This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Dynamic State-Control Modeling for Generalized Remote Sensing Image Super-Resolution

Chenyu Li
Southeast University
Nanjing, China
chenyuli.erh@gmail.com

Zhaojie Pan
Southeast University
Nanjing, China
panzhj99@163.com

Danfeng Hong*
AIR, Chinese Academy of Sciences
Beijing, China
hongdanfeng1989@gmail.com

Abstract

Image super-resolution (SR) is essential for overcoming sensor hardware limitations, yet most existing techniques primarily focus on resolution enhancement without effectively managing error propagation during intermediate stages. The Mamba model has shown promise by representing the reconstruction process as a sequence of states, enabling intermediate-level control. However, its fixed linear mapper, with a limited receptive field, constrains its ability to capture fine-grained details. To address these limitations, we introduce MambaX, a novel spectrum-state sequence model that maps spectral bands into a latent state space while dynamically learning control parameters for SR tasks. MambaX advances beyond existing approaches in three key aspects: 1) It employs nonlinear dynamic learning to more effectively approximate control matrix parameters in the latent space. 2)It introduces a cross-state control framework, enabling a unified approach to both single-image super-resolution (SISR) and multimodal fusion for enhanced flexibility. 3) It leverages transitional learning to mitigate degradation effects in remote sensing images. Extensive experiments demonstrate that MambaX outperforms state-of-the-art models in both SISR and multimodal fusion-based SR tasks, highlighting its capability for advanced spectral modeling across diverse dimensions and modalities. To ensure reproducibility and foster further research, we will publicly release the datasets and code used in this study.

1. Introduction

High-resolution optical imaging is essential for applications in computer vision, remote sensing, target identification, fine-structure analysis, and scientific research. However, it faces inherent limitations due to diffraction, sensor sampling rates, and environmental noise [12]. These constraints and hardware limitations present significant technical challenges and costs in improving image resolution. Super-resolution (SR) techniques can be classified into two types based on the input modalities: single-image super-resolution (SISR) and multimodal super-resolution (MFSR).

Mathematically, SISR is an inherently challenging task to estimate the high-resolution image (HR) $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ from an observed low-resolution image (LR) $\mathbf{Y} \in \mathbb{R}^{h\times w\times C}$ that suffers severe information degradation, which can be expressed as:

$$\mathbf{Y} = \mathbf{KX} + \mathcal{N}, \tag{1}$$

where $\mathbf{K}$ represents the integrated degradation operator, which includes degradations such as blurring and downsampling [26], $\mathcal{N}$ represents generalized noise. In contrast to SISR, MFSR emphasizes integrating data from multiple sensors or imaging modalities as complementary references to enhance image resolution and quality. This can be expressed as:

$$\begin{cases} \mathbf{Y}_{\mathrm{H}} = \mathbf{XB} + \mathcal{N}_H, \\ \mathbf{Y}_{\mathrm{M}} = \mathbf{CX} + \mathcal{N}_M, \end{cases} \tag{2}$$

where $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ is the original image that, subjected to spatial degradation $\mathbf{B}$ and affected by noise $\mathcal{N}_H$, yields LR image $\mathbf{Y}_H$. Meanwhile, when it experiences spectral degradation $\mathbf{C}$ and is also impacted by noise $\mathcal{N}_M$, it produces a low spectral resolution image $\mathbf{Y}_M$.

Early SR methods primarily relied on the intrinsic mathematical and physical properties of images to develop optimization frameworks [8]. Notable techniques include sparse-based methods [13, 18] and dictionary-learning approaches [15]. In contrast to these traditional methods, deep learning (DL) techniques offer superior capabilities for extracting complex nonlinear features [4, 34]. To improve the performance of SISR, techniques often focus on preserving local details, modeling global context [6, 27], and leveraging deep hierarchical representations [1].

For MFSR, approaches commonly use pixel-level fusion via input stacking [17] and feature-level fusion using

*Corresponding author.

dual-branch feature interactions [11] to effectively integrate auxiliary information. The transformer architecture [5, 23], known for its attention mechanism, has become a dominant approach in recent research. Transformers have been utilized to model long-term temporal dependencies across spectral channels and spatial relationships [20, 21].

Recently, selective state space models (SSM), most typically Mamba, have demonstrated promising results in modeling sequences with linear complexity. The state passing process of mamba can be expressed as:

$$h_i = \bar{\mathbf{A}} h_{i-1} + \bar{\mathbf{B}} x_i, \tag{3}$$

$$y_i = \mathbf{C} h_i + \mathbf{D} x_i, \tag{4}$$

where, h_{i-1} denotes prior state, and h_i is the present state $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ represents $\mathbf{A}$ and $\mathbf{B}$ after $\triangle$ dispersion, respectively. $\mathbf{A}$ is constructed by HiPPO [9], which can ensure the mathematical approximation of all previous histories. x_i represents the input state vector, influenced by the input control matrix $\bar{\mathbf{B}}$. Similarly, the control variable $\mathbf{C}$ describes the mapping of the state h_i to the output y_i. Notice that $\mathbf{B}$, $\mathbf{C}$, and $\triangle$ all depend on x. The control variable $\mathbf{D}$ represents the direct impact of the x_i on y_i. While Mamba offers notable advantages, its linear generation of $\mathbf{B}$, $\mathbf{C}$, and $\triangle$ limits its ability to capture complex nonlinear relationships. Linear control matrices have a restricted receptive field, limiting feature representation. Furthermore, heterogeneity exists not only in spatial modalities but also in cross-domain interactions between images and sequences. Such complex cross-domain mappings exceed the approximation capacity of linear models, often leading to error accumulation, model non-convergence, or overfitting during training.

