MambaIRv2: Attentive State Space Restoration

Hang Guo[1,*] Yong Guo[2,*] Yaohua Zha[1,6] Yulun Zhang[3] Wenbo Li[4]
Tao Dai[5,†] Shu-Tao Xia[1,6] Yawei Li[7]
[1]Tsinghua University [2]Max Planck Institute for Informatics
[3]Shanghai Jiao Tong University [4]The Chinese University of Hong Kong
[5]Shenzhen University [6]Peng Cheng Laboratory [7]ETH Zürich

Abstract

The Mamba-based image restoration backbones have recently demonstrated significant potential in balancing global reception and computational efficiency. However, the inherent causal modeling limitation of Mamba, where each token depends solely on its predecessors in the scanned sequence, restricts the full utilization of pixels across the image and thus presents new challenges in image restoration. In this work, we propose MambaIRv2, which equips Mamba with the non-causal modeling ability similar to ViTs to reach the attentive state space restoration model. Specifically, the proposed attentive state-space equation allows to attend beyond the scanned sequence and facilitate image unfolding with just one single scan. Moreover, we further introduce a semantic-guided neighboring mechanism to encourage interaction between distant but similar pixels. Extensive experiments show our MambaIRv2 outperforms SRFormer by ***even 0.35dB*** *PSNR for lightweight SR even with* ***9.3% less*** *parameters and suppresses HAT on classic SR by* ***up to 0.29dB****. Code is available at* https://github.com/csguoh/MambaIR.

1. Introduction

Image restoration aims to recover high-quality images from low-quality observations, tackling various sub-problems such as image super-resolution, image denoising, and JPEG compression reduction, and others. With the advent of deep learning, state-of-the-art performance has been consistently achieved. Early works primarily utilized convolutional neural networks (CNNs) as the backbone [12, 13, 30, 45, 53]. Later, vision transformers (ViTs) [15] gained popularity for their superior performance [5, 7, 10, 28, 29]. More recently, the selective state-space model (Mamba) [17] has been explored, showing considerable potential as an alternative backbone for image restoration tasks [18, 36].

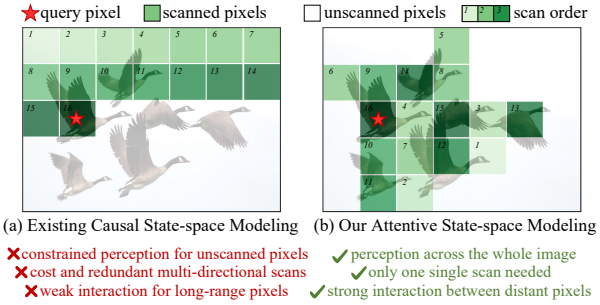

Figure 1. (a) The existing method [18] suffers from the adverse effects of the causal nature of Mamba (the multi-directional scans are not shown for presentation clarity). (b) The proposed MambaIRv2 can achieve attentive state-space modeling that embeds ViT-like non-causal properties into Mamba.

Despite its potential, existing Mamba-based methods face significant challenges, particularly due to their reliance on causal state-space modeling. Specifically, existing methods [18] unfold the 2D image with a predefined scanning rule to generate the 1D token sequence. However, in Mamba, each pixel is modeled based solely on its preceding pixels in the scanned sequence, *i.e.*, the causal property, which results in several detrimental effects for non-causal image restoration tasks. **First**, as shown in Fig. 1(a), the query pixel can only capture information from its preceding ones and cannot perceive subsequent pixels, which results in under-utilization of helpful pixels across the image. **Second**, the inherent causal property leads to the necessity of multi-directional scans, which is widely adopted by existing approaches [18, 36, 39] for mitigating information loss. Yet, this multi-scanning inevitably increases the computational complexity, particularly for high-resolution inputs. Furthermore, an empirical investigation in Sec. 3 reveals that there is also substantial information redundancy among these multi-directional scans. **Third**, our findings in Sec. 3 demonstrate that Mamba [17] is prone to long-range decay in token interaction, meaning distant tokens in the sequence have diminished interactions. Consequently, even previously scanned pixels that are distant yet relevant cannot be effectively utilized by the query pixel.

*Equal contribution
†Corresponding author

In this work, we propose MambaIRv2, aiming to address the adverse effect of causal state-space modeling. Since the ViTs [15, 37] naturally support non-causal processing, our key idea is to integrate ViT-like non-causal modeling into the Mamba-based methods. To this end, we begin by delving deep into the connection between attention and state space for valuable insights. Our in-depth analysis in Sec. 4.1 reveals that the output matrix of the state-space equation resembles the query in the attention mechanism. This similarity inspires us to utilize the output matrix to **"query"** relevant pixels in the unscanned sequence. Benefiting from attending beyond the scanned sequences, this strategy also naturally eliminates the need for multi-directional scanning. Moreover, to encourage the interaction between distant but relevant pixels, we propose to restructure the image to place similar pixels spatially closer within the 1D sequence. In this way, it allows for semantic rather than spatial sequence modeling, thus mitigating the impact of long-range decay. Since the proposed method allows the Mamba to behave similarly to the attention, we thus refer to it as **"attentive state-space restoration"**.

Overall, we make three key contributions: **I.** We propose the **Attentive State-space Equation (ASE)**, which leverages the prompt learning [22] within the original state space equation of Mamba to query semantically similar pixels beyond the scanned sequences. In detail, the prompts are designed to represent sets of pixels that are similar across the entire image, and we then incorporate the representative prompts into the output matrix of the state-space equation through residual addition to derive our ASE. As the core component, the proposed ASE not only alleviates the causal nature of Mamba for improved performance but also enables single-pass scanning for boosted efficiency. **II.** We further develop the **Semantic Guided Neighboring (SGN)** to encourage strong interaction between distant yet similar pixels. Specifically, we first assign the corresponding semantic label to each pixel. Then we restructure the image based on these labels to generate the semantic-neighboring 1D sequence, where semantically similar pixels are also spatially close to each other. Thanks to the mitigation of the long-range decay of Mamba, SGN facilitates effective interaction between pixels that are distant in the original image. **III.** Integrating the two core modules and other auxiliary parts, we present MambaIRv2, an attentive state-space restoration method that equips Mamba's state-space modeling with ViT-like non-causal capabilities. Extensive experiments demonstrate that MambaIRv2 significantly improves **both effectiveness and efficiency**. In particular, MmabaIRv2 outperforms state-of-the-art Transformer-based baseline SRFormer [58] by 0.35dB on Urban100 dataset for $2\times$ lightweight SR with 9.3% less parameters, and HAT [7] by 0.29dB for $2\times$ classical SR on the Manga109 dataset.

