MambaIR: A Simple Baseline for Image Restoration with State-Space Model

Hang Guo[1,⋆], Jinmin Li[1,⋆], Tao Dai[2],
Zhihao Ouyang[3], Xudong Ren[1], and Shu-Tao Xia[1,4]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University
[2] College of Computer Science and Software Engineering, Shenzhen University
[3] Bytedance Inc.
[4] Peng Cheng Laboratory
{cshguo, daitao.edu}@gmail.com, {ljm22,rxd21}@mails.tsinghua.edu.cn
zhihao.ouyang@bytedance.com, xiast@sz.tsinghua.edu.cn

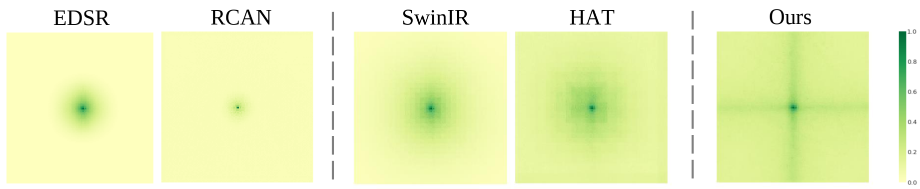
Abstract. Recent years have seen significant advancements in image restoration, largely attributed to the development of modern deep neural networks, such as CNNs and Transformers. However, existing restoration backbones often face the dilemma between global receptive fields and efficient computation, hindering their application in practice. Recently, the Selective Structured State Space Model, especially the improved version Mamba, has shown great potential for long-range dependency modeling with linear complexity, which offers a way to resolve the above dilemma. However, the standard Mamba still faces certain challenges in low-level vision such as local pixel forgetting and channel redundancy. In this work, we introduce a simple but effective baseline, named MambaIR, which introduces both local enhancement and channel attention to improve the vanilla Mamba. In this way, our MambaIR takes advantage of the local pixel similarity and reduces the channel redundancy. Extensive experiments demonstrate the superiority of our method, for example, MambaIR outperforms SwinIR by up to 0.45dB on image SR, using similar computational cost but with a global receptive field. Code is available at https://github.com/csguoh/MambaIR.

Keywords: Image Restoration · State Space Model · Mamba

1 Introduction

Image restoration, aiming to reconstruct a high-quality image from a given low-quality input, is a long-standing problem in computer vision and further has a wide range of sub-problems such as super-resolution, image denoising, *etc*. With the introduction of modern deep learning models such as CNNs [12,14,37,70,76] and Transformers [8,9,11,34,36], state-of-the-art performance has continued to be refreshed in the past few years.

⋆ Equal Contribution



(a) CNN-based Methods (b) Transformer-based Methods (c) Mamba-based Method

Fig. 1: The Effective Receptive Field (ERF) visualization [40] for EDSR [37], RCAN [76], SwinIR [36], HAT [9], and the proposed MambaIR. A larger ERF is indicated by a more extensively distributed dark area. Only the proposed MambaIR achieves a significant global effective receptive field.

To some extent, the increasing performance of deep restoration models largely stems from the increasing network receptive field. First, a large receptive field allows the network to capture information from a wider region, enabling it to refer to more pixels to facilitate the reconstruction of the anchor pixel. Second, with a larger receptive field, the restoration network can extract higher-level patterns and structures in the image, which can be crucial for some structure preservation tasks such as image denoising. Finally, Transformer-based restoration methods which possess larger receptive fields experimentally outperform CNN-based methods, and the recent art [9] also points out that activating more pixels usually leads to better restoration results.

Despite possessing many attractive properties, it appears that there exists an inherent choice dilemma between global receptive fields and efficient computation for current image restoration backbones. For CNN-based restoration networks [37, 76], although the effective receptive field is limited (as shown in Fig. 1(a)), it is appropriate for resource-constrained device deployments due to the favorable efficiency of convolution parallel operations. By contrast, Transformer-based image restoration methods usually set the number of tokens to the image resolution [8, 9, 36], therefore, despite the global receptive field, directly using the standard Transformer [57] will come at an unacceptable quadratic computational complexity. Moreover, employing some efficient attention techniques such as shifted window attention [39] for image restoration, usually comes at the expense of a globally effective receptive field (as shown in Fig. 1(b)), and does not intrinsically escape out of the trade-off between a global receptive field and efficient computation.

Recently, structured state-space sequence models (S4), especially the improved version Mamba, have emerged as an efficient and effective backbone for constructing deep networks [17,19,21,46,55]. This development hints at a potential resolution to balancing global receptive field and computational efficiency in image restoration. In detail, the discretized state space equations in Mamba can be formalized into a recursive form and can model very long-range dependencies when equipped with specially designed structured reparameterization [20]. This means that Mamba-based restoration networks can naturally activate more pixels, thus improving the reconstruction quality. Furthermore, the parallel scan algorithm [19] renders Mamba to process each token in a parallel fashion, facilitating efficient training on modern hardware such as GPU. The above promising properties motivate us to explore the potential of Mamba to achieve efficient long-range modeling for image restoration networks.

However, the standard Mamba [19], which is designed for 1D sequential data in NLP, is not a natural fit for image restoration scenarios. First, since Mamba processes flattened 1D image sequences in a recursive manner, it can result in spatially close pixels being found at very distant locations in the flattened sequences, resulting in the problem of local pixel forgetting. Second, due to the requirement to memorize the long sequence dependencies, the number of hidden states in the state space equations is typically large, which can lead to channel redundancy, thus hindering the learning of critical channel representations.

To address the above challenges, we introduce MambaIR, a simple but very effective benchmark model, to adapt Mamba for image restoration. MambaIR is formulated with three principal stages. Specifically, the **1)Shallow Feature Extraction** stage employs a simple convolution layer to extract the shallow feature. Then the **2)Deep Feature Extraction** stage performs with several stacked Residual State Space Blocks (RSSBs). As the core component of our MambaIR, the RSSB is designed with local convolution to mitigate local pixel forgetting when applying the vanilla Mamba to 2D images, and it is also equipped with channel attention to reduce channel redundancy caused by the excessive hidden state number. We also employ the learnable factor to control the skip connection within each RSSB. Finally, the **3)High-Quality Image Reconstruction** stage aggregates both shallow and deep features to produce a high-quality output image. Through possessing both a global effective receptive field as well as linear computational complexity, our MambaIR serves as a new alternative for image restoration backbones.

