FOVEAMAMBA: STRUCTURE-CONDITIONED STATE EVOLUTION FOR REMOTE SENSING IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote-sensing image super-resolution requires recovering fine geometric details while modeling long-range spatial dependencies. An HR-reference-based structure-stratified analysis reveals substantial SSIM degradation from low- to high-structure regions across representative SR methods. We propose FoveaMamba, a structure-conditioned state-space network that incorporates multi-scale two-dimensional context into selective state evolution while retaining fixed directional traversal. Specifically, its Context-Guided State Space Module (CGSSM) uses the extracted context to modulate the parameterization of the discretization step Δ. A stabilized high-frequency residual pathway provides complementary local detail compensation alongside long-range state-space propagation. Under a unified AID-training protocol, FoveaMamba achieves the highest PSNR on all six evaluated remote-sensing datasets and the highest SSIM on five. Structure-stratified comparisons indicate larger paired SSIM gains over selected baselines in high-structure regions, while internal mechanism analyses reveal greater network-averaged discretization-step modulation in these regions. Following DIV2K fine-tuning, FoveaMamba also attains the highest PSNR on four of the five evaluated natural-image benchmarks. Source code and trained models will be made publicly available.

1 INTRODUCTION

Remote sensing image super-resolution (RSISR) requires recovering fine geometric details across large spatial extents, where homogeneous regions coexist with dense structures such as roads, building boundaries, and agricultural patterns. CNN-based methods exploit local inductive biases (Dong et al., 2016; Lim et al., 2017; Zhang et al., 2018a), while Transformer-based approaches model long-range interactions through self-attention (Liang et al., 2021; Chen et al., 2023). Recently, state space models (SSMs), particularly Mamba (Gu & Dao, 2024), have provided efficient long-range modeling and been adopted in image restoration (Guo et al., 2024; 2025; Choi et al., 2026; Li et al., 2025). Nevertheless, reconstructing geometrically complex regions remains challenging.

To characterize this limitation, we investigate how reconstruction performance varies with local structural complexity through an HR-reference-based structure-stratified analysis. As illustrated in Fig. 1(a), representative HR patches exhibit progressively higher Sobel energy, ranging from relatively homogeneous regions with sparse gradient responses to regions containing dense geometric structures and pronounced gradients. Using identical HR-defined structural partitions for all methods, Fig. 1(b) quantifies the corresponding reconstruction disparity. Across representative CNN-, Transformer-, and SSM-based methods, SSIM decreases by approximately 0.19–0.20 from Low- to High-Structure regions, corresponding to a relative reduction of 21%–23% with respect to Low-Structure SSIM. These results reveal a consistent association between higher structural energy and lower reconstruction quality across the evaluated architectural families.

The persistence of this disparity in SSM-based methods raises the question of how two-dimensional context can inform recurrent state evolution. Existing restoration-oriented SSMs employ directional scanning, semantic sequence organization, recurrent modulation, and spatial–frequency interaction (Guo et al., 2024; 2025; Choi et al., 2026; Li et al., 2025). Rather than modifying the scanning

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

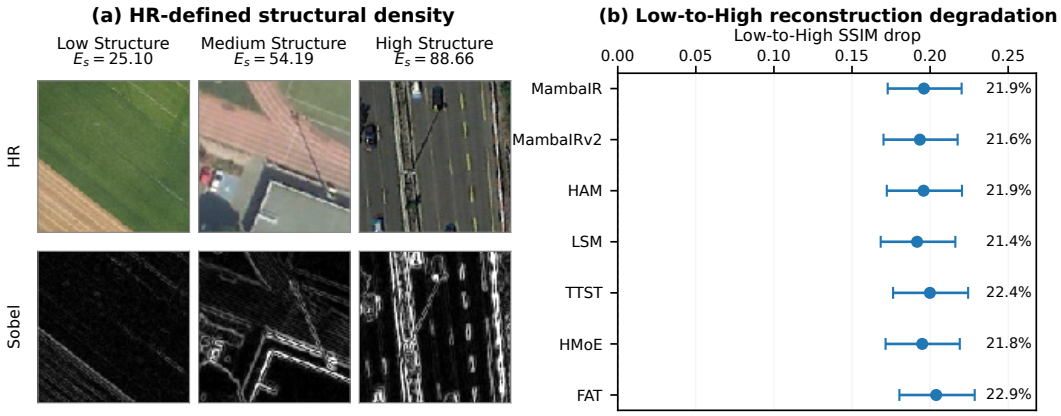


Figure 1: **Structure-stratified reconstruction analysis.** (a) Representative HR patches with increasing Sobel structural energy and their corresponding gradient responses. (b) Low-to-High SSIM degradation across representative SR methods. Error bars denote 95% image-level bootstrap confidence intervals; percentages indicate degradation relative to Low-Structure SSIM.

strategy, we investigate the state-transition parameterization. In a discretized SSM, the transition contains $\exp(\Delta_t \mathbf{A})$, where Δ_t directly influences state evolution. This motivates conditioning the discretization-step parameterization on multi-scale two-dimensional context while retaining fixed directional traversal.

To this end, we propose **FoveaMamba**, a structure-conditioned state-space network for RSISR. Pyramid Context Dilation (PCD) extracts spatially aligned multi-scale context through parallel dilated depthwise convolutions. The Context-Guided State Space Module (CGSSM) uses this context to modulate the input-dependent step parameter under fixed four-direction scanning. In parallel, MHI-Stable provides local high-frequency compensation through Laplacian-guided spatial modulation, global channel gating, and learnable residual scaling. We evaluate reconstruction performance across structural-density levels and examine the corresponding internal discretization-step modulation.

Our contributions are summarized as follows:

- We establish a structure-stratified analysis based on ground-truth HR references and quantify the consistent degradation of representative SR methods as local structural density increases.
- We propose FoveaMamba, which uses multi-scale two-dimensional context to condition selective state evolution under fixed directional traversal, complemented by a parallel high-frequency compensation pathway.
- Across six remote-sensing datasets, FoveaMamba achieves the highest PSNR on all six and the highest SSIM on five. Structure-stratified and mechanism analyses further show larger relative gains over selected baselines and stronger network-averaged step modulation in structure-dense regions.

2 Related Work

2.1 Image Super-Resolution and Remote Sensing SR

Single-image super-resolution has progressed from convolutional to Transformer-based architectures. CNN methods improve local representation through residual, dense, and channel-attention mechanisms (Dong et al., 2016; Lim et al., 2017; Zhang et al., 2018b;a), while Transformer approaches extend long-range interaction through self-attention (Liang et al., 2021; Chen et al., 2023). Remote sensing image super-resolution (RSISR) additionally involves heterogeneous scenes, repeated geometric patterns, sharp boundaries, and large spatial extents. Recent methods introduce se-

lective token interaction, frequency-aware processing, and heterogeneous expert routing for RSISR (Xiao et al., 2024; Yan et al., 2025; Chen et al., 2025). These developments motivate models that jointly preserve local geometric information and capture long-range spatial dependencies.

2.2 Visual State Space Modeling

State space models provide linear-complexity sequence modeling for long-range representation. Mamba introduces selective state-space dynamics by making recurrent parameters, including the discretization step Δ, input dependent (Gu & Dao, 2024). Visual adaptations extend this formulation to two-dimensional data through different sequence constructions: VMamba employs multi-directional selective scanning, while Vim uses bidirectional state-space modeling (Liu et al., 2024; Zhu et al., 2024). Hierarchical and multi-scale variants further improve spatial interaction (Qiao et al., 2025). Despite these developments, recurrent computation is still performed over ordered representations derived from feature maps, leaving how spatially aligned two-dimensional context should participate in recurrent state evolution as an important design consideration.

2.3 Adaptive State-Space Modeling for Image Restoration

State-space and recurrent models have recently been adapted to image restoration through different forms of feature and recurrence modulation. MambaIR combines selective SSMs with local enhancement and channel attention, while MambaIRv2 further introduces attentive state-space modeling and semantic-guided neighboring (Guo et al., 2024; 2025). For RSISR, FMSR and HAM incorporate frequency-aware and hierarchical spatial–frequency modeling (Xiao et al., 2025; Li et al., 2025). LSM is particularly related to our work because it conditions LRU recurrence using semantic modulation together with spatial categorization and prototype-based enhancement (Choi et al., 2026). The two methods, however, differ in both conditioning information and intervention point: LSM uses learned semantic prototypes to modulate LRU recurrence, whereas FoveaMamba uses spatially aligned multi-scale two-dimensional context to condition the discretization dynamics of a selective state-space model while retaining fixed directional traversal.