2. Related Work

Recent years have witnessed great advancements in the domain of image restoration [2, 20, 23]. Early attempts usually adopt the convolutional neural networks (CNNs), such as SRCNN [13] for image super-resolution, DnCNN [45] for image denoising, and ARCNN [14] for JPEG compression artifacts reduction. To further enhance the performance of CNN-based methods, various techniques have been introduced. For instance, EDSR [25] employ the residual connection strategy to allow the training of very deep neural networks, RDN [53] uses the dense connection to improve model representation ability. RCAN [52] introduces the channel attention for selecting salient channels, followed by SAN [12] which uses the second-order attention for performance improvement. Despite the great progress of CNN-based methods, the convolution operator inherently restricts the receptive field to the local kernel, preventing interaction between distant pixels.

As Transformer [37] has proven its effectiveness in multiple computer vision tasks, applying Transformer for image restoration thus appears to be promising. However, the direct application of vanilla self-attention, which exhibits quadratic computational complexity with the input size, is costly and impractical. To improve the efficiency of attention, a variety of techniques have been developed. For example, IPT [5] divides one image into several small patches and processes each patch independently with self-attention. After that, SwinIR [29] further introduces the shifted window self-attention [32] to improve the performance. ART [44] and OminiSR [38] utilize sparse attention to expand the receptive field by enlarging the attention windows. GRL [28] adopts the anchor attention to learn the local, regional, and global image hierarchies. Recently, the ATD [49] uses the adaptive token dictionary to store input-agnostic knowledge, thus allowing the attention to attend information out of the local window.

To balance the efficient computation and global receptive fields, the Mamba [17] has recently been explored in image restoration with promising results. MambaIR [18] is among the first to introduce Mamba for image restoration and addresses two specific challenges, *i.e.*, local pixel forgetting and channel redundancy. Since then, the Mamba model has been explored in various image restoration tasks. FreqMamba [56] uses the state space model in the Fourier domain for image deraining to perceive global degradation. MambaLLIE [39] improves the state space equation to allow the locality enhancement for low-light image enhancement tasks. Moreover, Mamba has also achieved promising results in image dehazing [57], debluring [16], and other tasks [3, 31, 34, 35, 40, 41, 50]. However, existing methods still struggle with the causal modeling nature of Mamba. Given image restoration as a non-causal task, this mismatch can lead to limited performance as well as inefficiency.

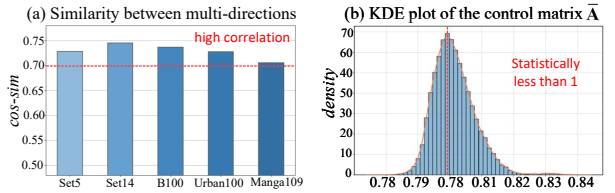

Figure 2. (a) We compute the cosine similarity of scanned features across all 4 directions and all layers in MambaIR [18]. (b) The kernel density estimation of the distribution of the control matrix in MambaIR [18].

3. Motivation

Mamba-based Image Restoration. The existing state-space restoration methods are mainly developed from the Mamba [17] architecture. Formally, the Mamba adopts the discrete state space equation to model the interaction among tokens:

$$\begin{aligned} h_i &= \overline{\mathbf{A}} h_{i-1} + \overline{\mathbf{B}} x_i, \\ y_i &= \mathbf{C} h_i + \mathbf{D} x_i, \end{aligned} \tag{1}$$

where the $\overline{\mathbf{A}} = \exp(\mathbf{\Delta A})$ is the control matrix, the $\overline{\mathbf{B}} = (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I})\mathbf{\Delta B} \approx \mathbf{\Delta B}$ is the input matrix, and the $\mathbf{C}$ is the output matrix. Eq. (1) indicates that the i-th token completely depends on its previous $i-1$ tokens, *i.e.*, the state-space modeling possesses causal properties. Although this causal nature is helpful for autoregressive tasks like NLP, it poses challenges for image restoration.

Challenges from Causal Modeling. Existing Mamba-based methods usually adopt a specific scanning strategy to unfold the 2D image into a 1D sequence for sequential modeling with Mamba. However, the i-the pixel can only see constrained $i-1$ pixels of the entire image, failing to globally utilize similar pixels. To this end, current methods typically employ multi-directional scans to allow for a broader receptive field, which is inevitably accompanied by an increase in computational complexity. Furthermore, as shown in Fig. 2(a), the similarity of different scanned sequences on all testing datasets reaches even above 0.7, indicating a high correlation with large redundancy. Moreover, the Mamba [17] itself also possesses long-range decay defects due to its causal nature. Specifically, the interaction between pixels can be quantitatively denoted by the power of control matrix $\overline{\mathbf{A}}^k$, where k is the pairwise distance (the proof is given in the *Suppl.*). In Fig. 2(b), we show that the learned $\overline{\mathbf{A}}$ is statistically less than 1. As a result, the interaction $\overline{\mathbf{A}}^k$ will become weak when two pixels are far apart, *i.e.*, a large k, indicating current methods fail to utilize distant but useful scanned pixels.

4. Attentive State Space Restoration

In the following, we aim to address the causal nature of Mamba through the proposed attentive state space restoration. To get started, we visit the mathematical connection between state space and attention in Sec. 4.1 for insights of subsequent design. Then, we detail the specific techniques of our attentive state space module in Sec. 4.2. In Sec. 4.3, we give the overall architecture of the proposed methods.

4.1. Bridging Attention and State-Space

As pointed out by [19], the state space has a strong relationship to attention, which may potentially offer insights to incorporate non-causal modeling ability into Mamba. In this section, we first reformulate attention and state space into the common form for comparison, followed by a detailed connection analysis.

Reformulation of Attention. Since the Mamba belongs to causal models with linear complexity, we adopt the corresponding causal linear attention [24] as its counterpart. Specifically, given the query, key and value matrix $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$, the output of linear attention is computed as follows:

$$y_i = \sum_{j=1}^{i} \frac{\mathbf{Q}_i \mathbf{K}_j^\top}{\sum_{t=1}^{i} \mathbf{Q}_i \mathbf{K}_t^\top} \mathbf{V}_j = \frac{\mathbf{Q}_i \left(\sum_{j=1}^{i} \mathbf{K}_j^\top \mathbf{V}_j \right)}{\mathbf{Q}_i \left(\sum_{t=1}^{i} \mathbf{K}_t^\top \right)}. \tag{2}$$

Denote the $\mathbf{S}_i = \sum_{j=1}^{i} \mathbf{K}_j^\top \mathbf{V}_j, \mathbf{Z}_i = \sum_{t=1}^{i} \mathbf{K}_t^\top$, then the formulation of the linear attention can be rewritten as:

$$y_i = \mathbf{Q}_i \mathbf{S}_i / \mathbf{Q}_i \mathbf{Z}_i, \tag{3}$$

where $\mathbf{S}_i = \mathbf{S}_{i-1} + \mathbf{K}_i^\top \mathbf{V}_i$, and $\mathbf{Z}_i = \mathbf{Z}_{i-1} + \mathbf{K}_i^\top$. To allow subsequent connection analysis, we further reformulate Eq. (3) to the common form as follows:

$$\begin{aligned} \mathbf{S}_i &= \mathbf{I}\mathbf{S}_{i-1} + \mathbf{K}_i^\top \mathbf{V}_i, \\ y_i &= \mathbf{Q}_i \mathbf{S}_i / \mathbf{Q}_i \mathbf{Z}_i + \mathbf{O} x_i, \end{aligned} \tag{4}$$

where the $\mathbf{I}$ and $\mathbf{O}$ denotes identity and zero matrix, respectively. x_i is the input token at the i-th step.

