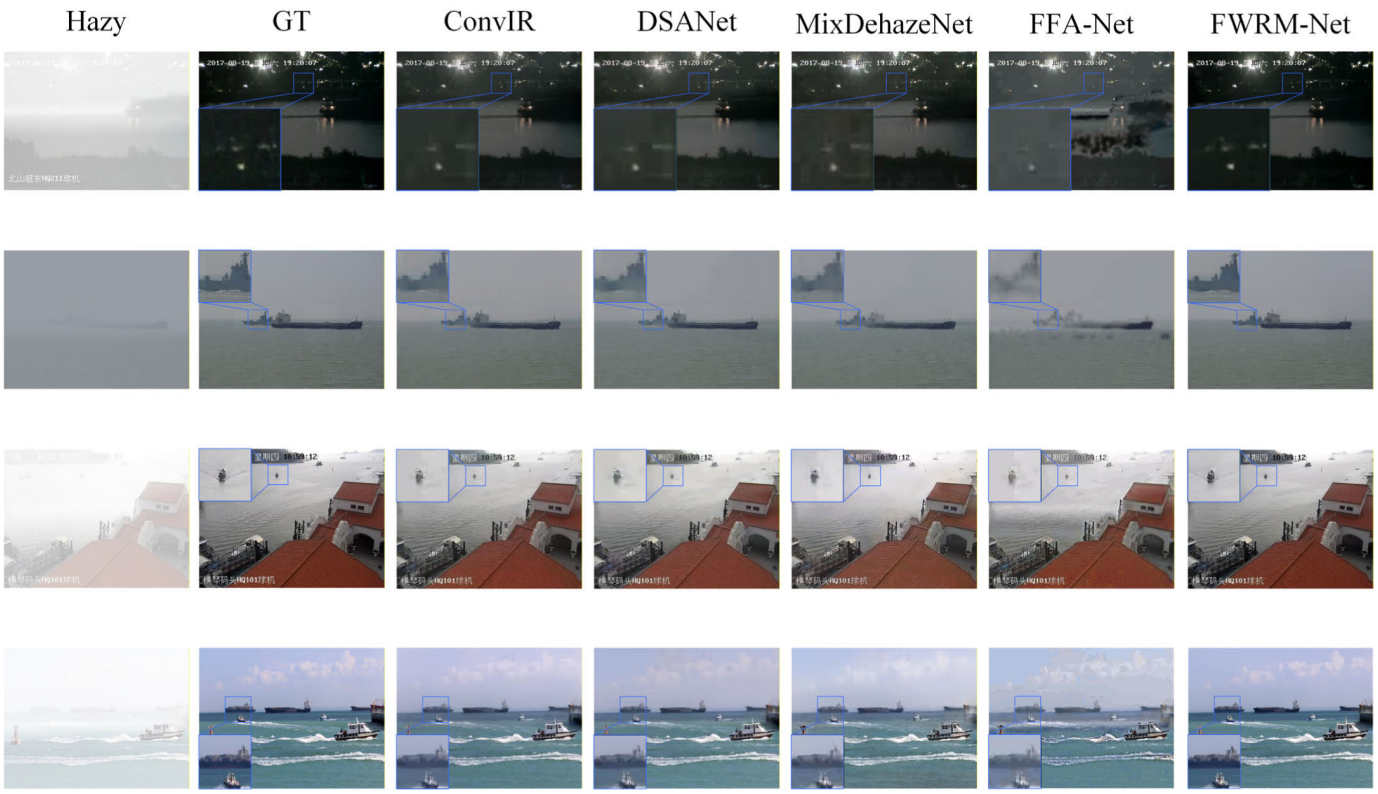


Fig. 5. Qualitative comparison of dehazing performance on maritime scenes using different methods.

Specifically, TWRM-NET achieves an LPIPS score of approximately 0.0880, which is significantly lower than that of other models such as ConvIR (around 0.1060). This indicates that the dehazed images generated by TWRM-NET are perceptually closest to the ground truth, exhibiting the highest level of perceptual consistency. Even more notably, TWRM-NET attains a FID score of only 23.65, substantially outperforming competing methods, with the second-best being around 26.32, and traditional models like FFA-Net reaching as high as 88.70. This large margin suggests that the distribution of TWRM-NET's output images is the most aligned with that of real haze-free images, reflecting superior visual authenticity.

