

VmambaIR: Visual State Space Model for Image Restoration

Yuan Shi^{1,3*}, Bin Xia^{2,3 **}, Xiaoyu Jin¹, Xing Wang³, Tianyu Zhao³, Xin Xia³,
Xuefeng Xiao³, and Wenming Yang¹

¹ Tsinghua Shenzhen International Graduate School, Tsinghua University

² Department of Computer Science and Engineering, Chinese University of Hong Kong

³ Bytedance Inc.
VmambaIR Project

Abstract. Image restoration is a critical task in low-level computer vision, aiming to restore high-quality images from degraded inputs. Various models, such as convolutional neural networks (CNNs), generative adversarial networks (GANs), transformers, and diffusion models (DMs), have been employed to address this problem with significant impact. However, CNNs have limitations in capturing long-range dependencies. DMs require large prior models and computationally intensive denoising steps. Transformers have powerful modeling capabilities but face challenges due to quadratic complexity with input image size. To address these challenges, we propose VmambaIR, which introduces State Space Models (SSMs) with linear complexity into comprehensive image restoration tasks. We utilize a Unet architecture to stack our proposed Omni Selective Scan (OSS) blocks, consisting of an OSS module and an Efficient Feed-Forward Network (EFFN). Our proposed omni selective scan mechanism overcomes the unidirectional modeling limitation of SSMs by efficiently modeling image information flows in all six directions. Furthermore, we conducted a comprehensive evaluation of our VmambaIR across multiple image restoration tasks, including image deraining, single image super-resolution, and real-world image super-resolution. Extensive experimental results demonstrate that our proposed VmambaIR achieves state-of-the-art (SOTA) performance with much fewer computational resources and parameters. Our research highlights the potential of state space models as promising alternatives to the transformer and CNN architectures in serving as foundational frameworks for next-generation low-level visual tasks.

Keywords: State space models · Mamba · Image restoration

1 Introduction

Image restoration refers to the process of restoring high-quality images from their low-quality counterparts. This encompasses a range of tasks, including image deblurring, super-resolution, and image deraining. These tasks play a crucial role in the low-level domain of computer vision and have garnered significant attention over the years.

* Completed during internship at ByteDance.

** Yuan Shi and Bin Xia contributed equally to this paper.

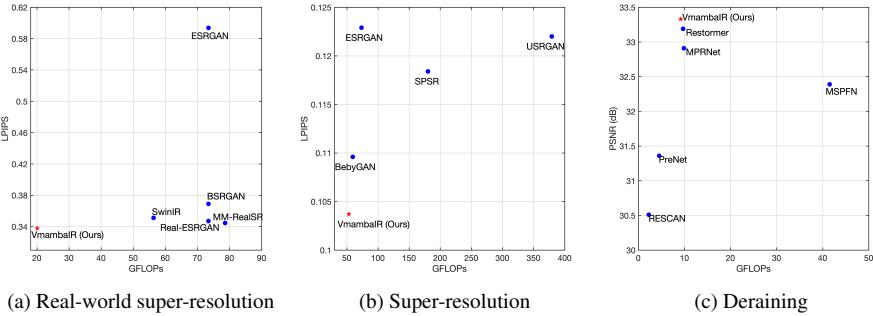


Fig. 1: Our VmambalIR demonstrates outstanding performance by achieving higher accuracy in image restoration tasks while requiring less computational cost. The GFLOPs are computed based on an input image size of 64×64 . In the real-world super-resolution task, VmambalIR achieves higher reconstruction accuracy with only **26%** of the computational cost.

In recent years, deep learning techniques have made remarkable strides in the field of image restoration. Deep learning models, including convolutional neural networks (CNNs) [7, 8, 47, 53, 60, 62, 64], generative adversarial networks (GANs) [16, 44], vision transformers [2, 3, 25, 54], and diffusion models (DMs) [26, 41, 48], have demonstrated exceptional capabilities in addressing complex image restoration tasks. These models leverage their ability to learn intricate patterns and features from extensive datasets, enabling them to effectively restore high-quality images from low-quality inputs.

However, CNNs often encounter limitations in modeling capabilities when dealing with large datasets and long-range dependencies. Diffusion models (DMs) in image restoration typically involve significant computational burden and time consumption due to the utilization of large prior models and intensive denoising processes. On the other hand, vision transformers exhibit a quadratic complexity in processing input sequences, which poses challenges in handling large-sized images which are commonly encountered in typical image restoration tasks.

Recently, state space model (SSM) [13, 15, 39], a novel approach originating from control systems, has garnered attention due to its linear complexity in processing input sequences. However, state space models suffer from the limitations of unidirectional modeling of input data and a lack of spatial awareness. These issues pose challenges when it comes to handling visual data. The development of a comprehensive data modeling mechanism that effectively harnesses the multidimensional information embedded in visual data flow, coupled with the design of streamlined and efficient network architectures tailored for state space models, is a challenging and unresolved task. This endeavor aims to fully exploit the inherent linear complexity and powerful high-frequency modeling capabilities of state space models in image restoration tasks.

To tackle these challenges, we present VmambalIR, a comprehensive image restoration network that leverages the state space model. To exploit the multi-scale features of images more effectively, we developed a network architecture that draws inspiration from the Unet framework [37]. Our architecture incorporates a novel Omni Selective Scan (OSS) block, consisting of an OSS module and an Efficient Feed-Forward Network (EFFN). These components work together to enable a comprehensive and efficient

modeling of the information flow. The omni selective scan, which serves as the core module of the OSS block, enables comprehensive modeling of the information flow, effectively addressing the limitation of unidirectional modeling in the mamba block [13].

We conducted extensive experiments on multiple image restoration tasks, including image deraining, single image super-resolution, and real-world image super-resolution. The experimental results, as shown in Fig. 1, demonstrate that our proposed VmambaIR surpasses the accuracy of the current baseline on all image restoration tasks while requiring less computational resources. Particularly, in the real-world super-resolution task, VmambaIR achieves higher reconstruction accuracy with only **26%** of the computational cost compared to the existing SOTA method.

Our contributions can be summarized as follows:

- We propose VmambaIR, a comprehensive image restoration model based on the state space model. VmambaIR incorporates our proposed OSS blocks into a Unet architecture, enabling effective handling of the multi-scale features of images.
- Diverging from recent vision Mamba blocks [27, 38, 65, 66], our designed OSS block comprises an OSS module and an EFFN. We have discovered that EFFN brings advantages to image restoration tasks, despite the fact that SSMs do not inherently require the handling of additional positional embedding bias.
- We propose Omni Selective Scan, which enables comprehensive pattern recognition and modeling of the information flow in image data by modeling it from six directions, while incurring minimal additional computational burden.
- We conducted experiments on comprehensive image restoration tasks, including single-image super-resolution, real-world image super-resolution, and image de-raining. Extensive experimental results demonstrate that our VmambaIR achieves better performance with lower computational and parameter requirements.

2 Related work

2.1 Image Restoration

Image restoration, a long-standing and well-established research area in computer vision, has witnessed notable advancements with the emergence of deep learning. Throughout the years, a multitude of remarkable models and works have emerged, continually pushing the boundaries of image restoration techniques.

Pioneering works, such as SRCNN [8] and ARCNN [7], have employed compact convolutional neural networks (CNNs) to achieve impressive performance in image restoration tasks, particularly in the domain of image denoising. Subsequently, CNN-based methods gained popularity over traditional image restoration approaches. Over time, researchers have explored CNNs from various perspectives, leading to the development of more elaborate network architectures and learning schemes. These include the integration of residual blocks [21, 60], the utilization of generative adversarial networks (GANs) [16, 44], the incorporation of attention mechanisms [6, 47, 53, 64], and the exploration of other innovative approaches [3, 12, 18, 20, 49, 50, 56].

Subsequently, the transformer, initially applied in natural language processing, has demonstrated exceptional performance in computer vision and image restoration domains. It has quickly become one of the fundamental model architectures for image

restoration tasks [2, 3, 25, 54]. The Transformer model outperforms CNNs by capturing global dependencies and modeling complex relationships. However, the computational complexity of self-attention in the transformer scales quadratically with the input image size. In image restoration tasks, where larger image sizes are often involved, this poses a challenge for the application of transformers.