3 Method

3.1 FoveaMamba Overview

As illustrated in Fig. 2(a), FoveaMamba consists of shallow feature extraction, deep feature modeling, and image reconstruction. Given a low-resolution image I_{LR}, the network extracts shallow features F_0 and progressively models them through N_2 stacked Fovea Residual Groups (FRGs). Each FRG comprises N_1 FoveaMamba Blocks (FMBs), followed by spatial feature aggregation and a group-level residual connection. The resulting deep features are integrated with F_0 through a global residual connection and passed to the reconstruction head to produce the super-resolved image I_{SR}.

Within each FMB, Pyramid Context Dilation (PCD) extracts multi-scale spatial context to condition state evolution in the Context-Guided State Space Module (CGSSM), while MHI-Stable provides a parallel pathway for local detail compensation. Given the normalized input $\widehat{X}_{l-1} = \mathrm{LN}_1(X_{l-1})$, the two branches produce responses Y_{ssm} and Y_{mhi}, respectively, which are fused with the input as

$$U_l = \widehat{X}_{l-1} + Y_{\mathrm{ssm}} + Y_{\mathrm{mhi}}. \tag{1}$$

The fused representation undergoes channel mixing and residual refinement, followed by a ConFFN branch. Figure 2(b) presents the detailed structures of PCD, CGSSM, and MHI-Stable, which are described in the following subsections.

3.2 Structure-Conditioned State Evolution

CGSSM incorporates multi-scale two-dimensional context into the discretization-step parameterization of a selective SSM while retaining fixed directional traversal. In a standard selective SSM, the discretization step for directional stream k is computed from an input-dependent parameter $d_t^{(k)}$

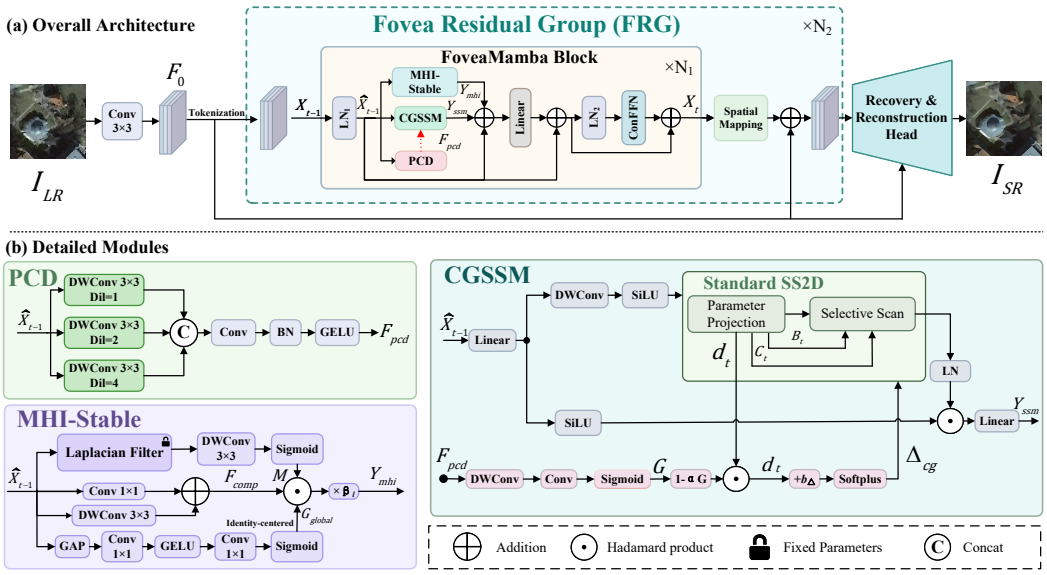


Figure 2: **Architecture of FoveaMamba.** (a) Overall architecture comprising N_2 FRGs, each containing N_1 FMBs, with hierarchical residual connections and an image reconstruction head. (b) Detailed structures of PCD, CGSSM, and MHI-Stable.

and a learned bias $b_\Delta^{(k)}$:

$$\Delta_{\text{base}}^{(k)} = \text{softplus}\left(d_t^{(k)} + b_\Delta^{(k)}\right). \tag{2}$$

Since the effective state transition depends on the discretization step through $\exp(\Delta_t \mathbf{A})$, its parameterization provides an intervention point for context-dependent state evolution. We introduce a conditioning signal derived from the original two-dimensional feature space without modifying the selective scanning operation.

Multi-scale spatial context. PCD extracts contextual features from $\widehat{X}_{l-1}$ using three parallel depthwise 3×3 convolutions with dilation rates $\{1, 2, 4\}$. Denoting their outputs by F_1, F_2, and F_4, the fused context is

$$F_{\text{pcd}} = \text{GELU}\left(\text{BN}\left(\text{Conv}_{1\times1}\left([F_1, F_2, F_4]\right)\right)\right). \tag{3}$$

The parallel branches aggregate neighborhoods at different spatial extents, while their fusion preserves the original two-dimensional feature layout. Rather than directly enhancing the input features, the resulting context provides a separate conditioning signal for CGSSM.

Context-guided step modulation. CGSSM projects $\widehat{X}_{l-1}$ into a state branch X_s and an output-gating branch Z. Following depthwise 3×3 convolution and SiLU activation, X_s is serialized into horizontal and vertical sequences and their reverse directions. For each directional stream k, selective projections generate $d_t^{(k)}$, $\mathbf{B}_t^{(k)}$, and $\mathbf{C}_t^{(k)}$.

In parallel, the PCD features are transformed into a conditioning gate:

$$G = \sigma\left(\text{Conv}_{1\times1}\left(\text{DWConv}_{3\times3}(F_{\text{pcd}})\right)\right). \tag{4}$$

The pointwise projection maps the context features to the SSM inner dimension. Let $G^\flat$ denote the raster-order flattening of G. In the evaluated implementation, this sequence is broadcast across all four directional streams without direction-specific reordering. The projected step parameters are modulated as

$$\widetilde{d}_t^{(k)} = d_t^{(k)} \odot \left(1 - \alpha_l G^\flat\right), \tag{5}$$

where $\odot$ denotes element-wise multiplication and α_l is a learnable scalar initialized to zero. The conditioned discretization step is then obtained as

$$\Delta_{\mathrm{cg}}^{(k)} = \mathrm{softplus}\left(\widetilde{d}_t^{(k)} + b_{\Delta}^{(k)}\right). \tag{6}$$

When $\alpha_l = 0$, Eq. 6 reduces to the unmodulated step parameterization in Eq. 2. Context conditioning is therefore introduced through learning without changing the initial discretization-step computation.

The conditioned steps are supplied to the standard selective scan together with $\mathbf{A}$, $\mathbf{B}_t^{(k)}$, $\mathbf{C}_t^{(k)}$, and $\mathbf{D}$. Outputs from the four directions are restored to a common spatial order and aggregated, followed by LayerNorm, output gating with $\mathrm{SiLU}(Z)$, and linear projection to obtain Y_{ssm}. Because modulation precedes the bias and Softplus transformation, it introduces context dependence without imposing a predetermined direction of change on the resulting discretization step.

3.3 Stable High-Frequency Compensation

While CGSSM incorporates multi-scale context into recurrent state evolution, MHI-Stable provides a parallel pathway for local detail compensation. Operating directly on $\widehat{X}_{l-1}$, this branch combines local feature transformations with spatial and channel modulation without modifying the state-space transition.

The compensation feature combines depthwise spatial filtering and pointwise channel projection:

$$F_{\mathrm{comp}} = \mathrm{DWConv}_{3\times3}(\widehat{X}_{l-1}) + \mathrm{Conv}_{1\times1}(\widehat{X}_{l-1}). \tag{7}$$

To regulate this feature using local second-order responses, a fixed signed Laplacian operator is followed by a learnable depthwise convolution and sigmoid activation:

$$M_l = \sigma\left(\mathrm{DWConv}_{3\times3}\left(\mathbf{K}_{\mathrm{lap}} * \widehat{X}_{l-1}\right)\right). \tag{8}$$

Here, $\mathbf{K}_{\mathrm{lap}}$ denotes the fixed Laplacian kernel, and M_l serves as a learned spatial modulation mask rather than an explicit boundary prediction.