To address these challenges, we introduce MambaX, an innovative super-resolution framework that utilizes dynamic nonlinear mappings to compute multistage differential coefficients within state models. To improve the transition from LR to HR images, we implement a progressive transition learning strategy. This strategy resizes LR inputs to align with the spatial dimensions of the final output before passing them through a Mamba-based encoder-decoder. To leverage the state-space model's ability to capture global dependencies, MambaX incorporates an enhanced Mamba mechanism that dynamically generates state control variables through structure and channel perception operators. For MFSR, a fusion strategy based on Mamba principles is applied, optimizing the integration of auxiliary information. The main contributions of this paper are as follows:

- A novel dynamic control learning approach for state-space models is introduced, enabling the adaptive learning and adjustment of state control parameters. This method ensures the accurate approximation of images across varying resolutions and modalities with high fidelity.
- A pioneering multimodal fusion concept is proposed, leveraging the robust and generalizable capabilities of the dynamic state-control model. By mapping the super-resolution process onto a latent state space, the method integrates cross-modal attribute fusion through control matrices applied to each state vector within the space.
- A progressive transition learning strategy is presented for degradation processes, offering an effective mechanism to model gradual image degradations with enhanced interpretability and precision.

2. Method

The proposed MambaX aims to model complex global dependencies between LR images and HR images, as illustrated in Figure 1. In this section, we first introduce the overall architecture of MambaX for the MFSR, followed by the dynamic control learning-based feature extraction block (nSPC), the cross-domain transition module (CDTM), and the state-level feature fusion method.

2.1. MambaX for MFSR

The primary differences between the structures of SISR and MFSR are as follows: 1) MFSR includes an additional supplementary feature extraction module (SFEM). 2) Within the basic encoder layer, the MFSR structure incorporates a shallow feature fusion block. 3) Most importantly, MambaX features a state-level fusion module designed to achieve comprehensive information fusion.

Let the LR image $L \in \mathbb{R}^{h \times w \times C}$ and auxiliary image $S \in \mathbb{R}^{H \times W \times c}$ be inputs, where (h, w, C) and (H, W, c) denote the height, width, and channel dimensions of the LR and auxiliary data respectively. The MambaX framework reconstructs the HR image $X \in \mathbb{R}^{H \times W \times C}$. A multi-layer CDTM first progressively upsamples L to generate a coarse SR result $X_c \in \mathbb{R}^{H \times W \times c_m}$ with spatial resolution aligned to X. This process can be effectively illustrated as

$$X_c = \mathrm{CDTM}(\cdots \mathrm{CDTM}\left(\mathrm{LeReLU}(\mathrm{Conv}_{3\times3}(L))\right)), \tag{5}$$

The CDTM performs iterative $2\times$ upsampling, requiring $\lfloor \log_2(H/h) \rfloor$ layers to align spatial resolutions. The coarse SR result $X_c \in \mathbb{R}^{H \times W \times c_m}$ is then refined by encoder-decoder architecture (MambaED), producing the final reconstruction $X_f \in \mathbb{R}^{H \times W \times c_m}$. Finally, MambaX fuses X_f and X_c through residual connection with convolutional feature transformation, formulated as:

$$X = \mathcal{F}_{\mathrm{out}}\left(X_f + X_c\right), \tag{6}$$

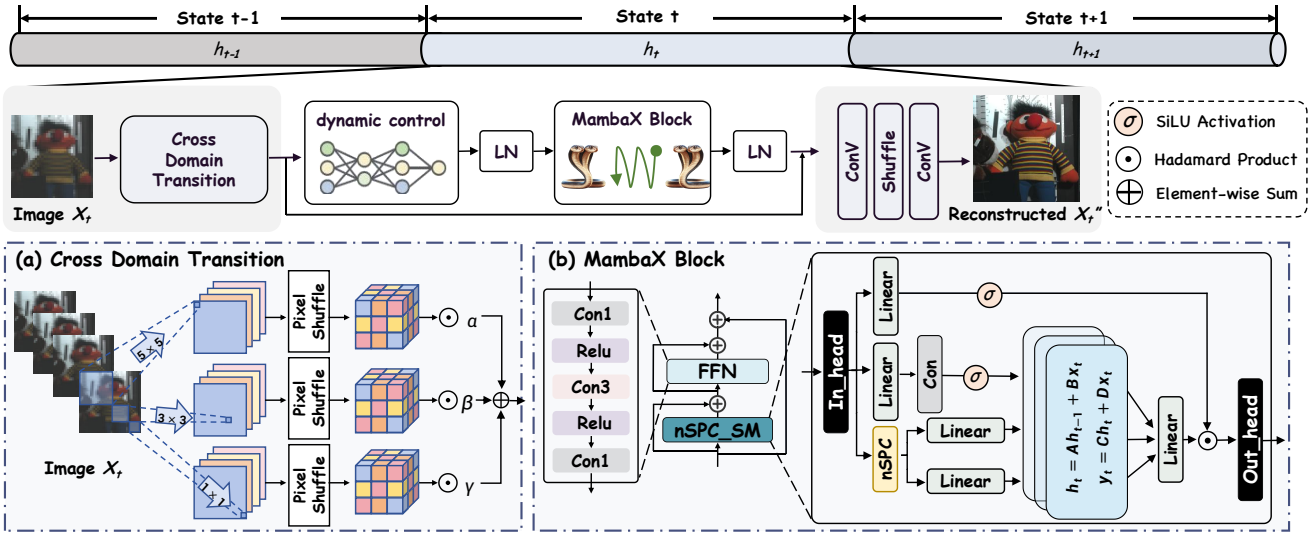


Figure 1. Overview of the proposed MambaX for SISR. The model employs an encoder-decoder architecture with two key components. First, a cross-domain transition module bridges the spatial scales between LR and HR images through image-to-state transformations. Then, a state transition network with dynamic control learning-based Mamba blocks establishes non-linear global relationships across resolution domains.

The MambaED architecture comprises three encoder modules (MambaE) and two decoder modules (MambaD). Each MambaE unit contains three principal components: 1) nSPC, 2) feedforward network (FFN), and 3) a downsampling operator. Conversely, the MambaD layer incorporates three sequential processing stages: upconvolution operation, nSPC transformation, and FFN refinement. Notably, the SFEM generates hierarchical auxiliary features $S_{i}^{3}{}_{i=1}$ through multi-resolution before feature propagation in the MambaED framework. This design guarantees scale-aware feature alignment with the encoder-decoder hierarchy in MambaED.

2.2. Cross-domain transition module

Conventional SR methods' reliance on single-scale operators limits their ability to capture complex multi-scale degradation. Our method adaptively fuses three local upsampling kernels to simulate degradation blur kernels. Specifically, we first obtain different feeling fields X_{i,b_i} using various mapping operators:

$$X_{i,b_i} = \mathcal{F}_{\mathrm{PS}}(\mathrm{Con}(X_{i-1,c})), \tag{7}$$

where $\mathcal{F}_{\mathrm{PS}}(\cdot)$ is the pixelshuffle function, and $X_{i-1,c}$ is the output of the $i-1$-th layer. Subsequently, to facilitate a more flexible transform from image space into the state space domain, we introduce learning weighting factors that adaptively fuse the three bases. This mechanism can be formally expressed as:

$$X_{i,c} = \sum_{i=1}^{3} (\mathrm{softmax}([\alpha, \beta, \gamma])_i \odot X_{i,b_i}), \tag{8}$$

where α, β and γ are adaptive weights. Moreover, the adaptive weights for each layer are optimized independently during training, greatly enhancing the model's ability to capture complex degradation patterns and blur kernels.

2.3. nSPC in Mamba

In SSM, control parameters are crucial for propagating state information and integrating auxiliary data into reconstructed images, as shown in Sec. 2.4. The original Mamba architecture's linear parameterization cannot capture global dependencies in complex SR tasks. To address this, we propose a dynamic state-space formulation. The mathematical implementations are:

$$\mathrm{DSC}_B(S_i) = \mathrm{Linear}_B((S_i)) \tag{9}$$

$$\begin{aligned} C_j^{Dyna} &:= \mathrm{DSC}_C(X_{f,s}), \\ &= \sigma\left[K_2(K_1 X_{f,s}^i + b_1) + b_2\right], j \in \{1, \cdots, T\}. \end{aligned} \tag{10}$$

$$\begin{aligned} \Delta_j^{Dyna} &:= \mathrm{DSC}_\Delta(X_{f,s}), \\ &= softplus\left\{\sigma[K_2(K_1 X_{f,s}^i + b_1) + b_2]\right\}, \\ &j \in \{1, \cdots, T\}. \end{aligned} \tag{11}$$

K and b are the weights and biases learned by the convolutional network with appropriate dimensions. Specifically, the first convolutional layer (K_1, b_1) operates only in the spatial dimension, while the second layer (K_2, b_2) operates on the channel dimension. This approach enables fine-grained feature extraction and fusion. Assuming the input features to the nSPC are denoted as $X_i \in R^{H \times W \times c_m}$. This

architecture can be formally expressed through the following mathematical formulation:

$$\mathrm{DSC}_{\Delta}(X_i) = W^{\Delta}_{c_r \times c_d}(W^{\Delta}_{c_d \times c_r}(C_l(S_l(W_{c_m \times c_d} X_i)))) \tag{12}$$

$$s.t. \begin{cases} S_l(\cdot) = \mathrm{BN}(\mathrm{Con}_{7\times7,d}(\mathrm{Con}_{5\times5,d}(\cdot))) \\ C_l(\cdot) = \mathcal{F}_{c,3}\Big(\mathcal{F}_{c,2}\big(\mathcal{F}_{c,1}(X_{i,S}, S_{i,l}), C^i_{spa}\big), X_{i,S}\Big) \end{cases} \tag{13}$$

where $W_{c_m \times c_d} X_i$ with channel c_d is used to generate the control variable. In addition, $W^{\Delta}_{c_d \times c_r}$ and $W^{\Delta}_{c_r \times c_d}$ are sequentially applied to compress and expand the feature dimensions, thereby preserving critical information, where $c_d \gg c_r$.