The LPIPS and FID metrics reveal distinct aspects of model performance at local and global levels, respectively. LPIPS primarily assesses the perceptual similarity of local image regions, demonstrating heightened sensitivity to fine-grained textual and structural variations while exhibiting considerable robustness against global color shifts and illumination changes. In contrast, FID quantifies the distance between generated and real image distributions in a deep feature space, thereby reflecting macroscopic distributional consistency.

Although MixDehazeNet achieves a level of local texture restoration comparable to DSANet, resulting in similar LPIPS scores, it introduces noticeable deficiencies in global color fidelity and illumination consistency. These global distortions, while not adequately captured by the LPIPS metric, substantially widen the discrepancy between the generated images and the true haze-free image distribution at a holistic level, which is directly manifested in its significantly elevated FID score. This evidence indicates that MixDehazeNet remains inadequate in preserving the overall distributional characteristics of natural images.

B. Qualitative Comparison

The proposed TWRM-NET model exhibits superior dehazing capability across various challenging scenarios, significantly outperforming competing methods (see Figure 5). Whether in bright nighttime scenes, low-contrast distant sea surfaces, or complex near-field environments with diverse lighting and color conditions, TWRM-NET consistently restores natural colors and structural details. In the first row, depicting a nighttime riverside scene, TWRM-NET accurately preserves the relationship between water reflections and bright point light sources, avoiding the over-enhancement and detail loss issues observed in DSANet and FFA-Net. In the fourth row sea surface image, TWRM-NET effectively removes haze and reconstructs clear ship contours and water tones, whereas MixDehazeNet and ConvIR still suffer from residual fog and blurring. On the shore scene of the red roof (third row), TWRM-NET produces sharp edges and maintains color fidelity, unlike FFA-Net and DSANet, which introduce color distortions and regional blur. In the second row, TWRM-NET successfully reconstructs the sea-sky boundary and fine textures on the water surface, achieving results that are not only perceptually close to the ground truth but also visually aligned with human perception. These observations highlight the advantages of the dual residual attention mechanism in

TABLE II

QUANTITATIVE COMPARISON OF VARIOUS DEBLURRING METHODS TRAINED ON MARITIME DATASETS

Model	Evaluation indicators			
	PSNR↑	SSIM↑	LPIPS↓	FID↓
TWRM-Net	35.07	0.9735	0.0880	23.65
-AFM	33.83	0.9729	0.0902	25.75
-DRAB-AFM	33.28	0.9717	0.0910	26.86
-2*DRAB-AFM	33.01	0.9707	0.0911	29.24

enhancing multi-scale feature selection, and demonstrate the effectiveness of the modulation fusion module in handling complex and spatially varied haze. As a result, TWRM-NET consistently generates natural, sharp, and highly realistic dehazed images across a wide range of hazy scenes.

The computational complexity and parameter count of the proposed model were evaluated using a standard FLOPs profiling tool. Under an input resolution of 640×480, the model achieves approximately 70 GFLOPs for a single forward pass, with a total parameter count of around 29 million. This level of computational cost is moderate for contemporary image restoration networks and remains feasible for deployment on high-performance edge devices or embedded GPU platforms. The model maintains a good balance between restoration accuracy and computational efficiency, making it suitable for real-time or near–real-time maritime perception applications.

C. Quantitative Comparison

To thoroughly evaluate the effectiveness of the proposed TWRM-Net, we conducted an extensive quantitative comparison on maritime image datasets using four evaluation metrics: LPIPS, FID, PSNR, and SSIM (see Figure 6). The results demonstrate that TWRM-Net consistently achieves the lowest LPIPS and FID scores among all compared methods, indicating its superior performance in generating dehazed images with enhanced perceptual quality and a closer distribution alignment to real haze-free images. Specifically, its notable advantage in the LPIPS metric underscores its exceptional capability in preserving fine-grained texture details, while the reduced FID values highlight its proficiency in producing structurally realistic outputs that adhere to natural image statistics.