Reformulation of State Space. Starting with state-space equation in Eq. (1), note that $\overline{\mathbf{B}} x_i \approx \mathbf{\Delta B} x_i = \mathbf{B}(\mathbf{\Delta} x_i)$, Eq. (1) can then be reformulated to the common form as:

$$\begin{aligned} h_i &= \overline{\mathbf{A}} h_{i-1} + \mathbf{B}(\mathbf{\Delta} x_i), \\ y_i &= \mathbf{C} h_i / \mathbf{I} + \mathbf{D} x_i. \end{aligned} \tag{5}$$

Connection Analysis. By comparing Eq. (4) and Eq. (5), it can be seen there is a close mathematical similarity between attention and state space, *i.e.*, $h_i \sim \mathbf{S}_i$, $\mathbf{B} \sim \mathbf{K}^\top$, and $\mathbf{C} \sim \mathbf{Q}$. It should be noted that the $\mathbf{Q}_i \mathbf{Z}_i$ in Eq. (4) is to ensure the attention score is normalized with summation 1 and can be approximately ignored if we relax the normalization constraints. Therefore, the above observation motivates us to delve deep into the output matrix $\mathbf{C}$ in Eq. (1), which plays a role similar to the query in the attention mechanism. The core idea is to integrate the information of the unscanned sequence into $\mathbf{C}$, thus allowing $\mathbf{C}$ to attentively "query" unseen pixels to facilitate the restoration of x_i.

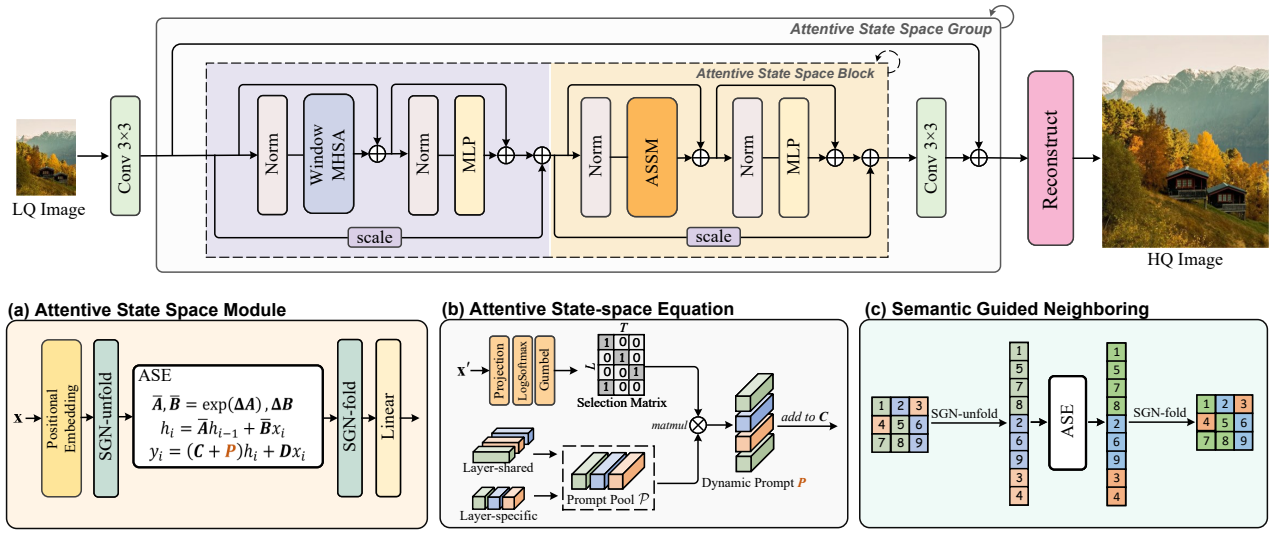

Figure 3. The overall architecture of our proposed MambaIRv2, as well as the (a) Attentive State Space Module (ASSM), (b) Attentive State-space Equition (ASE), and (c) Semantic Guided Neighboring (SGN).

4.2. Attentive State Space Module

In this section, we introduce the attentive state-space module (ASSM), which acts as the core block of our MambaIRv2 to enable non-causal modeling with Mamba. As shown in Fig. 3(a), given the input feature $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$, where H and W are the height and width, respectively, and C is the channel dimension, we first apply the positional encoding [11] on $\mathbf{x}$ to preserve the original structure information. After that, we propose the Semantic Guided Neighboring (SGN) to unfold the 2D image into 1D sequences for subsequent Attentive State-space Equation (ASE) modeling. Finally, another SGN is employed as the inverse operator of the previous one to fold the sequence back to the image followed by a linear projection to obtain the block output. More details are given below.

Attentive State-space Equation. As analyzed in Sec. 4.1, we aim to modify the output matrix $\mathbf{C} \in \mathbb{R}^{L\times d}$, where $L = HW$ is the flattened image sequence length and d is the number of hidden states in Mamba, to globally query related pixels across the image. To this end, we propose the Attentive State-space Equation (ASE) which develops from the original state-space equation of Mamba but possesses a non-causal nature. As shown in Fig. 3(b), the proposed ASE incorporates prompts, which learn to represent a certain set of pixels with similar semantics, into $\mathbf{C}$ to supplement the missing information of the unseen pixels. Specifically, we first construct the prompt pool $\mathcal{P} \in \mathbb{R}^{T\times d}$, where T is the number of prompts in $\mathcal{P}$. For parameterization of $\mathcal{P}$, we employ the semantic decoupling for better interpretability:

$$\mathcal{P} = \mathbf{MN}, \quad \mathbf{M} \in \mathbb{R}^{T\times r}, \quad \mathbf{N} \in \mathbb{R}^{r\times d}, \tag{6}$$

where $\mathbf{N}$ is shared across different blocks, $\mathbf{M}$ is block-specific, and r is the inner rank with $r \ll \min\{T, d\}$. The main idea behind the semantic decoupling is that we want different blocks to share similar feature space, *i.e.*, $\mathbf{N}$ is shared, while the combination coefficients of the shared features can be varying for different blocks *i.e.*, $\mathbf{M}$ is specific.

