Transformer-Progressive Mamba Network for Lightweight Image Super-Resolution

Sichen Guo, Wenjie Li, Yuanyang Liu, Guangwei Gao, *Senior Member, IEEE*, Jian Yang, *Member, IEEE*, and Chia-Wen Lin, *Fellow, IEEE*

***Abstract*—Recently, Mamba-based super-resolution (SR) methods have demonstrated the ability to capture global receptive fields with linear complexity, addressing the quadratic computational cost of Transformer-based SR approaches. However, existing Mamba-based methods lack fine-grained transitions across different modeling scales, which limits the efficiency of feature representation. In this paper, we propose T-PMambaSR, a lightweight SR framework that integrates window-based self-attention with Progressive Mamba. By enabling interactions among receptive fields of different scales, our method establishes a fine-grained modeling paradigm that progressively enhances feature representation without introducing additional computational cost. Furthermore, we introduce an Adaptive High-Frequency Refinement Module (AHFRM) to recover high-frequency details lost during Transformer and Mamba processing. Extensive experiments demonstrate that T-PMambaSR progressively enhances the model's receptive field and expressiveness, achieving competitive performance with recent Transformer- or Mamba-based methods while incurring lower computational cost. The code is available at https://github.com/IVIPLab/T-PMambaSR.**

***Index Terms*—Lightweight super-resolution, Mamba-based, receptive field.**

Received 6 November 2025; revised 23 April 2026 and 17 June 2026; accepted 26 July 2026. Date of publication 12 August 2026; date of current version 14 August 2026. This work was supported in part by the Foundation of the State Key Laboratory of Integrated Services Networks of Xidian University under Grant ISN27-4 and in part by the National Natural Science Foundation of China under Grant U24A20330 and Grant 62361166670. The associate editor coordinating the review of this article and approving it for publication was Dr. Abd-Krim Seghouane. *(Sichen Guo and Wenjie Li contributed equally to this work.) (Corresponding author: Guangwei Gao.)*

Sichen Guo and Yuanyang Liu are with the Bell Honors School, Nanjing University of Posts and Telecommunications, Nanjing 210023, China, and also with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: q22010218@njupt.edu.cn; q22010108@njupt.edu.cn).

Wenjie Li is with the Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100080, China (e-mail: lewj2408@gmail.com).

Guangwei Gao is with the PCA Laboratory, Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information, Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China, and also with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: gwgao@njust.edu.cn).

Jian Yang is with the PCA Laboratory, Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information, Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: csjyang@njust.edu.cn).

Chia-Wen Lin is with the Department of Electrical Engineering and the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 300044, Taiwan (e-mail: cwlin@ee.nthu.edu.tw).

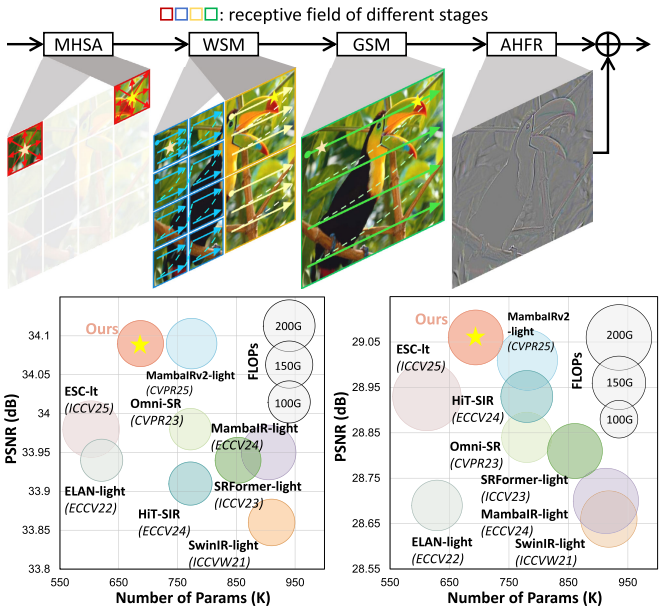
Digital Object Identifier 10.1109/TIP.2026.3721353