In short, our main contributions can be summarized as follows:

- We are the first to work to adapt state space models for low-level image restoration via extensive experiments to formulate MambaIR, which acts as a simple but effective alternative for CNN- and Transformer-based methods.
- We propose the Residue State Space Block (RSSB) which can boost the power of the standard Mamba with local enhancement and channel redundancy reduction.
- Extensive experiments on various tasks demonstrate our MambaIR outperforms other strong baselines to provide a powerful and promising backbone solution for image restoration.

2 Related Work

2.1 Image Restoration

Image restoration has been significantly advanced since the introduction of deep learning by several pioneering works, such as SRCNN [14] for image super-resolution, DnCNN [70] for image denoising, ARCNN [13] for JPEG compression artifact reduction, *etc*. Early attempts usually elaborate CNNs with techniques such as residual connection [6, 29, 67], dense connection [60, 78] and others [12,18,31,62,76] to improve model representation ability. Despite the success, CNN-based restoration methods typically face challenges in effectively modeling global dependencies. As Transformers have been proven to be a strong competitor for CNNs in multiple computer vision tasks [5,15,39], using Transformers for image restoration appears promising. Despite the global receptive field, it still faces specific challenges from the quadratic computational complexity of the self-attention mechanism [57]. To address this, IPT [8] divides one image into several small patches and processes each patch independently with self-attention. SwinIR [36] further introduces shifted window attention [39] to improve the performance. In addition, progress continues to be made in designing efficient attention mechanisms for restoration [9–11, 32, 63, 66]. Nonetheless, efficient attention design usually comes at the expense of global effective receptive fields, and the dilemma of the trade-off between efficient computation and global modeling is not essentially resolved.

2.2 State Space Models

State Space Models (SSMs) [21, 22, 55], stemming from classics control theory [28], are recently introduced to deep learning as a competitive backbone for state space transforming. The promising property of linearly scaling with sequence length in long-range dependency modeling has attracted great interest from searchers. For example, the Structured State-Space Sequence model (S4) [21] is a pioneer work for the deep state-space model in modeling the long-range dependency. Later, S5 layer [55] is proposed based on S4 and introduces MIMO SSM and efficient parallel scan. Moreover, H3 [17] achieves promising results that nearly fill the performance gap between SSMs and Transformers in natural language. [46] further improve S4 with gating units to obtain the Gated State Space layer to boost the capability. More recently, Mamba [19], a data-dependent SSM with selective mechanism and efficient hardware design, outperforms Transformers on natural language and enjoys linear scaling with input length. Moreover, there are also pioneering works that adopt Mamba to vision tasks such as image classification [38, 81], video understanding [58], biomedical image segmentation [42] and others [26, 50]. In this work, we explore the potential of Mamba to image restoration with restoration-specific designs to serve as a simple but effective baseline for future work.

3 Methodology

3.1 Preliminaries

The recent advancements of the class of structured state-space sequence models (S4) are largely inspired by the continuous linear time-invariant (LTI) systems, which maps a 1-dimensional function or sequence $x(t) \in \mathbb{R} \rightarrow y(t) \in \mathbb{R}$ through an implicit latent state $h(t) \in \mathbb{R}^N$. Formally, this system can be formulated as a linear ordinary differential equation (ODE):

$$
\begin{aligned}
h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\
y(t) &= \mathbf{C}h(t) + \mathbf{D}x(t),
\end{aligned} \tag{1}
$$

where N is the state size, $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times N}$, and $\mathbf{D} \in \mathbb{R}$.

After that, the discretization process is typically adopted to integrate Eq. (1) into practical deep learning algorithms. Specifically, denote Δ as the timescale parameter to transform the continuous parameters $\mathbf{A}$, $\mathbf{B}$ to discrete parameters $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$. The commonly used method for discretization is the zero-order hold (ZOH) rule, which is defined as follows:

$$
\begin{aligned}
\overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}), \\
\overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}.
\end{aligned} \tag{2}
$$

After the discretization, the discretized version of Eq. (1) with step size Δ can be rewritten in the following RNN form:

$$
\begin{aligned}
h_k &= \overline{\mathbf{A}}h_{k-1} + \overline{\mathbf{B}}x_k, \\
y_k &= \mathbf{C}h_k + \mathbf{D}x_k.
\end{aligned} \tag{3}
$$

Furthermore, the Eq. (3) can also be mathematically equivalently transformed into the following CNN form:

$$
\begin{aligned}
&\overline{\mathbf{K}} \triangleq (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{AB}}, \cdots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}), \\
&\mathbf{y} = \mathbf{x} \circledast \overline{\mathbf{K}},
\end{aligned} \tag{4}
$$

where L is the length of the input sequence, $\circledast$ denotes convolution operation, and $\overline{\mathbf{K}} \in \mathbb{R}^L$ is a structured convolution kernel.

The recent advanced state-space model, Mamba [19], have further improved $\overline{\mathbf{B}}$, $\mathbf{C}$ and Δ to be input-dependent, thus allowing for a dynamic feature representation. The intuition of Mamba for image restoration lies in its development on the advantages of S4 model. Specifically, Mamba shares the same recursive form of Eq. (3), which enables the model to memorize ultra-long sequences so that more pixels can be activated to aid restoration. At the same time, the parallel scan algorithm [19] allows Mamba to enjoy the same advantages of parallel processing as Eq. (4), thus facilitating efficient training.