Furthermore, TWRM-Net surpasses all baseline models in terms of both PSNR and SSIM, attaining higher reconstruction accuracy and superior structural fidelity. In contrast to traditional CNN-based models like FFA-Net and DSANet, which exhibit sharp performance fluctuations and significant degradation in complex scenes, TWRM-Net maintains highly stable and superior performance across all test samples. This remarkable robustness provides further validation for the significant advantages of our proposed dual-residual attention mechanism and modulation fusion strategy in handling diverse haze conditions and challenging maritime scenarios.

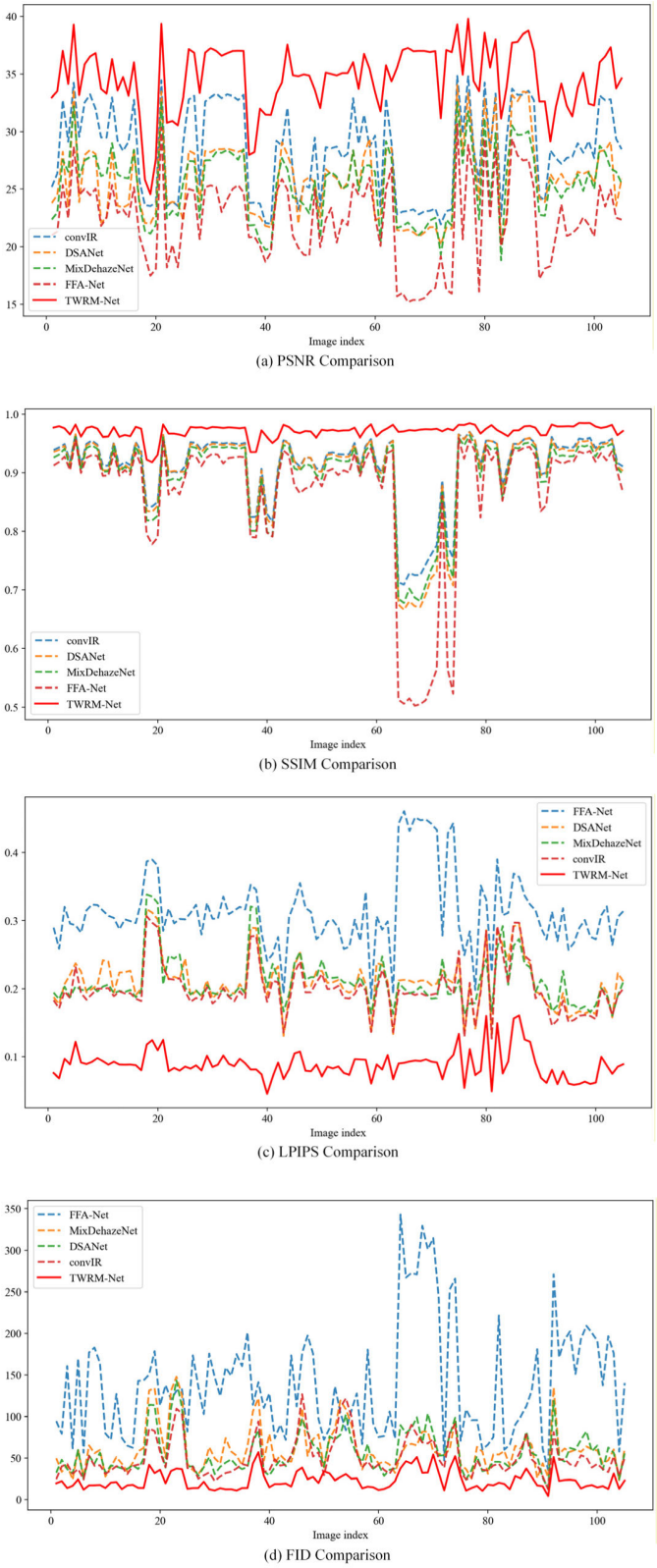


Fig. 6. Quantitative evaluation on maritime images using four metrics: (a) PSNR, (b) SSIM, (c) LPIPS, and (d) FID.