After that, we develop routing strategies to select from $\mathcal{P}$ to obtain L instance-specific prompts $\mathbf{P} \in \mathbb{R}^{L\times d}$, which will be added into $\mathbf{C}$ to include information of unscanned pixels. In detail, given the flattened input feature $\mathbf{x}' \in \mathbb{R}^{L\times C}$, we employ a linear layer to project the channel dimension of $\mathbf{x}'$ from C to T, followed by the $\mathrm{LogSoftmax}$ to predict the log probability, which indicates the probability of each prompt in $\mathcal{P}$ being sampled by $\mathbf{x}'_i, i = 1, 2, \cdots L$. After that, we introduce the gumbel-softmax [21] trick on the log probability to allow differentiable prompt selection to obtain the one-hot routing matrix $\mathbf{R} \in \mathbb{R}^{L\times T}$. Then, the instance-specific prompt $\mathbf{P} \in \mathbb{R}^{L\times d}$ is generated through the matrix multiplication as $\mathbf{P} = \mathbf{R}\mathcal{P}$. Finally, we incorporate $\mathbf{P}$ into $\mathbf{C}$ through residual addition to formulate our attentive state-space equation:

$$\begin{aligned} h_i &= \overline{\mathbf{A}} h_{i-1} + \overline{\mathbf{B}} x_i, \\ y_i &= (\mathbf{C} + \mathbf{P}) h_i + \mathbf{D} x_i. \end{aligned} \tag{7}$$

The learned prompts allow for the attention-like capability to query pixels across the whole image, and we also present a visualization on the attentive map in Fig. 5. By injecting the prompts that represent the set of similar pixels, the proposed ASE can effectively alleviate constrained perception for unscanned pixels. As another advantage, it allows scanning with only one single direction, eliminating the high computational cost and redundancy of multi-directional scans in existing methods.

Semantic Guided Neighboring. As pointed out in Sec. 3, the causal modeling property of Mamba leads to the detrimental effects of long-range decay. In existing Mamba-based image restoration methods, pixels that are distant in the original image are usually still far apart in the unfolded sequence, causing the weak utilization of the query pixel for the already scanned pixels which are spatially distant but similar. To this end, we propose the Sematic Guided Neighboring (SGN) as shown in Fig. 3(c). Our key insight is that different from the autoregressive language modeling, the image restoration is a non-causal task and all pixels are observable at once, therefore we can re-define the token neighborhood to enable semantically similar tokens to be spatially closer in the unfolded sequence. Following this idea, we first determine the semantic label of each pixel. Note that the routing matrix $\mathbf{R}$ in the ASE, which has learned the prompt category of each pixel, we thus employ this off-the-shelf semantics to restructure the image. Specifically, we propose the SGN-unfold which groups pixels with the i-th prompt category together to form the i-th semantic group and then combines different groups according to the category value i to generate the semantic-neighbored sequence. After that, we feed this sequence into the proposed ASE for state-space modeling. At last, we employ the SGN-fold which performs as the inverse transformation of SGN-unfold to reshape the semantic-space sequence back to the spatial-space feature map to obtain the output.

4.3. Overall Network Architecture

The proposed ASSM can efficiently capture the global dependencies using the Mamba model. We then further consider modeling local interactions, which have been shown crucial for Mamba-based approaches [18, 39]. Since the scanning only once in the ASSM provides more parameter budgets, we thus opt for the powerful window multi-head self-attention (MHSA) [32] to enhance local interactions within the window, which together with the ASSM compose the basic elements of our MambaIRv2. As shown in Fig. 3, given a low-quality image as the input, we first utilize the 3×3 convolution layer to extract shallow features. Then the shallow features are fed into several Attentive State Space Groups (ASSGs), where each groups contain multiple Attentive State Space Blocks (ASSBs). For each ASSB, we consider progressive local-to-global modeling to form the image hierarchy [28]. We use Norm and Token Mixer, followed by Norm and FFN to form the template, and employ window MHSA and ASSM as the instantiations for the Token Mixer of the local and global parts, respectively. In addition, two residual connections with learnable scales are introduced [9, 18]. After the ASSGs, we utilize the task-specific reconstruction modules, *e.g.*, pixelshuffle for super-resolution, and convolution for denoising, to obtain the high-quality image output.

Table 1. Ablation on the effectiveness of different components.

MHSA	ASE	SGN	Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM
✔			32.89	0.9343	39.11	0.9772
✔	✔		32.94	0.9351	39.20	0.9780
✔	✔	✔	**32.97**	**0.9355**	**39.24**	**0.9784**

Table 2. Ablation experiments on different injection positions of the learnable prompts in the ASE.

positions	Set14		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
B	33.97	0.9215	32.96	**0.9356**	39.23	0.9781
Δ	33.92	0.9211	32.93	0.9350	39.19	0.9779
y	33.97	0.9210	32.94	0.9351	39.21	0.9782
C	**33.99**	**0.9216**	**32.97**	0.9355	**39.24**	**0.9784**

5. Experiments

Following previous image restoration works [18, 29], we conduct experiments on three representative image restoration tasks, *i.e.*, image super-resolution including classic SR and lightweight SR, JPEG compression artifact reduction (JPEG CAR), and Gaussian color image denoising.

5.1. Experimental Settings

In accordance with previous works, we perform data augmentation by applying horizontal flips and random rotations of $90°, 180°$, and $270°$. Additionally, we crop the original images into 64×64 patches for image SR and 128×128 patches for image denoising during training. For image SR, we use the pre-trained weights from the $2\times$ model to initialize those of $3\times$ and $4\times$ and halve the learning rate and total training iterations to reduce training time [30]. To ensure a fair comparison, we adjust the training batch size to 32 for image SR and 8 for image denoising and JPEG CAR. We employ the Adam [26] as the optimizer for training our MambaIRv2 with $\beta_1 = 0.9, \beta_2 = 0.999$. Similar to previous training protocol [29], we use the L_1 loss for image SR, and Charbonnier loss [4] for denoising and JPEG CAR. The initial learning rate is set at 2×10^{-4} and is halved when the training iteration reaches specific milestones. For classic image SR, we provide three variants with different parameters including the small, base, and large versions (MambaIRv2-S, MambaIRv2-B, MambaIRv2-L). Due to the page limit, more details are provided in the *Suppl.*

5.2. Ablation Study

We conduct ablations with MambaIRv2-light $2\times$ SR model trained for 250K iterations on the DIV2K dataset.