MHI-Stable additionally employs global channel modulation. Global average pooling and two channel projections produce

$$Q_l = \sigma\left(W_{g,2}\left(\mathrm{GELU}\left(W_{g,1}\left(\mathrm{GAP}(\widehat{X}_{l-1})\right)\right)\right)\right), \tag{9}$$

which is converted into an identity-centered channel gate:

$$G_{\mathrm{global},l} = 1 + \eta(2Q_l - 1), \qquad \eta = 0.5. \tag{10}$$

The final compensation response combines spatial and channel modulation with learnable residual scaling:

$$Y_{\mathrm{mhi}} = \beta_l \left(G_{\mathrm{global},l} \odot M_l \odot F_{\mathrm{comp}}\right). \tag{11}$$

The final projection in the global gating branch is zero-initialized, yielding $Q_l = 0.5$ and $G_{\mathrm{global},l} = 1$ initially. The residual scale β_l is also initialized to zero, so MHI-Stable initially contributes no additional residual response and its contribution is learned during optimization.

3.4 Training Objective

FoveaMamba is optimized using an L_1 reconstruction loss and an auxiliary Laplacian consistency loss:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{pix}} + \lambda_{\mathrm{edge}}\mathcal{L}_{\mathrm{edge}}. \tag{12}$$

The pixel reconstruction loss is defined as

$$\mathcal{L}_{\mathrm{pix}} = \frac{1}{N}\left\|I_{\mathrm{SR}} - I_{\mathrm{HR}}\right\|_1, \tag{13}$$

while the Laplacian consistency loss measures differences between the second-order responses of the reconstructed and reference images:

$$\mathcal{L}_{\mathrm{edge}} = \frac{1}{N}\left\|\mathbf{K}_{\mathrm{lap}} * I_{\mathrm{SR}} - \mathbf{K}_{\mathrm{lap}} * I_{\mathrm{HR}}\right\|_1. \tag{14}$$

Here, N denotes the number of scalar image elements used for loss normalization, and $\lambda_{\mathrm{edge}} = 0.1$. The Laplacian term provides auxiliary supervision for local second-order consistency as part of the full training objective.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

The main FoveaMamba model is trained on four NVIDIA A100 GPUs. AID (Xia et al., 2017) is the sole training dataset, while DOTA (Xia et al., 2018), NWPU-RESISC45 (Cheng et al., 2017), UCMerced LandUse (Yang & Newsam, 2010), WHU-RS19 (Dai & Yang, 2011), and Optimal-31 (Wang et al., 2019) are evaluated without fine-tuning. LR images are generated using bicubic downsampling (Keys, 1981) at $\times 4$. All learned baselines use the same AID split and validation protocol, with PSNR and SSIM as the primary metrics. FoveaMamba uses $C = 180$, $d_{\text{state}} = 16$, and six FRGs with six FMBs per group. Training uses Adam (Kingma & Ba, 2015) for 300K iterations with an initial learning rate of 4×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.99)$, zero weight decay, 32×32 LR patches, and random flipping and rotation. Detailed dataset partitions, training configurations, and evaluation protocols are provided in Appendix A.

4.2 OVERALL COMPARISON

We compare convolutional methods EDSR (Lim et al., 2017), RDN (Zhang et al., 2018b), and RCAN (Zhang et al., 2018a); Transformer methods SwinIR (Liang et al., 2021) and HAT (Chen et al., 2023); remote-sensing methods TTST (Xiao et al., 2024), FAT (Yan et al., 2025), and HMOE (Chen et al., 2025); and state-space/recurrent methods MambaIR (Guo et al., 2024), MambaIRv2 (Guo et al., 2025), HAM (Li et al., 2025), and LSM (Choi et al., 2026). Bicubic interpolation is included as a non-learned reference. All learned methods use the AID split and evaluation protocol described in Section 4.1.

Table 1: Remote-sensing $\times 4$ SR results. All learned methods use the same AID split; external datasets are evaluated without adaptation. Each cell reports PSNR/SSIM. Best and second-best results are **bold** and underlined, respectively.

Method	AID	DOTA	NWPU	UCM	WHU	Opt.-31
Non-SSM methods						
Bicubic	27.58/0.7260	30.49/0.7913	26.32/0.6855	25.54/0.6879	28.08/0.7453	26.06/0.6831
EDSR	29.26/0.7875	32.62/0.8475	27.86/0.7499	27.03/0.7475	29.64/0.7990	27.76/0.7541
RDN	29.26/0.7869	32.66/0.8473	27.85/0.7488	27.09/0.7489	29.62/0.7976	27.78/0.7536
RCAN	29.31/0.7896	32.72/0.8500	27.89/0.7516	27.12/0.7514	29.69/0.8009	27.82/0.7564
SwinIR	29.34/0.7901	32.82/0.8520	27.93/0.7525	27.16/0.7530	29.72/0.8020	27.87/0.7577
HAT	29.42/0.7934	32.89/0.8536	28.00/0.7558	27.29/0.7577	29.82/0.8048	27.94/0.7609
TTST	29.12/0.7803	32.57/0.8499	27.75/0.7420	26.95/0.7410	29.46/0.7937	27.67/0.7466
FAT	28.96/0.7773	32.32/0.8461	27.64/0.7403	26.82/0.7382	29.31/0.7909	27.55/0.7450
HMOE	29.37/0.7916	32.81/0.8520	27.94/0.7534	27.18/0.7544	29.75/0.8027	27.88/0.7587
SSM/recurrent methods						
MambaIR	29.40/0.7924	32.82/0.8522	27.97/0.7548	27.23/0.7551	29.79/0.8039	27.92/0.7597
HAM	29.36/0.7910	32.82/0.8521	27.96/0.7534	27.28/0.7564	29.74/0.8027	27.90/0.7583
MambaIRv2	29.40/0.7932	32.87/0.8536	27.95/0.7545	27.19/0.7534	29.80/0.8049	27.91/0.7602
LSM	29.44/0.7938	32.88/0.8536	28.00/0.7563	27.29/**0.7582**	29.82/0.8051	27.96/0.7613
FoveaMamba (Ours)	**29.45/0.7940**	**32.92/0.8539**	**28.04/0.7570**	**27.33**/0.7570	**29.85/0.8058**	**27.99/0.7622**

As shown in Table 1, FoveaMamba obtains the highest PSNR on all six remote-sensing datasets and the highest SSIM on five. It records higher PSNR than both MambaIRv2 and LSM on each of the five unseen datasets, while LSM obtains higher SSIM on UCMerced. These comparisons describe the measured reconstruction performance under the common AID training split and cross-dataset evaluation protocol. Network-level inference measurements and supplementary SCC and ERGAS results are provided in Appendices E and C, respectively.

As shown in Fig. 3, FoveaMamba reconstructs building boundaries in AID, fine linear structures in WHU-RS19, and vehicle contours in Optimal-31 with improved structural fidelity, achieving the highest local SSIM among the compared methods in all three examples. These observations complement the quantitative comparisons in Table 1. Additional qualitative results are provided in Appendix G.

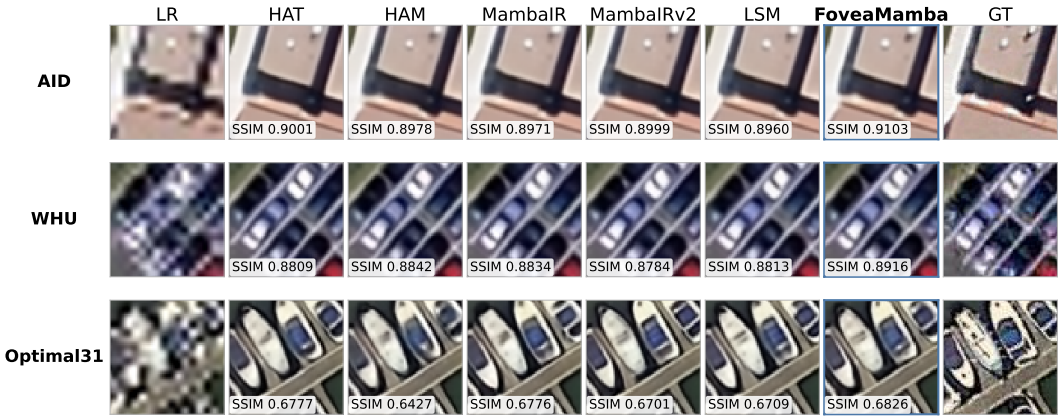


Figure 3: **Qualitative comparison for $\times 4$ remote-sensing SR.** Representative crops from AID, WHU-RS19, and Optimal-31, with local SSIM values reported for each reconstruction.

4.3 Ablation Study

We conduct controlled ablations to examine multi-scale context utilization, context-conditioned state evolution, MHI-Stable, and Laplacian supervision. All variants share the same backbone, network scale, AID training split, and optimization schedule, and are trained from scratch for 80K iterations. Full FoveaMamba is also retrained under this budget for consistent within-study comparisons.