Recently, the application of diffusion models in the field of image restoration has garnered attention [26, 41, 48]. These models demonstrate powerful data-fitting capabilities, and pre-trained diffusion models possess extensive prior knowledge, enabling them to generate visually appealing restored images of high quality. However, the substantial computational resource requirements and limitations in fidelity have hindered the widespread adoption of diffusion models in the field of image restoration.

2.2 State Space Models (SSMs)

SSMs have recently gained prominence in deep learning, particularly in the domain of state space transformation [15, 39]. These models draw inspiration from continuous state space models in control systems and show promise in addressing long-range dependency issues, as demonstrated by LSSL [15]. To mitigate the computational complexity associated with LSSL, S4 [14] proposes parameter normalization using a diagonal structure, offering an alternative to CNNs and Transformers while specifically focusing on modeling long-range dependencies. Consequently, multiple structured state space models have emerged. S5 [39] introduces the MIMO SSM and efficient parallel scan into the S4 layer, presenting the new S5 layer. And the Gated State Space layer [33] enhances expressivity by incorporating additional gating units into the existing S4 layer.

More recently, a data-dependent SSM layer and a generic language model backbone called Mamba have been proposed [13]. Mamba surpasses Transformers in terms of performance on large-scale real data, showcasing its effectiveness across various sizes and demonstrating linear scalability in sequence length. The advantages of Mamba, particularly its computational efficiency for large-scale image processing, make it highly significant for research in the yet unexplored field of image restoration.

3 Preliminaries: State Space Models

SSMs, which draw inspiration from continuous systems, are linear time-invariant systems that map the input stimulation $x(t) \in \mathbb{R}^L$ to the output response $y(t) \in \mathbb{R}^L$. Mathematically, SSMs can be formulated as linear ordinary differential equations (ODEs),

$$\dot{h}(t) = Ah(t) + Bx(t) \quad (1)$$

$$y(t) = Ch(t) + Dx(t) \quad (2)$$

where $h(t) \in \mathbb{R}^N$ is a hidden state, $A \in \mathbb{R}^{N \times N}$, $B \in \mathbb{R}^N$ and $C \in \mathbb{R}^N$ are the parameters for a state size N, and $D \in \mathbb{R}^1$ represents the skip connection.

Following that, discrete versions of SSMs were proposed, which transformed the original continuous-time nature and made them applicable in machine learning. The

discretization process converts the ODE into a discrete function and aligns the model with the sample rate of the underlying signal present in the input data $x(t) \in \mathbb{R}^{L \times D}$.

The ODE (Eq. 1) can be discretized using the zeroth-order hold (ZOH) rule, which incorporates a timescale parameter Δ to convert the continuous parameters \mathbf{A}, \mathbf{B} into discrete parameters $\bar{\mathbf{A}}, \bar{\mathbf{B}}$, which can be defined as follows:

$$\mathbf{h}'_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}x_t \quad (3)$$

$$y_t = \mathbf{C}\mathbf{h}_t + \mathbf{D}x_t \quad (4)$$

$$\bar{\mathbf{A}} = e^{\Delta \mathbf{A}} \quad (5)$$

$$\bar{\mathbf{B}} = (\Delta \mathbf{A})^{-1}(e^{\Delta \mathbf{A}} - I) \cdot \Delta \mathbf{B} \quad (6)$$

where $\Delta \in \mathbb{R}^D$ and $\mathbf{B}, \mathbf{C} \in \mathbb{R}^{D \times N}$.

In contrast to previous linear time-invariant (LTI) SSMs, our proposed Omni Mamba leverages the selection mechanism introduced in Mamba [13]. This empowers it to effectively capture the characteristics of long-sequence signals using a straightforward architecture, resulting in notable computational efficiency and accuracy.

4 Method

In this section, we initially present the inter-block structures and overall network pipeline of our designed VmambaIR. Subsequently, we present the fundamental block of our model, the OSS block, which consists of an OSS module and an EFFN. This block effectively utilizes mamba’s high-frequency modeling capability to capture information flow. Lastly, we introduce our proposed omni selective scan, which overcomes the limitations of unidirectional modeling in the state space model by efficiently modeling image information flow from all three dimensions.

4.1 Model Architecture

Common inter-block structures in image restoration models include plain stacking structure [25], multi-stage architecture [4], multi-scale fusion architecture [5], and UNet architecture [45, 54]. To ensure that VmambaIR is capable of capturing image features at different scales and remains robust to inputs of various sizes, we implemented a multi-scale UNet architecture based on our proposed OSS block, following [37, 54]. Specifically, the model architecture is depicted in Fig. 2.

Given a low-quality input image $I_{in} \in \mathbb{R}^{H \times W \times 3}$, a convolution is first applied to obtain the shallow feature embeddings $E_1 \in \mathbb{R}^{H \times W \times C}$. Then the hierarchical encoding is achieved through three layers of OSS blocks and downsampling, applied to the features E_1 . The features E_2, E_3 , and E_4 , processed by each OSS block at different layers, are downsampled to sizes $\frac{H}{2} \times \frac{W}{2}, \frac{H}{4} \times \frac{W}{4}$, and $\frac{H}{8} \times \frac{W}{8}$ respectively. The features E_i from the i-th layer of the encoder are concatenated with the features D_{i+1} from the previous layer of the decoder through skip connections. After three upsampling and decoder layers, we obtain the output feature D_1 from the decoder. Subsequently, D_1 is refined through M_F OSS blocks to obtain the feature $F_R \in \mathbb{R}^{H \times W \times 2C}$.

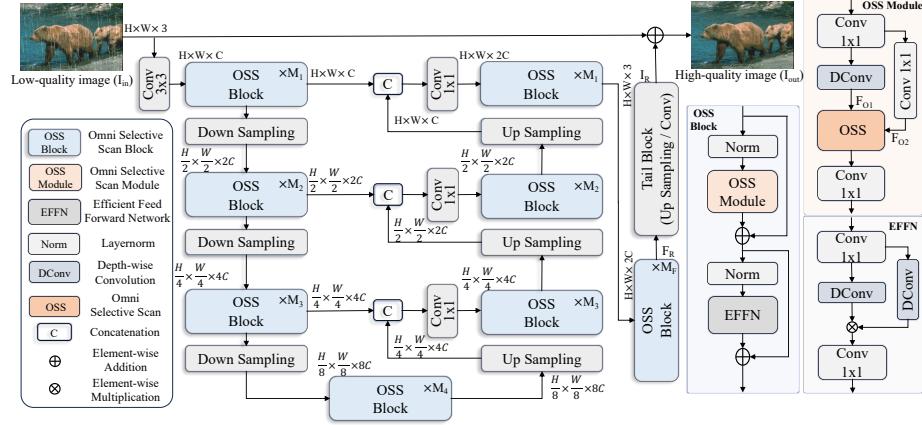


Fig. 2: Overview of our VmambaIR. The low-quality image undergoes an initial convolutional processing step to extract shallow features. These features are then fed into a Unet architecture, which is constructed using our proposed OSS block, enabling the extraction and reconstruction of features at various scales. The reconstructed features are subsequently refined through multiple iterations of OSS blocks. Finally, the refined features are passed through a tail block, typically involving convolution or gradual upsampling, to reconstruct the final high-quality image.

For image restoration tasks such as image super-resolution that involve upsampling, we utilize convolution and pixel shuffle within the tail block to upsample the features F_R . On the other hand, for tasks like image de-raining or deblurring that do not require a change in image size, we directly process the feature F_R using vanilla convolution to obtain the residual of the final output image $I_R \in \mathbb{R}^{H \times W \times 3}$.

4.2 OSS Block

The OSS block is employed to extract and model information flows from the omni domain of feature embeddings, as depicted in Figure 2. In contrast to the recent design of the Vision Mamba block [27, 38, 65, 66], which heavily relied on selection scanning mechanisms and linear mapping, our proposed OSS block introduces an OSS module capable of modeling information flows from diverse feature dimensions. Additionally, it incorporates an Efficient Feed-Forward Network (EFFN).

OSS Module The input of the OSS module is initially processed by a convolutional layer, generating two information flows. One flow undergoes refinement through depth-wise convolution and a Silu activation, capturing intricate patterns. Simultaneously, the other flow is processed with a Silu activation. The two flows enter the core OSS mechanism, which models information across all feature dimensions. Subsequently, the two flows are fused within the OSS, merging the refined features with complementary information. After passing through a 1x1 convolution, the output of OSS generates the final output of the OSS block, offering a comprehensive representation of the input with improved feature extraction and modeling capabilities.