Effectiveness of Different Components. As the core module of MambaIRv2, the Attentive State Space Module (ASSM), which contains the Attentive State-space Equation (ASE) and the Semantic Guided Neighboring (SGN), plays an important role in Mamba-based global modeling. In this ablation, we design three settings to verify the role of the different components. In the first setup, we directly re-

Table 3. Quantitative comparison on ***lightweight image super-resolution*** with state-of-the-art methods. The best and the second best results are in red and blue.

Method	scale	#param	MACs	Set5		Set14		BSDS100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CARN [1]	2×	1,592K	222.8G	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
LatticeNet [33]	2×	756K	169.5G	38.13	0.9610	33.78	0.9193	32.25	0.9005	32.43	0.9302	-	-
SwinIR-light [29]	2×	910K	244.2G	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
MambaIR-light [18]	2×	905K	334.2G	38.13	0.9610	33.95	0.9208	32.31	0.9013	32.85	0.9349	39.20	0.9782
ELAN [51]	2×	621K	203.1G	38.17	0.9611	33.94	0.9207	32.30	0.9012	32.76	0.9340	39.11	0.9782
SRFormer-light [58]	2×	853K	236.3G	38.23	0.9613	33.94	0.9209	32.36	0.9019	32.91	0.9353	39.28	0.9785
MambaIRv2-light	2×	774K	286.3G	38.26	0.9615	34.09	0.9221	32.36	0.9019	33.26	0.9378	39.35	0.9785
CARN [1]	3×	1,592K	118.8G	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
LatticeNet [33]	3×	765K	76.3G	34.53	0.9281	30.39	0.8424	29.15	0.8059	28.33	0.8538	-	-
SwinIR-light [29]	3×	918K	111.2G	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624	33.98	0.9478
MambaIR-light	3×	913K	148.5G	34.63	0.9288	30.54	0.8459	29.23	0.8084	28.70	0.8631	34.12	0.9479
ELAN [51]	3×	629K	90.1G	34.61	0.9288	30.55	0.8463	29.21	0.8081	28.69	0.8624	34.00	0.9478
SRformer-light [58]	3×	861K	105.4G	34.67	0.9296	30.57	0.8469	29.26	0.8099	28.81	0.8655	34.19	0.9489
MambaIRv2-light	3×	781K	126.7G	34.71	0.9298	30.68	0.8483	29.26	0.8098	29.01	0.8689	34.41	0.9497
CARN [1]	4×	1,592K	90.9G	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
LatticeNet [33]	4×	777K	43.6G	32.30	0.8962	28.68	0.7830	27.62	0.7367	26.25	0.7873	-	-
SwinIR-light [29]	4×	930K	63.6G	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
MambaIR-light [18]	4×	924K	84.6G	32.42	0.8977	28.74	0.7847	27.68	0.7400	26.52	0.7983	30.94	0.9135
ELAN [51]	4×	640K	54.1G	32.43	0.8975	28.78	0.7858	27.69	0.7406	26.54	0.7982	30.92	0.9150
SRformer-light [58]	4×	873K	62.8G	32.51	0.8988	28.82	0.7872	27.73	0.7422	26.67	0.8032	31.17	0.9165
MambaIRv2-light	4×	790K	75.6G	32.51	0.8992	28.84	0.7878	27.75	0.7426	26.82	0.8079	31.24	0.9182

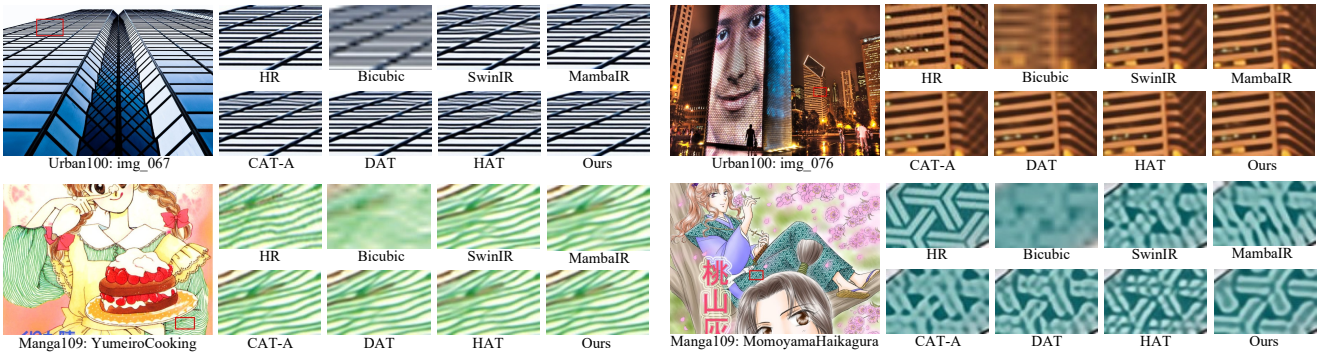

Figure 4. Qualitative comparison of our MambaIRv2 with different methods on 4× classic image SR.

move the ASSM, leading to a pure Attention variant. In the second setup, we add the proposed ASSM but remove the SGN. The third setup corresponds to our proposed method. As shown in Tab. 1, only using window attention limits the receptive field to the local window, limiting the performance. Further, the addition of the ASE, which allows for the query of similar pixels across images, improves the performance by 0.05 dB/0.09 dB in Urban100/Manga109. Finally, after introducing SGN which can effectively overcome the long-range decay of Mamba, the final model achieves the best performance of 32.97/39.24 dB PSNR on Urban100/Manga109. The experiments validate the effectiveness of different components in the proposed method.

Ablation on Attentive State-space Equation. In Sec. 4.1, we pointed out that the matrix $\mathbf{C}$ in the state-space equation in fact behaves similarly to the Query in the attention, following which we propose the ASE to insert global prompts into $\mathbf{C}$ by residual addition. In Tab. 2, we explore other inserting positions of the prompts in ASE. Since $\mathbf{B}$ can be analogized to the Key in attention, inserting the prompts into $\mathbf{B}$ also yields decent results, but slightly inferior to adding to $\mathbf{C}$. We attribute this to the fact that $\mathbf{C}$ is closer to the output end in the state-space equation Eq. (1), leading to a greater impact on the final result. Further, inserting prompts to the discrete time-step $\mathbf{\Delta}$ and the output $\mathbf{y}$ both fail to give satisfactory performance, which justifies our focus on the matrix $\mathbf{C}$ in the proposed ASE. Due to the page limit, we provide more ablation experiments in the *Suppl.*.

5.3. Comparison on Image Super-Resolution

Lightweight Image Super-Resolution. Following previous works [33, 58], we also report the number of parameters (#param) and MACs (upscaling a low-resolution image to 1280×720 resolution) as the efficiency metric. The results in Tab. 3 show that our MambaIRv2 outperforms the state-of-the-art method SRFormer [58] using even significantly fewer parameters. For instance, our MambaIRv2-light outperforms SRFormer-light by 0.35dB on 2× Urban100 with

Table 4. Quantitative comparison on ***classic image super-resolution*** with state-of-the-art methods.

