Visual Intelligence

RESEARCH Open Access

DehazeMamba: large multi-modal model guided single image dehazing via mamba

Ruikun Zhang[1], Zhiyuan Yang[1] and Liyuan Pan[1*]

Abstract

Deep neural networks have achieved significant success in image dehazing. However, existing backbones face an irreconcilable trade-off between the global receptive field and computational efficiency, hindering further applications. State space models, such as Mamba, offer a potential solution to this conflict by modeling long-range dependencies with linear complexity. Although Mamba is well-suited for sequential tasks (e.g., natural language processing), it still encounters challenges when applied to low-level vision tasks. In this work, we propose a large multi-modal model (LMM) guided, Mamba-based image dehazing method (DehazeMamba). It enhances the standard Mamba architecture by incorporating image quality priors provided by the LMM and a channel attention mechanism. Additionally, we present a challenging image dehazing dataset and conduct new benchmark studies based on the LMM, evaluating hazy images and dehazing results by simulating human perception. Our experimental results demonstrate that our dataset exhibits superior haze quality, and our method outperforms current state-of-the-art (SOTA) dehazing methods by achieving a performance improvement of more than 5% on both the O-Haze and Dense-Haze datasets.

Keywords: Image dehazing, State space model, Large multi-modal model (LMM), Mamba

1 Introduction

The goal of image dehazing is to restore clean images from hazy images, which is crucial for both computational photography and various advanced applications [1–3].

In recent years, dehazing methods based on deep neural networks have shown remarkable advancements [4–6]. Typically, hazy images are generated via the atmospheric scattering model (ASM) [7], which can be formulated as

$$\begin{aligned} \boldsymbol{I}(x) &= \boldsymbol{J}(x)\boldsymbol{T}(x) + \boldsymbol{A}(1-\boldsymbol{T}(x)), \\ \boldsymbol{T}(x) &= \mathrm{e}^{-\beta \boldsymbol{D}(x)}, \end{aligned} \tag{1}$$

where $\boldsymbol{I}$, $\boldsymbol{J}$, $\boldsymbol{A}$, $\boldsymbol{T}$, and $\boldsymbol{D}$ represent the hazy image, the haze-free image, the global atmospheric light intensity, the medium transmission map, and the depth map, respectively. The variable x refers to the pixel location, and β denotes the atmospheric scattering coefficient.

To some degree, the improvement in the performance of deep neural networks for image dehazing is attributed primarily to the expansion of the receptive field [8]. A larger receptive field allows the network to extract information from a broader image region, abstracting higher-level structures and features. This enables the network to retain the original image structure while utilizing more pixel information to reconstruct images, leading to better restoration results. Despite many advantages, current image dehazing backbones face an irreconcilable trade-off between receptive field size and computational efficiency.

Image dehazing methods based on convolutional neural networks (CNNs) [9, 10] can be deployed with relatively low computational costs because of the efficiency of convolutional parallel operations. However, their limited receptive field size restricts performance to some extent. In contrast, Transformer-based image dehazing methods typically have a global receptive field, but since the atten-

*Correspondence: liyuan.pan@bit.edu.cn
[1]School of Computer Science and Technology, Beijing Institute of Technology (BIT), Beijing, China

Springer

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

tion mechanism requires interaction between all tokens in the sequence, directly using the standard Transformer [11] results in quadratic computational complexity. Even when employing efficient attention techniques (e.g., shifted window attention [12]) for image dehazing, a global effective receptive field is often sacrificed, and a trade-off between receptive field size and computational efficiency still exists.

State space models (SSMs), i.e., Mamba [13], have emerged as an efficient alternative backbone for deep networks [13–17]. Specifically, the state space equations in Mamba are capable of modeling long-range dependencies with linear complexity, offering a feasible solution to the aforementioned trade-off. However, the standard Mamba is designed for sequential tasks (e.g., natural language processing [14, 16, 17]) and still faces challenges when applied to low-level vision tasks, i.e., local pixel forgetting [18].

In recent years, pre-trained large multi-modal models (LMMs) have garnered significant attention from researchers because of their powerful image perception capacity and generalizability [19–22]. However, exploration of their application to image dehazing remains limited. In this paper, we propose DehazeMamba, which aims to combine LMM with the image dehazing task. We leverage the LMM to obtain the quality prior of hazy images and design a dehazing network based on the Mamba architecture. Our DehazeMamba has three key components: 1) the Mamba group, 2) the Mamba block, and 3) the LMM guided image quality priors. Specifically: 1) We replace the standard feedforward network in Transformer with Mamba blocks, reducing computational complexity while maintaining a global receptive field, thereby improving the restoration performance and adaptability to complex scenes. 2) By designing Mamba blocks based on a 2D-selective scan module (2D-SSM), we adapt Mamba to the image dehazing task. 3) We utilize the LMM to obtain image quality priors, and then fuse the priors with the features extracted by the Mamba blocks, which enhances the model's ability to perceive local details and mitigates the issue of local pixel forgetting.

We conduct a comprehensive comparison with existing state-of-the-art (SOTA) dehazing methods. In addition to commonly used metrics, such as peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM), we also employ an LMM-based evaluation metric to obtain results that more closely align with human visual perception. Furthermore, we assess the performance of SOTA methods across various downstream tasks, highlighting their practical applicability. Notably, our findings indicate that the SOTA methods face challenges in achieving satisfactory results on datasets with high-intensity haze conditions. Therefore, we introduce a new dehazing dataset, EDHaze, to evaluate the performance of SOTA methods in challenging scenarios.

Our contributions can be summarized as follows:

1) We propose DehazeMamba, an LMM guided, Mamba-based image dehazing method. Our extensive experiments demonstrate the competitive performance of DehazeMamba in both qualitative and quantitative evaluations, including high-level vision tasks, such as segmentation, detection, and captioning.
2) We introduce a challenging image dehazing dataset, EDHaze, and conduct a new benchmark study on real-world dehazing datasets based on the LMM, evaluating hazy images and dehazing results by simulating human perception.

2 Related work

Image dehazing Early image dehazing methods were predominantly based on prior knowledge and relied on manually designed rules, utilizing statistical differences between hazy and haze-free images to remove haze. For example, He et al.[23] introduced the dark channel prior (DCP), a widely recognized method that became foundational in the field of image dehazing. Numerous studies have refined and extended the application of DCP, explored its limitations, and enhanced its effectiveness across different scenarios [24–26]. Zhu et al.[27] proposed the color attenuation prior (CAP), which exploited the relationship between pixel brightness and saturation to estimate the thickness of haze. Similarly, Berman et al.[28] proposed a non-local prior based on their observation that colors in a haze-free image form tight clusters in RGB space. However, their performance tends to degrade in real-world scenarios that deviate from their original assumptions [6, 29–37].

In recent years, models based on CNNs have achieved significant success in low-level vision tasks [4–6, 38–44]. For example, Ren et al. [42] introduced a multi-scale deep neural network for single image dehazing, which combined a coarse-scale network for global transmission map prediction with a fine-scale network for local refinement. Jiang et al. [39] designed PCNet, a progressive coupled network with a coupled representation module (CRM) that enabled efficient image deraining and enhanced performance in downstream vision tasks. Ren et al. [43] developed an efficient end-to-end neural network for image dehazing, leveraging a fusion-based strategy that integrated white balance, contrast enhancing, and Gamma correction, along with a multi-scale approach to mitigate halo artifacts. Jiang et al. [44] introduced a hierarchical dense connection network (HDN) for image super-resolution, incorporating hierarchical dense residual blocks to improve feature representation while optimizing memory usage.

With the emergence of the Transformer architecture [11] in natural language processing, attention mechanisms have been adapted for dehazing tasks, with promising results [45–50]. The integration of different backbones has also become a key trend. For instance, Xiao et al. [40]

designed the top-k token selective Transformer (TTST), which integrated a residual token selective group (RTSG) to filter out irrelevant tokens and a multi-scale feedforward layer (MFL) to enhance feature aggregation, leading to more efficient and effective image super-resolution. Chen et al.[50] incorporated convolution operations with attention mechanisms to address the inherent limitations of standard CNN architectures, thereby strengthening their ability to focus on relevant features. Deng et al.[29] applied a multi-model fusion network that integrated multiple dehazing models, enabling better adaptation to diverse haze conditions. In our work, we integrate Mamba [13] into the standard Transformer backbone, reducing computational complexity while preserving a global receptive field, which enhances the restoration performance and adaptability to complex scenarios.

State space model The development of SSM has its origins in classical control theory [51, 52]. SSM is initially used in engineering for tasks, such as signal processing [53], navigation [54], and control [55–57], with the Kalman filter [58] being one of the most notable applications. Over time, the framework of the SSM was extended to accommodate more complex and nonlinear systems, i.e., the extended Kalman filter [59] and the unscented Kalman filter [60]. These advancements enabled the integration of Kalman filters into neural networks, enhancing learning and prediction in dynamic systems [61, 62].