It is noteworthy that, in the design of the OSS block, we opt for convolutional layers instead of linear layers to map the feature dimensions. This deliberate decision aims to minimize the operations conducted on feature shapes throughout the entire OSS block design. By leveraging convolutional layers, we not only enhance computational efficiency but also ensure a higher level of network-wide consistency.

Efficient Feed-Forward Network (EFFN) After the modeling process with the OSS module, the features are subjected to layer normalization to mitigate pattern collapse. Subsequently, the normalized features are passed through an efficient feed-forward network. Consistent with [54], the EFFN structure, illustrated in Figure 2, incorporates a 1x1 convolution to map the features into a high-dimensional space. The hidden layer features are then processed with depth-wise convolution and a gated mechanism. Lastly, a 1x1 convolution is employed to map the features back to their original dimension.

We have observed that the EFFN plays a crucial role in governing the information flow across the hierarchical levels in our pipeline, even though our omni selective scan does not necessitate positional encoding like self-attention [25, 28] does. Through information flow regulation, the EFFN facilitates synchronized operation, enabling each level to contribute its specialized expertise and leverage the strengths of other levels. This ensures that each level focuses on capturing complementary fine details.

4.3 Omni Selective Scan Mechanism

Mamba (S6) [13] is an auto regressive model widely recognized for its effectiveness in temporal and causal sequence modeling. It excels at capturing sequence dependencies in a unidirectional manner with high efficiency. However, the causal processing of input data in Mamba limits its ability to capture information beyond the scanned portion. In contrast to Transformers, Mamba encounters difficulties in modeling non-causal relationships, such as those found in image data.

To address the unidirectional modeling limitation of Mamba, one straightforward approach is to process the input data simultaneously in both forward and backward directions [66]. Alternatively, similar to vmamba [27], images can be scanned in a two-dimensional plane. However, these methods fail to fully integrate channel dimension information, which is crucial for comprehensive image modeling and important in image processing. UVM-Net [65] proposes a method that scans both the two-dimensional information and channel information of images. However, the scanning remains unidirectional, and the network requires a fixed input image size. This significantly limits the model's performance and convenience.

To enhance the multidimensional modeling capability of Mamba in image processing, we introduce the Omni Selective Scan (OSS) mechanism, illustrated in Fig. 3.

We utilize the two information streams, $F_{O1}, F_{O2} \in \mathbb{R}^{B \times C \times H \times W}$ in OSS module, as inputs for the OSS. First, we perform bidirectional scanning in the longitudinal and transverse directions on F_{O1} to capture the planar two-dimensional information of the features. "H Forward Scan" and "H Backward Scan" in Fig. 3 refer to scanning from the top left to the bottom right and from the bottom right to the top left of the two-dimensional features, respectively. While "W forward scan" represents scanning the image from the bottom left to the top right. Afterward, we stack and reshape the features

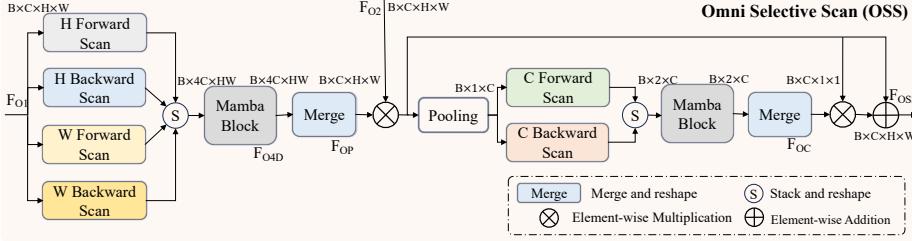


Fig.3: The architecture of one of our core designs, Omni Selective Scan. In the figure, "H Forward Scan", "W Forward Scan", and "C Forward Scan" indicate scanning from the top left to the bottom right, scanning from the bottom left to the top right on the two-dimensional image plane, and scanning the feature channels from front to back, respectively. The term "Backward" denotes the reverse direction of scanning. For the sake of simplicity in representation, operations on the feature dimensions, such as reshape and permute, have been omitted.

obtained from the four scans and model them using the Mamba block, resulting in $F_{O4D} \in \mathbb{R}^{B \times 4C \times HW}$. The features processed by the Mamba block are merged back into a single feature for planar modeling by splitting and adding the features from the four directions, resulting in the feature $F_{OP} \in \mathbb{R}^{B \times C \times H \times W}$.

After multiplying F_{OP} and F_{O2} , we first perform a pooling operation, and then scan the features in both forward and backward directions along the channel dimension. Compared to the channel modeling approach in [65], our utilization of pooling operations not only significantly reduces computational complexity while maintaining nearly unchanged performance but also eliminates the requirement for fixed image input sizes in the model. Subsequently, we continue to employ bidirectional scanning to ensure comprehensive modeling of channel information. After undergoing sequential processing, including stacking, the Mamba block, and merging, we obtain the feature $F_{OC} \in \mathbb{R}^{B \times C \times 1 \times 1}$, which models the channel dimension.

We employ a residual-based approach to effectively fuse the information from channel modeling and planar modeling, without the need for additional information streams. This strategy not only reduces computational costs but also enhances the stability of the model. Finally, we obtain the feature $F_{OSS} \in \mathbb{R}^{B \times C \times H \times W}$, which represents the comprehensive scanning and modeling of the input information flow by omni selective scan.

Compared to the self-attention mechanism in Transformers and recent works related to vision mamba block, our proposed Omni Selective Scan (OSS) enables comprehensive modeling of image features from six directions with minimal additional computational burden. The process of scanning, modeling, and processing the information flow through our omni selective scan can be illustrated in Fig. 4.

While self-attention models operate within the two-dimensional plane and exhibit quadratic complexity with respect to the input sequence, our proposed omni selective scan leverages Mamba's long-range modeling capability to comprehensively capture three-dimensional image features while maintaining linear complexity.

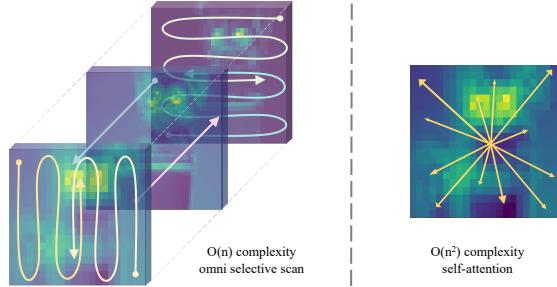


Fig. 4: The comparison between our proposed omni selective scan and self-attention reveals that Omni selective scan enables the modeling of image features from six directions and possesses linear computational complexity.

5 Experiments and Analysis

5.1 Experiment Settings

We validated the effectiveness of our proposed method on the following three image restoration tasks: (1) single image super-resolution, (2) real-world image super-resolution, and (3) image deraining.

We employed a four-layer symmetric encoder-decoder network structure for the aforementioned three tasks. The number of blocks in each layer varied as follows: [14, 1, 1, 1] for single image super-resolution, [6, 2, 2, 1] for real-world image super-resolution, and [4, 4, 6, 8] for deraining. Additionally, we integrated a refinement block with counts of 14, 6, and 2 for each respective task. The feature dimensions of each layer in the network were set as [48, 96, 192, 384]. Following the training settings in [54], we employed a progressive training strategy to handle input images of different sizes in the image deraining task. In image super-resolution tasks, the size of the ground-truth images is 256×256 . We train models with Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.99$).

To achieve better visual reconstruction quality, we employed perceptual loss and GAN loss for image super-resolution tasks. For the other image restoration tasks, we utilized the L1 loss function as the training objective.

5.2 Single Image Super-Resolution Results

We trained our VmambaIR model on the DIV2K dataset (800 images) [1] and Flickr2K dataset (2650 images) [40]. The batch sizes are set to 64, and the low-quality (LQ) patch size are 64×64 for $4 \times$ super-resolution. We compared our method with existing GAN-based image super-resolution methods on various datasets, including Set14 [57], General100 [9], Urban100 [17], Manga109 [32], and DIV2K100 [40]. To comprehensively assess the visual quality and fidelity of the generated images, we evaluated the performance of our model using LPIPS and PSNR metrics.