D. Ablation Experiments

To systematically validate the effectiveness of the proposed network architecture, we first trained five distinct network models in the ablation study. These models were meticulously designed to isolate the contributions of key components, specifically including: a single-layer dual residual attention

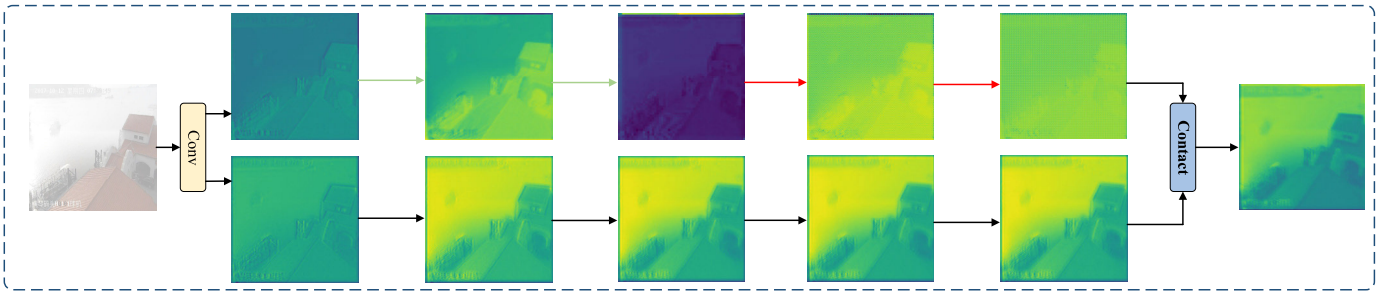


Fig. 7. Visualization of DRAB feature maps where GAM highlights haze localization and DEM captures textures and structures.

network, a two-layer dual-residual attention block, a TWRM-Net variant employing only the adaptive fusion module, a TWRM-Net variant without the residual attention network, and the complete TWRM-Net model. This progressive architectural comparison enables precise evaluation of each module's contribution to overall performance. Comprehensive quantitative results are documented in Table II.

1) Dual-Residual Attention Block (DRAB): As shown in the table, the performance of TWRM-Net improves across all evaluation metrics after incorporating DRAB. Moreover, the enhancement becomes more pronounced as the number of DRAB layers increases. The original TWRM-Net achieves a PSNR of 33.01 and an SSIM of 0.9707, with LPIPS and FID scores of 0.0911 and 29.24, respectively. After introducing a single DRA block, PSNR increases to 33.28 and SSIM rises to 0.9717, while the perceptual metric LPIPS decreases to 0.0910 and FID drops to 26.86, indicating improvements in both image quality and perceptual consistency. When stacking two DRA blocks, PSNR further improves to 33.83, SSIM reaches 0.9729, LPIPS reduces to 0.0902, and FID decreases to 25.75, demonstrating enhanced capability in detail modeling and structural restoration, with more realistic and natural output. Overall, DRAB mechanism effectively strengthens feature representation, and the consistent performance gains with additional layers highlight the module's scalability and practical value.

We illustrate the intermediate feature maps generated within DRAB (see Figure 7). The upper stream corresponds to the GAM, which focuses on position-aware representation by enhancing haze-concentrated regions and edge boundaries, thereby providing spatial cues for haze localization. The lower stream corresponds to the DEM, which emphasizes content-aware representation, capturing fine-grained textures, structural contours, and semantic details. As features are progressively refined through multiple stacked blocks, the two streams complement each other, and their concatenation produces a unified representation that integrates both spatial localization and semantic restoration. This joint design enables the network to generate dehazed outputs that are not only structurally accurate but also perceptually natural.