In recent years, SSM has gained increasing attention in deep learning, particularly because of its ability to model long-range dependencies in sequential data. The state space sequence model (S4) [15, 63] pioneers improvements in the efficiency and scalability of feature extraction from sequential data. Models based on S4 have demonstrated exceptional performance across various domains, including natural language processing [14, 16, 17] and computer vision [64, 65].

Further innovations have expanded the application of SSM across multiple fields [66–70]. The Mamba model [13] extends the capabilities of SSM by incorporating data-dependent and selective mechanisms along with hardware optimizations, significantly enhancing their efficiency in processing large-scale datasets. For instance, Xiao et al. [41] introduced the FMSR framework for remote sensing image super-resolution, which integrated the vision state space model (Mamba) to efficiently capture long-range dependencies, using multi-level fusion and spatial-frequency correlations for improved performance with reduced computational complexity. These advancements have proven transformative not only in natural language processing [15] but also in computer vision tasks, including image classification [71], object detection [72], and image restoration [18, 73]. We replace the standard Transformer feedforward network with Mamba blocks, reducing computational complexity while maintaining a global receptive field.

Image quality assessment Image quality assessment (IQA) focuses on how distortions in images affect human perception. Early IQA methods typically relied on manual annotations and were evaluated via statistical prior knowledge [74, 75]. However, as image distortion types diversify and the complex of visual content increases, data-driven end-to-end neural networks have become increasingly prevalent in the IQA domain, exemplified by models such as NIMA [76], DBCNN [77], and HyperIQA [78]. Despite their success, the performance of these methods significantly deteriorates when the style of test images shifts, revealing limitations in their generalizability. In contrast, the LMM provides robust background knowledge across a wide spectrum of visual and linguistic tasks. These models not only understand high-level visual content [79, 80] but also effectively perceive low-level visual attributes [81] and possessed reasoning skills enabled by their advanced language decoders [82]. These strengths establish a solid foundation for creating more accurate and stable visual evaluators. We employ an LMM-based evaluation metric to assess hazy images and dehazing results, providing an evaluation that not only captures the semantic quality of images but also better aligns with human visual perception.

3 Method

3.1 Pre-liminaries

Recent developments in state-space models have been significantly influenced by continuous linear time-invariant systems. These systems transform a one-dimensional sequence or function $x(s) \in \mathbb{R}$ into an output $y(s) \in \mathbb{R}$ via an implicit latent state $h(s) \in \mathbb{R}^N$, where s denotes current time. This transformation can be formally described as a linear ordinary differential equation:

$$\begin{aligned} h'(s) &= \boldsymbol{A}h(s) + \boldsymbol{B}x(s)\,, \\ y(s) &= \boldsymbol{C}h(s) + \boldsymbol{D}x(s)\,, \end{aligned} \tag{2}$$

where N is the size of the states, $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, $\boldsymbol{B} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{C} \in \mathbb{R}^{1\times N}$, and $\boldsymbol{D} \in \mathbb{R}$, denoting the model parameter matrices.

Following this, a discretization process is generally applied to integrate Eq. (2) into deep learning networks. Specifically, let Δ represent the timescale parameter used to convert the continuous parameters $\boldsymbol{A}$ and $\boldsymbol{B}$ into discrete counterparts $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{B}}$. A widely used approach for discretization is the zero-order hold (ZOH) rule, defined as

$$\begin{aligned} \bar{\boldsymbol{A}} &= \exp(\Delta \boldsymbol{A})\,, \\ \bar{\boldsymbol{B}} &= (\Delta \boldsymbol{A})^{-1}(\exp(\boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta \boldsymbol{B}\,, \end{aligned} \tag{3}$$

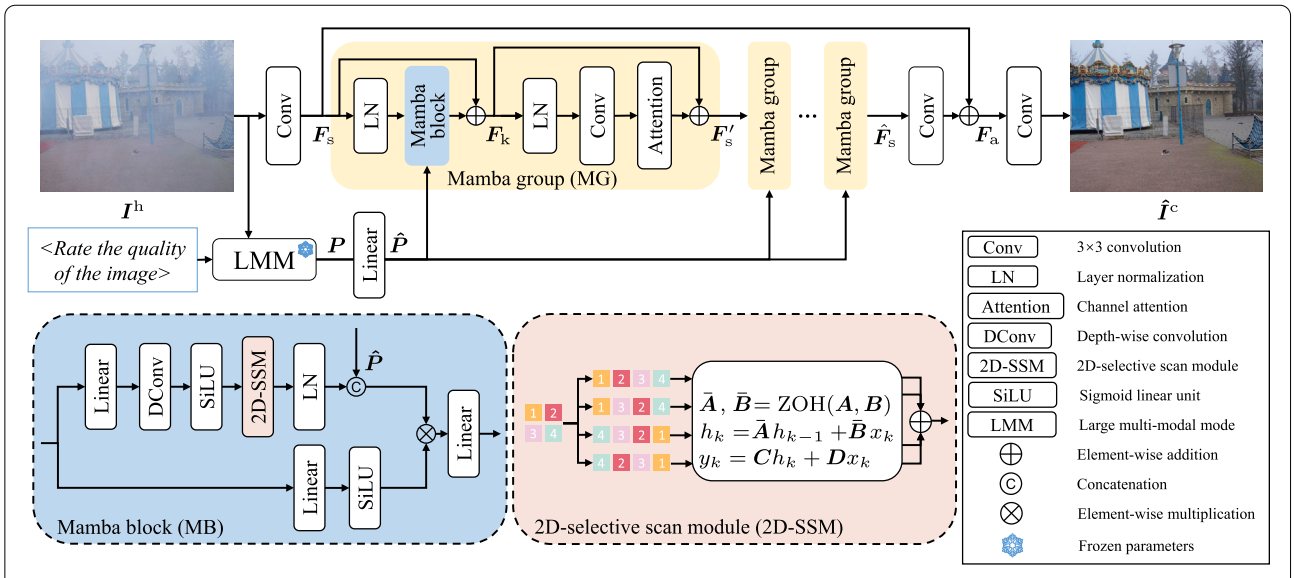


Figure 1 Overview of our DehazeMamba framework. Given a hazy image $\boldsymbol{I}^{\mathrm{h}}$, we first feed the text prompt and the hazy image $\boldsymbol{I}^{\mathrm{h}}$ into the large multi-modal model (LMM) to query the image probability distribution $\boldsymbol{P}$. $\boldsymbol{P}$ is subsequently projected to the image quality prior $\hat{\boldsymbol{P}}$ through a mapping network. We restore $\boldsymbol{I}^{\mathrm{h}}$ to the clean image $\hat{\boldsymbol{I}}^{\mathrm{c}}$ through the following steps. Starting with $\boldsymbol{I}^{\mathrm{h}}$, a 3×3 convolution layer is first applied to extract the shallow feature $\boldsymbol{F}_{\mathrm{s}}$. By utilizing both $\boldsymbol{F}_{\mathrm{s}}$ and $\hat{\boldsymbol{P}}$, multiple Mamba groups are employed to extract the refined feature $\boldsymbol{F}_{\mathrm{k}}$, $\boldsymbol{F}'_{\mathrm{s}}$ and $\hat{\boldsymbol{F}}_{\mathrm{s}}$. These features are then aligned via element-wise addition, resulting in $\boldsymbol{F}_{\mathrm{a}}$. Finally, $\hat{\boldsymbol{I}}^{\mathrm{c}}$ is reconstructed from $\boldsymbol{F}_{\mathrm{a}}$ through another 3×3 convolution layer. ZOH denotes the zero-order hold rule and $\boldsymbol{A}, \boldsymbol{B}, \bar{\boldsymbol{A}}, \bar{\boldsymbol{B}}, \boldsymbol{C}$, and $\boldsymbol{D}$ denote the model parameter matrices. h_k, h_{k-1}, x_k, and y_k are the discrete hidden states at time step $t = k \times \Delta$, where Δ is timescale parameter. The images are sourced from the O-Haze [32] dataset

where $\exp(\cdot)$ represents the operation of exponentiating a matrix, and $\boldsymbol{I}$ denotes identity matrix. After discretization, Eq. (2) can be reformulated with a step size Δ in the following recurrent neural network structure:

$$\begin{aligned} h_k &= \bar{\boldsymbol{A}} h_{k-1} + \bar{\boldsymbol{B}} x_k \,, \\ y_k &= \boldsymbol{C} h_k + \boldsymbol{D} x_k \,, \end{aligned} \tag{4}$$

where h_k, h_{k-1}, x_k, and y_k are the discrete hidden states at time step $t = k \times \Delta$. Additionally, Eq. (4) can be mathematically converted into an equivalent convolutional neural network form,

$$\begin{aligned} \boldsymbol{K} &\triangleq (\boldsymbol{C}\bar{\boldsymbol{B}}, \boldsymbol{C}\bar{\boldsymbol{A}}\bar{\boldsymbol{B}}, \ldots, \boldsymbol{C}\bar{\boldsymbol{A}}^{L-1}\bar{\boldsymbol{B}}) \,, \\ y &= Conv(x, \boldsymbol{K}) \,, \end{aligned} \tag{5}$$

where L represents the length of the input sequence, $\boldsymbol{K}$ denotes a structured convolutional kernel, and $Conv(\cdot, \cdot)$ is the convolution operation.