The quantitative experimental results are presented in Table 1. It is evident that our VmambaIR outperforms existing SOTA methods in terms of both PSNR and LPIPS on

Table 1: Quantitative comparison (LPIPS/PSNR) for $4\times$ **Single image super-resolution** on benchmarks. Best and second best performance are marked in bold and underlined, respectively.

| Method | Manga109 [32] | | Set14 [57] | | General100 [9] | | Urban100 [17] | | DIV2K100 [40] | |
|------------------|----------------------|--------------|-------------------|--------------|-----------------------|--------------|----------------------|--------------|----------------------|--------------|
| | LPIPS ↓ | PSNR ↑ | LPIPS ↓ | PSNR ↑ | LPIPS ↓ | PSNR ↑ | LPIPS ↓ | PSNR ↑ | LPIPS ↓ | PSNR ↑ |
| SFTGAN [43] | 0.0716 | 28.17 | 0.1313 | 26.74 | 0.0947 | 29.16 | 0.1343 | 24.34 | 0.1331 | 28.09 |
| SRGAN [22] | 0.0707 | 28.11 | 0.1327 | 26.84 | 0.0964 | 29.33 | 0.1439 | 24.41 | 0.1257 | 28.17 |
| ESRGAN [44] | 0.0649 | 28.41 | 0.1241 | 26.59 | 0.0879 | 29.43 | 0.1229 | 24.37 | 0.1154 | 28.18 |
| USRGAN [59] | 0.0630 | 28.75 | 0.1347 | 27.41 | 0.0937 | 30.00 | 0.1330 | 24.89 | 0.1325 | 28.79 |
| SPSR [31] | 0.0672 | 28.56 | 0.1207 | 26.86 | 0.0862 | 29.42 | 0.1184 | 24.80 | 0.1099 | 28.18 |
| BebyGAN [23] | <u>0.0529</u> | <u>29.19</u> | <u>0.1157</u> | <u>27.09</u> | <u>0.0778</u> | <u>29.95</u> | <u>0.1096</u> | <u>25.23</u> | <u>0.1022</u> | <u>28.62</u> |
| VmambalIR (Ours) | 0.0496 | 29.99 | 0.1097 | 27.64 | 0.0762 | 30.34 | 0.1037 | 25.71 | 0.0995 | 29.18 |



Fig. 5: Visual comparison of **single image super-resolution** methods. Zoom-in for better details.

all test datasets, demonstrating its superior performance in single-image super-resolution tasks. Specifically, on the Urban100 dataset, our method achieves a PSNR improvement of **0.48 dB** over BebyGAN [23] while maintaining a **lower LPIPS score**. To further validate the visual quality of the generated images by VmambalIR, the generated images from different methods are showcased in Fig. 5.

From Figure 5, it can be observed that the images generated by VmambalIR exhibit better fidelity and finer details, which aligns with the modeling capability of mamba in capturing high-frequency components in the information flow. In Figure 5, our method is the only one that accurately generates the eyes and nose of the person without introducing any artifacts. Furthermore, the water surfaces, boats, and bridges in the images generated by our method also exhibit enhanced details and fewer artifacts.

5.3 Real-World Image Super-Resolution Results

To further validate the image processing capabilities of our VmambalIR, we conducted experiments on more challenging real-world super-resolution tasks. We train our model on the DIV2K [1], Flickr2K [40] and OST v2 [43] dataset for $4\times$ real-world image super-resolution. we adopt the high-order degradation model following [42, 61] to generate degraded images. The low-quality (LQ) patch sizes are 64×64 in $4\times$ super-

Table 2: Quantitative comparison for $4\times$ **Real-World Image Super-Resolution** on benchmark datasets. Best and second best performance are marked in bold and underlined, respectively. The GFLOPs are computed based on an input image size of 64×64 .

| Method | Parms FLOPs | | NTIRE2020 [29] | | | AIM2019 [30] | | |
|------------------|--------------|-------------|----------------|--------------|---------------|---------------|--------------|---------------|
| | (M) | (G) | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ |
| ESRGAN [61] | 16.69 | 73.4 | 0.5938 | 21.14 | 0.3119 | 0.5558 | 23.17 | 0.6192 |
| BSRGAN [61] | 16.69 | 73.4 | 0.3691 | <u>26.75</u> | 0.7386 | 0.4048 | 24.20 | 0.6904 |
| Real-ESRGAN [42] | 16.69 | 73.4 | 0.3471 | 26.40 | 0.7431 | 0.3956 | 23.89 | 0.6892 |
| SwinIR [25] | 11.56 | 56.3 | 0.3512 | 26.39 | <u>0.7414</u> | 0.3980 | 23.88 | 0.6905 |
| MM-RealSR [34] | 26.13 | 78.6 | <u>0.3446</u> | 25.19 | 0.7404 | 0.3948 | 23.05 | 0.6889 |
| Vmamba-IR (Ours) | 10.50 | 20.5 | 0.3379 | 27.06 | 0.7501 | 0.3891 | <u>23.90</u> | 0.6972 |

resolution. Similarly, we employ a GAN approach to train our model in this task, aiming to attain improved visual quality. And we evaluate our VmambaIR on two benchmark datasets (NTIRE2020 [29] and AIM2019 [30]) using LPIPS [63], SSIM [46] and PSNR. The quantitative results are shown in Table 2.

The quantitative results demonstrate that our VmambaIR surpasses existing SOTA methods in comprehensive image quality evaluation metrics. Our VmambaIR consistently achieves lower LPIPS scores, and higher PSNR and SSIM scores across all test datasets, indicating its outstanding performance in real-world super-resolution tasks.

Furthermore, our VmambaIR model showcases its robust modeling capacity in image restoration tasks by utilizing approximately **26%** of the computational resources and achieving the **lowest** parameter count compared to the existing state-of-the-art methods. Importantly, our approach does not rely on additional techniques such as distillation or pruning. Without a doubt, our VmambaIR showcases the immense potential of state space models in the realm of image restoration tasks. To provide a more intuitive comparison, qualitative results of existing methods are presented in Fig. 6.

The qualitative results of the real-world super-resolution task further highlight the powerful modeling capability of our VmambaIR in capturing high-frequency details in images. In the restored images by VmambaIR, the leaves and fruits of the coconut trees exhibit improved details. And our images are the only ones where both coconuts are recognized. Additionally, the high-rise building lights generated by VmambaIR exhibit fewer artifacts and do not suffer from excessive enhancement.

5.4 Image Deraining Results

We trained and validate our VmambaIR on the Rain13K [19], Rain100H [51], Rain100L [51], Test1200 [58], and Test2800 [11] dataset for image deraining task. Following the previous works [19, 35, 54], we evaluated the performance of our model in terms of PSNR and SSIM metrics on the Y channel in the YCbCr color space. And we employed a progressive learning strategy to train our model. The quantitative experimental results of our VmambaIR on image deraining are presented in Table 3.

Clearly, our VmambaIR exhibits higher accuracy compared to existing SOTA methods, despite the inherent instability in image deraining tasks. In comparison to the ex-

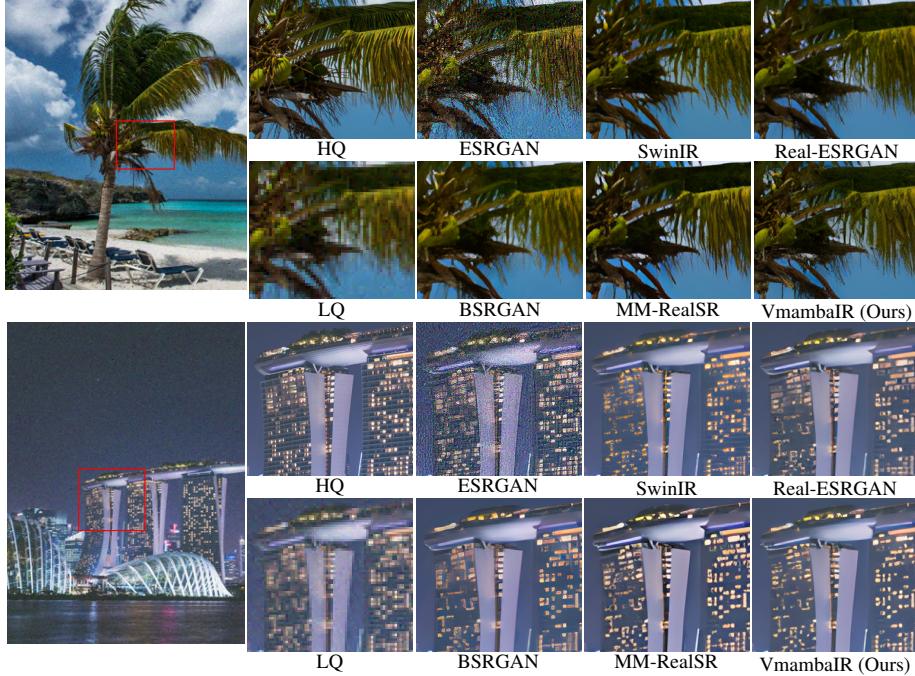


Fig. 6: Visual comparison of **real-world image super-resolution** methods. Zoom-in for better details. Our VmambaIR achieves superior high-frequency details and fidelity simultaneously.