2.4. Multimodal Fusion in Intermediate State-Level

To guide LR image reconstruction, we leverage the auxiliary image to generate the key control variable B, while maintaining matrix A's original initialization for state stability. Accounting for the distinct roles of B, C, and Δ in state/output equations, we devise separate learning strategies for each parameter. This framework is formally expressed as:

$$\bar{B}^i = \mathrm{D}_{B^i}(A^i, \mathrm{Linear}_B(S_i), \triangle), \tag{14}$$

$$C^i = \mathrm{DSC}_{C^i}(X^i_{f,s}), \tag{15}$$

$$\triangle^i = \mathrm{DSC}_{\triangle^i}(X^i_{f,s}), \tag{16}$$

The simple generation method for B stabilizes $\bar{B}$, thereby deliberately constraining the fused state's modeling space and ensuring overall stability.

Before formally using the state space model to fuse $X^i_{f,s}$ and S_i, it is necessary to Conert the control variables and $X^i_{f,s}$ into sequence form $\bar{B}^i = \left[\bar{B}^i_1, \bar{B}^i_2, \cdots \bar{B}^i_k, \cdots\right]$, $C^i = \left[C^i_1, C^i_2, \cdots C^i_k, \cdots\right]$, $\bar{A}^i = \left[\bar{A}^i_1, \bar{A}^i_2, \cdots \bar{A}^i_k, \cdots\right]$, and $X^i_{f,s} = \left[x^i_1, x^i_2, \cdots x^i_k, \cdots\right]$. Subsequently, the fusion process at each position in the sequence can be represented as:

$$\begin{aligned} \begin{bmatrix} h^s_k \\ y^i_k \end{bmatrix} &= \begin{bmatrix} \bar{A}^i_k & \bar{B}^i_k \\ \bar{C}^i_k \bar{A}^i_k & \bar{C}^i_k \bar{B}^i_k + \bar{D}^i_k \end{bmatrix} \begin{bmatrix} h^s_{k-1} \\ x^i_k \end{bmatrix}, \\ &= \begin{bmatrix} \bar{A}^i_k & \mathcal{F}_{\bar{B}^i_k}(S_i, X^i_{f,s}) \\ \mathcal{F}_{\bar{C}^i_k}(X^i_{f,s})\bar{A}^i_k & \mathcal{F}_{\bar{C}^i_k \bar{B}^i_k}(S_i, X^i_{f,s}) + \bar{D}^i_k \end{bmatrix} \begin{bmatrix} h^s_{k-1} \\ x^i_k \end{bmatrix}, \end{aligned} \tag{17}$$

where h^s_k represents the intermediate state, which also indicates the initial fusion result of the supplementary information S_i and $X^i_{f,s}$. The y^i_k denotes the refined fusion result, wherein the original image $X^i_{f,s}$ is further integrated with the initial fusion outcome to calibrate the final fusion output. Subsequently, the outputs of the state space equations are merged and rearranged to obtain Y_i. Finally, to prevent the model from forgetting spectral information, the fused result is further guided by $X^i_{f,s}$. It can be represented as

$$X_{i,S} = \mathrm{Linear}(\mathrm{LN}(Y_i) \odot \mathrm{SiLU}(\mathrm{Linear}(X^i_f))), \tag{18}$$

where $X_{i,S}$ represents the output of the dynamic state control state space model, and $\mathrm{SiLU}(\cdot)$ represents the sigmoid linear unit.

3. Experiments

3.1. Datasets and Evaluation

Single image super-resolution datasets: The CAVE dataset [1], collected by Columbia University's Computer Vision Lab, contains 31 indoor hyperspectral images (512 × 512 resolution) with 31 spectral bands (400–700 nm, 10 nm intervals). For experiments, 10 scenes are randomly selected for testing, and the rest for training. SR is performed at scale factors of 2, 4, and 8, with input resolutions of 64 × 64, 32 × 32, and 16 × 16, and a fixed output resolution of 128 × 128. The Pavia dataset [7], captured with the ROSIS sensor, includes 102 spectral bands (430–860 nm) after removing 13 noisy ones. A low-quality region is excluded, resulting in a reference image of 896 × 672 × 102, which is split into 12 patches (224 × 224 × 102). Nine patches are used for training, and three for testing. For SR, input resolutions for scale factors of 2, 4, and 8 are 56 × 56, 28 × 28, and 28 × 28, while output resolutions are 112 × 112, 112 × 112, and 224 × 224, respectively.

Image fusion-based super-resolution datasets: We perform MFSR experiments on the WorldView-III (WV3) dataset containing urban roads and cityscapes. With no ground truth available, the dataset was generated following the Wald protocol [31], comprising 20 reduced-resolution test samples. Each training sample contains PAN/LRMS/GT triplets sized 64×64×1, 16×16×8, and 64×64×8, respectively. Test samples feature larger PAN/LRMS/GT triplets of 256×256×1, 64×64×8, and 256×256×8. All data originates from the PanCollection [3].

Evaluation Metrics: To comprehensively assess the model's performance from both spatial and spectral perspectives, we employ four widely used metrics across all experiments: peak signal-to-noise ratio (PSNR) [14], structural similarity index (SSIM) [33], spectral angle mapper (SAM) [37], and the relative global dimensional synthesis error (ERGAS) [30].