The recent SSM in computer vision, i.e., Mamba [13], retains the same recursive structure as Eq. (4), allowing it to handle ultra-long sequences. Simultaneously, the parallel scan algorithm [13] enables Mamba to leverage the parallel processing benefits as shown in Eq. (5), thereby supporting efficient training.

3.2 Overview

Given a hazy image $\boldsymbol{I}^{\mathrm{h}}$ affected by haze, the objective is used to restore the clean image $\hat{\boldsymbol{I}}^{\mathrm{c}}$. As illustrated in Fig. 1, our method restores the clean image $\hat{\boldsymbol{I}}^{\mathrm{c}}$ by leveraging an image quality prior $\hat{\boldsymbol{P}}$, which is extracted from the pre-trained LMM. To obtain $\hat{\boldsymbol{P}}$, which encodes the image quality representation, we first design a text prompt Q. Then, both the hazy image $\boldsymbol{I}^{\mathrm{h}}$ and Q are fed into the LMM to query the image probability distribution $\boldsymbol{P}$. This distribution $\boldsymbol{P}$ is subsequently projected to the image quality prior $\hat{\boldsymbol{P}}$ through a mapping network.

We restore $\boldsymbol{I}^{\mathrm{h}}$ to $\hat{\boldsymbol{I}}^{\mathrm{c}}$ through the following steps. Starting with the hazy image $\boldsymbol{I}^{\mathrm{h}}$, a 3×3 convolution layer is first applied to extract the shallow feature $\boldsymbol{F}_{\mathrm{s}}$. By utilizing both $\boldsymbol{F}_{\mathrm{s}}$ and the image quality prior $\hat{\boldsymbol{P}}$, multiple Mamba groups are employed to extract the refined feature $\hat{\boldsymbol{F}}_{\mathrm{s}}$. These features are then aligned via element-wise addition, resulting in $\boldsymbol{F}_{\mathrm{a}}$. Finally, a clean image $\hat{\boldsymbol{I}}^{\mathrm{c}}$ is reconstructed from $\boldsymbol{F}_{\mathrm{a}}$ through another 3×3 convolution layer.

3.3 DehazeMamba

LMM-based image quality prior. We leverage the powerful low-level visual perception capabilities of the LMM (e.g., Q-Align [19]) by designing a prompt Q to assess the quality of hazy images. We follow the design for low-quality images in Q-Align [19] as Q = <*Rate the quality of the image*>. To preserve the representation capabilities of

the LMM, we use the embedding before the output layer as $\boldsymbol{P}$. Then we align $\boldsymbol{P}$ to the embedding space with size C by using a linear projection network. The entire process can be denoted as

$$\hat{\boldsymbol{P}} = Linear\big(M(\boldsymbol{I}^{\mathrm{h}}, Q)\big), \quad \hat{\boldsymbol{P}} \in \mathbb{R}^{L\times C}, \tag{6}$$

where $M(\cdot,\cdot)$ denotes the process of the LMM, L is the length of all tokens in the LMM, $Linear(\cdot)$ is the linear perception layer, and $\hat{\boldsymbol{P}}$ is the outputting image quality prior.

Mamba group. The image quality prior $\hat{\boldsymbol{P}}$ is then fed to the Mamba group to align with the shallow feature $\boldsymbol{F}_{\mathrm{s}}$, thereby achieving quality-aware image dehazing as

$$\boldsymbol{F}_{\mathrm{k}} = MB\big(LN(\boldsymbol{F}_{\mathrm{s}}), \hat{\boldsymbol{P}}\big) + \delta \cdot \boldsymbol{F}_{\mathrm{s}}, \tag{7}$$

$$\boldsymbol{F}'_{\mathrm{s}} = Attention\Big(Conv\big(LN(\boldsymbol{F}_{\mathrm{k}})\big)\Big) + \delta' \cdot \boldsymbol{F}_{\mathrm{k}}, \tag{8}$$

where $\boldsymbol{F}_{\mathrm{s}}$ is the feature obtained by the convolution layer from $\boldsymbol{I}^{\mathrm{h}}$, $\boldsymbol{F}_{\mathrm{k}}$ represents the intermediate feature, and δ and δ' are learnable scale factors. $MB(\cdot,\cdot)$ denotes the processing of Mamba blocks. $LN(\cdot)$, $Conv(\cdot)$, $Attention(\cdot)$ are LayerNorm, convolution and channel attention layers, respectively. After passing through H Mamba groups, we obtain the refined feature $\hat{\boldsymbol{F}}_{\mathrm{s}}$.

Mamba block. Given $\boldsymbol{F}_{\mathrm{s}}$, we first pass them through a LayerNorm, following Ref. [71], and resulting in the intermediate feature $\boldsymbol{F}_{\mathrm{ln}}$. Based on the standard Mamba [13] mechanism, we design a Mamba-based feedforward process to model long-range dependencies, which can be denoted as

$$\boldsymbol{X} = SiLU\Big(DConv\big(LN(\boldsymbol{F}_{\mathrm{ln}})\big)\Big), \tag{9}$$

$$\boldsymbol{Y} = 2D - SSM\big(LN(\boldsymbol{X})\big), \tag{10}$$

$$\boldsymbol{F}_{\mathrm{b}} = Linear\Big(Cat(\boldsymbol{Y}, \hat{\boldsymbol{P}}) \otimes SiLU\big(Linear(\boldsymbol{F}_{\mathrm{ln}})\big)\Big), \tag{11}$$

where $DConv(\cdot)$, $SiLU(\cdot)$, and $Cat(\cdot,\cdot)$ are the depthwise convolution layer, $SiLU(\cdot)$ activation function and concatenation, respectively. $2D-SSM(\cdot)$ denotes the 2D-selective scan module. As shown in Fig. 1, the 2D-SSM first flattens 2D features into 1D sequences by scanning in four directions. The long-range dependencies in each sequence are then captured according to Eq. (4). All the sequences are combined through element-wise addition, followed by a reshaping operation to restore the original 2D structure.

Finally, a clean image $\hat{\boldsymbol{I}}^{\mathrm{c}}$ is reconstructed from $\hat{\boldsymbol{F}}_{\mathrm{s}}$ and $\boldsymbol{F}_{\mathrm{s}}$ through another 3×3 convolution layer using Eq. (12):

$$\hat{\boldsymbol{I}}^{\mathrm{c}} = Conv\big(Conv(\hat{\boldsymbol{F}}_{\mathrm{s}}) + \boldsymbol{F}_{\mathrm{s}}\big). \tag{12}$$

Loss function. We adopt the Charbonnier loss [83] to optimize our network, which is calculated using Eq. (13):

$$\mathcal{L} = \sqrt{\|\boldsymbol{I}^{\mathrm{c}} - \hat{\boldsymbol{I}}^{\mathrm{c}}\|^2 + \varepsilon^2}, \tag{13}$$

where $\boldsymbol{I}^{\mathrm{c}}$ and $\hat{\boldsymbol{I}}^{\mathrm{c}}$ denote the ground truth and clean images, respectively. Here, ε is a constant that is empirically set to 1.0×10^{-3}.

EDHaze dataset. To evaluate the performance of baseline dehazing methods under challenging conditions, we collect EDHaze, an additional test set of real-world high-resolution hazy/haze-free image pairs. As shown in Fig. 2, we set up a capture system to simulate real-world haze scenes.

The entire setup consists of a digital single-lens reflex (DSLR) camera, a stable tripod, a wireless shutter, and a professional haze generator, which can swiftly forces smoke oil through a heated tube at high temperatures, resulting in a fine white gaseous mist. We collect multiple indoor and outdoor scenes. After the haze generator is activated, fans are used to evenly distribute the haze throughout the space. To create more challenging conditions, we generate denser haze to simulate thick hazy conditions. The entire capture process is conducted in a windless, static environment to alleviate motion blur. Finally, we collect EDHaze, an extensive test set containing 200 challenging real-world hazy images.

Quality assessment score metric. Based on the premise that the LMM can perceive hazy in images and obtain image quality priors, we also design an evaluation metric based on the LMM, called the quality assessment score (QAS), to provide a more comprehensive and perceptually relevant evaluation of dehazed image quality.