The Shared-Backbone SSM retains residual connections, channel mixing, ConFFN, and standard four-direction selective scanning without the proposed modules or edge supervision. PCD-Input uses multi-scale context for input enhancement, whereas CGSSM uses the same context to modulate the pre-Softplus step parameter (Eq. 5). CGSSM+MHI further introduces MHI-Stable, and Full FoveaMamba additionally employs Laplacian supervision with $\lambda_{\text{edge}} = 0.1$. Baseline+Edge evaluates the effect of this supervision on the shared backbone. We report overall reconstruction accuracy and performance on the top 30% HR Sobel-energy patches defined in the motivation analysis, using identical evaluation patches across variants.

Table 2: Ablation results on AID under a unified 80K training budget. Overall and High-Struct. report image-level and high-structure patch-level PSNR (dB)/SSIM, respectively.

Variant	PCD	MHI	Edge	Overall	High-Struct.
Shared-Backbone SSM	–			29.3342/0.7896	23.9559/0.682346
PCD-Input	Input			29.3212/0.7894	23.9575/0.681898
CGSSM	Cond.			29.3280/0.7899	23.9544/0.683164
CGSSM+MHI	Cond.	✓		29.3106/0.7895	23.9501/0.682033
Baseline+Edge	–		✓	29.3341/0.7900	23.9610/0.683315
Full FoveaMamba	Cond.	✓	✓	**29.3588/0.7904**	**24.0034/0.683924**

Table 2 shows configuration-dependent differences under the fixed 80K AID training budget. With pixel-only supervision, CGSSM obtains higher overall and high-structure SSIM than both the shared backbone and PCD-Input, although its PSNR is lower than that of the shared backbone. Adding MHI-Stable to CGSSM under the same pixel-only objective reduces both metrics at the image and high-structure levels. With Laplacian supervision, the full configuration obtains higher values for all four metrics than CGSSM+MHI without edge supervision and Baseline+Edge with the same edge-supervision setting. It also obtains the highest overall and high-structure PSNR/SSIM among the evaluated variants. These comparisons characterize the tested architectural and supervision combinations; the isolated effect of MHI-Stable under matched edge supervision is not evaluated.

4.4 Structure and Mechanism Analysis

Following Fig. 1, we analyze reconstruction performance across structural-density levels using Sobel energy computed from HR references:

$$E_{\mathrm{s}}(p) = \frac{1}{|p|} \sum_{i \in p} \|\nabla_{\mathrm{Sobel}} Y_{\mathrm{HR}}(i)\|_2 \,. \tag{15}$$

Within each dataset, patches are divided into Low-, Medium-, and High-Structure subsets using the 30th and 70th percentiles, with identical partitions applied to all methods. For baseline b and structural subset $s \in \{L, M, H\}$, the SSIM gain is defined as $\Delta_{\mathrm{SSIM}}^{(b,s)} = \mathrm{SSIM}_{\mathrm{FoveaMamba}}^{(s)} - \mathrm{SSIM}_{b}^{(s)}$. To quantify gain amplification across structural levels, we define $A^{(b,s)} = \Delta_{\mathrm{SSIM}}^{(b,s)} / \Delta_{\mathrm{SSIM}}^{(b,L)}$. Statistical significance is assessed using the unnormalized High-minus-Low contrast $C_{\mathrm{H-L}}^{(b)} = \Delta_{\mathrm{SSIM}}^{(b,H)} - \Delta_{\mathrm{SSIM}}^{(b,L)}$. Detailed statistical procedures are provided in Appendix B.

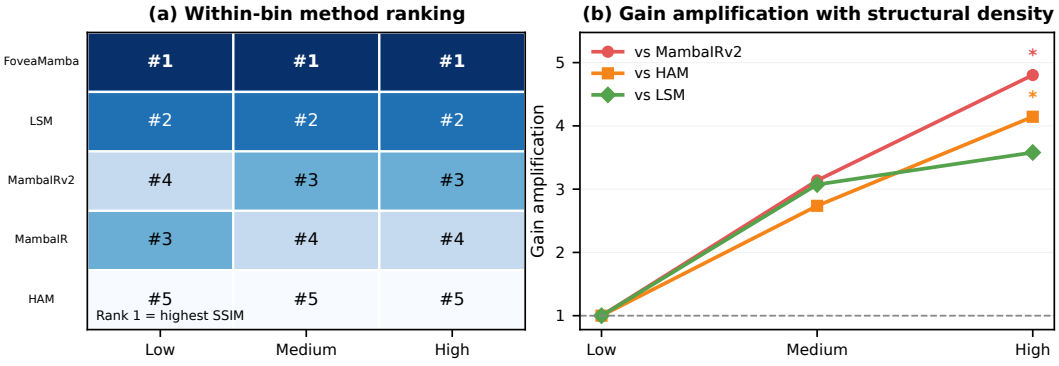


Figure 4: **Structure-stratified reconstruction analysis.** (a) Within-subset SSIM rankings among five representative methods. (b) SSIM-gain amplification relative to the Low-Structure subset, normalized separately for each baseline. Asterisks indicate that the image-level 95% bootstrap confidence interval of the unnormalized High-minus-Low contrast excludes zero.

Figure 4(a) shows that FoveaMamba ranks first in SSIM across all three structural subsets, followed by LSM. As shown in Fig. 4(b), its relative gains over MambaIRv2, HAM, and LSM increase from Low- to High-Structure regions. The unnormalized High-minus-Low contrasts are positive for all three baselines, with 95% bootstrap confidence intervals excluding zero for MambaIRv2 and HAM. The corresponding interval for LSM includes zero, indicating that the increase in relative gain is not statistically established against every baseline.

To investigate whether CGSSM exhibits structure-dependent state modulation, we examine how the context gate affects discretization steps across different structural regions. Using the same trained FoveaMamba, we compute the conditioned step $\Delta_{\mathrm{cg}}^{(k)}$ and its unmodulated counterpart $\Delta_{\mathrm{base}}^{(k)}$, obtained by disabling gate modulation without changing the learned parameters. Their absolute difference, $D_{\Delta}^{(k)} = |\Delta_{\mathrm{cg}}^{(k)} - \Delta_{\mathrm{base}}^{(k)}|$, quantifies the magnitude of context-induced step modulation, irrespective of its direction. We aggregate this measure across spatial positions, feature dimensions, scan directions, and network blocks to determine whether the overall modulation magnitude differs among the HR-defined structural subsets.

As shown in Fig. 5(a), High-Structure regions exhibit greater network-averaged discretization-step modulation than Low- and Medium-Structure regions. Figure 5(b) further shows a positive High-minus-Low contrast of 1.110×10^{-3}, with its 95% bootstrap confidence interval excluding zero, whereas the Medium-minus-Low contrast is smaller and its interval includes zero. These results provide evidence that CGSSM introduces structure-dependent differences in internal state modulation, with a greater aggregate response in structurally complex regions. Together with the reconstruction comparisons in Fig. 4, they establish complementary trends in reconstruction gains and internal modulation across structural levels, without implying a causal relationship between them. Additional spatial-response and layer-wise analyses are provided in Appendix D.

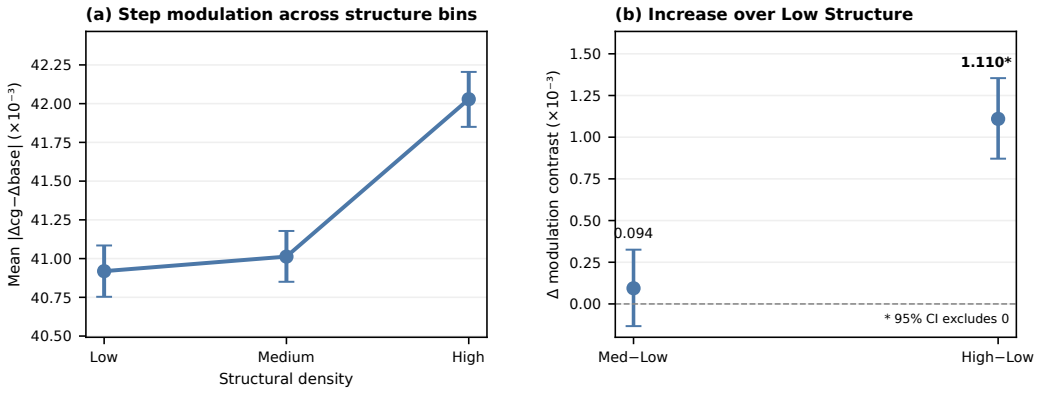


Figure 5: **Discretization-step modulation.** (a) Network-averaged modulation across structural subsets. (b) Modulation contrasts with 95% bootstrap confidence intervals; * indicates an interval excluding zero.