3.2. Compared Methods and Implementation

Single-modal Comparison: We compare our method with state-of-the-art image SR approaches: 3D convolution-based bidirectional 3D quasi-recurrent neural network (3DQRNN) [6], Mixed convolution-based MCNet [22],

[1] http://www.cs.columbia.edu/CAVE/databases/

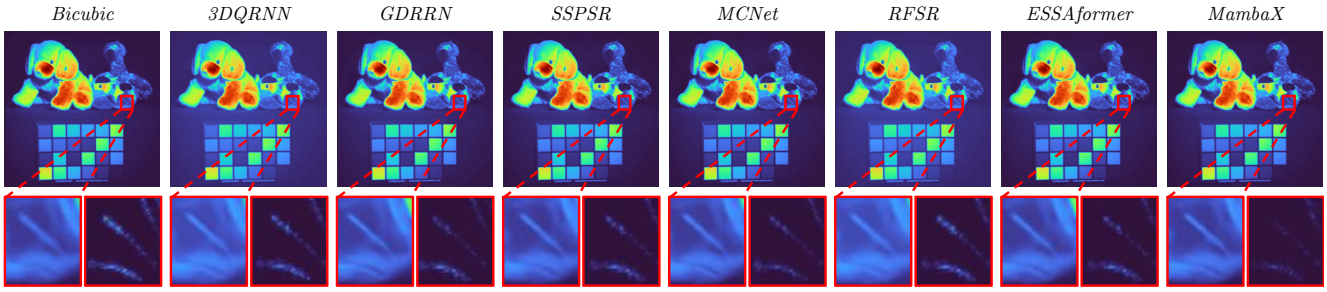


Figure 2. Visual quality comparison on CAVE datasets. The first row shows the pseudo-color image (6, 14, 25) of the 2x super-resolution result. The second row shows the local enlargement of the mean squared error (MSE) results.

Table 1. Quantitative performance on testing samples of the CAVE dataset at different scale factors. **Bold** represents the best result and underline represents the second best.

Comparison	2 ×				4 ×				8 ×			
Methods	PSNR	SSIM	SAM	ERGAS	PSNR	SSIM	SAM	ERGAS	PSNR	SSIM	SAM	ERGAS
Bicubic	39.81	0.9658	-	2.59	34.55	0.9151	-	4.50	30.48	0.8469	2.44	6.92
3DQRNN	40.56	0.9694	1.37	2.36	35.72	0.9246	2.39	3.97	31.61	0.8600	3.07	6.13
GDRRN	41.84	0.9726	1.49	2.12	36.18	0.9295	2.16	3.93	31.69	0.8623	2.99	6.17
SSPSR	42.57	0.9742	1.25	1.98	37.25	0.9378	2.01	3.45	32.58	0.8775	2.68	5.58
MCNet	42.55	0.9726	4.42	2.46	37.27	0.9373	1.81	3.47	32.64	0.8768	2.47	**5.54**
RFSR	42.24	0.9730	1.18	2.06	36.93	0.9343	1.62	3.59	32.63	0.8768	**2.20**	5.55
ESSAformer	42.40	0.9742	1.31	2.00	37.21	0.9373	1.81	3.47	32.48	0.8745	3.80	5.57
MambaX	**43.35**	**0.9755**	**1.10**	**1.87**	**37.39**	**0.9391**	**1.56**	**3.46**	**32.73**	**0.8783**	2.41	5.57

Residual networks-based grouped deep recursive residual network (GDRRN) [24], Spatial-spectral prior network (SSPSR) [16], Recurrent feedback network (RFSR) [32] employing structure-channel priors, ESSAformer [38] with improved transformer architecture. Bicubic interpolation serves as the baseline.

Multi-modal Comparison: For the MFSR, the representative algorithms for comparison includes SR-guided progressive pansharpening neural network (SRPPNN) [2], parallel multiscale attention constraint network (PMAC) [25], local-context adaptive convolution kernels-based LAGNet [17], multiscale dual-domain guidance network (MSDDN) [11], U2Net [28] with a dual-stream CNN encoder-decoder structure, flow-Based Deep Network (PanflowNet) [35] and memory-augmented deep conditional unfolding network (MDCUN) [36], and the Mamba-based methods PanMamba [10] and FusionMamba [29].

For all tasks, the Adam optimizer [19] is used to train MambaX with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the loss function follows the configuration in [32]. In SISR, both datasets are trained for 120 epochs. The initial learning rate is set to 0.0003 for the Pavia dataset and 0.0001 for the CAVE dataset, with a decay factor of 0.5 applied at epochs $[5, 15, 30, 45, 75]$. For MFSR, the model is trained on the WV3 dataset for 500 epochs with an initial learning rate of 0.0007. The MultiStepLR scheduler is used, decaying the learning rate at epochs $[50, 100, 150, 200, 250, 300, 350, 400, 450]$ with a decay rate of 0.9. All experiments are implemented in PyTorch and conducted on eight NVIDIA GeForce GTX 4090 GPUs (24 GB each).