I. Introduction

LIGHTWEIGHT image super-resolution (SR) [1] aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) input with limited computational cost, and has broad applications in image processing [2], video enhancement [3], and security surveillance [4]. Early CNN-based methods [5], [6] are constrained by their inherently local receptive fields. Even with increased depth, CNNs still struggle to capture long-range dependencies, which are critical for accurately restoring global structures and fine textures. With the development of Vision Transformers [7], Transformer-based SR methods [8], [9], [10], [11], [12] have addressed this limitation by leveraging self-attention to model long-range dependencies. However, the quadratic computational complexity of the above Transformer-based SR methods makes them prohibitively expensive for lightweight SR scenarios.

Mamba-based lightweight SR methods [15], [16] have shown that state space models (SSMs) [17] can effectively capture global context with a complexity linear to the spatial resolution, surpassing Transformer-based baselines [8], [10], [12]. However, existing Mamba-based approaches [15], [16] often integrated components with different receptive fields in a coarse-grained manner. For instance, MambaIRv2 [16] employed window attention and SSM modules as separate stages, resulting in hierarchical discontinuity: the model transitions abruptly from the fine-grained local details captured by window attention to the broad global context modeled by Mamba. This lacks a progressive mechanism for cross-scale feature fusion, limiting the ability to synergize local precision with global scope and ultimately leading to additional computational overhead.

To address this challenge, we introduce Transformer-Progressive Mamba (T-PMambaSR), a lightweight SR architecture that effectively combines multi-level feature modeling. As illustrated at the top of Fig. 1, the design integrates standard Window Multi-Head Self-Attention (W-MHSA) with two specialized state-space modules. First, the Window Scan Mamba Layer (WSML) enables both intra-window modeling and long-range inter-window interactions through multi-scale window scanning, progressively expanding the modeling windows and enriching the feature representation with multi-scale information. Subsequently, the Global Scan Mamba Layer (GSML) captures global image features from multiple directions via a computationally efficient multi-head mechanism, maintaining high performance without degradation. The MHSA → WSML



(1) Trade-off on ×2 Set14 [13] test set. (2) Trade-off on ×3 Urban100 [14] test set.

Fig. 1. (**Top**): Our design rationale is based on progressively exploiting internal interactions within Window Multi-Head Self-Attention (W-MHSA), combined with a Window Scan Mamba (WSM) and a Global Scan Mamba (GSM) operation. This hierarchical structure facilitates the gradual expansion of receptive fields, ensuring comprehensive information exchange both within and across windows. (**Bottom**): Leveraging our design, our method strikes an optimal balance across Params, FLOPs, and PSNR, surpassing existing Transformer- and Mamba-based methods.

→ GSML framework forms a hierarchical network that incrementally expands the receptive field, effectively overcoming the locality constraints of window attention and ensuring smooth information flow across local, regional, and global scales.

Moreover, to mitigate the loss of high-frequency details in SR, we introduce an Adaptive High-Frequency Refinement Module (AHFRM). Prior studies [18], [19] have demonstrated that although Transformer- and Mamba-based architectures excel at capturing low-frequency information, they are inherently limited in modeling high-frequency representations. As features propagate through these low-pass-filter-like modules, high-frequency information is gradually attenuated. To address this issue, AHFRM employs a high-frequency filter to extract high-frequency features from both the input and output of the Transformer–Mamba block. The preserved pre-processed features are then used to guide the recovery of degraded post-processed content. By embedding high-frequency recovery into the network, our T-PMambaSR enhances the sharpness of SR images without compromising efficiency. Based on the above model architectures, as illustrated at the bottom of Fig. 1, our T-PMambaSR achieves highly efficient SR, offering favorable trade-offs among model size and reconstruction quality.