3.2 Overall Architecture

As shown in Fig. 2, our MambaIR consists of three stages: shallow feature extraction, deep feature extraction, and high-quality reconstruction. Given a low-quality (LQ) input image $I_{LQ} \in \mathbb{R}^{H\times W\times 3}$, we first employ a 3×3 convolution layer from the shallow feature extraction to generate the shallow feature $F_S \in \mathbb{R}^{H\times W\times C}$, where H and W represent the height and width of the input image, and C is the number of channels. Subsequently, the shallow feature F_S undergoes the deep feature extraction stage to acquire the deep feature

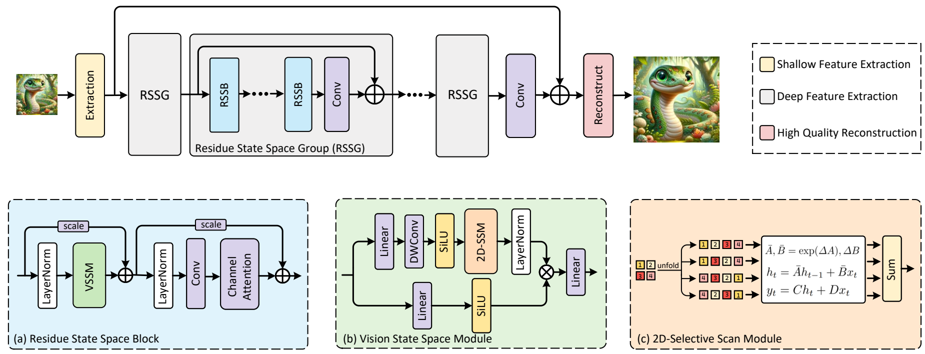


Fig. 2: The overall network architecture of our MambaIR, as well as the (a) Residual State-Space Block (RSSB), the (b) Vision State-Space Module (VSSM), and the (c) 2D Selective Scan Module (2D-SSM).

$F_D^l \in \mathbb{R}^{H\times W\times C}$ at the l-th layer, $l \in \{1, 2, \cdots L\}$. This stage is stacked by multiple Residual State-Space Groups (RSSGs), with each RSSG containing several Residue State-Space Blocks (RSSBs). Moreover, an additional convolution layer is introduced at the end of each group to refine features extracted from RSSB. Finally, we use the element-wise sum to obtain the input of the high-quality reconstruction stage $F_R = F_D^L + F_S$, which is used to reconstruct the high-quality (HQ) output image I_{HQ}.

3.3 Residual State-Space Block

The block design in previous Transformer-based restoration networks [9, 11, 36, 66] mainly follow the `Norm` $\rightarrow$ `Attention` $\rightarrow$ `Norm` $\rightarrow$ `MLP` flow. Although Attention and SSM can both model global dependencies, however, we find these two modules behave differently (see *supplementary material* for more details) and simply replacing Attention with SSM can only obtain sub-optimal results. Therefore, it is promising to tailor a brand-new block structure for Mamba-based restoration networks.

To this end, we propose the Residual State-Space Block (RSSB) to adapt the SSM block for restoration. As shown in Fig. 2(a), given the input deep feature $F_D^l \in \mathbb{R}^{H\times W\times C}$, we first use the LayerNorm (LN) followed by the Vision State-Space Module (VSSM) [38] to capture the spatial long-term dependency. Moreover, we also use learnable scale factor $s \in \mathbb{R}^C$ to control the information from skip connection:

$$Z^l = \mathrm{VSSM}(\mathrm{LN}(F_D^l)) + s \cdot F_D^l. \tag{5}$$

Furthermore, since SSMs process flattened feature maps as 1D token sequences, the number of neighborhood pixels in the sequence is greatly influenced by the flattening strategy. For example, when employing the four-direction unfolding strategy of [38], only four nearest neighbors are available to the anchor

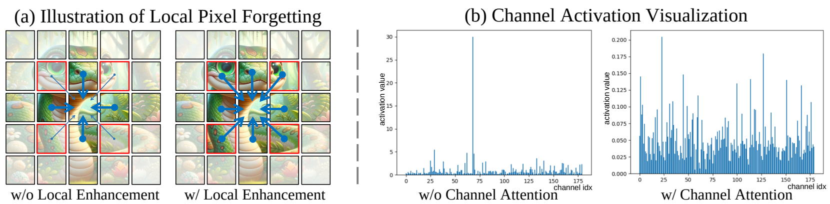


Fig. 3: (a) Without using local enhancement will cause spatially close pixels (area in the red box) get forgotten in the flattened 1D sequence due to the long distance. (b) We use RELU and global average pooling on the VSSM outputs from the last layer to get the channel activation values. Most channels are not activated (*i.e.*, channel redundancy) when channel attention is not used.

pixel (see Fig. 3(a)), *i.e.*, some spatially close pixels in 2D feature map are instead distant from each other in the 1D token sequence, and this over-distance can lead to local pixel forgetting. To this end, we introduce an additional local convolution after VSSM to help restore the neighborhood similarity. Specifically, we employ LayerNorm to first normalize the Z^l and then use convolution layers to compensate for local features. In order to maintain efficiency, the convolution layer adopts the bottleneck structure, *i.e.*, the channel is first compressed by a factor γ to obtain features with the shape $\mathbb{R}^{H\times W\times \frac{C}{\gamma}}$, then we perform channel expansion to recover the original shape.

In addition, SSMs typically introduce a larger number of hidden states to memorize very long-range dependencies, and we visualize the activation results for different channels in Fig. 3(b) and find notable channel redundancy. To facilitate the expressive power of different channels, we introduce the Channel Attention (CA) [23] to RSSB. In this way, SSMs can focus on learning diverse channel representations after which the critical channels are selected by subsequent channel attention, thus avoiding channel redundancy. At last, another tunable scale factor $s' \in \mathbb{R}^C$ is used in residual connection to acquire the final output F_D^{l+1} of the RSSB. The above process can be formulated as:

$$F_D^{l+1} = \mathrm{CA}(\mathrm{Conv}(\mathrm{LN}(Z^l))) + s' \cdot Z^l. \tag{6}$$

3.4 Vision State-Space Module

To maintain efficiency, the Transformer-based restoration networks usually divide input into small patches [8] or adopt shifted window attention [36], hindering the interaction at the whole-image level. Motivated by the success of Mamba in long-range modeling with linear complexity, we introduce the Vision State-Space Module to image restoration.

The Vision State-Space Module (VSSM) can capture long-range dependencies with the state space equation, and the architecture of VSSM is shown in Fig. 2(b). Following [38], the input feature $X \in \mathbb{R}^{H\times W\times C}$ will go through two parallel branches. In the first branch, the feature channel is expanded to λC by a linear layer, where λ is a pre-defined channel expansion factor, followed by a depth-wise convolution, SiLU [54] activation function, together with the 2D-SSM layer and LayerNorm. In the second branch, the features channel is also expanded to λC with a linear layer followed by the SiLU activation function. After that, features from the two branches are aggregated with the Hadamard product. Finally, the channel number is projected back to C to generate output X_{out} with the same shape as input:

$$\begin{aligned} X_1 &= \mathrm{LN}(\text{2D-SSM}(\mathrm{SiLU}(\mathrm{DWConv}(\mathrm{Linear}(X))))), \\ X_2 &= \mathrm{SiLU}(\mathrm{Linear}(X)), \\ X_{out} &= \mathrm{Linear}(X_1 \odot X_2), \end{aligned} \tag{7}$$

where DWConv represents depth-wise convolution, and $\odot$ denotes the Hadamard product.