isting method [54], our VmambaIR achieves a PSNR improvement of **over 0.1 dB** on the **Rain100H** [51], **Rain100L** [51], and **Test1200** [58] datasets, while maintaining less parameter and computational complexity. In addition, our method has consistently achieved higher SSIM scores on all datasets, indicating a higher visual quality of the restored images. To further validate the deraining performance of VmambaIR, we provide a qualitative comparison of the restored images from different methods in Fig. 7.

Compared to previous methods, our VmambaIR is capable of generating images in the image deraining task that are nearly indistinguishable from the ground truth. VmambaIR demonstrates almost perfect restoration results on the majority of the test images, highlighting the outstanding performance and generalization ability of our method.

5.5 Ablation Studies

We conducted ablation experiments to investigate the effects of our proposed omini selective scan (OSS) mechanism, bidirectional channel scanning, efficient feed-forward network (EFFN), and OSS module detail improvements. We trained models with different network configurations using the L1 loss function on $4\times$ real-world image super-resolution tasks, and the results are presented in Table 4.

Effects of omini selective scan In VmambaIR-V1, we assessed the influence of the proposed Omni Selective Scan on the network. During training, we employed one-way

Table 3: Quantitative comparison (PSNR/SSIM) for **Image Deraining** on five benchmark datasets. Best and second best performance are marked in bold and underlined, respectively.

| Method | Rain100H [51] | | Rain100L [51] | | Test2800 [11] | | Test1200 [58] | |
|-----------------|----------------------|---------------|----------------------|---------------|----------------------|---------------|----------------------|---------------|
| | PNSR ↑ SSIM ↑ | PNSR ↑ SSIM ↑ | PNSR ↑ SSIM ↑ | PNSR ↑ SSIM ↑ | PNSR ↑ SSIM ↑ | PNSR ↑ SSIM ↑ | PNSR ↑ SSIM ↑ | PNSR ↑ SSIM ↑ |
| DerainNet [10] | 14.92 | 0.592 | 27.03 | 0.884 | 24.31 | 0.861 | 23.38 | 0.835 |
| UMRL [52] | 26.01 | 0.832 | 29.18 | 0.923 | 29.97 | 0.905 | 30.55 | 0.910 |
| RESCAN [24] | 26.36 | 0.786 | 29.80 | 0.881 | 31.29 | 0.904 | 30.51 | 0.882 |
| PreNet [36] | 26.77 | 0.858 | 32.44 | 0.950 | 31.75 | 0.916 | 31.36 | 0.911 |
| MSPFN [19] | 28.66 | 0.860 | 32.40 | 0.933 | 32.82 | 0.930 | 32.39 | 0.916 |
| MPRNet [55] | 30.41 | 0.890 | 36.40 | 0.965 | 33.64 | 0.938 | 32.91 | 0.916 |
| SPAIR [35] | 30.95 | 0.892 | 36.93 | 0.969 | 33.34 | 0.936 | 33.04 | 0.922 |
| Restormer [54] | 31.46 | 0.904 | 38.99 | 0.978 | 34.18 | 0.944 | 33.19 | 0.926 |
| VmambaIR (Ours) | 31.66 | 0.909 | 39.09 | 0.979 | <u>34.01</u> | 0.944 | 33.33 | 0.926 |

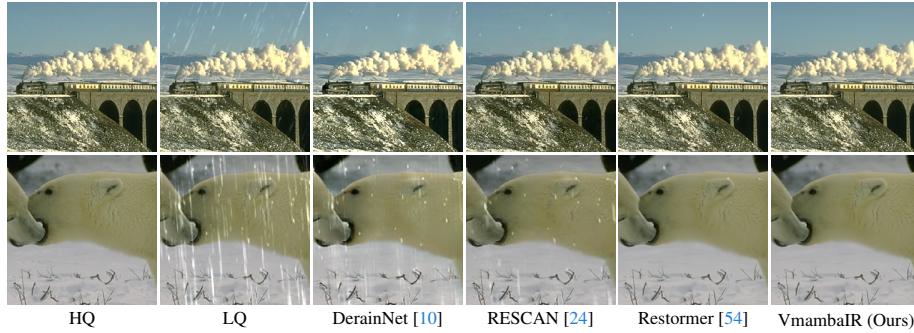


Fig. 7: Visual comparison of **image deraining** methods. Zoom-in for better details.

scanning in the image features instead of our omni selective scan. As shown in Table 4, the network’s computational complexity reduced by approximately 7%, while the network’s accuracy experienced a decline of approximately **0.43 dB**. This observation underscores the substantial limitations imposed by the one-way modeling of the state space model in processing two-dimensional image information.

Effects of bidirectional channel scanning In VmambaIR-V2, we employed 2D selective scan instead of Omni Selective Scan to validate the impact of the channel scanning mechanism in Omni Selective Scan. By removing the bidirectional channel scanning, the computational complexity of the network remained almost the same, but the network’s accuracy decreased by approximately **0.14 dB**, which demonstrates that the channel scanning we designed is very efficient and effective.

Effects of efficient feed-forward network In VmambaIR-V3, we abandoned the EFFN module in the OSS block and solely utilized the OSS module for network modeling. To maintain comparable network scale, we increased the number of OSS blocks. With similar computational complexity, VmambaIR with EFFN achieved higher accuracy,

Table 4: The influence of different network configurations on our model. The PSNR results are evaluated on NTIRE2020 Track1 [29] for real-world image super-resolution. The performance and GFLOPs are measured on an LQ size of 64×64 .

| Method | GFLOPs | Plane scan | Channel scan | EFFN | Conv1x1 | PSNR |
|-----------------|--------|------------|--------------|------|---------|-------|
| VmambaIR (Ours) | 20.45 | ✓ | ✓ | ✓ | ✓ | 28.50 |
| VmambaIR-V1 | 18.93 | ✗ | ✗ | ✓ | ✓ | 28.07 |
| VmambaIR-V2 | 20.43 | ✓ | ✗ | ✓ | ✓ | 28.36 |
| VmambaIR-V3 | 20.03 | ✓ | ✓ | ✗ | ✓ | 28.39 |
| VmambaIR-V4 | 20.45 | ✓ | ✓ | ✓ | ✗ | 28.47 |

demonstrating the effectiveness of our OSS block and EFFN. Additionally, due to significant data type and dimension conversions in the current selective scan operations, the speed of selective scan operations with the same computational complexity is slower than vanilla convolution. Therefore, we employed EFFN to enhance both the computational accuracy and efficiency of our proposed VmambaIR.

Improvements in OSS module In VmambaIR-V4, we conducted a validation study on the influence of specific detailed design choices within the OSS module on the network. In VmambaIR-V4, we opted to use linear layers and additional reshape operations for feature dimension transformations, as opposed to 1x1 convolutions. Despite maintaining comparable computational complexity, parameter count, and accuracy, our designed VmambaIR demonstrated an approximate **8.6%** enhancement in computational speed.