Method	scale	#param	Set5		Set14		BSDS100		Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [30]	2×	42.6M	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
RCAN [52]	2×	15.4M	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN [12]	2×	15.7M	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
IPT [5]	2×	115M	38.37	-	34.43	-	32.48	-	33.76	-	-	-
SwinIR [29]	2×	11.8M	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
EDT [27]	2×	11.5M	38.45	0.9624	34.57	0.9258	32.52	0.9041	33.80	0.9425	39.93	0.9800
MambaIR [18]	2×	20.4M	38.57	0.9627	34.67	0.9261	32.58	0.9048	34.15	0.9446	40.28	0.9806
CAT-A [8]	2×	16.5M	38.51	0.9626	34.78	0.9265	32.59	0.9047	34.26	0.9440	40.10	0.9805
DAT [10]	2×	14.8M	38.58	0.9629	34.81	0.9297	32.61	0.9051	34.37	0.9458	40.33	0.9807
HAT [7]	2×	20.6M	38.63	0.9630	34.86	0.9274	32.62	0.9053	34.45	0.9466	40.26	0.9809
MambaIRv2-S	2×	9.6M	38.53	0.9627	34.62	0.9256	32.59	0.9048	34.24	0.9454	40.27	0.9808
MambaIRv2-B	2×	22.9M	38.65	0.9631	34.89	0.9275	32.62	0.9053	34.49	0.9468	40.42	0.9810
MambaIRv2-L	2×	34.2M	38.65	0.9632	34.93	0.9276	32.62	0.9053	34.60	0.9475	40.55	0.9807
EDSR [30]	3×	42.6M	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
RCAN [52]	3×	15.4M	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
SAN [12]	3×	15.7M	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
IPT [5]	3×	115M	34.81	-	30.85	-	29.38	-	29.49	-	-	-
SwinIR [29]	3×	11.8M	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
EDT [27]	3×	11.5M	34.97	0.9316	30.89	0.8527	29.44	0.8142	29.72	0.8814	35.13	0.9534
MambaIR [18]	3×	20.4M	35.08	0.9323	30.99	0.8536	29.51	0.8157	29.93	0.8841	35.43	0.9546
CAT-A [8]	3×	16.5M	35.06	0.9326	31.04	0.8538	29.52	0.8160	30.12	0.8862	35.80	0.9546
DAT [10]	3×	14.8M	35.16	0.9331	31.11	0.8550	29.55	0.8169	30.18	0.8886	35.59	0.9554
HAT [7]	3×	20.6M	35.07	0.9329	31.08	0.8555	29.54	0.8167	30.23	0.8896	35.53	0.9552
MambaIRv2-S	3×	9.8M	35.09	0.9326	31.07	0.8547	29.51	0.8157	30.08	0.8871	35.44	0.9549
MambaIRv2-B	3×	23.1M	35.18	0.9334	31.12	0.8557	29.55	0.8169	30.28	0.8905	35.61	0.9556
MambaIRv2-L	3×	34.2M	35.16	0.9334	31.18	0.8564	29.57	0.8175	30.34	0.8912	35.72	0.9561
EDSR [30]	4×	43.0M	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
RCAN [52]	4×	15.6M	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN [12]	4×	15.7M	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
IPT [5]	4×	116M	32.64	-	29.01	-	27.82	-	27.26	-	-	-
SwinIR [29]	4×	11.9M	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
EDT [27]	4×	11.6M	32.82	0.9031	29.09	0.7939	27.91	0.7483	27.46	0.8246	32.05	0.9254
MambaIR [18]	4×	20.4M	33.03	0.9046	29.20	0.7961	27.98	0.7503	27.68	0.8287	32.32	0.9272
CAT-A [8]	4×	16.6M	33.08	0.9052	29.18	0.7960	27.99	0.7510	27.89	0.8339	32.39	0.9285
DAT [10]	4×	14.8M	33.08	0.9055	29.23	0.7973	28.00	0.7515	27.87	0.8343	32.51	0.9291
HAT [7]	4×	20.8M	33.04	0.9056	29.23	0.7973	28.00	0.7517	27.97	0.8368	32.48	0.9292
MambaIRv2-S	4×	9.8M	32.99	0.9037	29.23	0.7965	27.97	0.7502	27.73	0.8307	32.33	0.9276
MambaIRv2-B	4×	23.1M	33.14	0.9057	29.23	0.7975	28.00	0.7511	27.89	0.8344	32.57	0.9295
MambaIRv2-L	4×	34.2M	33.19	0.9062	29.29	0.7982	28.01	0.7521	28.07	0.8383	32.66	0.9304

Table 5. Complexity comparison with state-of-the-art methods on the 2× classic SR with output size 256 × 256.

Models	#param	MACs	Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM
CAT-A [8]	16.6M	350.7G	34.26	0.9440	40.10	0.9805
DAT [10]	14.8M	265.7G	34.37	0.9458	40.33	0.9807
HAT [7]	20.8M	514.9G	34.45	0.9466	40.26	0.9809
MambaIRv2-S	9.6M	192.9G	34.24	0.9454	40.27	0.9808
MambaIRv2-B	22.9M	445.8G	34.49	0.9468	40.42	0.9810
MambaIRv2-L	34.2M	664.5G	34.60	0.9475	40.55	0.9807

79K fewer #param. This experiment validates the efficiency and effectiveness of the proposed method.

Classic Image Super-Resolution. Tab. 4 gives the comparison results of our MambaIRv2 with different model sizes to existing state-of-the-art classic SR methods. Thanks to the attentive state space modeling, our proposed method achieves the best performance across most five benchmark datasets and up-sample scales. For example, our MambaIRv2-B exceeds HAT [7] by 0.16dB on the 2× Manga109 dataset. Interestingly, even the MambaIRv2-S with 9.6M #param, outperforms the previous 20.4M MambaIR [18] by 0.09dB PSNR on the 2× Urban100 dataset, further demonstrating our MambaIRv2 serves as an elegant balance of performance and efficiency. Finally, our method also shows promising scaling-up capabilities. When we scale up the #param to 34.2M to obtain the MambaIRv2-L model, this larger model achieved steady performance gains compared to its base counterpart, *e.g.*, 0.18dB PSNR gains on 4× Urban100. The visual comparisons are shown in Fig. 4, and our method can facilitate the reconstruction of sharp edges and natural textures.

Model Complexity Comparison. As shown in Tab. 5, our MambaIRv2-S model, which uses only 55.0% of the MACs compared to CAT-A [8], outperforms CAT-A by 0.17dB PSNR on Manga109. Additionally, our MambaIRv2-B model, which roughly matches the #param of HAT [7], achieves a 13.4% reduction in MACs, while delivering 0.04/0.16dB PSNR improvements on Urban100/Manga109. The above results demonstrate that our MambaIRv2 strikes a sweet spot between performance and efficiency.