3.5 2D Selective Scan Module

The standard Mamba [19] causally processes the input data, and thus can only capture information within the scanned part of the data. This property is well suited for NLP tasks that involve a sequential nature but poses significant challenges when transferring to non-causal data such as images. To better utilize the 2D spatial information, we follow [38] and introduce the 2D Selective Scan Module (2D-SSM). As shown in Fig. 2(c), the 2D image feature is flattened into a 1D sequence with scanning along four different directions: top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right. Then the long-range dependency of each sequence is captured according to the discrete state-space equation. Finally, all sequences are merged using summation followed by the reshape operation to recover the 2D structure.

3.6 Loss Function

To make a fair comparison with previous works [36, 66, 76], we optimize our MambaIR with L_1 loss for image SR, which can be formulated as:

$$\mathcal{L} = ||I_{HQ} - I_{LQ}||_1, \tag{8}$$

where $||\cdot||_1$ denotes the L_1 norm. For image denoising, we utilize the Charbonnier loss [7] with $\epsilon = 10^{-3}$:

$$\mathcal{L} = \sqrt{||I_{HQ} - I_{LQ}||^2 + \epsilon^2}. \tag{9}$$

4 Experiences

4.1 Experimental Settings

Dataset and Evaluation. Following the setup in previous works [36, 66], we conduct experiments on various image restoration tasks, including image super-resolution (*i.e.*, classic SR, lightweight SR, and real SR) and image denoising (*i.e.*, Gaussian color image denoising and real-world denoising). We employ DIV2K [56] and Flickr2K [37] to train classic SR models and use only

Table 1: Ablation experiments for different design choices of RSSB.

settings	#param	MACs	Set5	Set14	B100	Urban100	Manga109
(0)baseline	16.7M	439G	38.57	34.67	32.58	34.15	40.28
(1)+w/o Conv	11.9M	316G	38.48	34.54	32.56	34.04	40.20
(2)+w/o Conv+CA	11.8M	315G	38.55	34.64	32.57	34.06	40.14
(3)+replace with MLP	14.3M	379G	38.55	34.68	32.59	34.22	40.13

DIV2K to train lightweight SR models. Moreover, we use Set5 [4], Set14 [65], B100 [44], Urban100 [24], and Manga109 [45] to evaluate the effectiveness of different SR methods. For gaussian color image denoising, we utilize DIV2K [56], Flickr2K [37], BSD500 [3], and WED [43] as our training datasets. Our testing datasets for guassian color image denoising includes BSD68 [44], Kodak24 [16], McMaster [74], and Urban100 [24]. For real image denoising, we train our model with 320 high-resolution images from SIDD [1] datasets, and use the SIDD test set and DND [52] dataset for testing. Following [36, 76], we denote the model as MambaIR+ when self-ensemble strategy [37] is used in testing. The performance is evaluated using PSNR and SSIM on the Y channel from the YCbCr color space. In addition, we also include other restoration tasks such as JPEG compression artifact reduction, more experimental results can be seen in the *supplementary material*.

Training Details. In accordance with previous works [9,36,66], 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 $\times 2$ model to initialize those of $\times 3$ and $\times 4$ and halve the learning rate and total training iterations to reduce training time [37]. To ensure a fair comparison, we adjust the training batch size to 32 for image SR and 16 for image denoising. We employ the Adam [30] as the optimizer for training our MambaIR with $\beta_1 = 0.9, \beta_2 = 0.999$. The initial learning rate is set at 2×10^{-4} and is halved when the training iteration reaches specific milestones. Our MambaIR model is trained with 8 NVIDIA V100 GPUs.

4.2 Ablation Study

Effects of different designs of RSSB. As the core component, the RSSB can improve Mamba with restoration-specific priors. In this section, we ablate different components of the RSSB. The results, presented in Tab. 1, indicate that (1) directly using the standard Mamba to process flattened sequences can lead to local pixel forgetting, and the utilization of simple convolution layers can effectively enhance the local interaction. (2) Without using additional convolution and channel attention, *i.e.*, directly employing off-the-shelf Mamba for restoration, can only obtain sub-optimal results, which also supports our previous analysis. (3) Replacing Conv+ChanelAttention with MLP, whose resulted structure will be similar to Transformer, also leads to unfavorable results, indicating that although both SSMs and Attention have the global modeling ability,

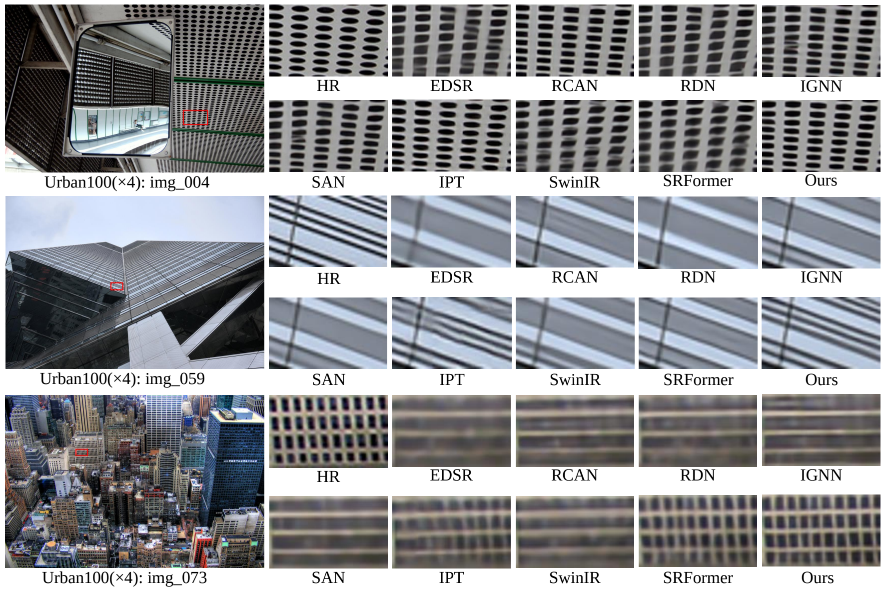


Fig. 4: Qualitative comparison of our MambaIR with CNN and Transformer based methods on **classic image SR** with scale $\times 4$.