In summary, the contributions of this paper are as follows:

- Our T-PMambaSR gradually expands the receptive field, enabling hierarchical feature modeling to transition from local to global in a fine-grained manner, thereby improving the efficiency in capturing features at different levels.
- Our AHFRM counteracts the low-pass filtering tendencies inherent in both Transformer and Mamba modules, specifically recovering high-frequency contexts and edges that are lost to low-pass filtering.
- Extensive experiments show that our T-PMambaSR achieves competitive performance compared with most existing lightweight Transformer- and Mamba-based SR methods, both qualitatively and quantitatively, with less computational cost.

II. Related Work

A. Lightweight SR Methods

The development of lightweight SR [20] has evolved through several architectural paradigms. Early efforts were dominated by Convolutional Neural Networks (CNNs), where foundational works established key principles for efficiency. For instance, FSRCNN [21] pioneered processing features in the LR space to reduce computational cost, while FDIWN [5] further refined network design with sophisticated block-level techniques such as information distillation. These models excelled at local feature representation but were inherently limited by their local receptive fields. To capture long-range dependencies, Vision Transformers were adapted for SR. For example, SwinIR [8] introduced a highly effective window-based approach to manage the quadratic complexity of self-attention. To overcome the limited receptive field of fixed windows in window-based SR methods [22], [23], SRFormer [10] introduced permuted self-attention to enable larger window sizes without prohibitive cost, while HiT-SIR [11] employed a hierarchical design with progressively expanding windows to capture multi-scale context. However, to mitigate the quadratic computational complexity associated with increasing window sizes, these approaches often compromise spatial structural information: HiT-SIR relies on lossy spatial downsampling, while SRFormer alters spatial structures via its permuted mechanism. Furthermore, these Transformer-based methods generally lack effective long-range inter-window interactions.

More recently, works based on State Space Models (SSMs) [15] have emerged as a powerful alternative for SR, offering global receptive fields with linear complexity. Recent works like MambaIRv2 [16] demonstrated the powerful performance of the hybrid Transformer-Mamba architecture. Similarly, SRMamba-T [24] explored Mamba-Transformer hybridization, while HiMamba [25] introduced hierarchical state-space modeling. However, these architectures typically stack modules in a coarse-grained manner and lack mechanisms to smoothly bridge discrete features at different scales, which can lead to abrupt receptive field transitions.

In contrast, our T-PMambaSR addresses the limitations of both paradigms by establishing a strictly **fine-grained transition** that progresses smoothly from local to global scales without any spatial downsampling. By altering window-based selective scan paths via our WSML and GSML modules rather than enlarging attention windows, we achieve efficient inter-window communication and progressive receptive field expansion without incurring additional computational overhead.

B. Frequency-Based Restoration Methods

Another line of research explicitly targets high-frequency restoration for fine textures that are often lost in spatial backbones. For instance, CRAFT [18] enhanced high-frequency features to compensate for the Transformer model's bias toward low-frequency components. ESRT [26] incorporated high-frequency filtering to preserve texture details, while WFEN [27] reduced texture loss during downsampling by separating high-frequency features. AFFNet [28] used adaptive frequency filters to dynamically mix tokens in the frequency domain, improving long-range context recovery. FourierSR [29] modulated the frequency space using the Fourier frequency domain theorem to expand the receptive field of super-resolution. FDSR [30] reconstructed images step-by-step using Fourier-based frequency band separation to restore high-frequency details. DMNet [12] utilized the wavelet transform to explore the relationship between high-frequency and other frequency domain features in the SR process. CFSR [31] improved image reconstruction by integrating gradient-based edge priors into its feed-forward network to preserve and strengthen high-frequency information. PoolNet [32] employed average and max pooling to implicitly separate and modulate low- and high-frequency features, using the enhanced high-frequency cues to refine spatial details. CAN [33] strengthened robustness through frequency-domain augmentation that adjusts high-frequency components while preserving key frequency details. Furthermore, while existing approaches often rely on complex frequency-domain transformations or implicit feature separation, it is crucial to recognize that the successive processing in modern Transformer and Mamba modules inherently causes high-frequency degradation due to their low-pass filtering tendencies.