3.3. Single Image Super Resolution

3.3.1 Quantitative Comparison

Tab. 1 compares MambaX with existing methods on the CAVE dataset using three scaling factors. Convolution-based methods like 3DQRNN and GDRRN struggle with local feature extraction, limiting their effectiveness in hyperspectral super-resolution. In contrast, spatial-spectral prior methods (SSPSR, RFSR, ESSAformer) perform better, highlighting the benefits of joint spectral-spatial modeling. MambaX surpasses existing solutions by 0.5–1.2 dB in PSNR, leveraging two key architectural strengths: its state-space modeling paradigm, which enhances spatial-spectral integration for finer detail recovery, and Mamba's selective scanning mechanism, which adaptively expands the receptive field beyond conventional convolution and transformer-based attention. These features improve spectral fidelity and spatial resolution, as reflected in superior SAM and SSIM

Table 2. Quantitative performance on Worldview III dataset. **Bold** represents the best result and underline represents the second best.

Method	Worldview III			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓
SRPPNN	36.53	0.9673	4.32	2.42
PMAC	36.76	0.9686	4.22	2.36
PanflowNet	36.82	0.9695	4.16	2.34
LAGNet	36.96	0.9697	4.13	2.30
MDCUN	36.97	0.9707	4.06	2.29
MSDDN	37.07	0.9707	4.06	2.29
U2Net	37.30	0.9726	3.91	2.20
PanMamba	37.28	0.9727	3.93	2.21
FusionMamba	37.40	0.9731	3.87	2.17
MambaX	**37.87**	**0.9755**	**3.66**	**2.06**

metrics.

3.3.2 Visual Comparison

Fig. 2 shows MambaX's 2× super-resolution performance on the CAVE dataset, where it achieves the lowest MSE values alongside ESSAFormer while better preserving fine edges and textures. This highlights how its global perception mechanism enhances brightness consistency and spatial detail recovery. Spectral analysis in Fig. 2 further confirms its high spectral fidelity with minimal band-wise errors. These results demonstrate MambaX's effectiveness in spatial-spectral reconstruction, leveraging its state-space framework to optimize both local details and global spectral relationships, which is crucial for extreme upscaling tasks.

3.4. Image Fusion Super Resolution

3.4.1 Quantitative Comparison

Tab. 2 presents the quantitative experimental results of various algorithms on the image fusion-based super-resolution datasets, WV3. In the WV3 dataset, SRPPNN and Panflownet have relatively high parameter counts but show poor performance. LAGNet and MDCUN, with fewer parameters, still fall behind the Mamba-based methods, such as PanMamba and FusionMamba, in terms of PSNR and SSIM. In contrast, MambaX not only achieves the highest accuracy but also has a moderate parameter count. Specifically, in the WV3 dataset, MambaX outperforms other algorithms, achieving the highest PSNR and SSIM values, with accuracy improvements ranging from 1.34 dB to 0.40 dB and 0.0082 to 0.0024, respectively. These results effectively demonstrate MambaX's exceptional performance in maintaining image quality. Additionally, MambaX achieves the lowest SAM and ERGAS values, indicating superior performance in spectral fidelity and global error control.

Table 3. Quantitative comparisons of variants of State-Level Fusion on WV3 dataset. **Bold** represents the best result and underline represents the second best.

Variant	Worldview III			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓
B, C and $\triangle \Rightarrow X$	37.73	0.9748	3.75	2.10
B and $C \Rightarrow S$	37.58	0.9743	3.79	2.12
B, C and $\triangle \Rightarrow S$	37.81	0.9752	3.71	2.07
$B \Rightarrow S$	**37.87**	**0.9755**	**3.66**	**2.06**

3.4.2 Visual Comparison

Fig. 3 demonstrates MambaX's coastal scene reconstruction on the WV dataset. In these challenging scenarios with complex wave textures, our method outperforms competitors in both visual quality (land-cover details) and spectral accuracy (minimum reconstruction errors). This consistent performance across varied scenes and scaling factors confirms MambaX's robustness, particularly its capacity to handle irregular spatial-spectral patterns - a critical requirement for practical remote sensing applications.

3.5. Ablation Study

3.5.1 Effect of State-Level Fusion

To evaluate different state-level fusion strategies, we compared the model's performance when generating control variables using either X or auxiliary images separately, as shown in Tab. 3. The results indicate that generating all control variables from X yields inferior performance compared to using S, highlighting the crucial role of auxiliary information in state-space-based image fusion.

Notably, the model achieves optimal performance when S is used to generate B. This can be attributed to B's indirect modulation of the intermediate state through X, which introduces less disruption compared to direct control via C. Furthermore, the multiplication of B and X serves as an initial fusion step, which is subsequently refined through state transitions and output adjustments, ensuring global spatial consistency. This approach enhances the model's ability to integrate and leverage auxiliary information effectively.

3.5.2 Effect of Cross-domain transition module

In this subsection, we compare different upsampling methods for super-resolution (see Tab. 4). For 2× and 4× degradations, quadratic spline interpolation and convolution with PixelShuffle perform similarly. However, for 8× degradation, quadratic spline yields better accuracy, likely because deep learning methods struggle with high degradation due to limited receptive fields. Bicubic interpolation, by con-