Table 2: Ablation experiments for different scan modes in VSSM.

scan mode	#param	MACs	Set5	Set14	B100	Urban100	Manga109
one-direction	16.7M	439G	38.53	34.63	32.58	34.06	40.31
two-direction	16.7M	439G	38.56	34.60	32.56	33.96	40.14
baseline	16.7M	439G	38.57	34.67	32.58	34.15	40.28

the behavior of these two modules is different and thus accustomed block structure should be considered for further improvements.

Effects of Different Scan Modes in VSSM. To allow Mamba to process 2D images, the feature map needs to be flattened before being iterated by the state-space equation. Therefore, the unfolding strategy is particularly important. In this work, we follow [38] which uses scans in four different directions to generate scanned sequences. Here, we ablate different scan modes to study the effects, the results are shown in Tab. 2. Compared with one-direction(top-left to bottom-right) and two-direction(top-left to bottom-right, bottom-right to top-left), using four directions of scanning allows the anchor pixel to perceive a wider range of neighborhoods, thus achieving better results. Surprisingly, using more scanning directions brings an almost negligible increase in the number of parameters and computational complexity. This benefit may facilitate future exploration of more effective scanning mechanisms for the Mamba-based image restoration networks. We also include other ablation experiments, such as the layer number of RSSBs, please see *supplementary material* for more analysis.

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

Method	scale	Set5		Set14		BSDS100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [37]	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
RCAN [76]	×2	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN [12]	×2	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
HAN [51]	×2	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
IGNN [79]	×2	38.24	0.9613	34.07	0.9217	32.41	0.9025	33.23	0.9383	39.35	0.9786
CSNLN [48]	×2	38.28	0.9616	34.12	0.9223	32.40	0.9024	33.25	0.9386	39.37	0.9785
NLSA [47]	×2	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
ELAN [75]	×2	38.36	0.9620	34.20	0.9228	32.45	0.9030	33.44	0.9391	39.62	0.9793
IPT [8]	×2	38.37	-	34.43	-	32.48	-	33.76	-	-	-
SwinIR [36]	×2	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
SRFormer [80]	×2	38.51	0.9627	34.44	0.9253	32.57	0.9046	34.09	0.9449	40.07	0.9802
MambaIR	×2	38.57	0.9627	34.67	0.9261	32.58	0.9048	34.15	0.9446	40.28	0.9806
MambaIR+	×2	38.60	0.9628	34.69	0.9260	32.60	0.9048	34.17	0.9443	40.33	0.9806
EDSR [37]	×3	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
RCAN [77]	×3	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
SAN [12]	×3	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
HAN [51]	×3	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
IGNN [79]	×3	34.72	0.9298	30.66	0.8484	29.31	0.8105	29.03	0.8696	34.39	0.9496
CSNLN [48]	×3	34.74	0.9300	30.66	0.8482	29.33	0.8105	29.13	0.8712	34.45	0.9502
NLSA [47]	×3	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
ELAN [75]	×3	34.90	0.9313	30.80	0.8504	29.38	0.8124	29.32	0.8745	34.73	0.9517
IPT [8]	×3	34.81	-	30.85	-	29.38	-	29.49	-	-	-
SwinIR [36]	×3	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
SRformer [80]	×3	35.02	0.9323	30.94	0.8540	29.48	0.8156	30.04	0.8865	35.26	0.9543
MambaIR	×3	35.08	0.9323	30.99	0.8536	29.51	0.8157	29.93	0.8841	35.43	0.9546
MambaIR+	×3	35.13	0.9326	31.06	0.8541	29.53	0.8162	29.98	0.8838	35.55	0.9549
EDSR [37]	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
RCAN [76]	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN [12]	×4	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
HAN [51]	×4	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
IGNN [79]	×4	32.57	0.8998	28.85	0.7891	27.77	0.7434	26.84	0.8090	31.28	0.9182
CSNLN [48]	×4	32.68	0.9004	28.95	0.7888	27.80	0.7439	27.22	0.8168	31.43	0.9201
NLSA [47]	×4	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
ELAN [75]	×4	32.75	0.9022	28.96	0.7914	27.83	0.7459	27.13	0.8167	31.68	0.9226
IPT [8]	×4	32.64	-	29.01	-	27.82	-	27.26	-	-	-
SwinIR [36]	×4	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
SRFormer [80]	×4	32.93	0.9041	29.08	0.7953	27.94	0.7502	27.68	0.8311	32.21	0.9271
MambaIR	×4	33.03	0.9046	29.20	0.7961	27.98	0.7503	27.68	0.8287	32.32	0.9272
MambaIR+	×4	31.13	0.9054	29.25	0.7971	28.01	0.7510	27.80	0.8303	32.48	0.9281

4.3 Comparison on Image Super-Resolution

Classic Image Super-Resolution. Tab. 3 shows the quantitative results between MambaIR and state-of-the-art super-resolution methods. Thanks to the significant global receptive field, our proposed MambaIR achieves the best performance on almost all five benchmark datasets for all scale factors. For example, our Mamba-based baseline outperforms the Transformer-based benchmark model SwinIR by 0.41dB on Manga109 for ×2 scale, demonstrating the prospect

Table 4: Model size comparisons. We compare the complexity on the $\times 4$ classic SR model of different methods, the output size is set to $3 \times 640 \times 640$ for MACs.

Method	EDSR [37]	RCAN [76]	SRFBN [35]	HAN [51]	CSNLN [48]	SwinIR [36]	MambaIR (Ours)
Params (M)	43.1	15.6	3.6	16.1	7.2	11.9	16.7
Mult-Adds (G)	1,286	407	498	420	103,640	336	439
PSNR on Urban100 (dB)	26.64	26.82	26.60	26.85	27.22	27.45	27.68
PSNR on Manga109 (dB)	31.02	31.22	31.15	31.42	31.43	32.03	32.32

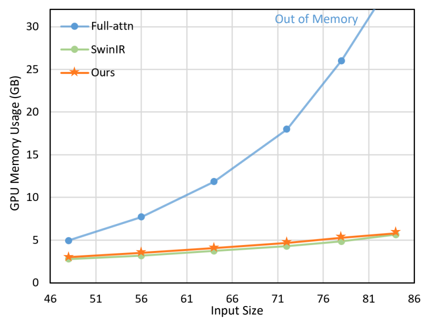


Fig. 5: Computational complexity comparison with different input scales. We set the standard attention [57] which has a global receptive field as baseline, and denote it as "Full-attn". We adjust the model to ensure the GPU usage is roughly similar at the beginning, and then scale the input resolution from 48×48 to 84×84.

of Mamba for image restoration. We also give visual comparisons in Fig. 4, and it can be seen that our method can facilitate the reconstruction of sharp edges and natural textures.