4.5 Cross-Domain Adaptation to Natural Images

To examine applicability beyond remote-sensing imagery, we adapt AID-pretrained models to natural-image SR using the complete DIV2K training set (Agustsson & Timofte, 2017) under a unified 30K-iteration adaptation budget. We compare FoveaMamba with TTST (Xiao et al., 2024), MambaIRv2 (Guo et al., 2025), and LSM (Choi et al., 2026). All methods follow the same target-domain data selection, degradation settings, and adaptation budget while retaining their respective training objectives. Performance is evaluated on Set5, Set14, BSD100, Urban100, and Manga109 using PSNR and SSIM. The complete adaptation and evaluation protocols are provided in Appendix F.

Table 3: Cross-domain adaptation on five natural-image SR benchmarks ($\times 4$). Each cell reports PSNR/SSIM. Best and second-best results are **bold** and underlined, respectively.

Method	Set5	Set14	BSD100	Urban100	Manga109
TTST (Xiao et al., 2024)	31.9405/0.8927	28.4616/0.7800	27.5179/0.7354	26.0459/0.7859	30.0052/0.9029
MambaIRv2 (Guo et al., 2025)	32.1096/0.8954	28.5780/0.7818	27.5812/0.7372	26.3392/0.7953	30.3412/0.9070
LSM (Choi et al., 2026)	32.1314/0.8953	28.6003/**0.7822**	27.5959/**0.7373**	**26.4214/0.7976**	30.3541/0.9070
FoveaMamba (Ours)	**32.1794/0.8958**	**28.6036**/0.7820	**27.6015**/0.7368	26.4029/0.7952	**30.5416/0.9082**

Table 3 shows that FoveaMamba obtains higher PSNR than MambaIRv2 on all five natural-image benchmarks and the highest PSNR among the evaluated methods on four. It obtains the highest PSNR and SSIM on Set5 and Manga109. On Urban100, LSM records higher values for both metrics; LSM also records higher SSIM on Set14 and BSD100. These results characterize the performance of the evaluated AID-initialized models after adaptation using the complete DIV2K training split and a common 30K-iteration budget.

5 Conclusion

We presented FoveaMamba, which conditions selective state evolution on multi-scale two-dimensional context under fixed directional traversal, complemented by parallel local high-frequency compensation. Under unified AID training, FoveaMamba achieves the highest PSNR on all six remote-sensing datasets and the highest SSIM on five, while demonstrating competitive performance after natural-image adaptation. Structure-stratified and mechanism analyses reveal larger relative reconstruction gains over selected baselines in structure-dense regions, accompanied by stronger network-averaged discretization-step modulation. These findings support context-conditioned state evolution as an approach to structure-aware image super-resolution.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript organization, language editing, LaTeX drafting, and brainstorming the organization of experiments and analyses. All AI-assisted material was reviewed and independently verified by the authors, including experimental protocols, code, numerical results, citations, and technical claims. The authors take full responsibility for the final content of the manuscript.

REPRODUCIBILITY STATEMENT

The architecture and principal equations are described in Section 3, and the primary training and evaluation protocols are summarized in Section 4.1. The ablation study is presented in Section 4.3. Additional implementation details, dataset partitioning and statistical protocols, computational measurements, mechanism analyses, natural-image adaptation settings, qualitative comparisons, and supplementary evaluation results are provided in the appendix.

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1122–1131, 2017. doi: 10.1109/CVPRW.2017.150.

Bowen Chen, Keyan Chen, Mohan Yang, Zhengxia Zou, and Zhenwei Shi. Heterogeneous mixture of experts for remote sensing image super-resolution. *IEEE Geoscience and Remote Sensing Letters*, 22:1–5, 2025.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22367–22377, 2023.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mingyu Choi, Woo Kyoung Han, Sunghoon Im, and Kyong Hwan Jin. Linear recurrent unit with semantic modulation for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 4950–4960, 2026.

Dengxin Dai and Wen Yang. Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Geoscience and Remote Sensing Letters*, 8(1):173–176, 2011.

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.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. MambaIR: A simple baseline for image restoration with state-space model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 222–241, 2024.

Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. MambaIRv2: Attentive state space restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28124–28133, 2025.

Robert G. Keys. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1153–1160, 1981.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Mingyue Li, Chengyi Xiong, Zhirong Gao, and Jiayi Ma. HAM: Hierarchical attention Mamba with spatial–frequency fusion for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using Swin Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1833–1844, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 136–144, 2017.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Junbo Qiao, Jincheng Liao, Wei Li, Yulun Zhang, Yong Guo, Jiao Xie, Jie Hu, and Shaohui Lin. HiMamba: Hierarchical Mamba for efficient image super-resolution. *IEEE Transactions on Image Processing*, 34:8461–8473, 2025.

Qi Wang, Shaoteng Liu, Jocelyn Chanussot, and Xuelong Li. Scene classification with recurrent attention of VHR remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):1155–1167, 2019.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3974–3983, 2018.

Yi Xiao, Qiangqiang Yuan, Kui Jiang, Jiang He, Chia-Wen Lin, and Liangpei Zhang. TTST: A top-k token selective Transformer for remote sensing image super-resolution. *IEEE Transactions on Image Processing*, 33:738–752, 2024.

Yi Xiao, Qiangqiang Yuan, Kui Jiang, Yuzeng Chen, Qiang Zhang, and Chia-Wen Lin. Frequency-assisted Mamba for remote sensing image super-resolution. *IEEE Transactions on Multimedia*, 27:1783–1796, 2025.

Xinyu Yan, Jiuchen Chen, Qizhi Xu, and Wei Li. Mitigating texture bias: A remote sensing super-resolution method focusing on high-frequency texture reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–18, 2025.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 270–279, 2010.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286–301, 2018a.

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2472–2481, 2018b.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 62429–62442, 2024.

Table 4: Training configuration for the 300K FoveaMamba main model.

Configuration	Setting
Training iterations	300K
Optimizer	Adam
Initial learning rate	4×10^{-4}
Adam betas	$(0.9, 0.99)$
Weight decay	0
LR scheduler	MultiStepLR
Milestones	250K, 300K
Decay factor	0.5
Warm-up	2K iterations
GPUs	$4 \times$ NVIDIA A100
Batch size	10/GPU (40 total)
HR / LR crop size	128^2 / 32^2
Augmentation	Flipping, rotation
Training loss	$\mathcal{L}_{\text{pix}} + 0.1\mathcal{L}_{\text{edge}}$
Training seed	10
Validation interval	1K iterations
Checkpoint interval	5K iterations

A Implementation and Reproducibility Details

A.1 Dataset Splits and Preprocessing

Source-domain partition. All remote-sensing datasets are obtained from their official sources. AID is the sole source-domain training dataset. Within each scene category, 50% of the original images are sampled with seed 42; these images are then divided into training, testing, and validation subsets at an approximate 7:2:1 ratio. Cumulative allocation limits rounding discrepancies across categories, and the remaining AID images are excluded.

Cross-dataset evaluation. Five additional datasets are reserved for evaluation without target-domain adaptation. For NWPU-RESISC45, UCMerced LandUse, WHU-RS19, and Optimal-31, approximately class-balanced subsets are sampled with seed 42, targeting roughly 1,000 images per dataset. Per-class quotas are rounded to multiples of ten and capped by data availability. The original WHU-RS19 training and validation partitions are pooled before sampling. For DOTA, random 512×512 HR crops are extracted from the original images, targeting 1,000 evaluation patches.

Degradation protocol. All remote-sensing experiments use bicubic degradation at $\times 4$. LR images are generated from their HR references using OpenCV `INTER_CUBIC`, reducing each spatial dimension by a factor of four. TIFF images are converted to PNG where necessary.

A.2 Training and Evaluation Protocol

Training protocol. FoveaMamba is trained from scratch on four NVIDIA A100 GPUs. Paired LR–HR images undergo random cropping, flipping, and rotation. A dataset enlargement ratio of 100 enables repeated sampling without adding unique training images. Table 4 specifies the optimization and training settings.

Model selection and evaluation. The checkpoint with the highest PSNR on the held-out AID validation subset is selected once and evaluated on the AID test set and all five external datasets, without target-domain adaptation or dataset-specific checkpoint selection. FoveaMamba employs overlapping tiled inference with reflection padding and non-overlapping assembly of output regions. PSNR and SSIM are calculated on BT.601 luminance after excluding a four-pixel border; dataset-level results average image-level measurements.