2) Adaptive Fusion Module (AFM): From the table results, it is evident that incorporating AFM into the TWRM-Net with two-layer dual-residual attention blocks (2×DRAB) yields a significant performance boost. The PSNR improves from 33.83 to 35.07, indicating enhanced reconstruction accuracy,

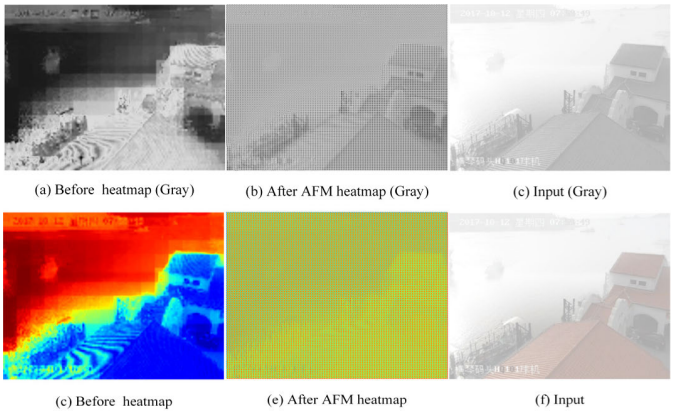


Fig. 8. Visualization of AFM feature modulation showing clearer haze edges in grayscale input maps and more compact informative responses in original input images.

while SSIM rises to 0.9735, reflecting more complete structural recovery. Additionally, the LPIPS decreases to 0.0880, suggesting higher perceptual similarity to the ground truth, and the FID drops to 23.65, outperforming both the original model (29.24) and the 2×DRAB variant (25.75), indicating more realistic and distributionally aligned outputs. AFM achieves this by introducing cross-branch feature guidance and attention-based weighting, which enhances the integration of spatial and semantic information across different scales. This not only mitigates redundancy and bias introduced by attention stacking but also enables more effective multi-scale feature collaboration. Overall, the inclusion of AFM significantly strengthens the model's representation and fusion capabilities, leading to a notable improvement in dehazing performance and confirming its effectiveness as a structural enhancement to the TWRM-Net architecture.

We illustrate the effect of AFM on feature modulation using both grayscale input images and the original input images (see Figure 8). For the grayscale inputs, before AFM (a) the activations appear scattered and noisy, where haze-related structures are mixed with redundant responses. After AFM (b), salient regions such as haze edges and object contours become clearer, while irrelevant background responses are effectively suppressed. The corresponding grayscale hazy input is shown in (c). For the original input images, the heatmap before AFM (d) presents dispersed and noise-like responses, whereas after AFM (e) the activations are more compact and concentrated on informative structures. The original hazy image is provided

in (f) for reference. These results demonstrate that AFM selectively modulates features for both grayscale and original inputs, suppressing irrelevant responses while enhancing discriminative components, thereby ensuring global structural consistency and sharper local details under dense and non-uniform haze conditions.

VI. Conclusion

In this study, we addressed the critical challenge of haze degradation in maritime surveillance imagery and proposed a Transformer-based dehazing framework, TWRM-Net, tailored for intelligent maritime transportation. The network integrates shallow feature extraction, a multi-stage Transformer encoder for global semantic modeling, and a cross-scale feature enhancement module. Furthermore, the incorporation of a Dual-Residual Attention Block (DRAB) and an Adaptive Fusion Module (AFM) significantly improves haze-region awareness and edge detail recovery. Through extensive evaluations across nearshore, offshore, and night-time fog scenarios, TWRM-Net has demonstrated superior performance with a PSNR of 35.07 dB, SSIM of 0.9735, LPIPS of 0.0880, and FID of 23.65, consistently outperforming existing methods. These results validate that the proposed framework achieves both structural accuracy and perceptual realism, demonstrating consistent performance across varied maritime haze conditions.

TWRM-Net offers significant practical value for real-world maritime transportation. Its ability to restore visibility and structural detail under dense, non-uniform haze provides reliable perception for tasks such as ship detection, collision avoidance, and autonomous navigation. Due to its modular hybrid architecture combining convolutional and Transformer mechanisms, TWRM-Net can be readily adapted for onboard and edge deployment, ensuring real-time operation in intelligent maritime transportation systems. We plan to extend TWRM-Net toward cross-weather generalization by incorporating domain adaptation techniques to handle adverse maritime conditions such as fog, rain, and sea spray. Furthermore, lightweight model compression and optimization strategies will be explored to facilitate efficient deployment on shipborne embedded devices.