To explicitly counteract this issue, rather than implicitly estimating lost details, our proposed AHFRM introduces a direct reference-guided restoration mechanism. It leverages pristine high-frequency information extracted from an unprocessed feature map to explicitly guide the restoration of attenuated high-frequency content in deeply processed features, thereby ensuring robust and sharp texture reconstruction.

III. Proposed Method

In this section, we start by introducing the preliminaries of state space models before presenting a general architecture of our T-PMambaSR. Then, we examine the architectural details of its core components: the Window Interaction State Space Module (WISSM) and the Multi-head Global State Space Module (MGSSM). Finally, we discuss the design of our Adaptive High-Frequency Refinement Module (AHFRM).

A. Preliminaries

State Space Models (SSMs) [17] are a class of models that map a 1-D input sequence $x(t) \in \mathbb{R}$ to an N-D latent state $h(t) \in \mathbb{R}^N$ before projecting it to a 1-D output $y(t) \in \mathbb{R}$. Originating from classical control theory, they are particularly adept at modeling long-range dependencies. Specifically, a continuous-time SSM can be formulated as a linear Ordinary Differential Equation (ODE):

$$\begin{aligned} \frac{dh(t)}{dt} &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t) + \mathbf{D}x(t), \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the state transition matrix, which describes how the internal state evolves, and $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times N}$, $\mathbf{D} \in \mathbb{R}$ are projection matrices that map the input to state, the state to output, and the input to output, respectively.

To be applied to discrete sequential data, such as text or flattened image patches, this continuous system must be discretized. A common method is the zero-order hold (ZOH), which transforms the continuous parameters $(\mathbf{A}, \mathbf{B})$ into discrete counterparts $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ using a timescale parameter Δ. The discretized SSM is then formulated as

$$\begin{aligned} h_k &= \bar{\mathbf{A}}h_{k-1} + \bar{\mathbf{B}}x_k, \\ y_k &= \mathbf{C}h_k + \mathbf{D}x_k, \end{aligned} \tag{2}$$

where $\bar{\mathbf{A}} = \exp(\Delta\mathbf{A})$ and $\bar{\mathbf{B}} = (\exp(\Delta\mathbf{A}) - \mathbf{I})\mathbf{A}^{-1}\mathbf{B}$. This discrete formulation can be unrolled and computed efficiently as a global convolution, allowing it to be integrated into deep neural networks.

While the Structured State Space Sequence Model (S4) made SSMs computationally efficient by structuring the $\mathbf{A}$ matrix, its parameters remain static and input-invariant. Mamba [17] enhances this by introducing a selection mechanism that renders key parameters ($\mathbf{B}$, $\mathbf{C}$, Δ) input-dependent. This enables Mamba to selectively modulate information flow based on content, achieving state-of-the-art performance with linear complexity. Overall, our work aims to leverage this efficient and content-aware modeling for SR.

B. General Architecture

As shown in Fig. 2, given a low-resolution (LR) input image $\boldsymbol{I}_{LR} \in \mathbb{R}^{H\times W\times 3}$, an initial 3×3 convolutional layer first maps the input into a shallow feature $\boldsymbol{F}_S \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represent the height, width, and channel dimensions of the input image. This shallow feature $\boldsymbol{F}_S$ is then fed into several Transformer-Progressive Mamba (T-PM) Groups to extract deep features $\boldsymbol{F}_D \in \mathbb{R}^{H\times W\times C}$. Each T-PM Group consists of N stacked Transformer-Window Scan Mamba (T-WSM) Blocks that progressively refine the features. A Global Scan Mamba Layer (GSML) is then applied to expand the receptive field for global modeling, followed by an Adaptive High-Frequency Refinement Module (AHFRM) to restore high-frequency details. An additional convolutional layer is applied at the end of each group for further feature enhancement. Finally, a long skip connection combines the shallow and deep features through element-wise addition $\boldsymbol{F}_R = \boldsymbol{F}_D + \boldsymbol{F}_S$, and the resulting fused feature map $\boldsymbol{F}_R$ is upsampled to generate the HR output $\boldsymbol{I}_{HR}$.