6 Conclusion

In this paper, we propose VmambaIR, a novel image restoration network, leveraging the linear complexity and high-frequency modeling capabilities of the mamba block [13]. We employ the UNet [37] architecture to enable our proposed OSS block to effectively model and utilize images at different scales. The OSS block consists of an OSS module and an EFFN module. The OSS module leverages mamba’s long-range modeling capability to comprehensively and efficiently model image features. The EFFN further maps and modulates the image information flow, enhancing the accuracy and efficiency of the network. Our model’s capabilities are evaluated on various image restoration tasks, including image deraining, image super-resolution, and real-world image super-resolution. Extensive experimental results demonstrate that our designed model achieves state-of-the-art performance. Furthermore, our network is intentionally designed without the utilization of elaborate techniques such as distillation, teacher networks, hybrid network structures, and others. Our primary objective is to establish a simple yet effective mamba image restoration baseline that can serve as both inspiration and motivation for future research endeavors. Through the demonstration of the potential of state space models in the field of image restoration, our work endeavors to make a substantial and valuable contribution to the progress and development of this domain.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: CVPR workshops (2017) 9, 10
2. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: CVPR (2021) 2, 4
3. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: ECCV. Springer (2022) 2, 3, 4
4. Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: Hinet: Half instance normalization network for image restoration. In: CVPR (2021) 5
5. Cho, S.J., Ji, S.W., Hong, J.P., Jung, S.W., Ko, S.J.: Rethinking coarse-to-fine approach in single image deblurring. In: ICCV (2021) 5
6. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: CVPR (2019) 3
7. Dong, C., Deng, Y., Loy, C.C., Tang, X.: Compression artifacts reduction by a deep convolutional network. In: ICCV (2015) 2, 3
8. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE TPAMI **38**(2) (2015) 2, 3
9. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: ECCV. Springer (2016) 9, 10
10. Fu, X., Huang, J., Ding, X., Liao, Y., Paisley, J.: Clearing the skies: A deep network architecture for single-image rain removal. IEEE TIP (2017) 13, 20, 23
11. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: CVPR (2017) 11, 13
12. Fu, X., Zha, Z.J., Wu, F., Ding, X., Paisley, J.: Jpeg artifacts reduction via deep convolutional sparse coding. In: ICCV (2019) 3
13. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023) 2, 3, 4, 5, 7, 14, 19
14. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021) 4
15. Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., Ré, C.: Combining recurrent, convolutional, and continuous-time models with linear state space layers. NeurIPS (2021) 2, 4
16. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. NeurIPS (2017) 2, 3
17. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: CVPR (2015) 9, 10
18. Jia, X., Liu, S., Feng, X., Zhang, L.: Focnet: A fractional optimal control network for image denoising. In: CVPR (2019) 3
19. Jiang, K., Wang, Z., Yi, P., Chen, C., Huang, B., Luo, Y., Ma, J., Jiang, J.: Multi-scale progressive fusion network for single image deraining. In: CVPR (2020) 11, 13
20. Jin, X., Shi, Y., Xia, B., Yang, W.: Llmra: Multi-modal large language model based restoration assistant. arXiv preprint arXiv:2401.11401 (2024) 3
21. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: CVPR (2016) 3
22. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017) 10, 19
23. Li, W., Zhou, K., Qi, L., Lu, L., Lu, J.: Best-buddy gans for highly detailed image super-resolution. In: AAAI. vol. 36 (2022) 10, 19

24. Li, X., Wu, J., Lin, Z., Liu, H., Zha, H.: Recurrent squeeze-and-excitation context aggregation net for single image deraining. In: ECCV (2018) [13](#), [20](#), [23](#)
25. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: ICCV (2021) [2](#), [4](#), [5](#), [7](#), [11](#), [19](#)
26. Lin, X., He, J., Chen, Z., Lyu, Z., Fei, B., Dai, B., Ouyang, W., Qiao, Y., Dong, C.: Diffbir: Towards blind image restoration with generative diffusion prior. arXiv preprint arXiv:2308.15070 (2023) [2](#), [4](#)
27. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166 (2024) [3](#), [6](#), [7](#)
28. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021) [7](#)
29. Lugmayr, A., Danelljan, M., Timofte, R.: Ntire 2020 challenge on real-world image super-resolution: Methods and results. In: CVPR Workshops (2020) [11](#), [14](#)
30. Lugmayr, A., Danelljan, M., Timofte, R., Fritsche, M., Gu, S., Purohit, K., Kandula, P., Suin, M., Rajagopalan, A., Joon, N.H., et al.: Aim 2019 challenge on real-world image super-resolution: Methods and results. In: ICCVW. IEEE (2019) [11](#)
31. Ma, C., Rao, Y., Cheng, Y., Chen, C., Lu, J., Zhou, J.: Structure-preserving super resolution with gradient guidance. In: CVPR (2020) [10](#)
32. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications **76** (2017) [9](#), [10](#)
33. Mehta, H., Gupta, A., Cutkosky, A., Neyshabur, B.: Long range language modeling via gated state spaces. arXiv preprint arXiv:2206.13947 (2022) [4](#)
34. Mou, C., Wu, Y., Wang, X., Dong, C., Zhang, J., Shan, Y.: Metric learning based interactive modulation for real-world super-resolution. In: ECCV. Springer (2022) [11](#), [19](#)
35. Purohit, K., Suin, M., Rajagopalan, A., Boddeti, V.N.: Spatially-adaptive image restoration using distortion-guided networks. In: ICCV (2021) [11](#), [13](#)
36. Ren, D., Zuo, W., Hu, Q., Zhu, P., Meng, D.: Progressive image deraining networks: A better and simpler baseline. In: CVPR (2019) [13](#)
37. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. Springer (2015) [2](#), [5](#), [14](#)
38. Ruan, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491 (2024) [3](#), [6](#)
39. Smith, J.T., Warrington, A., Linderman, S.W.: Simplified state space layers for sequence modeling. arXiv preprint arXiv:2208.04933 (2022) [2](#), [4](#)
40. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: CVPR workshops (2017) [9](#), [10](#)
41. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. arXiv preprint arXiv:2305.07015 (2023) [2](#), [4](#)
42. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: ICCV (2021) [10](#), [11](#), [18](#), [19](#)
43. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: CVPR (2018) [10](#)
44. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: ECCV workshops (2018) [2](#), [3](#), [10](#)
45. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: CVPR (2022) [5](#)
46. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP (2004) [11](#)

47. Xia, B., Hang, Y., Tian, Y., Yang, W., Liao, Q., Zhou, J.: Efficient non-local contrastive attention for image super-resolution. In: AAAI (2022) [2](#), [3](#)
48. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Van Gool, L.: Diffir: Efficient diffusion model for image restoration. arXiv preprint arXiv:2303.09472 (2023) [2](#), [4](#)
49. Xia, B., Zhang, Y., Wang, Y., Tian, Y., Yang, W., Timofte, R., Van Gool, L.: Knowledge distillation based degradation estimation for blind super-resolution. arXiv preprint arXiv:2211.16928 (2022) [3](#)
50. Xia, B., Zhang, Y., Wang, Y., Tian, Y., Yang, W., Timofte, R., Van Gool, L.: Basic binary convolution unit for binarized image restoration network. ICLR (2023) [3](#)
51. Yang, W., Tan, R.T., Feng, J., Liu, J., Guo, Z., Yan, S.: Deep joint rain detection and removal from a single image. In: CVPR (2017) [11](#), [12](#), [13](#)
52. Yasarla, R., Patel, V.M.: Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In: CVPR (2019) [13](#)
53. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: ICCV (2019) [2](#), [3](#)
54. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR (2022) [2](#), [4](#), [5](#), [7](#), [9](#), [11](#), [12](#), [13](#), [19](#), [20](#), [23](#)
55. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: CVPR (2021) [13](#)
56. Zeng, Y., Fu, J., Chao, H., Guo, B.: Aggregated contextual transformations for high-resolution image inpainting. IEEE Transactions on Visualization and Computer Graphics (2022) [3](#)
57. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7. Springer (2012) [9](#), [10](#)
58. Zhang, H., Patel, V.M.: Density-aware single image de-raining using a multi-stream dense network. In: CVPR (2018) [11](#), [12](#), [13](#)
59. Zhang, K., Gool, L.V., Timofte, R.: Deep unfolding network for image super-resolution. In: CVPR (2020) [10](#)
60. Zhang, K., Li, Y., Zuo, W., Zhang, L., Van Gool, L., Timofte, R.: Plug-and-play image restoration with deep denoiser prior. IEEE TPAMI (2021) [2](#), [3](#)
61. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: ICCV (2021) [10](#), [11](#), [19](#)
62. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE TIP (2017) [2](#)
63. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) [11](#)
64. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: ECCV (2018) [2](#), [3](#)
65. Zheng, Z., Wu, C.: U-shaped vision mamba for single image dehazing. arXiv preprint arXiv:2402.04139 (2024) [3](#), [6](#), [7](#), [8](#)
66. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024) [3](#), [6](#), [7](#)

7 Appendix

7.1 More Training Details on Real-World Image Super-Resolution

To generate data and train the network, we adopted a configuration identical to that of Real-ESRGAN [42]. Specifically, we implemented a dual degradation process to produce low-quality images. This involved applying random resizing, Gaussian noise, gray noise, and blur techniques to introduce various forms of degradation. Throughout this process, we maintained consistency with the degradation parameter settings used in Real-ESRGAN [42] to ensure comparable results.