While TWRM-Net effectively handles haze-induced degradation, its current design is grounded in the atmospheric scattering model and therefore does not explicitly account for other adverse maritime weather conditions such as rain streaks, sea spray droplets, or snow scatter. These degradations exhibit different physical characteristics from haze, including occlusion-like artifacts and directional streak patterns. As a result, the present model may not generalize optimally to such conditions without retraining or additional weather-specific modules. Future work will explore cross-weather training, domain adaptation strategies, and joint degradation modeling to enhance robustness under mixed and multi-type maritime weather.

In future work, we plan to further extend the framework to address other adverse weather conditions, such as rain, snow, and sea spray, and to explore domain adaptation techniques for cross-regional generalization. We will also investigate lightweight model compression and deployment on edge devices to facilitate real-time maritime monitoring and onboard navigation systems.

References

[1] N. P. Ventikos and K. Louzis, "Risk dynamics for marine systems: Towards a bio-inspired framework for dynamic risk assessment," *Transp. Saf. Environ.*, vol. 4, no. 3, Sep. 2022, Art. no. tdac018.

[2] B. Goyal et al., "Recent advances in image dehazing: Formal analysis to automated approaches," *Inf. Fusion*, vol. 104, Apr. 2024, Art. no. 102151.

[3] D. Qiao, X. Kong, L. Kong, J. Liu, W. Mi, and S. Meng, "Prior-combined dehazing network based on mutual learning," *Signal, Image Video Process.*, vol. 17, no. 5, pp. 1935–1943, Jul. 2023.

[4] H. Wu et al., "Contrastive learning for compact single image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10546–10555.

[5] S. Zhang, L. Zhao, K. Hu, S. Feng, E. Fan, and L. Zhao, "Deep guided transformer dehazing network," *Sci. Rep.*, vol. 13, no. 1, p. 15333, Sep. 2023.

[6] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[7] H. Xu, J. Dong, W. Hu, R. Su, J. Xue, and S. Dong, "Vision transformer with pyramid feature extraction and efficient channel attention for classifying images," in *Proc. IEEE 14th Data Driven Control Learn. Syst. (DDCLS)*, May 2025, pp. 961–966.

[8] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5718–5729.

[9] Y. Li et al., "Efficient predictive control strategy for mitigating the overlap of EV charging demand and residential load based on distributed renewable energy," *Renew. Energy*, vol. 240, Feb. 2025, Art. no. 122154.

[10] S. Liu, X. Wang, M. Weiszer, and J. Chen, "Extracting multi-objective multigraph features for the shortest path cost prediction: Statistics-based or learning-based?," *Green Energy Intell. Transp.*, vol. 3, no. 1, Feb. 2024, Art. no. 100129.

[11] Y. Su, N. Wang, Z. Cui, Y. Cai, C. He, and A. Li, "Real scene single image dehazing network with multi-prior guidance and domain transfer," *IEEE Trans. Multimedia*, vol. 27, pp. 5492–5506, 2025.

[12] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 617–624.

[13] J. Tong, "AOD-Net: A lightweight real-time fruit detection algorithm for agricultural automation," *J. Food Meas. Characterization*, vol. 19, no. 4, pp. 2818–2830, Apr. 2025.

[14] X. Zhang, T. Wang, W. Luo, and P. Huang, "Multi-level fusion and attention-guided CNN for image dehazing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4162–4173, Nov. 2021.

[15] Q. Xu, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-Net: Feature fusion attention network for single image dehazing," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11908–11915.

[16] H. Zhou, W. Dong, Y. Liu, and J. Chen, "Breaking through the haze: An advanced non-homogeneous dehazing method based on fast Fourier convolution and ConvNeXt," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 1895–1904.

[17] Z. Chen, Z. He, and Z.-M. Lu, "DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention," *IEEE Trans. Image Process.*, vol. 33, pp. 1002–1015, 2024.