Table 6. Quantitative comparison on ***JPEG compression artifact reduction*** under different quality factors q.

Dataset	q	RNAN [54]		RDN [55]		DRUNet [48]		SwinIR [29]		MambaIR [18]		Ours	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Classic5	10	29.96	0.8178	30.00	0.8188	30.16	0.8234	30.27	0.8249	30.27	0.8256	30.37	0.8269
	30	33.38	0.8924	33.43	0.8930	33.59	0.8949	33.73	0.8961	33.74	0.8965	33.81	0.8970
	40	34.27	0.9061	34.27	0.9061	34.41	0.9075	34.52	0.9082	34.53	0.9084	34.64	0.9093
LIVE1	10	29.63	0.8239	29.67	0.8247	29.79	0.8278	29.86	0.8287	29.88	0.8301	29.91	0.8301
	30	33.45	0.9149	33.51	0.9153	33.59	0.9166	33.69	0.9174	33.72	0.9179	33.73	0.9179
	40	34.47	0.9299	34.51	0.9302	34.58	0.9312	34.67	0.9317	34.70	0.9320	34.73	0.9323

Table 7. Quantitative comparison of PSNR on ***gaussian color image denoising*** $\sigma = 15$ with state-of-the-art methods.

Method	CBSD68	Kodak24	McMaster	Urban100
IRCNN [46]	33.86	34.69	34.58	33.78
FFDNet [47]	33.87	34.63	34.66	33.83
DnCNN [45]	33.90	34.60	33.45	32.98
DRUNet [48]	34.30	35.31	35.40	34.81
SwinIR [29]	34.42	35.34	35.61	35.13
Restormer [42]	34.40	35.35	35.61	35.13
Xformer [43]	34.43	35.39	35.68	35.29
MambaIR [18]	34.48	35.42	35.70	35.37
MambaIRv2	34.48	35.43	35.73	35.42

Table 8. Comparison to MambaIR with different scanning modes on $2\times$ lightweight SR. The "MambaIR-n" indicates the MambaIR [18] variant with n scanning directions.

settings	#param	MACs	**Urban100**		**Manga109**	
			PSNR	SSIM	PSNR	SSIM
MambaIR-1 [18]	987K	291G	32.82	0.9348	39.19	0.9776
MambaIR-2 [18]	1.11M	383G	32.86	0.9450	39.26	0.9778
MambaIR-4 [18]	1.36M	568G	32.92	0.9356	39.31	0.9779
MambaIRv2	**774K**	**286G**	**33.26**	**0.9378**	**39.39**	**0.9786**

5.4. Comparison on JPEG CAR

Tab. 6 gives the results on JPEG compression reduction. It can be seen that the proposed MambaIRv2 achieves the best performance on all testing datasets across all quality factors. For example, our MambaIRv2 suppresses MambaIR [18] by 0.11dB PSNR with $q = 40$ on the Classic5 dataset, demonstrating the effectiveness of our MambaIRv2 on other restoration tasks.

5.5. Comparison on Image Denoising

We further include the gaussian color image denoising task for further validation. Notably, to maintain architectural consistency across different restoration tasks, we retain the straight-through structure and avoid using the UNet architecture, which has been shown to be particularly advantageous for denoising tasks [6]. The results presented in Tab. 7 demonstrate that MambaIRv2 outperforms all other models across the datasets. In particular, it surpasses U-shaped Restormer [42] by even 0.29dB PSNR on the Urban100 dataset. This experiment validates our MambaIRv2 serves as a generalized image restoration backbone.

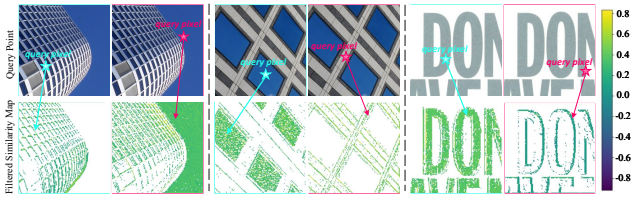

Figure 5. The visualization of the attentive state space. We compute the cosine similarity between the prompt corresponding to the query pixel and the matrix $\mathbf{C}$. We filter out low-similarity points for presentation clarity. More examples are provided in the *Suppl.*.

5.6. Discussion

Benefits from Reduced Scan Directions. Compared to the previous MambaIR [18], which performs 4 scans in pixel space, a significant advantage of our MambaIRv2 is that it requires only a single scan in the semantic space. As shown in Tab. 8, our MambaIRv2 is not only efficient but also boosts performance. For example, MambaIRv2 reduces even **43%** of the number of parameters and **50%** computational burden compared to the standard MambaIR, while still suppresses by **0.34dB** PSNR on $2\times$ Urban100.

Visualization of Attentive State Space. In the proposed attentive state space equation, the prompts play an important role in representing similar pixels across the whole image to facilitate the query pixel seeing out of the scanned sequence. As shown in Fig. 5, it can be seen that the query pixel is empowered to attend to its corresponding semantic part in the image through the prompt, thus enabling global information aggregation similar to attention mechanism.

6. Conclusion

In this work, we introduce MambaIRv2 to enhance state-space restoration models by addressing the causal modeling nature of Mamba. We propose the attentive state-space equation that incorporates prompt learning for enlarged token perception as well as scanning only once. Additionally, we introduce semantic guided neighboring which positions similar pixels closer to handle the long-range decay. These innovations enable MambaIRv2 to integrate ViT-like non-causal abilities into Mamba-based models to implement the attentive state space restoration. Extensive experiments confirm our MambaIRv2 as an efficient, high-performing backbone for image restoration.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China, under Grant (62302309, 62171248), Shenzhen Science and Technology Program (JCYJ20220818101014030, JCYJ20220818101012025), and the PCNL KEY project (PCL2023AS6-1).