Following Ref. [19], we link the embedding $\boldsymbol{P}$ with rating levels. First, we construct a candidate set of rating levels, denoted as $\mathcal{G}$. This set includes the standard ratings {bad, poor, fair, good, and excellent}, as defined by the International Telecommunication Union (ITU). In addition, to better align with the dehazing task, we incorporate extra reference ratings that reflect haze intensities, such as {clear, fog, and mist}. The resulting level set $\mathcal{G}$ consists of N distinct rating levels. Next, we define a mapping function $G(\cdot)$ that maps the text-defined rating levels to the corresponding probability indices, i.e., $G:\mathcal{G}_i \rightarrow j$, where $i\in[1,N]$ and $j\in[1,L]$. A closed-set softmax operation is then applied to $\boldsymbol{P}$ to obtain the degradation prior using Eq. (14):

$$\hat{\boldsymbol{P}}(i) = Softmax\Big(\boldsymbol{P}\big(G(\mathcal{G}_i)\big)\Big), \tag{14}$$

where $\hat{\boldsymbol{P}}(i)$ is the probability of the i-th rating level.

The QAS is denoted as

$$\mathrm{QAS} = \sum_{i=1}^{N} \boldsymbol{P}_{\mathcal{G}_i} \times G(\mathcal{G}_i) = \sum_{j=1}^{L} \hat{\boldsymbol{P}}_j \times j, \tag{15}$$

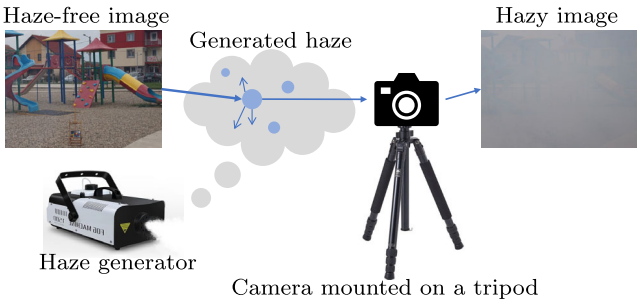


Figure 2 Capture system for obtaining paired hazy/haze-free images, including a photography tripod, a camera, and a professional haze generator. The camera's shutter is remotely controlled via a smartphone. The images are sourced from the Dense-Haze [34] dataset

where $\boldsymbol{P}_{\mathcal{G}_i}$ and $\hat{\boldsymbol{P}}_j$ refer to the value at the corresponding indices $\mathcal{G}_i$ and j, respectively. The QAS metric ranges from 1 to 5, with higher values indicating better image quality.

Compared with traditional IQA metrics, such as unsupervised image quality estimation (UNIQUE) and learning-based image quality estimation (LIQE), QAS has significant advantages in evaluating image quality. It can automatically learn complex image features, offering stronger generalization capabilities. QAS is able to adapt to various image styles and degradations without relying on manual feature design, and is capable of handling more complex image degradations.

4 Experiments and analysis

4.1 Experiments settings

Implementation details. DehazeMamba adopts a four-stage encoder-decoder architecture, where the Mamba groups are distributed across the four stages as {4,6,6,8}. DehazeMamba is trained with a batch size of 2 for 150,000 iterations, utilizing the AdamW optimizer. The learning rate follows a cosine annealing schedule, starts at 2.0×10^{-4} and gradually decreases to 1.0×10^{-5}. For training, the images are randomly cropped to size 256×256, whereas no cropping is applied during the testing phase. The LMM of our model utilizes Q-Align [19], which is further fine-tuned based on mPLUG-Owl2 [80]. We will release our dataset and code online to facilitate future research and comparisons.

Datasets. We evaluate three datasets, i.e., the O-Haze [32], Dense-Haze [34], and our EDHaze datasets. As outlined in Table 1, the O-Haze dataset consists of 45 pairs of hazy/haze-free images, where 40 pairs are used for training and 5 pairs are used for testing. Dense-Haze comprises 55 hazy/haze-free image pairs which include 50 training pairs and 5 testing pairs. The EDHaze dataset is our collected evaluation set for challenging scenarios only. Statistical information about the dataset can be found in Table 1.

Evaluation metrics. For evaluation, we utilize the peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) and learned perceptual image patch similarity (LPIPS) as our primary metrics. In addition, we assess the dehazed images with the LMM and the incorporate outputting QAS as supplementary evaluation criteria, following Ref. [19].

Table 1 Dataset statistics. "I&O" denotes datasets that include both indoor scenes and outdoor scenes. "GT" refers to datasets that include ground truth information for each scene

Dataset	Scale	Resolution	I&O	GT
O-Haze [32]	45	2000 ∼ 4000	O	✓
Dense-Haze [34]	55	2000 ∼ 4000	I&O	✓
EDHaze	200	5472 × 3648	I&O	✗

Baseline methods. We compare our method with 9 SOTA dehazing methods, including DCP [23], DehazeNet [4], GridDehazeNet [38], FFANet [5], DehazeFormer [49], SLP [33], DEANet [50], MambaIR [18], and the all-in-one method LDR [45]. With the exception of DCP and SLP, all other deep learning models have been retrained on the O-Haze dataset and Dense-Haze dataset, respectively.

4.2 Experimental results

Comparison with SOTA methods. We compare our method with 9 SOTA image dehazing methods on the O-Haze [32] and Dense-Haze [34] datasets in Table 2. On the O-Haze dataset, our method outperforms the second-best method, LDR [45], with an increase of 3.8% in PSNR and 1.3% in SSIM. Compared with the SOTA Mamba architecture method MambaIR [18], we achieve improvements of 5.0%/10.5% in PSNR and SSIM. For the Dense-Haze dataset, we outperform the second-best method, FFANet [5], with gains of 2.2% in PSNR and 5.6% in SSIM. Compared with the SOTA Mamba architecture method MambaIR [18], we achieve improvements of 5.9%/28% in PSNR and SSIM. These results highlight the advantages of our method. Additionally, we achieve the highest score in the LMM-based QAS evaluation metric, indicating that the images dehazed by our method have superior visual quality.

Table 2 Comparisons of state-of-the-art dehazing methods on the O-Haze and Dense-Haze datasets. The peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), quality assessment score (QAS), and learned perceptual image patch similarity (LPIPS) are evaluation metrics, where ↑ denotes that higher values indicate better quality, and ↓ denotes that lower values are preferred. The best performance is highlighted in bold

		O-Haze				Dense-Haze			
Method	Venue	PSNR ↑	SSIM ↑	QAS ↑	LPIPS ↓	PSNR ↑	SSIM ↑	QAS ↑	LPIPS ↓
DCP [23]	TPAMI'10	16.78	0.653	2.619	0.312	12.72	0.442	2.547	0.632
DehazeNet [4]	TIP'16	17.57	0.770	2.631	0.251	13.84	0.430	2.539	0.521
GridDehazeNet [38]	ICCV'19	18.92	0.672	2.682	0.228	14.62	0.443	2.693	0.517
FFANet [5]	AAAI'20	22.12	0.770	2.854	0.196	15.70	0.549	2.712	0.558
DehazeFormer [49]	TIP'23	20.53	0.676	2.838	0.135	15.45	0.509	2.744	0.448
SLP [33]	TIP'23	15.54	0.630	2.673	0.303	10.92	0.399	2.548	0.640
DEANet [50]	TIP'24	20.05	0.768	2.933	0.137	14.80	0.430	2.794	0.472
LDR [45]	CVPR'24	24.45	0.772	2.990	0.122	15.39	0.448	2.936	0.427
MambaIR [18]	ECCV'24	24.18	0.708	2.924	0.128	15.14	0.452	2.973	0.429
Ours	-	**25.38**	**0.782**	**3.158**	**0.111**	**16.04**	**0.580**	**3.014**	**0.420**

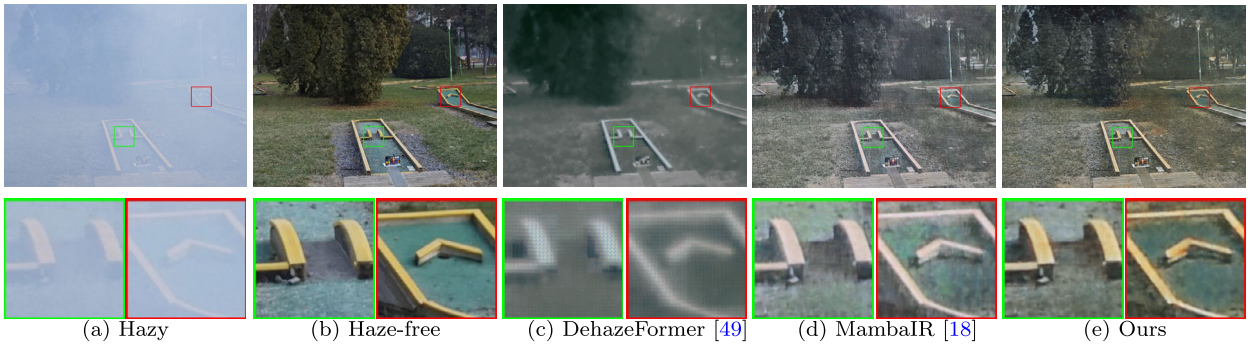


Figure 3 Example results on the Dense-Haze dataset. (a) and (b) are the hazy image and corresponding haze-free image. (c)-(e) show the dehazing results of different state-of-the-art (SOTA) dehazing models. Our model achieves the best dehazing performance while restoring the most accurate color details. Best viewed in color on the screen

The qualitative results are given in Figs. 3-5. Our method achieves superior dehazing performance while preserving the most accurate color details.