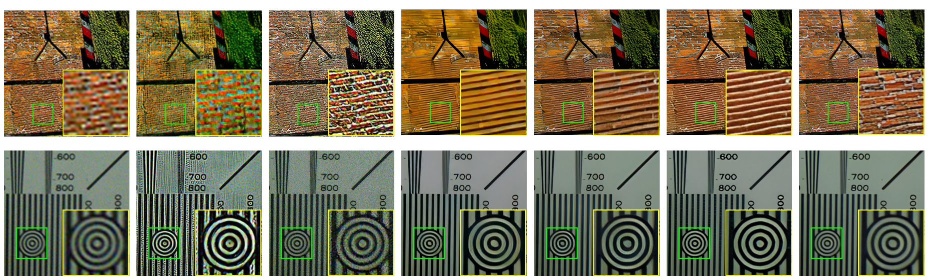
Model Complexity Comparison. Tab. 4 provides the comparison of different methods on parameters and MACs. Our MambaIR has comparable parameters and MACs compared with SwinIR, but can achieve a global receptive field and better PSNR performance. We further give comparison results of computational complexity with input sizes in Fig. 5. As one can see, our method is far more efficient than the full-attention baseline [57] and exhibits linear complexity with input resolution which is similar to the efficient attention techniques such as SwinIR. These observations above suggest that out MambaIR has similar scale properties as shifted window attention, while possessing a global receptive field similar to standard full attention.

Lightweight Image Super-Resolution. To demonstrate the scalability of our method, we train the Mamba-light model and compare it with state-of-the-art lightweight image SR methods. Following previous works [41, 80], we also report the number of parameters (#param) and MACs (upscaling a low-resolution image to 1280×720 resolution). Tab. 5 shows the results. It can be seen that our MambaIR-light outperforms SwinIR-light [36] by up to 0.17dB PSNR on the $\times 4$ scale Manga109 dataset with similar parameters and MACs. The performance results demonstrate the scalability and efficiency of our method.

Real-world Image Super-resolution. We also investigate the performance of the network for real-world image restoration. We follow the training protocol in

Table 5: Quantitative comparison on **lightweight image super-resolution** with state-of-the-art methods.

Method	scale	#param	MACs	Set5		Set14		BSDS100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CARN [2]	×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
IMDN [25]	×2	694K	158.8G	38.00	0.9605	33.63	0.9177	32.19	0.8996	32.17	0.9283	38.88	0.9774
LAPAR-A [33]	×2	548K	171.0G	38.01	0.9605	33.62	0.9183	32.19	0.8999	32.10	0.9283	38.67	0.9772
LatticeNet [41]	×2	756K	169.5G	38.15	0.9610	33.78	0.9193	32.25	0.9005	32.43	0.9302	-	-
SwinIR-light [36]	×2	878K	195.6G	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
SRFormer-light [80]	×2	853K	236G	38.23	0.9613	33.94	0.9209	32.36	0.9019	32.91	0.9353	39.28	0.9785
Ours	×2	859K	198.1G	38.16	0.9610	34.00	0.9212	32.34	0.9017	32.92	0.9356	39.31	0.9779
CARN [2]	×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
IMDN [25]	×3	703K	71.5G	34.36	0.9270	30.32	0.8417	29.09	0.8046	28.17	0.8519	33.61	0.9445
LAPAR-A [33]	×3	544K	114.0G	34.36	0.9267	30.34	0.8421	29.11	0.8054	28.15	0.8523	33.51	0.9441
LatticeNet [41]	×3	765K	76.3G	34.53	0.9281	30.39	0.8424	29.15	0.8059	28.33	0.8538	-	-
SwinIR-light [36]	×3	886K	87.2G	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624	33.98	0.9478
SRFormer-light [80]	×3	861K	105G	34.67	0.9296	30.57	0.8469	29.26	0.8099	28.81	0.8655	34.19	0.9489
Ours	×3	867K	88.7G	34.72	0.9296	30.63	0.8475	29.29	0.8099	29.00	0.8689	34.39	0.9495
CARN [2]	×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
IMDN [25]	×4	715K	40.9G	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838	30.45	0.9075
LAPAR-A [33]	×4	659K	94.0G	32.15	0.8944	28.61	0.7818	27.61	0.7366	26.14	0.7871	30.42	0.9074
LatticeNet [41]	×4	777K	43.6G	32.30	0.8962	28.68	0.7830	27.62	0.7367	26.25	0.7873	-	-
SwinIR-light [36]	×4	897K	49.6G	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
SRFormer-light [80]	×4	873K	62.8G	32.51	0.8988	28.82	0.7872	27.73	0.7422	26.67	0.8032	31.17	0.9165
Ours	×4	879K	50.6G	32.51	0.8993	28.85	0.7876	27.75	0.7423	26.75	0.8051	31.26	0.9175



LR RealSR ESRGAN BSRGAN Real-ESRGAN SwinIR Ours

Fig. 6: Qualitative comparison with RealSR [27], ESRGAN [60], BSRGAN [69], Real-ESRGAN [59], and SwinIR [36] on **real image super-resolution** with scale ×4.

[9] to train our MambaIR-real model. Since there are no ground-truth images for this task, only the visual comparison is given in Fig. 6. Compared with the other methods, our MambaIR exhibits a notable advancement in resolving fine details and texture preservation, demonstrating the robustness of our method.

4.4 Comparison on Image Denoising

Gaussian Color Image Denoising. The results of gaussian color image denoising are shown in Tab. 6. Following [67,70], the compared noise levels include 15, 25 and 50. As one can see, our model achieves better performance than all compared methods. In particular, it surpasses the SwinIR [36] by up to 0.25dB with σ=50 on the Urban100 dataset. We also give a visual comparison in Fig. 7. Thanks to the global receptive field, our MambaIR can achieve better structure preservation, leading to clearer edges and natural shapes.