C. Window Interaction State Space Module

As a core component of our Window Scan Mamba Layer (WSML) designed to facilitate long-range communication across windows and progressively expand the receptive field,

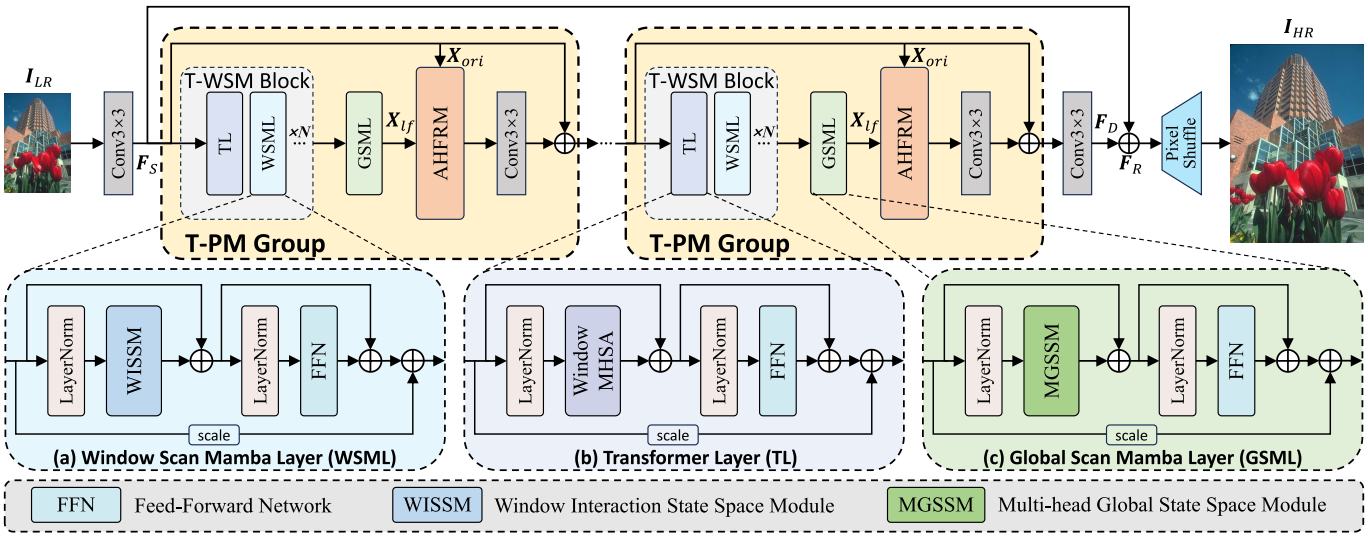


Fig. 2. The network architecture of our T-PMambaSR, as well as the framework of the (a) Window Scan Mamba Layer (WSML), (b) Transformer Layer (TL), and (c) Global Scan Mamba Layer (GSML).