Quality of the EDHaze dataset. To evaluate the quality of the dehazing datasets in terms of haze quality and degradation level while avoiding the introduction of subjective factors, we employ a pretrained LMM to assess the hazy images. We design two prompts: T_h = <*rate the realism of the haze in this image*> and T_q = <*rate the quality of this image*>. We choose the SOTA image quality assessment model Q-Align [19] to assess each hazy image from existing real-world dehazing datasets using the two prompts, obtaining QAS_h and QAS_q, respectively. The results are shown in Figs. 6 and 7. The images in our EDHaze dataset have the best haze quality and most severe image degradation.

4.3 Ablation studies and discussions

Model architecture. We study the effectiveness of using the Mamba architecture and image quality priors in our DehazeMamba. As shown in Table 3, the best performance is achieved when all the modules are used.

Downstream validation. One key objective of image dehazing is to increase the performance of downstream high-level vision tasks. As a result, it is crucial to assess the performance of dehazing methods in these applications. For this purpose, we select three representative tasks, and use SOTA methods as baselines. Specifically, object detection with DEQDet [84], image segmentation with segment anything model (SAM) [85], and image captioning with LLaVA [20].

Figure 8 presents example results for the detection task. We observe that even the SOTA detection model [84] experiences significant performance deterioration under the image degradation caused by haze, manifesting as an increase in false detection and missed detection as well as lower detection confidence. In contrast, images dehazed by our DehazeMamba show substantial improvements in accuracy and detection confidence. Figure 9 shows example results for the segmentation task, utilizing the

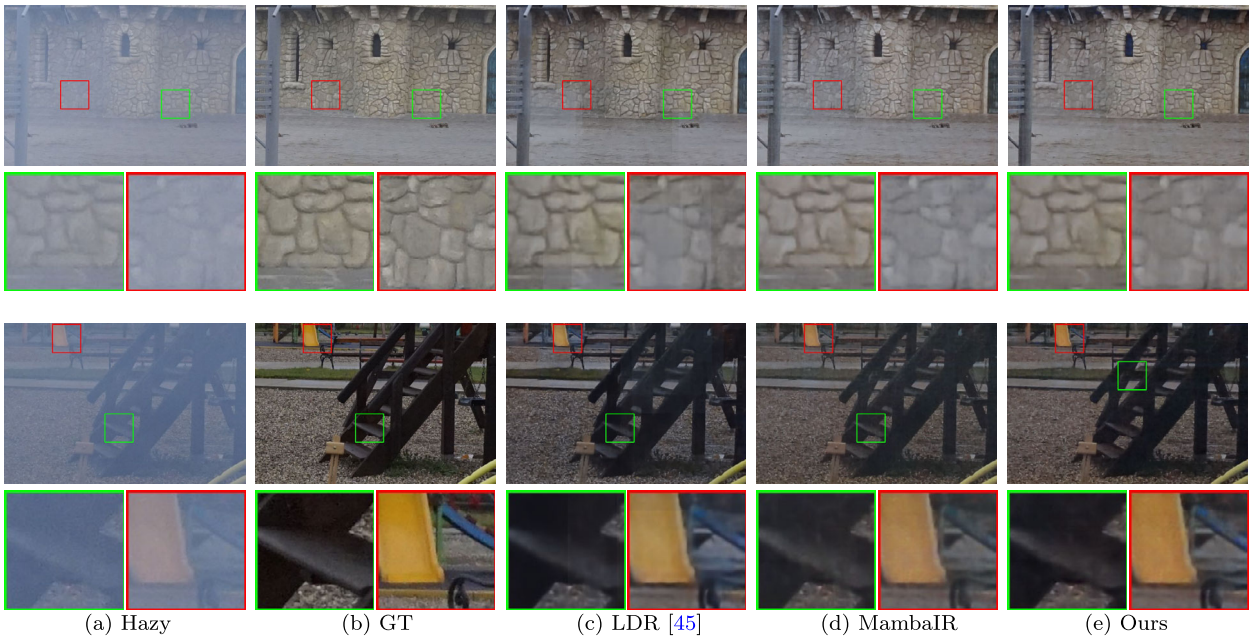


Figure 4 Example results on the O-Haze dataset. (a) and (b) are the hazy image and corresponding haze-free image. GT refers to ground truth. (c)-(e) show the dehazing results of different SOTA dehazing models. Our model achieves the best dehazing performance while restoring the most accurate color details. Best viewed in color on the screen

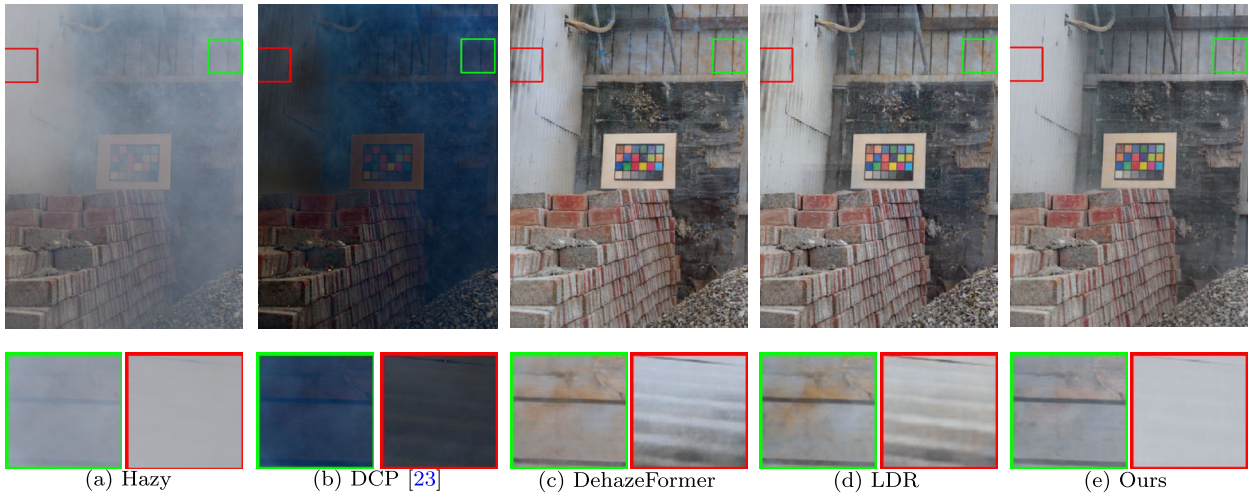


Figure 5 Example dehazing results on our EDHaze test set. (a) is the hazy image. (b)-(e) show the dehazing results of different SOTA dehazing methods. Our method achieves the best dehazing performance while restoring the most accurate color details. Best viewed in color on the screen

SOTA SAM for image segmentation. Our DehazeMamba achieves significantly higher segmentation accuracy on dehazed images than on the original hazy images, indicating that DehazeMamba, as a preprocessing step, effectively enhances the performance of the downstream segmentation task. Finally, Fig. 10 shows the image captions generated via the LLaVa model [20]. The results demonstrate that the LLaVa model generates more detailed and accurate descriptions of our dehazed images, further validating the effectiveness of DehazeMamba.

Furthermore, we conduct a quantitative comparison on 3 downstream tasks, with the results shown in Table 4. Our method demonstrates the most significant performance

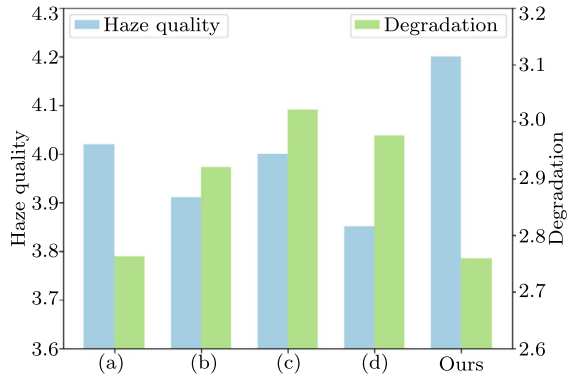


Figure 6 Mean quality assessment score (QAS) values on different real-world dehazing datasets. From (a) to (d) are Dense-Haze [34], I-Haze [36], NH-Haze [35], and O-Haze [32], respectively. The left vertical axis represents the mean haze quality (QAS_h) under prompt T_h, where a higher value indicates hazy images that are closer to real haze. The right vertical axis represents the mean degradation degree (QAS_q) under prompt T_q, where a lower value indicates more severe image degradation. The images in our EDHaze dataset have the best haze quality and the most severe image degradation