Table 5: Architecture and implementation parameters of FoveaMamba.

Configuration	Setting
Embedding dimension	180
FRGs / FMBs per FRG	6 / 6
State dimension	16
SSM expansion ratio	2
SSM inner dimension	360
Discretization rank	12
Initial step-size range	$[0.001, 0.1]$
Directional scans	4
SSM convolution	DWConv 3×3
PCD dilation rates	$\{1, 2, 4\}$
Feature groups	1
FFN expansion ratio	2
ConFFN gate reduction	4
MHI gate reduction	8
MHI global-gate strength	0.5
CGSSM α_l initialization	0
MHI β_l initialization	0
Maximum drop-path rate	0.1
Tokenization patch size	1
Residual-group projection	Conv 3×3
Upsampling	PixelShuffle $\times 4$

A.3 Detailed FoveaMamba Configuration

The architecture follows Section 3, with principal parameters listed in Table 5. RGB inputs are mean-centered using $[0.4488, 0.4371, 0.4040]$ and an image range of 1.0. The reconstruction head projects deep features to 64 channels and performs $\times 4$ upsampling through two successive $\times 2$ PixelShuffle stages.

PCD combines three parallel dilated depthwise convolutions using pointwise projection, BatchNorm, and GELU. CGSSM retains four-direction selective scanning and modulates the step parameter before adding its bias and applying Softplus. In the evaluated implementation, the PCD-derived gate is projected to the SSM inner dimension, flattened in raster order, and broadcast across all four scan streams without direction-specific reordering. . MHI-Stable uses the fixed signed Laplacian kernel $\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ to construct its spatial mask. Its global channel gate comprises two pointwise projections with GELU and sigmoid activations; the final projection is zero-initialized to yield an identity-centered gate initially. The compensation residual scale β_l is initialized to zero. ConFFN additionally applies global channel gating after its convolutional feed-forward transformation.

A.4 Baseline Implementations

All learned baselines retain their publicly released network architectures and are retrained on the same AID training split as FoveaMamba. For each model, checkpoint selection uses only PSNR on the held-out AID validation subset. Selected checkpoints are evaluated on the identical AID test set and five external test sets, without target-domain adaptation. The common evaluation procedure is described in Appendix A.2.

B Structure Partition and Statistical Protocol

B.1 Structure-Energy Computation and Partition

Structural energy is defined exclusively by HR references, independently of any reconstruction. Each eligible image is divided into non-overlapping 128×128 patches, discarding incomplete boundary regions. Images are converted to BT.601 luminance, and 3×3 Sobel gradients are calculated on the full image before patch extraction. Patch energy is the mean gradient magnitude defined

in Eq. 15. Dataset-specific 30th and 70th percentiles define Low-, Medium-, and High-Structure subsets; all methods are evaluated on the same HR-defined partitions. The examples in Fig. 1(a) are selected from the fixed AID diagnostic subset near the median energy of each structural category.

B.2 Statistical Analysis Protocol

The comparisons in Figs. 1(b) and 4 use a fixed set of 360 patches: 20 per structural category from each of six datasets. Each method is evaluated eight times per patch with a fixed checkpoint and predefined inference seeds. Patch-level SSIM is averaged across runs, aggregated by source image, and then averaged with equal dataset weights. The reported 95% confidence intervals use 5,000 image-level bootstrap replicates, resampling source images separately within each dataset and structural category; interval bounds are the 2.5th and 97.5th percentiles.

For Fig. 1(b), Low-to-High degradation is $D_m = \mathrm{SSIM}_m^{(L)} - \mathrm{SSIM}_m^{(H)}$. Confidence intervals are obtained by changing the signs and reversing the bounds of the bootstrap High-minus-Low contrasts. Relative degradation is defined as $100D_m/\mathrm{SSIM}_m^{(L)}$.

For Fig. 4, FoveaMamba–baseline SSIM differences are computed on identical patches and aggregated by source image and dataset. The High-minus-Low contrast compares the corresponding structural-group gains. Its confidence intervals resample source images independently within the Low- and High-Structure subsets of each dataset, while retaining the method pairing for every sampled patch.

C Supplementary Remote-Sensing Evaluation

To complement the PSNR and SSIM comparisons in Table 1, we report the spatial correlation coefficient (SCC) and relative global error in synthesis (ERGAS). SCC is the mean channel-wise Pearson correlation between reconstructed and reference RGB images. ERGAS is calculated from RGB images after excluding a four-pixel boundary. Tables 6 and 7 report the complete results across six datasets.

Table 6: SCC on six remote-sensing datasets ($\times 4$); higher is better. Best and second-best results are shown in **bold** and underlined, respectively. Ties are determined at the reported precision.

Method	AID	DOTA	NWPU	UCM	WHU	Opt.-31
Bicubic	0.9009	0.9045	0.8643	0.8661	0.9067	0.8560
EDSR	0.9293	0.9370	0.9000	0.8924	0.9304	0.8959
RDN	0.9292	0.9370	0.8996	0.8929	0.9299	0.8963
RCAN	0.9301	0.9384	0.9008	0.8945	0.9313	0.8976
SwinIR	0.9302	0.9396	0.9016	0.8956	0.9313	0.8984
HAT	0.9317	**0.9406**	0.9033	0.8982	0.9326	0.9002
TTST	0.9273	0.9382	0.8976	0.8904	0.9285	0.8940
FAT	0.9244	0.9342	0.8944	0.8881	0.9261	0.8907
HMOE	0.9310	0.9397	0.9019	0.8960	0.9318	0.8987
MambaIR	0.9316	0.9397	0.9029	0.8961	0.9325	0.8997
HAM	0.9310	0.9397	0.9025	0.8977	0.9320	0.8992
MambaIRv2	0.9314	0.9400	0.9022	0.8951	0.9327	0.8997
LSM	0.9321	**0.9406**	0.9035	**0.8983**	0.9330	0.9005
FoveaMamba	**0.9323**	0.9402	**0.9042**	0.8971	**0.9332**	**0.9013**

Relative to LSM, FoveaMamba records higher SCC on four of the six datasets and lower ERGAS on five. It attains the highest SCC on AID, NWPU-RESISC45, WHU-RS19, and Optimal-31, and the lowest ERGAS on these datasets and UCMerced LandUse. LSM retains higher SCC on UCMerced LandUse and, jointly with HAT at the reported precision, on DOTA; it also obtains the lowest ERGAS on DOTA. These supplementary RGB-based measures provide additional context

Table 7: ERGAS on six remote-sensing datasets ($\times 4$); lower is better. Best and second-best results are shown in **bold** and underlined, respectively.

Method	AID	DOTA	NWPU	UCM	WHU	Opt.-31
Bicubic	4.1009	3.7422	4.7560	3.9533	3.9313	4.9328
EDSR	3.3820	2.9540	4.0009	3.3719	3.3099	4.0753
RDN	3.3794	2.9378	4.0053	3.3530	3.3236	4.0669
RCAN	3.3598	2.9121	3.9835	3.3353	3.2913	4.0440
SwinIR	3.3517	2.8886	3.9674	3.3240	3.2809	4.0238
HAT	3.3148	2.8618	3.9328	3.2646	3.2493	3.9864
TTST	3.4252	2.9322	4.0448	3.4079	3.3455	4.1169
FAT	3.5008	3.0139	4.1074	3.4610	3.4052	4.1847
HMOE	3.3354	2.8814	3.9589	3.3052	3.2734	4.0193
MambaIR	3.3229	2.8843	3.9436	3.2929	3.2563	4.0009
HAM	3.3398	2.8742	3.9474	3.2628	3.2719	4.0086
MambaIRv2	3.3256	2.8741	3.9520	3.3084	3.2500	3.9994
LSM	3.3080	**2.8603**	3.9284	3.2588	3.2438	3.9810
FoveaMamba	**3.3001**	2.8701	**3.9132**	**3.2536**	**3.2394**	**3.9647**

for the luminance-based PSNR/SSIM comparisons in Table 1, without implying identical rankings across metrics.

D Additional Mechanism Analysis

D.1 Discretization-Step Diagnostics

The mechanism analysis in Fig. 5 uses 1,800 patches across six remote-sensing datasets, with 100 patches per structural category from each dataset. The Low-, Medium-, and High-Structure groups follow the HR-reference partition in Appendix B.1.

For each scan direction, conditioned and unmodulated discretization steps are calculated using the same trained checkpoint, with the latter omitting gate modulation from the step parameterization. Their absolute differences are averaged over feature channels, spatial positions, directions, and network blocks. This analysis measures the trained model's internal step modulation; it is not an ablation of reconstruction performance after removing CGSSM.