The ground-truth images were cropped to a size of 256×256 . And the size of training images were kept same in the training process for image super-resolution tasks. During the training process, we utilized 8 V100 GPUs, with each GPU processing a batch size of 9. To optimize the network, we employed the Adam optimizer with specific parameter values, namely $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate was set to $1 \times e^{-4}$, and a weight decay of 0 was applied. To control the learning rate, we employed a MultiStepLR strategy with a decay factor γ of 0.5. The model was trained for 300,000 iterations. On average, the training process for the $4\times$ real-world GAN-based image super-resolution model took approximately 5 days.

For training our VmambaIR, we incorporated GAN loss, perceptual loss, and L1 loss. These losses were given equal weights during the entire training process. To serve as the discriminator in our network, we adopted a U-Net architecture with a feature dimension of 64. The optimizer settings used for the generator were also applied to the Unet discriminator, ensuring consistency across both components.

7.2 More Training Details on Single Image Super-Resolution

For the GAN-based single image super-resolution task, we generated low-quality training images by employing bicubic downsampling. Following prior works, we utilized the DF2K dataset, which consists of 3450 images, to train our model.

For single image super-resolution task, the ground-truth images were cropped to a size of 256×256 . and the low-quality images were 64×64 for $4\times$ super-resolution. During the training process, we utilized 8 V100 GPUs, with each GPU processing a batch size of 8. To optimize the network, we employed the Adam optimizer with specific parameter values, namely $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate was set to $2 \times e^{-4}$, and a weight decay of 0 was applied. To control the learning rate, we employed a MultiStepLR strategy with a decay factor γ of 0.5. The model was trained for 300,000 iterations. On average, the training process for the $4\times$ GAN-based image super-resolution model took approximately 5 days.

Similar to the settings in real-world image super-resolution, during the entirety of the training process, we incorporated GAN loss, perceptual loss, and L1 loss into our model. To ensure equitable impact from each loss function, we assigned them equal weights. And the learning rate for the discriminator Unet was configured to $1 \times e^{-4}$.

7.3 More Training Details on Image Deraining

Unlike in image super-resolution tasks, in order to capture deeper features of images for the image de-raining task, the number of blocks per layer in the network was set to [4, 4, 6, 8], and the number of refinement blocks was set to 2.

In accordance with the progressive training approach outlined in Restormer [54], we followed specific settings during the training process. The input image sizes were sequentially set as [128, 160, 192, 256, 320, 384], while the corresponding batch sizes per GPU were [8, 5, 3, 2, 1, 1]. To train VmambaIR for the image deraining task, we utilized a total of 8 V100 GPUs. The training procedure spanned approximately 6 days, encompassing the complete training of our model.

We employed L1 loss as the objective function for training the image deraining task. The optimization was performed using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate for training was set to $3 \times e^{-4}$, and a weight decay of $1 \times e^{-4}$ was applied. To control the learning rate during training, we utilized the cosine annealing technique, following the settings in restormer [54].

7.4 More Visual Comparisons on Real-World Image Super-Resolution

In order to further demonstrate the enhanced fidelity and level of detail exhibited in the images generated by our proposed VmambaIR model for real-world image super-resolution tasks, we present additional visual comparisons in this section. Specifically, we compare the images generated by VmambaIR with those produced by previous SOTA methods such as ESRGAN [61], BSRGAN [61], SwinIR [25], Real-ESRGAN [42], and MM-RealSR [34], as illustrated in Figure 8. Compared to other methods, VmambaIR excels in real-world super-resolution tasks by producing clearer and more accurate high-frequency details while maintaining significantly lower computational overhead.

7.5 More Visual Comparisons on Single Image Super-Resolution

In order to further demonstrate the enhanced fidelity and level of detail exhibited in the images generated by our proposed VmambaIR model for single image super-resolution tasks, we present additional visual comparisons in this section. Specifically, we compare the images generated by VmambaIR with those produced by previous SOTA methods such as SRGAN [22], ESRGAN [61], and BebyGAN [23], as illustrated in Figure 9. In the single image super-resolution task, our VmambaIR exhibits remarkable proficiency in restoring fine details across a broader range of test images. This serves as a testament to the robust high-frequency modeling capability of mamba [13] in image restoration.

7.6 More Visual Comparisons on Image Deraining

In order to further demonstrate the enhanced fidelity and level of detail exhibited in the images generated by our proposed VmambaIR model for image deraining tasks, we present additional visual comparisons in this section. Specifically, we compare the images generated by VmambaIR with those produced by previous SOTA methods such

as Restormer [54], RESCAN [24], DerainNet [10], as illustrated in Figure 10. In the image deraining task, our VmambaIR consistently achieves near-perfect restoration results across a diverse set of images. When compared to previous state-of-the-art methods [24, 54], our VmambaIR showcases distinct visual advantages and delivers superior performance, solidifying its position as a leading approach in the field.

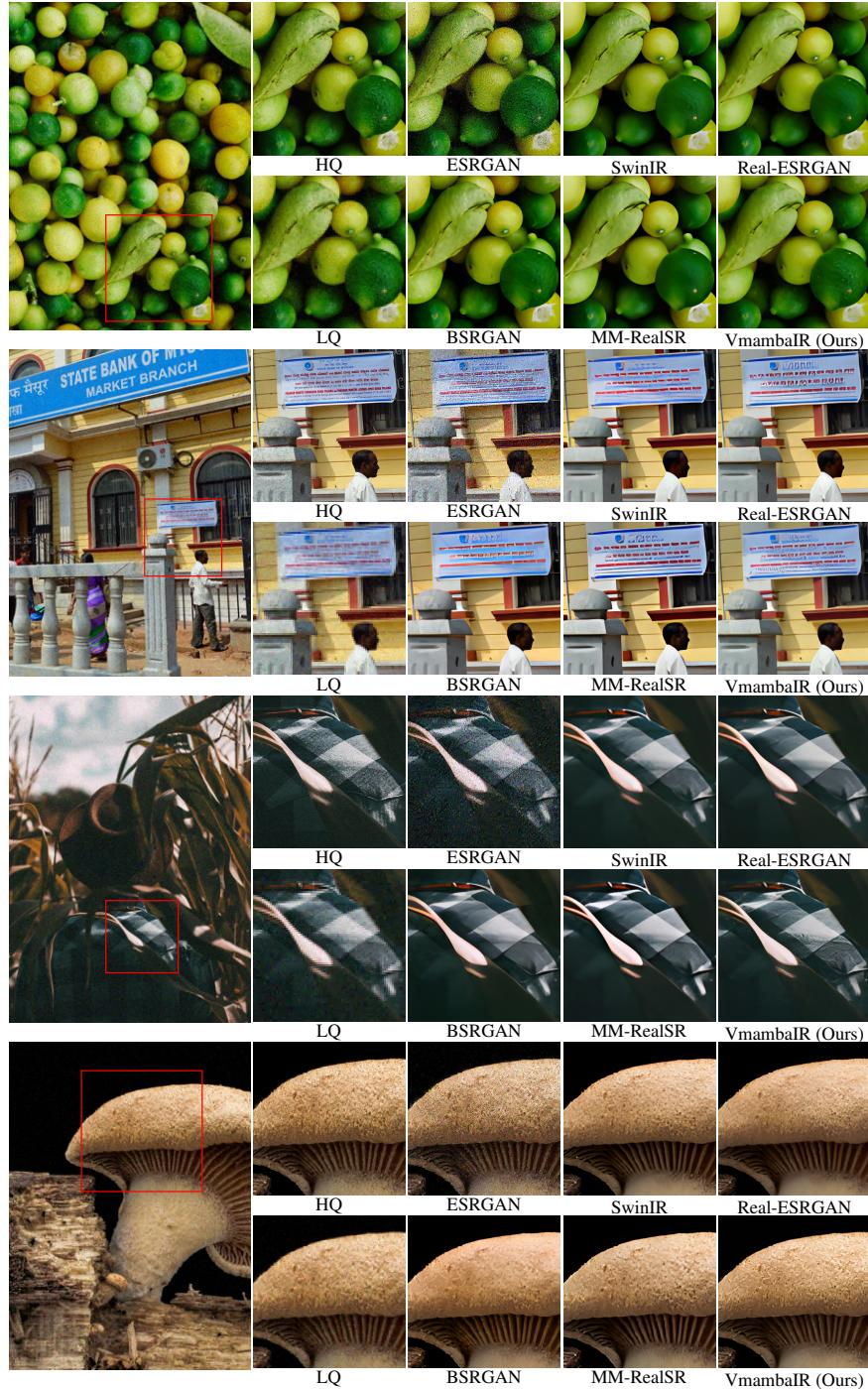


Fig. 8: Visual comparison of **real-world image super-resolution** methods. Zoom-in for better details. Our VmambaIR achieves superior high-frequency details and fidelity simultaneously.