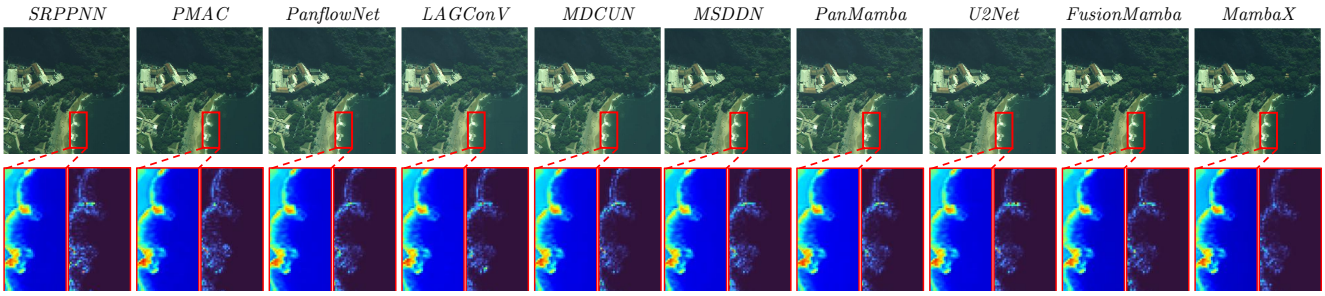


Figure 3. Visual quality comparison on WV3 dataset. The first row shows the true color images of different results. The second row shows the local visualization results of the band of the super-resolution results.

Table 4. Quantitative comparisons of some variants of the CDTM over the testing set on the Pavia dataset at the scale factor ×2, ×4, and ×8. **Bold** represents the best result and underline represents the second best.

Variant	×2			×4			×8		
	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓
Bicubic	35.42	0.9513	**3.63**	29.25	**0.7998**	**5.33**	25.82	**0.5756**	**7.12**
Conv+Pixelshuffle	35.42	0.9512	3.65	29.25	0.7988	5.42	25.80	0.5676	7.19
CDTM w/o adaptive weights	35.39	0.9511	3.65	29.24	0.7977	5.43	25.79	0.5655	7.19
CDTM w/ adaptive weights	**35.46**	**0.9514**	3.66	**29.28**	0.7990	5.42	**25.84**	0.5724	**7.12**

Table 5. Quantitative comparisons of variants of DSCP on WV3 dataset. **Bold** represents the best result and underline represents the second best.

Variant	Worldview III			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓
ALL Dynamic	37.79	0.9751	3.71	2.08
DSPC w/o spatial operator	37.85	**0.9755**	**3.68**	**2.06**
DSPC w/o spectral operator	37.83	0.9753	3.71	2.07
MambaX	**37.87**	**0.9755**	3.69	**2.06**

trast, remains stable on heavily degraded images. MambaX achieves the highest PSNR across all scales. When CDTM does not use adaptive weights to fuse upsampling bases, performance drops significantly, indicating that simply summing base transformations can hinder information recovery. These results underscore the need for adaptive fusion of base transformations. Notably, in multi-scale degradation scenarios, bicubic interpolation attains the best SAM, suggesting that MambaX's upsampling could be further refined to enhance spectral fidelity.

3.5.3 Effect of DSPC in Mamba

To evaluate this module's performance, we tested various state control models and examined how channel and structure information extraction affects control variable generation (see Tab. 5). The results show that generating all state control variables dynamically yields poorer performance than using other approaches, indicating that a purely dynamic method may not enhance model performance. In the WV3 dataset, the ablation study reveals that spectral operators are particularly important for multi-band image super-resolution, and the full MambaX structure consistently delivers the best results. This demonstrates that relying solely on linear methods limits the model's ability to fully leverage both spatial and spectral information, thereby validating our approach.

4. Conclusion

We introduce MambaX, a dynamic learning framework for nonlinear control matrices that uses a custom neural network to adaptively compute control coefficients. This improves the optimization of the blind SR process. MambaX extends the learnable scope of the control matrix, enhancing linearized dynamics in the latent space. To address domain and dimensional discrepancies, we propose a CDTM that reduces error during the spectral-to-state space transformation. Additionally, we present a cross-control mechanism for multimodal fusion, where latent space parameters guide optimization across modalities. We evaluate MambaX in various image SR tasks, showing performance similar to state-of-the-art stochastic SSMs. Although effective, MambaX remains a multi-stage linear system, and we plan to add uncertainty analysis for better adaptability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 42271350, and also supported by the International Partnership Program of the Chinese Academy of Sciences under Grant No.313GJHZ2023066FN.

References

[1] Saeed Anwar and Nick Barnes. Densely residual laplacian super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1192–1204, 2022. 1

[2] Jiajun Cai and Bo Huang. Super-resolution-guided progressive pansharpening based on a deep convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):5206–5220, 2021. 5

[3] Liang-jian Deng, Gemine Vivone, Mercedes E. Paoletti, Giuseppe Scarpa, Jiang He, Yongjun Zhang, Jocelyn Chanussot, and Antonio Plaza. Machine learning in pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine*, 10(3):279–315, 2022. 4

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. 1

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[6] Ying Fu, Zhiyuan Liang, and Shaodi You. Bidirectional 3d quasi-recurrent neural network for hyperspectral image super-resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2674–2688, 2021. 1, 4

[7] P. Gamba. A collection of data for urban area characterization. In *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium*, volume 1, page 72, 2004. 4

[8] Shangqi Gao and Xiahai Zhuang. Bayesian image super-resolution with deep modeling of image statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1405–1423, 2022. 1

[9] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Re. Hippo: Recurrent memory with optimal polynomial projections, 2020. 2