References

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*, pages 252–268, 2018. 6

[2] Anas M Ali, Bilel Benjdira, Anis Koubaa, Walid El-Shafai, Zahid Khan, and Wadii Boulila. Vision transformers in image restoration: A survey. *Sensors*, 23(5):2385, 2023. 2

[3] Jiesong Bai, Yuhao Yin, Qiyuan He, Yuanxian Li, and Xiaofeng Zhang. RetinexMamba: Retinex-based Mamba for low-light image enhancement. *arXiv preprint arXiv:2405.03349*, 2024. 2

[4] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICIP*, pages 168–172. IEEE, 1994. 5

[5] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, pages 12299–12310, 2021. 1, 2, 7

[6] Xiangyu Chen, Zheyuan Li, Yuandong Pu, Yihao Liu, Jiantao Zhou, Yu Qiao, and Chao Dong. A comparative study of image restoration networks for general backbone network design. *arXiv preprint arXiv:2310.11881*, 2023. 8

[7] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, pages 22367–22377, 2023. 1, 2, 7

[8] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Cross aggregation transformer for image restoration. *NeurIPS*, 35:25478–25490, 2022. 7

[9] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, and Xiaokang Yang. Recursive generalization transformer for image super-resolution. *arXiv preprint arXiv:2303.06373*, 2023. 5

[10] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *ICCV*, pages 12312–12321, 2023. 1, 7

[11] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021. 4

[12] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, pages 11065–11074, 2019. 1, 2, 7

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199. Springer, 2014. 1, 2

[14] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *ICCV*, pages 576–584, 2015. 2

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[16] Hu Gao, Bowen Ma, Ying Zhang, Jingfan Yang, Jing Yang, and Depeng Dang. Learning enriched features via selective state spaces model for efficient image deblurring. In *ACM MM*, pages 710–718, 2024. 2

[17] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 2, 3

[18] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. MambaIR: A simple baseline for image restoration with state-space model. In *ECCV*, pages 222–241. Springer, 2025. 1, 2, 3, 5, 6, 7, 8

[19] Dongchen Han, Ziyi Wang, Zhuofan Xia, Yizeng Han, Yifan Pu, Chunjiang Ge, Jun Song, Shiji Song, Bo Zheng, and Gao Huang. Demystify Mamba in vision: A linear attention perspective. *arXiv preprint arXiv:2405.16605*, 2024. 3

[20] Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *IEEE TPAMI*, 2025. 2

[21] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4

[22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727. Springer, 2022. 2

[23] Junjun Jiang, Zengyuan Zuo, Gang Wu, Kui Jiang, and Xianming Liu. A survey on all-in-one image restoration: Taxonomy, evaluation and future trends. *arXiv preprint arXiv:2410.15067*, 2024. 2

[24] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. pages 5156–5165. PMLR, 2020. 3

[25] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016. 2

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[27] Wenbo Li, Xin Lu, Shengju Qian, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer-based image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 7

[28] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *CVPR*, pages 18278–18289, 2023. 1, 2, 5

[29] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCVW*, pages 1833–1844, 2021. 1, 2, 5, 6, 7, 8

[30] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, pages 136–144, 2017. 1, 5, 7

[31] Wei-Tung Lin, Yong-Xiang Lin, Jyun-Wei Chen, and Kai-Lung Hua. PixMamba: Leveraging state space models in a dual-level architecture for underwater image enhancement. *arXiv preprint arXiv:2406.08444*, 2024. 2

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 2, 5

[33] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *ECCV*, pages 272–289. Springer, 2020. 6

[34] Junbo Qiao, Jincheng Liao, Wei Li, Yulun Zhang, Yong Guo, Yi Wen, Zhangxizi Qiu, Jiao Xie, Jie Hu, and Shaohui Lin. Hi-Mamba: Hierarchical Mamba for efficient image super-resolution. *arXiv preprint arXiv:2410.10140*, 2024. 2

[35] Shiyu Qin, Jinpeng Wang, Yimin Zhou, Bin Chen, Tianci Luo, Baoyi An, Tao Dai, Shutao Xia, and Yaowei Wang. MambaVC: Learned visual compression with selective state spaces. *arXiv preprint arXiv:2405.15413*, 2024. 2

[36] Yuan Shi, Bin Xia, Xiaoyu Jin, Xing Wang, Tianyu Zhao, Xin Xia, Xuefeng Xiao, and Wenming Yang. VmambaIR: Visual state space model for image restoration. *arXiv preprint arXiv:2403.11423*, 2024. 1

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2

[38] Hang Wang, Xuanhong Chen, Bingbing Ni, Yutian Liu, and Jinfan Liu. Omni aggregation networks for lightweight image super-resolution. In *CVPR*, pages 22378–22387, 2023. 2

[39] Jiangwei Weng, Zhiqiang Yan, Ying Tai, Jianjun Qian, Jian Yang, and Jun Li. MambaLLIE: Implicit retinex-aware low light enhancement with global-then-local state space. *arXiv preprint arXiv:2405.16105*, 2024. 1, 2, 5

[40] Hongtao Wu, Yijun Yang, Huihui Xu, Weiming Wang, Jinni Zhou, and Lei Zhu. RainMamba: Enhanced locality learning with state space models for video deraining. In *ACM MM*, pages 7881–7890, 2024. 2

[41] Xinyu Xie, Yawen Cui, Chio-In Ieong, Tao Tan, Xiaozhi Zhang, Xubin Zheng, and Zitong Yu. FusionMamba: Dynamic feature enhancement for multimodal image fusion with Mamba. *arXiv preprint arXiv:2404.09498*, 2024. 2

[42] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022. 8

[43] Jiale Zhang, Yulun Zhang, Jinjin Gu, Jiahua Dong, Linghe Kong, and Xiaokang Yang. Xformer: Hybrid x-shaped transformer for image denoising. *arXiv preprint arXiv:2303.06440*, 2023. 8

[44] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. In *ICLR*, 2023. 2

[45] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE TIP*, 26(7):3142–3155, 2017. 1, 2, 8

[46] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *CVPR*, pages 3929–3938, 2017. 8

[47] Kai Zhang, Wangmeng Zuo, and Lei Zhang. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE TIP*, 27(9):4608–4622, 2018. 8

[48] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE TPAMI*, 44(10):6360–6376, 2021. 8

[49] Leheng Zhang, Yawei Li, Xingyu Zhou, Xiaorui Zhao, and Shuhang Gu. Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary. In *CVPR*, pages 2856–2865, 2024. 2

[50] Mingjin Zhang, Longyi Li, Wenxuan Shi, Jie Guo, Yunsong Li, and Xinbo Gao. VmambaSCI: Dynamic deep unfolding network with mamba for compressive spectral imaging. In *ACM MM*, pages 6549–6558, 2024. 2

[51] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *ECCV*, pages 649–667. Springer, 2022. 6

[52] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018. 2, 7

[53] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, pages 2472–2481, 2018. 1, 2

[54] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 8

[55] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE TPAMI*, 43(7):2480–2495, 2020. 8

[56] Zou Zhen, Yu Hu, and Zhao Feng. FreqMamba: Viewing Mamba from a frequency perspective for image deraining. *arXiv preprint arXiv:2404.09476*, 2024. 2

[57] Zhuoran Zheng and Chen Wu. U-shaped vision Mamba for single image dehazing. *arXiv preprint arXiv:2402.04139*, 2024. 2

[58] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. SRFormer: Permuted self-attention for single image super-resolution. *arXiv preprint arXiv:2303.09735*, 2023. 2, 6