Patch-level measurements are aggregated by source image and structural category, with equal weighting across the six datasets. The network-level analysis uses 5,000 bootstrap replicates, resampling source images separately within each dataset and category. The supplementary layer-wise analysis uses 10,000 source-image cluster bootstrap replicates and retains source-image associations across structural categories and network stages.

D.2 Spatial Response Diagnostics

We inspect CGSSM and MHI-Stable responses in three fixed examples: an AID Low-Structure patch, an AID High-Structure patch, and a DOTA High-Structure patch. All maps are extracted from the third FMB of the fourth FRG, allowing direct comparison at a common network location.

CGSSM spatial responses. Figure 6 displays the native context gate and mean absolute discretization-step difference alongside each HR reference and Sobel magnitude.

The maps exhibit spatially varying modulation but do not show consistently edge-concentrated responses. Across all 1,800 patches, descriptive Sobel–modulation analysis at this module yields mean Spearman correlations of -0.349, -0.304, and -0.211 for the Low-, Medium-, and High-Structure groups, respectively. These negative spatial correlations are specific to the sampled network location and should not be conflated with the positive network-aggregated High-minus-Low modulation contrast.

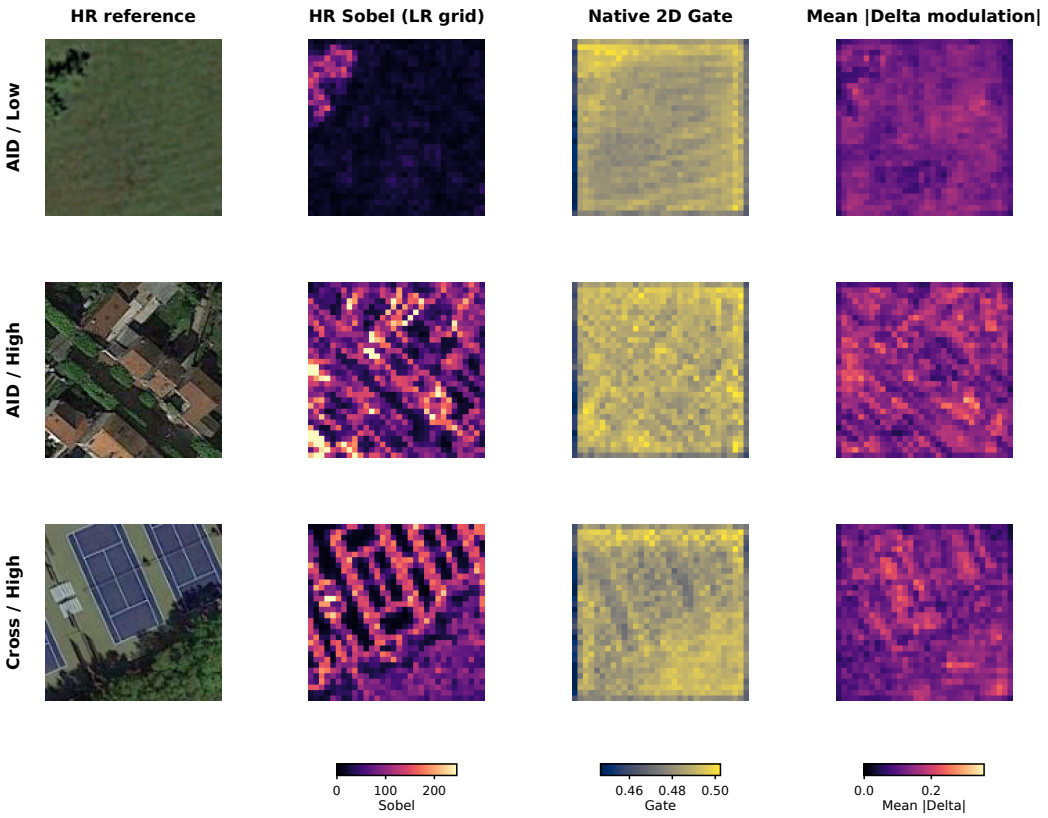


Figure 6: **CGSSM spatial response diagnostics.** HR references, Sobel magnitudes, native context gates, and mean absolute discretization-step differences for three fixed examples. All internal maps use the LR feature grid and shared color scales within each measurement.

In the evaluated implementation, the context gate is serialized in raster order and broadcast to all four scan streams without direction-specific reordering. Consequently, these diagnostics describe the implemented sequence-level modulation and do not establish pixelwise correspondence between the gate and every directional scan.

MHI-Stable spatial responses. Figure 7 compares the magnitude of the internal Laplacian response, the channel-averaged spatial mask, and the magnitude of the directly captured final compensation output.

The internal Laplacian magnitude correlates positively with HR Sobel responses in all three examples. By contrast, the channel-averaged spatial masks vary relatively little across space and correlate weakly with Sobel magnitude. The final compensation nevertheless exhibits nonuniform spatial responses, indicating that its output cannot be inferred from the channel-averaged mask alone.

The Sobel–compensation Spearman correlations are 0.191, 0.097, and 0.403 for AID Low, AID High, and DOTA High, respectively. These descriptive observations characterize the final output at the selected module, but do not isolate the causal contributions of individual MHI-Stable components to reconstruction quality.

D.3 Layer-wise Modulation Analysis

We examine how discretization-step modulation varies across the six Fovea Residual Groups (FRGs). For each patch, the absolute step difference is averaged over the six FMBs within each FRG and aggregated using the protocol described above.

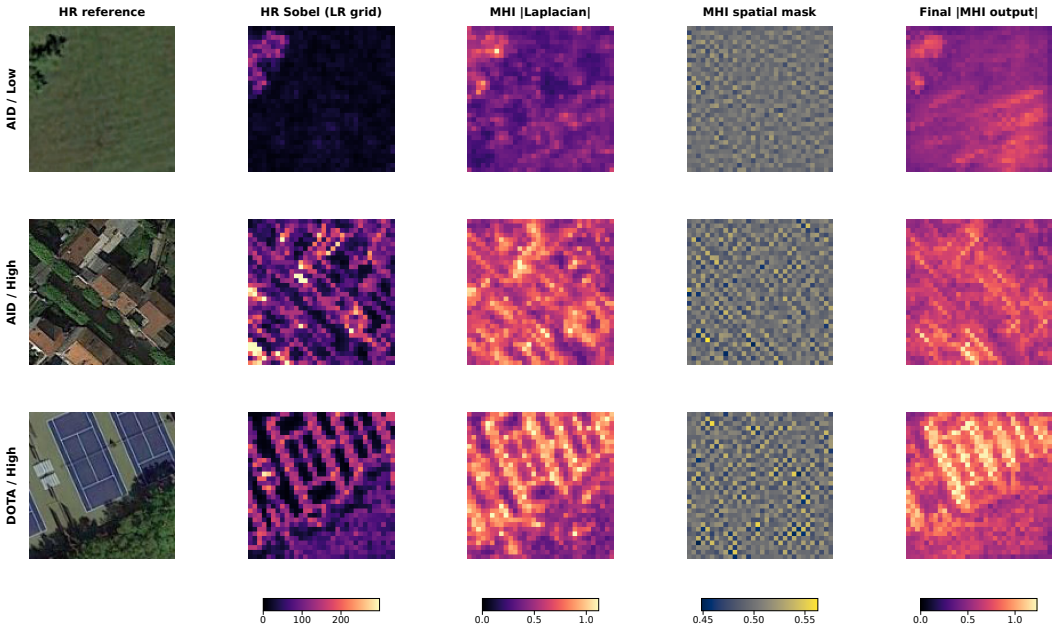


Figure 7: **MHI-Stable spatial response diagnostics.** HR references, Sobel magnitudes, absolute internal Laplacian responses, channel-averaged spatial masks, and final compensation magnitudes for the same examples. Compensation magnitudes average absolute output values across channels; each measurement shares a color scale.