[18] R. Zhu, Z. Tu, J. Liu, A. C. Bovik, and Y. Fan, "MWFormer: Multi-weather image restoration using degradation-aware transformers," *IEEE Trans. Image Process.*, vol. 33, pp. 6790–6805, 2024.

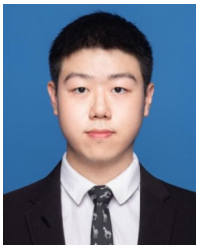
[19] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17662–17672.

[20] X. Song, J. Dai, Z. Fu, and D. Tan, "Optimization improvement of AECR-NET image dehazing network," in *Proc. IEEE 6th Int. Conf. Power, Intell. Comput. Syst. (ICPICS)*, Jul. 2024, pp. 1428–1436.

[21] X. Chen, Z. Wang, Q. Hua, W.-L. Shang, Q. Luo, and K. Yu, "AI-empowered speed extraction via port-like videos for vehicular trajectory analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 4, pp. 4541–4552, Apr. 2023.

[22] H. Yu, H. Sun, and Z. Pu, "GVDF-PIA: Group-based vehicular digital forensics and proxy-assisted integrity auditing for intelligent transportation," *IEEE Internet Things J.*, vol. 12, no. 24, pp. 55645–55659, Dec. 2025.
[23] X. Chen, Q. Ma, H. Wu, W. Shang, B. Han, and S. A. Biancardo, "Autonomous port traffic safety orientated vehicle kinematic information exploitation via port-like videos," *Transp. Saf. Environ.*, vol. 7, no. 3, p. 048, Oct. 2025.
[24] D. Wang and J. Yan, "Ship collision risk analysis in port waters integrating GRA algorithm and BPNN," *Transp. Saf. Environ.*, vol. 7, no. 1, 2025, Art. no. tdaf012.
[25] D. K. Prasad, D. Rajan, L. Rachmawati, E. Rajabally, and C. Quek, "Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 8, pp. 1993–2016, Aug. 2017.
[26] L. Lu, Q. Xiong, B. Xu, and D. Chu, "MixDehazeNet: Mix structure block for image dehazing network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2024, pp. 1–10.
[27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," presented at the *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–11.
[29] Y. Cui and A. Knoll, "Dual-domain strip attention for image restoration," *Neural Netw.*, vol. 171, pp. 429–439, Mar. 2024.

Xinqiang Chen (Member, IEEE) received the Ph.D. degree from Shanghai Maritime University, China, in 2018. From September 2015 to September 2016, he was a Visiting Student at the Smart Transportation Applications and Research Laboratory, University of Washington, USA. His research interests include transportation image processing and smart ship and maritime traffic situation awareness.



Zhengang Xin is currently pursuing the M.S. degree with the Institute of Logistics Science and Engineering, Shanghai Maritime University, China. His research interests include computer vision and cooperative vehicle infrastructure systems.

Han Zhang is currently pursuing the Ph.D. degree with the College of Ocean Science and Engineering, Shanghai Maritime University, China. His research interests include computer neural networks, big data of traffic information, and computer vision.

Yuzhen Wu received the B.S. degree in thermal power engineering from Shanghai Jiao Tong University, China. He is currently the Executive Vice General Manager of SPG Qingdao Port Group Company Ltd. His research interests include situation awareness, collision avoidance, vehicle position, and computer vision.

Chenxin Wei is currently pursuing the Ph.D. degree with the School of Transportation, Southeast University, China. His research interests include computer vision, traffic image processing, and intelligent transportation systems.

Octavian Postolache (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the Gheorghe Asachi Technical University of Iaşi, Iaşi, Romania, in 1999, and the Habilitation degree from the University of Lisbon, Lisbon, Portugal, in 2016. He is currently a Full Professor with the Iscte—Instituto Universitário de Lisboa and a Senior Researcher with the Instituto de Telecomunicações, Lisbon. He serves as a coordinator or a member of national and international research projects, particularly Portuguese and EU projects. He has authored or co-authored ten patents, ten books, 18 book chapters, and 400 papers in peer-reviewed international journals and conference proceedings. He has been listed among the top 2% of the Stanford–Elsevier most-cited scientists worldwide for five consecutive years.