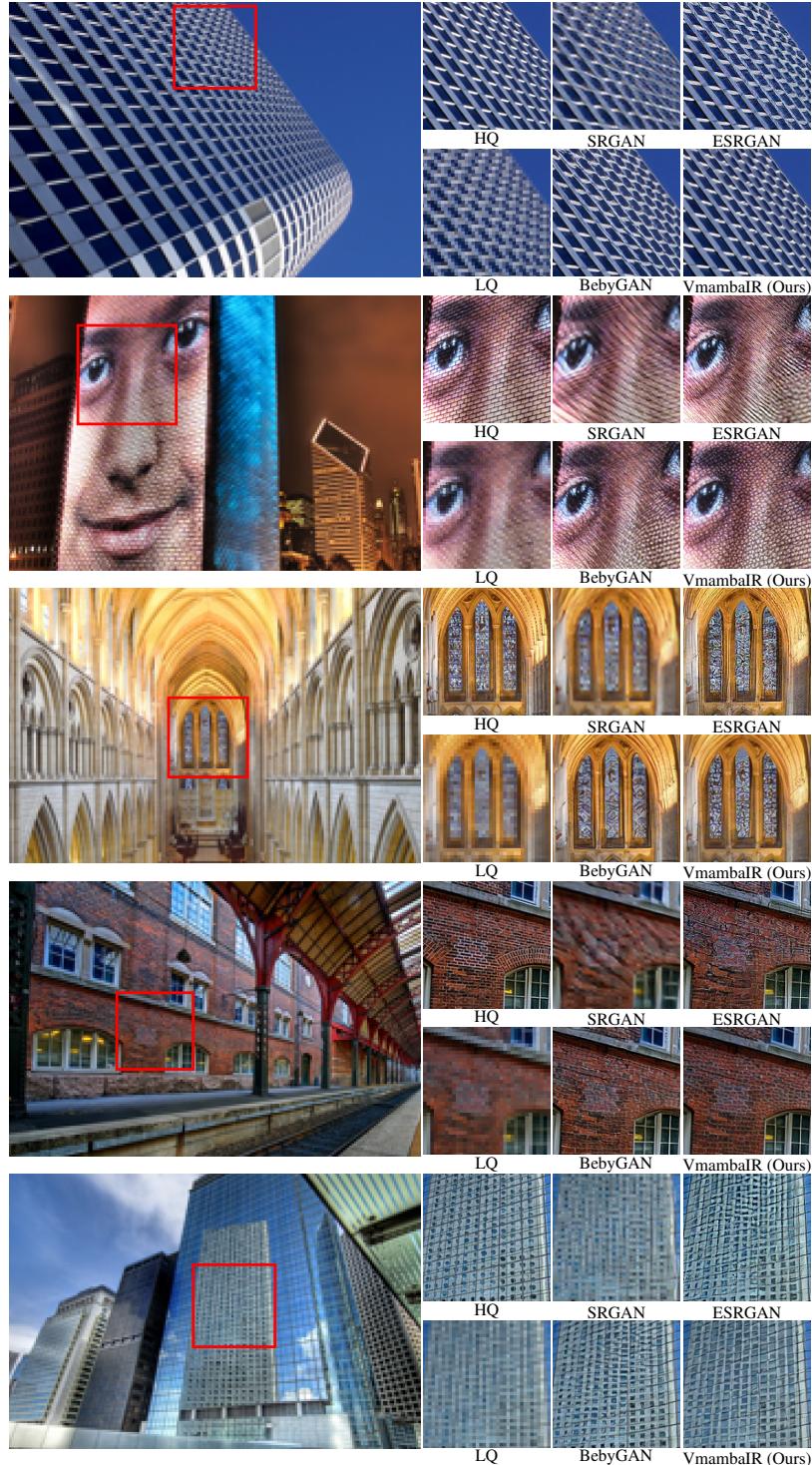


Fig. 9: Visual comparison of **single image super-resolution** methods. Zoom-in for better details.

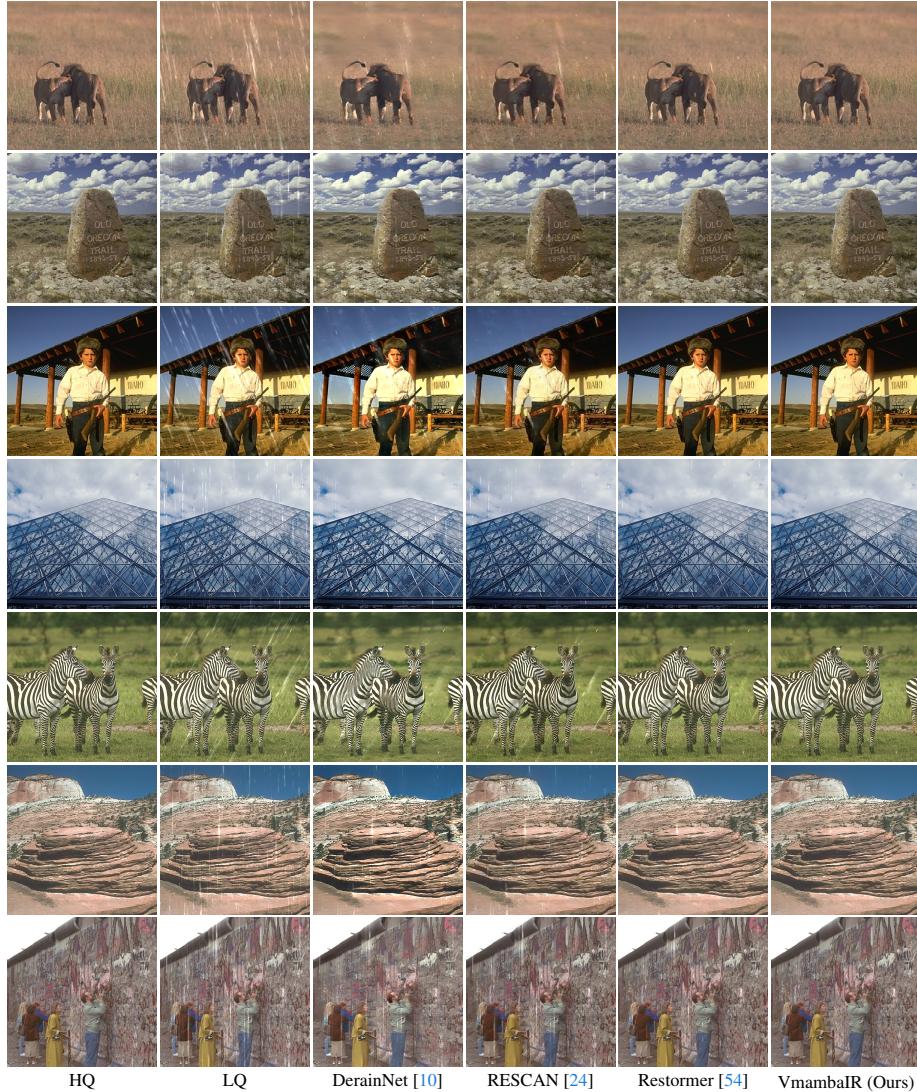


Fig. 10: Visual comparison of **image deraining** methods. Zoom-in for better details.