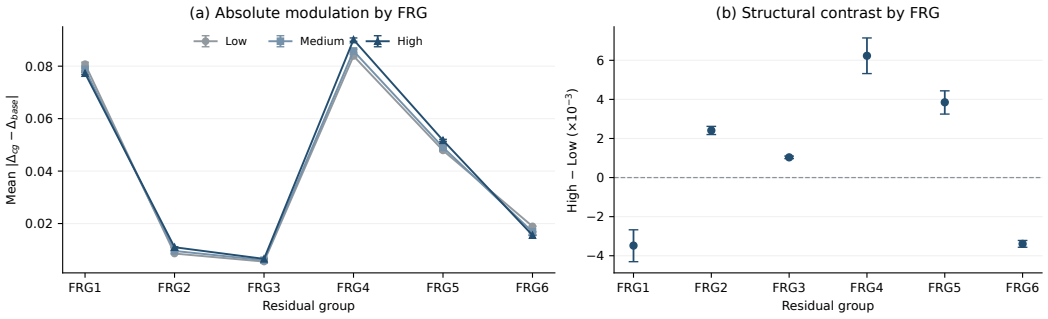


Figure 8: **Layer-wise discretization-step modulation.** (a) Mean absolute modulation across six FRGs for the three structural subsets. (b) High-minus-Low contrasts with 95% source-image cluster bootstrap confidence intervals; negative values indicate greater modulation in Low-Structure regions.

Figure 8 shows stage-dependent modulation. High-minus-Low contrasts are positive in FRGs 2–5 and negative in FRGs 1 and 6. Positive contrasts occur in FRGs 2–5 across all six datasets, whereas FRG6 is negative throughout; FRG1 is negative on five datasets and slightly positive on UCMerced LandUse.

All six FRG-level contrasts have Bonferroni-adjusted 99.17% bootstrap confidence intervals excluding zero. Thus, the greater network-averaged modulation in High-Structure regions coexists with contrasting responses across stages, rather than uniformly stronger modulation in every FRG.

Table 8: Measured inference cost for $\times 4$ SR. Memory denotes peak PyTorch GPU memory allocated during inference. All methods except FAT use 64×64 LR inputs; see protocol qualifications below.

Method	LR Size	Params (M)	Latency (ms)	Memory (MB)
EDSR	64^2	43.090	15.506	301.17
RDN	64^2	22.271	12.943	135.04
RCAN	64^2	15.592	31.198	93.77
SwinIR	64^2	11.900	42.124	108.40
HAT	64^2	20.773	39.396	248.90
TTST	64^2	18.367	71.024	155.42
HMOE	64^2	26.888	43.301	901.13
HAM‡	64^2	25.196	68.758	963.18
LSM	64^2	12.919	63.679	234.38
MambaIR‡	64^2	20.570	45.961	224.23
MambaIRv2‡	64^2	23.051	80.976	198.69
FoveaMamba‡	64^2	27.696	68.216	266.12
FAT†	60^2	19.591	29.017	182.15

Notes. † FAT uses its supported 60×60 LR input and a shortened protocol (3 warm-up iterations, 5 timed passes, and one round); its latency and memory are not directly comparable with standard-protocol measurements. ‡ Original CSV records for these four methods lack complete protocol metadata. Their recorded results are included with this qualification, pending verification against execution logs.

E Computational Efficiency Analysis

We assess the inference cost of FoveaMamba against representative CNN-, Transformer-, and state-space-based SR methods using parameter count, GPU latency, and peak allocated GPU memory. The standard benchmark uses a $3 \times 64 \times 64$ LR input at $\times 4$ magnification, batch size 1, and single-precision inference on an NVIDIA A100-SXM4-80GB GPU. Each method is instantiated using its original implementation and corresponding trained checkpoint. Checkpoint loading is verified by strict parameter-key and tensor-shape matching, and checkpoint hashes are retained for reproducibility.

For runs with complete protocol records, latency is measured with CUDA events after 50 warm-up iterations. We record 200 consecutive forward passes in each of three rounds and report the median of the three round-wise mean latencies. Peak allocated memory is measured separately using PyTorch CUDA memory statistics after warm-up. Both measurements exclude data loading, image saving, and metric computation. The earlier HAM, MambaIR, MambaIRv2, and FoveaMamba records use the intended benchmark configuration but lack complete original CSV protocol metadata; these entries are identified in Table 8.

Table 8 shows the inference-cost trade-offs of the evaluated implementations. FoveaMamba has 27.696M parameters, with recorded latency of 68.216 ms and peak allocated memory of 266.12 MB; MambaIRv2 has 23.051M parameters, 80.976 ms, and 198.69 MB, respectively. Thus, the recorded FoveaMamba run uses more parameters and allocated memory but has a lower reported latency. Because the original records of both models lack complete protocol metadata, the latency difference should be treated as descriptive rather than a verified speed advantage.

Parameter count does not directly predict runtime or allocated memory: feature dimensions, intermediate activations, custom operators, and implementation details also affect inference cost. Moreover, the models use their respective software implementations. The reported measurements therefore describe execution on the specified hardware and input configuration, not implementation-independent computational complexity.

These measurements cover network forward passes only; they exclude preprocessing, tiled or overlapping-window inference, and output postprocessing where applicable. We omit a unified FLOP/MAC table because the available profilers do not consistently account for the custom oper-

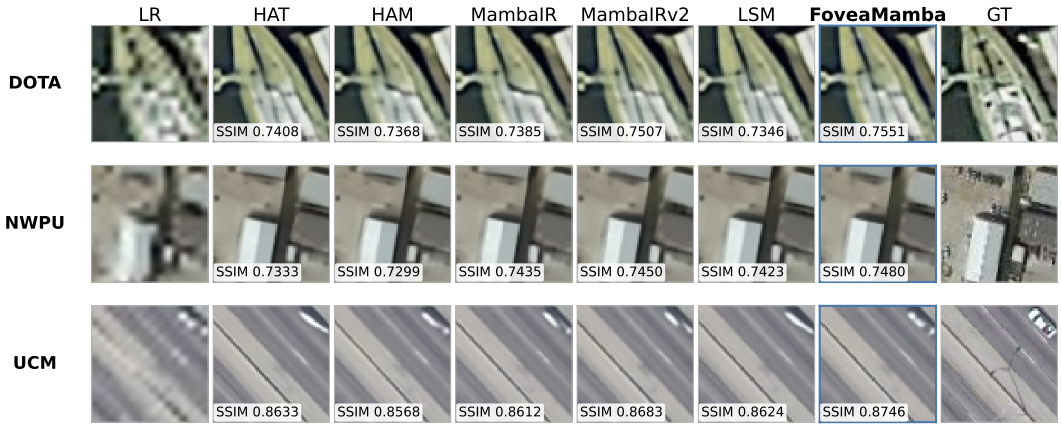


Figure 9: **Additional qualitative comparisons for remote-sensing SR ($\times 4$).** Examples from DOTA, NWPU-RESISC45, and UCMerced LandUse with LR inputs, GT references, and identical reconstruction crops across methods. Local SSIM is reported for each displayed crop.

ators used across all models, particularly selective state-space scanning. Incomplete counts would not provide a comparable measure of total computation.

F Natural-Image Adaptation Details

Datasets and initialization. To evaluate adaptation beyond remote sensing, FoveaMamba, TTST, MambaIRv2, and LSM are initialized from their respective AID-trained checkpoints and fine-tuned for natural-image SR. All methods use the complete 800-image DIV2K training split and its original validation split. The dataset partitions are unchanged, and corresponding LR images are generated by bicubic downsampling at $\times 4$. Adapted checkpoints are evaluated on Set5, Set14, BSD100, Urban100, and Manga109.

Adaptation protocol. Each method undergoes full-parameter fine-tuning for 30K iterations on two NVIDIA A100 GPUs, using four images per GPU. Training pairs comprise 256×256 HR and 64×64 LR crops, augmented by random horizontal flipping and rotation. Adam uses an initial learning rate of 1×10^{-5}, $(\beta_1, \beta_2) = (0.9, 0.99)$, and zero weight decay; MultiStepLR reduces the rate by 0.5 at 15K and 25K iterations. All runs use seed 10. The baselines retain their L1 reconstruction objectives, whereas FoveaMamba retains its original L1 loss with Laplacian supervision ($\lambda_{\text{edge}} = 0.1$). Thus, data, optimization settings, and training budgets are matched, while model-specific training objectives are preserved.

Validation and evaluation. Models are validated on the original DIV2K validation split every 1K iterations, and each model's best validation checkpoint is used for testing. All methods follow the same $\times 4$ degradation and evaluation protocol. PSNR and SSIM are computed on luminance with a four-pixel border exclusion, and the five-benchmark results are reported in Table 3.

G Additional Qualitative Results

Figure 9 supplements the main visual comparisons in Fig. 3 with examples from DOTA, NWPU-RESISC45, and UCMerced LandUse. All methods are evaluated on identical image regions, and local SSIM is computed on the displayed reconstruction crops. These examples illustrate reconstruction behavior across additional remote-sensing scenes without replacing the dataset-level quantitative comparisons.