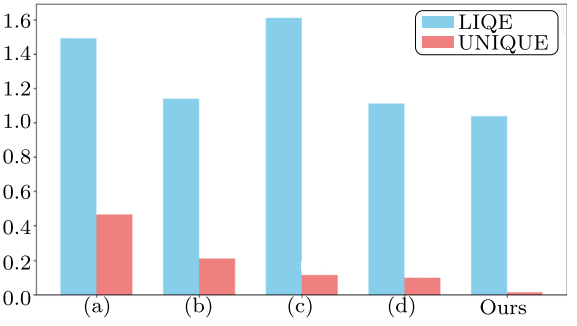


Figure 7 Mean unsupervised image quality estimation (UNIQUE) and learning-based image quality estimation (LIQE) values on different real-world dehazing datasets. From (a) to (d) are Dense-Haze, I-Haze, NH-Haze, and O-Haze. The blue and red bars represent the LIQE and UNIQUE scores, respectively, with lower values indicating more severe image degradation. Our EDHaze dataset exhibits the most severe degradation

Table 3 Ablation of the effectiveness of our model components on the O-Haze dataset. The Mamba architecture follows Ref. [18], and LMM denotes the large multi-modal model

Case	Baseline	Mamba	LMM	PSNR↑	SSIM↑
1	✓	✗	✗	22.32	0.674
2	✓	✓	✗	24.18	0.708
Ours	✓	✓	✓	25.38	0.782

improvement across all downstream tasks compared with the other SOTA methods.

Table 4 Quantitative comparisons of 3 downstream task evaluations. The metrics used include cosine similarity (Sim), mean intersection over union (mIoU), and mean average precision (mAP), with higher values indicating better performance

Method	Sim ↑	mIoU ↑	mAP ↑
Hazy	0.8534	0.0091	21.32
LDR	0.9724	0.6185	55.76
MambaIR	0.9694	0.6376	55.18
Ours	0.9928	0.7371	59.14

Table 5 Ablation study of Mamba blocks and layernorm configurations on the O-Haze dataset

Blocks	Layernorm	PSNR ↑	SSIM ↑
2,3,3,4	✗	17.10	0.602
4,6,6,8	✗	20.84	0.694
4,6,6,8	✓	25.38	0.782

Table 6 Cross-dataset comparison. Each method is trained on the O-Haze dataset and directly evaluated on the Dense-Haze dataset

Method	PSNR ↑	SSIM ↑
DEANet	11.29	0.417
MambaIR	13.51	0.455
Ours	15.46	0.566

Additionally, we conduct additional ablation experiments to explore the impact of the number of Mamba blocks and the use of layer normalization. Results are displayed in Table 5. The experimental results validate that the settings we have adopted are optimized for dehazing.

Cross-dataset comparison. We also conduct an additional experiment in which the model is trained on the O-Haze dataset and directly evaluated on the Dense-Haze dataset to assess the generalization performance of both the baseline methods and our proposed DehazeMamba. The results are presented in Table 6. Our DehazeMamba outperforms the baseline methods, demonstrating better generalization capabilities in cross-dataset evaluation.

Computational costs. Our DehazeMamba is trained on a single NVIDIA RTX 3090 GPU, with 13 million trainable parameters and 349.197 GFLOPs. For an image with a resolution of 1200×800, the average inference time of our model is 1.07 s.

Real-world results. To provide more comprehensive evaluation, we also conduct an experiment on a real-world dehazing test set without ground truth [86]. The results are demonstrated in Fig. 11. Our DehazeMamba removes the majority of the haze while preserving accurate color information, even on real-world hazy images.

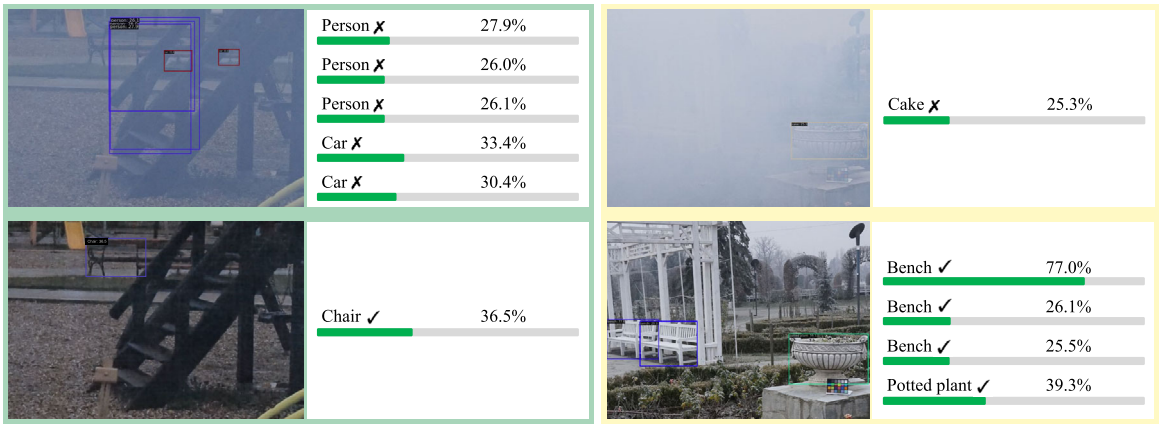


Figure 8 Example results on the detection task for the O-Haze and Dense-Haze datasets. The top images are the original hazy images, whereas the bottom images are the images dehazed by DehazeMamba. ✓ indicates a correct prediction, while ✗ indicates an incorrect prediction. The results demonstrate that DehazeMamba, as a preprocessing step, effectively enhances the performance of the downstream detection task. Best viewed in color on the screen



Figure 9 Example results on the segmentation task for the O-Haze and Dense-Haze datasets. The top images are the original hazy images, whereas the bottom images are the images dehazed by DehazeMamba. The results demonstrate that DehazeMamba significantly improves the segmentation accuracy. Best viewed in color on the screen



Figure 10 Example results on the captioning task for the O-Haze and Dense-Haze datasets. The top images are the original hazy images, whereas the bottom images are the images dehazed by DehazeMamba. The green and red colors highlight correct and incorrect descriptions, respectively. Images dehazed by DehazeMamba obtain more detailed and accurate captions. Best viewed in color on the screen

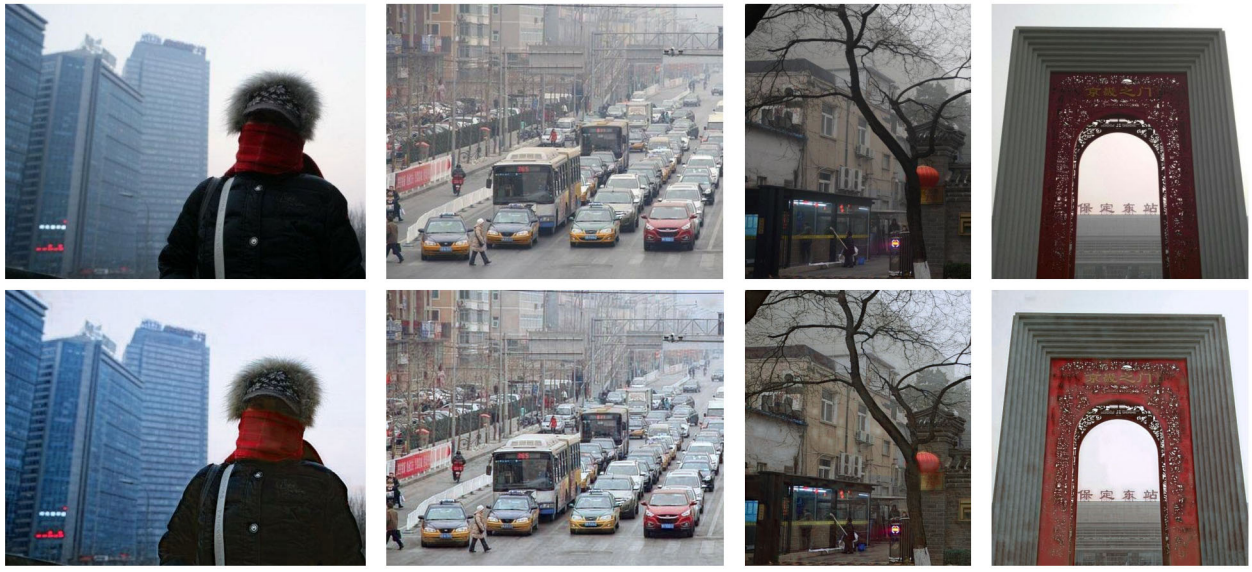


Figure 11 Example results on the real-world dehazing test set [86]. The top images are the unannotated hazy images, whereas the bottom images are the images dehazed by DehazeMamba. DehazeMamba removes the majority of the haze while preserving accurate color information. Best viewed in color on the screen

5 Conclusion and limitation

In this work, we propose an LMM guided image dehazing method, DehazeMamba, which is based on Mamba. By leveraging the image quality priors provided by the LMM and incorporating a channel attention mechanism, we improve the standard Mamba architecture. Additionally, we introduce a new evaluation metric based on the LMM, which simulates human perception to assess hazy images and dehazing results. To evaluate the dehazing performance of SOTA methods in challenging scenarios, we also propose a new image dehazing dataset, EDHaze. Experiments demonstrate that our dataset exhibits superior haze quality, and our method delivers better results than SOTA methods do.