[10] Xuanhua He, Ke Cao, Jie Zhang, Keyu Yan, Yingying Wang, Rui Li, Chengjun Xie, Danfeng Hong, and Man Zhou. Pan-mamba: Effective pan-sharpening with state space model. *Information Fusion*, 115:102779, 2025. 5

[11] Xuanhua He, Keyu Yan, Jie Zhang, Rui Li, Chengjun Xie, Man Zhou, and Danfeng Hong. Multiscale dual-domain guidance network for pan-sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. 2, 5

[12] Danfeng Hong, Chenyu Li, Bing Zhang, Naoto Yokoya, Jon Atli Benediktsson, and Jocelyn Chanussot. Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing big data in earth observation, 2024. 1

[13] Danfeng Hong, Naoto Yokoya, Jocelyn Chanussot, and Xiao Xiang Zhu. An augmented linear mixing model to address spectral variability for hyperspectral unmixing. *IEEE Transactions on Image Processing*, 28(4):1923–1938, 2018. 1

[14] Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44:800–801, 2008. 4

[15] Kui Jia, Xiaogang Wang, and Xiaoou Tang. Image transformation based on learning dictionaries across image spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):367–380, 2013. 1

[16] Junjun Jiang, He Sun, Xianming Liu, and Jiayi Ma. Learning spatial-spectral prior for super-resolution of hyperspectral imagery. *IEEE Transactions on Computational Imaging*, 6:1082–1096, 2020. 5

[17] Zi-Rong Jin, Tian-Jing Zhang, Tai-Xiang Jiang, Gemine Vivone, and Liang-Jian Deng. Lagconv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):1113–1121, Jun. 2022. 1, 5

[18] Kwang In Kim and Younghee Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1127–1133, 2010. 1

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 5

[20] Chenyu Li, Danfeng Hong, Bing Zhang, Yuxuan Li, Gustau Camps-Valls, Xiao Xiang Zhu, and Jocelyn Chanussot. Urbansam: Learning invariance-inspired adapters for segment anything models in urban construction. *arXiv preprint arXiv:2502.15199*, 2025. 2

[21] Chenyu Li, Bing Zhang, Danfeng Hong, Jun Zhou, Gemine Vivone, Shutao Li, and Jocelyn Chanussot. Casformer: Cascaded transformers for fusion-aware

computational hyperspectral imaging. *Information Fusion*, 108:102408, 2024. 2

[22] Qiang Li, Qi Wang, and Xuelong Li. Mixed 2d/3d convolutional network for hyperspectral image super-resolution. *Remote Sensing*, 12(10), 2020. 4

[23] Xuyang Li, Chenyu Li, Pedram Ghamisi, and Danfeng Hong. Fleximo: A flexible remote sensing foundation model. *arXiv preprint arXiv:2503.23844*, 2025. 2

[24] Yong Li, Lei Zhang, Chen Dingl, Wei Wei, and Yanning Zhang. Single hyperspectral image super-resolution with grouped deep recursive residual network. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pages 1–4, 2018. 5

[25] Yixun Liang, Ping Zhang, Yang Mei, and Tingqi Wang. Pmacnet: Parallel multiscale attention constraint network for pan-sharpening. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 5

[26] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind image super-resolution: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5461–5480, 2023. 1

[27] Zhaojie Pan, Chenyu Li, Antonio Plaza, Jocelyn Chanussot, and Danfeng Hong. Hyperspectral image classification with mamba. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[28] Siran Peng, Chenhao Guo, Xiao Wu, and Liang-Jian Deng. U2net: A general framework with spatial-spectral-integrated double u-net for image fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 3219–3227, New York, NY, USA, 2023. Association for Computing Machinery. 5

[29] Siran Peng, Xiangyu Zhu, Haoyu Deng, Liang-Jian Deng, and Zhen Lei. Fusionmamba: Efficient remote sensing image fusion with state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 5

[30] Lucien Wald. *Data Fusion. Definitions and Architectures - Fusion of Images of Different Spatial Resolutions*. Presses de l'Ecole, Ecole des Mines de Paris, Paris, France, 2002. ISBN 2-911762-38-X. 4

[31] Lucien Wald, Thierry Ranchin, and Marc Mangolini. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric engineering and remote sensing*, 63(6):691–699, 1997. 4

[32] Xinya Wang, Jiayi Ma, and Junjun Jiang. Hyperspectral image super-resolution via recurrent feedback embedding and spatial–spectral consistency regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 5

[33] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 4

[34] Xin Wu, Danfeng Hong, and Jocelyn Chanussot. Convolutional neural networks for multimodal remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–10, 2021. 1

[35] Gang Yang, Xiangyong Cao, Wen Xiao, Man Zhou, Aiping Liu, Xun Chen, and Deyu Meng. Panflownet: A flow-based deep network for pan-sharpening. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16811–16821, 2023. 5

[36] Gang Yang, Man Zhou, Keyu Yan, Aiping Liu, Xueyang Fu, and Fan Wang. Memory-augmented deep conditional unfolding network for pansharpening. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1787, 2022. 5

[37] Roberta H. Yuhas, Alexander F. H. Goetz, and Joe W. Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*, 1992. 4

[38] Mingjin Zhang, Chi Zhang, Qiming Zhang, Jie-Ru Guo, Xinbo Gao, and Jing Zhang. Essaformer: Efficient transformer for hyperspectral image super-resolution. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23016–23027, 2023. 5