Table 6: Quantitative comparison on **gaussian color image denoising** with state-of-the-art methods.

Method	BSD68			Kodak24			McMaster			Urban100		
	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50
IRCNN [72]	33.86	31.16	27.86	34.69	32.18	28.93	34.58	32.18	28.91	33.78	31.20	27.70
FFDNet [73]	33.87	31.21	27.96	34.63	32.13	28.98	34.66	32.35	29.18	33.83	31.40	28.05
DnCNN [71]	33.90	31.24	27.95	34.60	32.14	28.95	33.45	31.52	28.62	32.98	30.81	27.59
DRUNet [68]	34.30	31.69	28.51	35.31	32.89	29.86	35.40	33.14	30.08	34.81	32.60	29.61
SwinIR [36]	34.42	31.78	28.56	35.34	32.89	29.79	35.61	33.20	30.22	35.13	32.90	29.82
Restormer [63]	34.40	31.79	28.60	35.47	33.04	30.01	35.61	33.34	30.30	35.13	32.96	30.02
MambaIR	34.43	31.80	28.61	35.34	32.91	29.85	35.62	33.35	30.31	35.17	32.99	30.07
MambaIR+	34.44	31.82	28.64	35.35	32.92	29.87	35.63	33.36	30.32	35.17	32.99	30.06

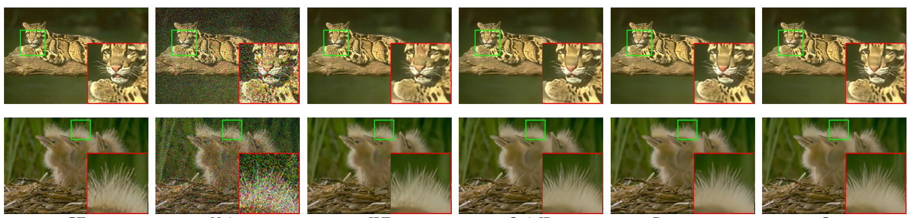


Fig. 7: Qualitative comparison of our MambaIR with other methods on **color image denoising** task with noise level level σ=50.

Table 7: Quantitative comparison on **real image denosing** with state-of-the-art methods.

Dataset	DeamNet [53]		MPRNet [64]		DAGL [49]		Uformer [61]		Restormer [63]		Ours	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SIDD	39.47	0.957	39.71	0.958	38.94	0.953	39.77	0.959	40.02	0.960	39.89	0.960
DND	39.63	0.953	39.80	0.954	39.77	0.956	39.96	0.956	40.03	0.956	40.04	0.956

Real Image Denoising. We further turn to the real image denoising task to evaluate the robustness of MambaIR when facing real-world degradation. Following [63], we adopt the progressive training strategy for fair comparison. The results, shown in Tab. 7, suggest that our method achieves comparable performance with existing state-of-the-art models Restormer [61] and outperforms other methods such as Uformer [61] by 0.12dB PSNR on SIDD dataset. The promising results indicate the ability of our method in real image denoising.

5 Conclusion

In this work, we explore for the first time the power of the recent advanced state space model, *i.e.*, Mamba, for image restoration, to help resolve the dilemma of trade-off between efficient computation and global effective receptive field. Specifically, we introduce the local enhancement to mitigate the neighborhood pixel forgetting problem from the flattening strategy and propose channel attention to reduce channel redundancy. Extensive experiments on multiple restoration tasks demonstrate our MambaIR serves as a simple but effective state-space model for image restoration.

References

1. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1692–1700 (2018) 9
2. Ahn, N., Kang, B., Sohn, K.A.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the European conference on computer vision (ECCV). pp. 252–268 (2018) 13
3. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE transactions on pattern analysis and machine intelligence **33**(5), 898–916 (2010) 9
4. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding (2012) 9
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) 4
6. Cavigelli, L., Hager, P., Benini, L.: Cas-cnn: A deep convolutional neural network for image compression artifact suppression. In: 2017 International Joint Conference on Neural Networks (IJCNN). pp. 752–759. IEEE (2017) 4
7. Charbonnier, P., Blanc-Feraud, L., Aubert, G., Barlaud, M.: Two deterministic half-quadratic regularization algorithms for computed imaging. In: Proceedings of 1st international conference on image processing. vol. 2, pp. 168–172. IEEE (1994) 8
8. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12299–12310 (2021) 1, 2, 4, 7, 11
9. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22367–22377 (2023) 1, 2, 4, 6, 9, 13
10. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X.: Recursive generalization transformer for image super-resolution. arXiv preprint arXiv:2303.06373 (2023) 4
11. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., Yu, F.: Dual aggregation transformer for image super-resolution. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12312–12321 (2023) 1, 4, 6
12. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11065–11074 (2019) 1, 4, 11
13. Dong, C., Deng, Y., Loy, C.C., Tang, X.: Compression artifacts reduction by a deep convolutional network. In: Proceedings of the IEEE international conference on computer vision. pp. 576–584 (2015) 3
14. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13. pp. 184–199. Springer (2014) 1, 3
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4

16. Franzen, R.: Kodak lossless true color image suite (2021), http://r0k.us/graphics/kodak/ 9
17. Fu, D.Y., Dao, T., Saab, K.K., Thomas, A.W., Rudra, A., Ré, C.: Hungry hungry hippos: Towards language modeling with state space models. arXiv preprint arXiv:2212.14052 (2022) 2, 4
18. Fu, X., Zha, Z.J., Wu, F., Ding, X., Paisley, J.: Jpeg artifacts reduction via deep convolutional sparse coding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2501–2510 (2019) 4
19. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023) 2, 3, 4, 5, 8
20. Gu, A., Dao, T., Ermon, S., Rudra, A., Ré, C.: Hippo: Recurrent memory with optimal polynomial projections. Advances in neural information processing systems **33**, 1474–1487 (2020) 2
21. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021) 2, 4
22. 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. Advances in neural information processing systems **34**, 572–585 (2021) 4
23. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018) 7
24. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5197–5206 (2015) 9
25. Hui, Z., Gao, X., Yang, Y., Wang, X.: Lightweight image super-resolution with information multi-distillation network. In: Proceedings of the 27th acm international conference on multimedia. pp. 2024–2032 (2019) 13
26. Islam, M.M., Hasan, M., Athrey, K.S., Braskich, T., Bertasius, G.: Efficient movie scene detection using state-space transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18749–18758 (2023) 4
27. Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., Huang, F.: Real-world super-resolution via kernel estimation and noise injection. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 466–467 (2020) 13
28. Kalman, R.E.: A new approach to linear filtering and prediction problems (1960) 4
29. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016) 4
30. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9
31. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 624–632 (2017) 4
32. Li, W., Lu, X., Qian, S., Lu, J., Zhang, X., Jia, J.: On efficient transformer-based image pre-training for low-level vision. arXiv preprint arXiv:2112.10175 (2021) 4
33. Li, W., Zhou, K., Qi, L., Jiang, N., Lu, J., Jia, J.: Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. Advances in Neural Information Processing Systems **33**, 20343–20355 (2020) 13