Limitation. Despite the improvements achieved by our method, we observed that small objects, which are often obscured by haze, can still be lost during the image restoration process. This issue remains challenging in the image dehazing task, and current baseline methods fail to address this problem effectively. While our DehazeMamba, guided by the LMM, provides some mitigation, it is not a complete solution. Future work will focus on further exploring and refining techniques to better recover small details and improve the accuracy of haze removal in such cases.

Abbreviations

CNN, convolutional neural network; IQA, image quality assessment; LMM, large multi-modal model; LPIPS, learned perceptual image patch similarity; PSNR, peak signal-to-noise ratio; QAS, quality assessment score; SSM, state space model; SSIM, structural similarity index measure; 2D-SSM, 2D-selective scan module.

Author contributions

RZ was responsible for designing the framework, collecting the data, and conducting the experiments. ZY handled the data preparation and survey analysis, and assisted with the experiments. The first draft of the manuscript was written by RZ and ZY. LP provided technical and theoretical support for the research. All the authors contributed to the study conception and design. All authors read and approved the final manuscript.

Funding information

This work was supported in part by the National Natural Science Foundation of China (No. 62302045) and the Beijing Institute of Technology Research Fund Program for Young Scholars, BIT Special-Zone.

Data availability

The datasets used in this study are publicly available at: O-Haze: https://data.vision.ee.ethz.ch/cvl/ntire18/o-haze/; Dense-Haze: https://data.vision.ee.ethz.ch/cvl/ntire19//dense-haze/; Real world test set: https://sites.google.com/view/reside-dehaze-datasets/reside-standard?authuser=3D0; EDHaze test set: https://drive.google.com/file/d/1PdjdR8JXkPG95YxLX7WiG8P5vDfYFOl8/view?usp=sharing.

Declarations

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

Received: 30 August 2024 Revised: 11 June 2025 Accepted: 11 June 2025
Published online: 01 July 2025

References

1. Sakaridis, C., Dai, D., & Van Gool, L. (2018). Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, *126*, 973–992.
2. Li, B., Peng, X., Wang, Z., Xu, J., & Feng, D. (2018). End-to-end united video dehazing and detection. In S. McIlraith & K. Weinberger (Eds.), *Proceedings of the 32nd AAAI conference on artificial intelligence* (pp. 7016–7023). Palo Alto: AAAI Press.

3. Lee, S., Son, T., & Kwak, S. (2022). FIFO: learning fog-invariant features for foggy scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18911–18921). Piscataway: IEEE.
4. Cai, B., Xu, X., Jia, K., Qing, C., & Tao, D. (2016). DehazeNet: an end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, *25*(11), 5187–5198.
5. Qin, X., Wang, Z., Bai, Y., Xie, X., & Jia, H. (2020). FFA-Net: feature fusion attention network for single image dehazing. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11908–11915). Palo Alto: AAAI Press.
6. Li, B., Peng, X., Wang, Z., Xu, J., & Feng, D. (2017). AOD-Net: all-in-one dehazing network. In *Proceedings of the IEEE international conference on computer vision* (pp. 4780–4788). Piscataway: IEEE.
7. McCartney, E. J. (1976). *Optics of the atmosphere: scattering by molecules and particles*. New York: Wiley.
8. Chen, X., Wang, X., Zhou, J., Qiao, Y., & Dong, C. (2023). Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22367–22377). Piscataway: IEEE.
9. Lim, B., Son, S., Kim, H., Nah, S., & Lee, K.M. (2017). Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 136–144). Piscataway: IEEE.
10. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Proceedings of the European conference on computer vision* (pp. 286–301). Cham: Springer.
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Proceedings of the 31st international conference on neural information processing systems* (pp. 5998–6008). Red Hook: Curran Associates.
12. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). SwinIR: image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1833–1844). Piscataway: IEEE.
13. Gu, A., & Dao, T. (2023). Mamba: linear-time sequence modeling with selective state spaces. arXiv preprint. arXiv:2312.00752.
14. Fu, D. Y., Dao, T., Saab, K. K., Thomas, A. W., Rudra, A., & Ré, C. (2023). Hungry hungry hippos: towards language modeling with state space models. In *Proceedings of the 11th international conference on learning representations*. Retrieved March 12, 2025, from https://openreview.net/forum?id=COZDy0WYGg.
15. Gu, A., Goel, K., & Ré, C. (2022). Efficiently modeling long sequences with structured state spaces. In *Proceedings of the 10th international conference on learning representations*. Retrieved March 12, 2025, from https://openreview.net/forum?id=uYLFoz1vlAC.
16. Mehta, H., Gupta, A., Cutkosky, A., & Neyshabur, B. (2023). Long range language modeling via gated state spaces. In *Proceedings of the 11th international conference on learning representations*. Retrieved March 12, 2025, from https://openreview.net/forum?id=5MkYIYCbva.
17. Smith, J. T. H., Warrington, A., & Linderman, S. (2023). Simplified state space layers for sequence modeling. In *Proceedings of the 11th international conference on learning representations*. Retrieved March 12, 2025, from https://openreview.net/forum?id=Ai8Hw3AXqks.
18. Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., & Xia, S.-T. (2024). MambaIR: a simple baseline for image restoration with state-space model. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Proceedings of the 18th European conference on computer vision* (pp. 222–241). Cham: Springer.
19. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., & Sun, W., et al. (2024). Q-ALIGN: teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st international conference on machine learning* (pp. 54015–54029). PMLR.
20. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Proceedings of the 37th conference on neural information processing systems* (pp. 1–25). Hook Red: Curran Associates.
21. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. arXiv preprint. arXiv:2303.08774.
22. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The Llama 3 herd of models. arXiv preprint. arXiv:2407.21783.
23. He, K., Sun, J., & Tang, X. (2011). Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(12), 2341–2353.
24. Kim, H., Jin, H., Hadap, S., & Kweon, I. (2013). Specular reflection separation using dark channel prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1460–1467). Piscataway: IEEE.
25. Pan, J., Sun, D., Pfister, H., & Yang, M.-H. (2016). Blind image deblurring using dark channel prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1628–1636). Piscataway: IEEE.
26. Yan, Y., Ren, W., Guo, Y., Wang, R., & Cao, X. (2017). Image deblurring via extreme channels prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4003–4011). Piscataway: IEEE.
27. Zhu, Q., Mai, J., & Shao, L. (2015). A fast single image haze removal algorithm using color attenuation prior. *IEEE Transactions on Image Processing*, *24*(11), 3522–3533.
28. Berman, D., Treibitz, T., & Avidan, S. (2016). Non-local image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1674–1682). Piscataway: IEEE.
29. Deng, Z., Zhu, L., Hu, X., Fu, C.-W., Xu, X., Zhang, Q., Qin, J., & Heng, P.-A. (2019). Deep multi-model fusion for single-image dehazing. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2453–2462). Piscataway: IEEE.
30. Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., & Ma, L. (2021). Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10551–10560). Piscataway: IEEE.
31. Jiang, H., Dang, Z., Gu, S., Xie, J., Salzmann, M., & Yang, J. (2023). Center-based decoupled point-cloud registration for 6D object pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3427–3437). Piscataway: IEEE.
32. Ancuti, C. O., Ancuti, C., Timofte, R., & De Vleeschouwer, C. (2018). O-haze: a dehazing benchmark with real hazy and haze-free outdoor images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 754–762). Piscataway: IEEE.
33. Ling, P., Chen, H., Tan, X., Jin, Y., & Chen, E. (2023). Single image dehazing using saturation line prior. *IEEE Transactions on Image Processing*, *32*, 3238–3253.
34. Ancuti, C. O., Ancuti, C., Sbert, M., & Timofte, R. (2019). Dense-haze: a benchmark for image dehazing with dense-haze and haze-free images. In *Proceedings of the IEEE international conference on image processing* (pp. 1014–1018). Piscataway: IEEE.
35. Codruta, C. O., Ancuti, C., & Timofte, R. (2020). Nh-haze: an image dehazing benchmark with non-homogeneous hazy and haze-free images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 444–445). Piscataway: IEEE.
36. Ancuti, C., Ancuti, C. O., Timofte, R., & De Vleeschouwer, C. (2018). I-haze: a dehazing benchmark with real hazy and haze-free indoor images. In J. Blanc-Talon, D. Helbert, W. Philips, D. C. Popescu, & P. Scheunders (Eds.), *Proceedings of the 19th advanced concepts for intelligent vision systems* (pp. 620–631). Cham: Springer.
37. Li, X., Jin, Y., Jin, X., Wu, Z., Li, B., Wang, Y., Yang, W., Li, Y., Chen, Z., Wen, B., et al. (2025). NTIRE 2025 challenge on day and night raindrop removal for dual-focused images: methods and results. arXiv preprint. arXiv:2504.12711.
38. Liu, X., Ma, Y., Shi, Z., & Chen, J. (2019). GridDehazeNet: attention-based multi-scale network for image dehazing. In *Proceedings of the IEEE/CVF international conference on computer vision*, Piscataway: IEEE.
39. Jiang, K., Wang, Z., Yi, P., Chen, C., Wang, Z., Wang, X., Jiang, J., & Lin, C.-W. (2021). Rain-free and residue hand-in-hand: a progressive coupled network for real-time image deraining. *IEEE Transactions on Image Processing*, *30*, 7404–7418.
40. Xiao, Y., Yuan, Q., Jiang, K., He, J., Lin, C.-W., & Zhang, L. (2024). TTST: a top-k token selective transformer for remote sensing image super-resolution. *IEEE Transactions on Image Processing*, *33*, 738–752.
41. Xiao, Y., Yuan, Q., Jiang, K., Chen, Y., Zhang, Q., & Lin, C.-W. (2024). Frequency-assisted Mamba for remote sensing image super-resolution. *IEEE Transactions on Multimedia*, *27*, 1783-1796.
42. Ren, W., Pan, J., Zhang, H., Cao, X., & Yang, M.-H. (2020). Single image dehazing via multi-scale convolutional neural networks with holistic edges. *International Journal of Computer Vision*, *128*, 240–259.

43. Ren, W., Ma, L., Zhang, J., Pan, J., Cao, X., Liu, W., & Yang, M.-H. (2018). Gated fusion network for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3253–3261). Piscataway: IEEE.
44. Jiang, K., Wang, Z., Yi, P., & Jiang, J. (2020). Hierarchical dense recursive network for image super-resolution. *Pattern Recognition*, *107*, 107475.
45. Yang, H., Pan, L., Yang, Y., & Liang, W. (2024). Language-driven all-in-one adverse weather removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 24902–24912). Piscataway: IEEE.
46. Qiu, Y., Zhang, K., Wang, C., Luo, W., Li, H., & Jin, Z. (2023). MB-TaylorFormer: multi-branch efficient transformer expanded by Taylor formula for image dehazing. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 12802–12813). Piscataway: IEEE.
47. Zheng, Y., Zhan, J., He, S., Dong, J., & Du, Y. (2023). Curricular contrastive regularization for physics-aware single image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5785–5794). Piscataway: IEEE.
48. Zhang, G., Fang, W., Zheng, Y., & Wang, R. (2024). SDBAD-Net: a spatial dual-branch attention dehazing network based on meta-former paradigm. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(1), 60–70.
49. Song, Y., He, Z., Qian, H., & Du, X. (2023). Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, *32*, 1927–1941.
50. Chen, Z., He, Z., & Lu, Z.-M. (2024). DEA-Net: single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Transactions on Image Processing*, *33*, 1002–1015.
51. Hamilton, J. D. (1994). State-space models. In R. F. Engle & D. L. McFadden (Eds.), *Handbook of econometrics* (pp. 3039–3080). Amsterdam: Elsevier.
52. Kitagawa, G. (1998). A self-organizing state-space model. *Journal of the American Statistical Association*, *93*(443), 1203–1215.
53. Riera, J. J., Watanabe, J., Kazuki, I., Naoki, M., Aubert, E., Ozaki, T., & Kawashima, R. (2004). A state-space model of the hemodynamic approach: nonlinear filtering of bold signals. *NeuroImage*, *21*(2), 547–567.
54. Howard, T. M., Green, C. J., Kelly, A., & Ferguson, D. (2008). State space sampling of feasible motions for high-performance mobile robot navigation in complex environments. *Journal of Field Robotics*, *25*(6–7), 325–345.
55. Fossen, T. I. (2005). A nonlinear unified state-space model for ship maneuvering and control in a seaway. *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, *15*(9), 2717–2746.
56. Raibert, M. H. (1977). *Motor control and learning by the state space model*. PhD thesis, Massachusetts Institute of Technology.
57. Kroutikova, N., Hernandez-Aramburo, C. A., & Green, T. C. (2007). State-space model of grid-connected inverters under current control mode. *IET Electric Power Applications*, *1*(3), 329–338.
58. Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*(1), 35–45.
59. Hoshiya, M., & Saito, E. (1984). Structural identification by extended Kalman filter. *Journal of Engineering Mechanics*, *110*(12), 1757–1770.
60. Julier, S. J., & Uhlmann, J. K. (1997). New extension of the Kalman filter to nonlinear systems. In *Proceedings of signal processing, sensor fusion, and target recognition* (pp. 182–193). Bellingham: SPIE.
61. Sum, J., Leung, C.-S., Young, G. H., & Kan, W.-K. (1999). On the Kalman filtering method in neural network training and pruning. *IEEE Transactions on Neural Networks*, *10*(1), 161–166.
62. Haykin, S. (2004). *Kalman filtering and neural networks*. New York: Wiley.
63. Gu, A., Dao, T., Ermon, S., Rudra, A., & Ré, C. (2020). Recurrent memory with optimal polynomial projections. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Proceedings of the 34th conference on neural information processing systems* (pp. 1474–1487). Red Hook: Curran Associates.
64. Wang, J., Zhu, W., Wang, P., Yu, X., Liu, L., Omar, M., & Hamid, R. (2023). Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6387–6397). Piscataway: IEEE.
65. Islam, M. M., Hasan, M., Athrey, K. S., Braskich, T., & Bertasius, G. (2023). Efficient movie scene detection using state-space transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18749–18758). Piscataway: IEEE.
66. Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O., & Matthiopoulos, J. (2008). State–space models of individual animal movement. *Trends in Ecology & Evolution*, *23*(2), 87–94.
67. Subramanian, K., Maravelias, C. T., & Rawlings, J. B. (2012). A state-space model for chemical production scheduling. *Computers & Chemical Engineering*, *47*, 97–110.
68. Kim, C.-J., & Nelson, C. R. (1999). *State-space models with regime switching: classical and Gibbs-sampling approaches with applications*. Cambridge: MIT Press.
69. Glickman, M. E., & Stern, H. S. (1998). A state-space model for national football league scores. *Journal of the American Statistical Association*, *93*(441), 25.
70. Zhang, R., Yang, H., Yang, Y., Fu, Y., & Pan, L. (2024). Lmhaze: intensity-aware image dehazing with a large-scale multi-intensity real haze dataset. In *Proceedings of the 6th ACM international conference on multimedia in Asia, Article No. 18*. New York: ACM.
71. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision Mamba: efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st international conference on machine learning* (pp. 62429–62442). PMLR.
72. Zhou, M., Li, T., Qiao, C., Xie, D., Wang, G., Ruan, N., Mei, L., & Yang, Y. (2024). DMM: disparity-guided multispectral Mamba for oriented object detection in remote sensing. arXiv preprint. arXiv:2407.08132.
73. Zhou, H., Wu, X., Chen, H., Chen, X., & He, X. (2024). Rsdehamba: lightweight vision Mamba for remote sensing satellite image dehazing. arXiv preprint. arXiv:2405.10030.
74. Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.
75. Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, *21*(12), 4695–4708.
76. Talebi, H., & Milanfar, P. (2018). NIMA: neural image assessment. *IEEE Transactions on Image Processing*, *27*(8), 3998–4011.
77. Zhang, W., Ma, K., Yan, J., Deng, D., & Wang, Z. (2018). Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(1), 36–47.
78. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., & Zhang, Y. (2020). Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3667–3676). Piscataway: IEEE.
79. Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 26296–26306). Piscataway: IEEE.
80. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. (2023). mPLUG-Owl: modularization empowers large language models with multimodality. arXiv preprint. arXiv:2304.14178.
81. Zhang, P., Dong, X., Wang, B., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., Duan, H., Zhang, S., Ding, S., et al. (2023). Internlm-xcomposer: a vision-language large model for advanced text-image comprehension and composition. arXiv preprint. arXiv:2309.15112.
82. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. (2023). Mmbench: is your multi-modal model an all-around player? arXiv preprint. arXiv:2307.06281.
83. Lai, W.-S., Huang, J.-B., Ahuja, N., & Yang, M.-H. (2018). Fast and accurate image super-resolution with deep Laplacian pyramid networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(11), 2599–2613.
84. Wang, S., Teng, Y., & Wang, L. (2023). Deep equilibrium object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6296–6306). Piscataway: IEEE.
85. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015–4026). Piscataway: IEEE.
86. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., & Wang, Z. (2018). Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, *28*(1), 492–505.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.