34. Li, Y., Fan, Y., Xiang, X., Demandolx, D., Ranjan, R., Timofte, R., Van Gool, L.: Efficient and explicit modelling of image hierarchies for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18278–18289 (2023) 1
35. Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W.: Feedback network for image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3867–3876 (2019) 12
36. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021) 1, 2, 4, 6, 7, 8, 9, 11, 12, 13, 14
37. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017) 1, 2, 8, 9, 11, 12
38. 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) 4, 6, 7, 8, 10
39. 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: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 2, 4
40. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. Advances in neural information processing systems **29** (2016) 2
41. Luo, X., Xie, Y., Zhang, Y., Qu, Y., Li, C., Fu, Y.: Latticenet: Towards lightweight image super-resolution with lattice block. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 272–289. Springer (2020) 12, 13
42. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024) 4
43. Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., Zhang, L.: Waterloo exploration database: New challenges for image quality assessment models. IEEE Transactions on Image Processing **26**(2), 1004–1016 (2016) 9
44. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. vol. 2, pp. 416–423. IEEE (2001) 9
45. 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**, 21811–21838 (2017) 9
46. Mehta, H., Gupta, A., Cutkosky, A., Neyshabur, B.: Long range language modeling via gated state spaces. arXiv preprint arXiv:2206.13947 (2022) 2, 4
47. Mei, Y., Fan, Y., Zhou, Y.: Image super-resolution with non-local sparse attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3517–3526 (2021) 11
48. Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T.S., Shi, H.: Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5690–5699 (2020) 11, 12
49. Mou, C., Zhang, J., Wu, Z.: Dynamic attentive graph learning for image restoration. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4328–4337 (2021) 14

50. Nguyen, E., Goel, K., Gu, A., Downs, G., Shah, P., Dao, T., Baccus, S., Ré, C.: S4nd: Modeling images and videos as multidimensional signals with state spaces. Advances in neural information processing systems **35**, 2846–2861 (2022) 4
51. Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., Zhang, K., Cao, X., Shen, H.: Single image super-resolution via a holistic attention network. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. pp. 191–207. Springer (2020) 11, 12
52. Plotz, T., Roth, S.: Benchmarking denoising algorithms with real photographs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1586–1595 (2017) 9
53. Ren, C., He, X., Wang, C., Zhao, Z.: Adaptive consistency prior based deep network for image denoising. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8596–8606 (2021) 14
54. Shazeer, N.: Glu variants improve transformer. arXiv preprint arXiv:2002.05202 (2020) 8
55. Smith, J.T., Warrington, A., Linderman, S.W.: Simplified state space layers for sequence modeling. arXiv preprint arXiv:2208.04933 (2022) 2, 4
56. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 114–125 (2017) 8, 9
57. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 2, 4, 12
58. Wang, J., Zhu, W., Wang, P., Yu, X., Liu, L., Omar, M., Hamid, R.: Selective structured state-spaces for long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6387–6397 (2023) 4
59. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1905–1914 (2021) 13
60. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: The European Conference on Computer Vision Workshops (ECCVW) (September 2018) 4, 13
61. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17683–17693 (June 2022) 14
62. Wei, Y., Gu, S., Li, Y., Timofte, R., Jin, L., Song, H.: Unsupervised real-world image super resolution via domain-distance aware training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13385–13394 (2021) 4
63. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022) 4, 14
64. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14821–14831 (2021) 14

65. 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. pp. 711–730. Springer (2012) 9
66. Zhang, J., Zhang, Y., Gu, J., Zhang, Y., Kong, L., Yuan, X.: Accurate image restoration with attention retractable transformer. In: ICLR (2023) 4, 6, 8, 9
67. Zhang, K., Li, Y., Zuo, W., Zhang, L., Van Gool, L., Timofte, R.: Plug-and-play image restoration with deep denoiser prior. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6360–6376 (2021) 4, 13
68. Zhang, K., Li, Y., Zuo, W., Zhang, L., Van Gool, L., Timofte, R.: Plug-and-play image restoration with deep denoiser prior. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6360–6376 (2021) 14
69. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4791–4800 (2021) 13
70. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE transactions on image processing **26**(7), 3142–3155 (2017) 1, 3, 13
71. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE transactions on image processing **26**(7), 3142–3155 (2017) 14
72. Zhang, K., Zuo, W., Gu, S., Zhang, L.: Learning deep cnn denoiser prior for image restoration. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3929–3938 (2017) 14
73. Zhang, K., Zuo, W., Zhang, L.: Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. IEEE Transactions on Image Processing **27**(9), 4608–4622 (2018) 14
74. Zhang, L., Wu, X., Buades, A., Li, X.: Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. Journal of Electronic imaging **20**(2), 023016–023016 (2011) 9
75. Zhang, X., Zeng, H., Guo, S., Zhang, L.: Efficient long-range attention network for image super-resolution. In: European Conference on Computer Vision. pp. 649–667. Springer (2022) 11
76. 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) 1, 2, 4, 8, 9, 11, 12
77. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018) 11
78. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2472–2481 (2018) 4
79. Zhou, S., Zhang, J., Zuo, W., Loy, C.C.: Cross-scale internal graph neural network for image super-resolution. In: Advances in Neural Information Processing Systems (2020) 11
80. Zhou, Y., Li, Z., Guo, C.L., Bai, S., Cheng, M.M., Hou, Q.: Srformer: Permuted self-attention for single image super-resolution. arXiv preprint arXiv:2303.09735 (2023) 11, 12, 13
81. 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) 4