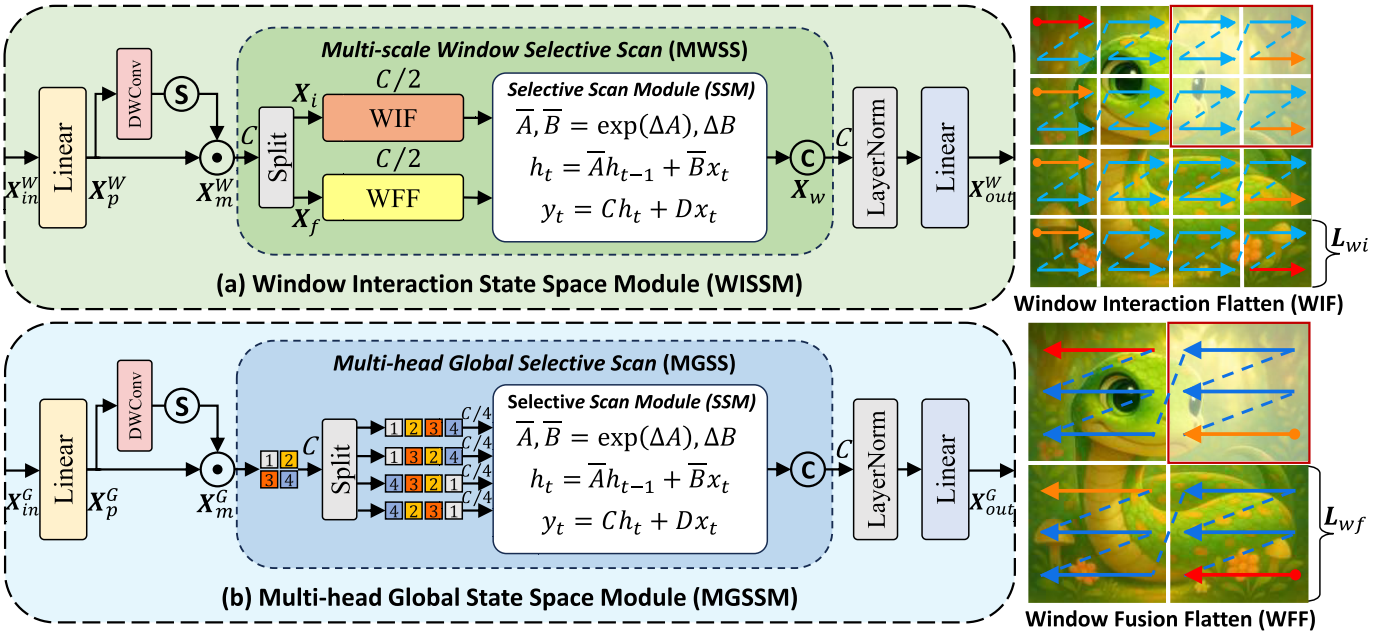


Fig. 3. The illustration of our (a) Window Interaction State Space Module (WISSM) with its two flatten mechanisms, Window Interaction / Fusion Flatten (WIF and WFF), and (b) Multi-head Global State Space Module (MGSSM).

we introduce the Window Interaction State Space Module (WISSM). As shown in Fig. 3 (a), WISSM first projects the input feature $\boldsymbol{X}_{in}^{W} \in \mathbb{R}^{H \times W \times C}$ with a linear layer. The projected feature is then divided into two parallel streams. The first stream acts directly as the content feature, while the second is passed through a gating mechanism to generate a spatial gate. This gate weight is then multiplied element-wise with the content feature to adaptively select and enhance important information while suppressing noise:

$$\begin{aligned} \boldsymbol{X}_{p}^{W} &= \text{Linear}(\boldsymbol{X}_{in}^{W}), \\ \boldsymbol{X}_{m}^{W} &= \boldsymbol{X}_{p}^{W} \odot \sigma(\text{DWConv}(\boldsymbol{X}_{p}^{W})), \end{aligned} \tag{3}$$

where $\odot$ denotes element-wise multiplication and σ is the sigmoid function. Subsequently, the feature $\boldsymbol{X}_{m}^{W}$ is processed by our Multi-scale Window Selective Scan (MWSS) to establish multi-scale and long-range window-based dependencies. The process can be described as:

$$\boldsymbol{X}_{out}^{W} = \text{Linear}(\text{LN}(\text{MWSS}(\boldsymbol{X}_{m}^{W}))). \tag{4}$$

Multi-scale Window Selective Scan (MWSS). Prior works based on SSM [15], [34], [35] typically employ the 2D Selective Scan for global modeling. However, in hybrid Transformer-Mamba architectures, this creates an abrupt leap from the local receptive field of window attention to the global scan, hindering a smooth multi-scale feature transition.

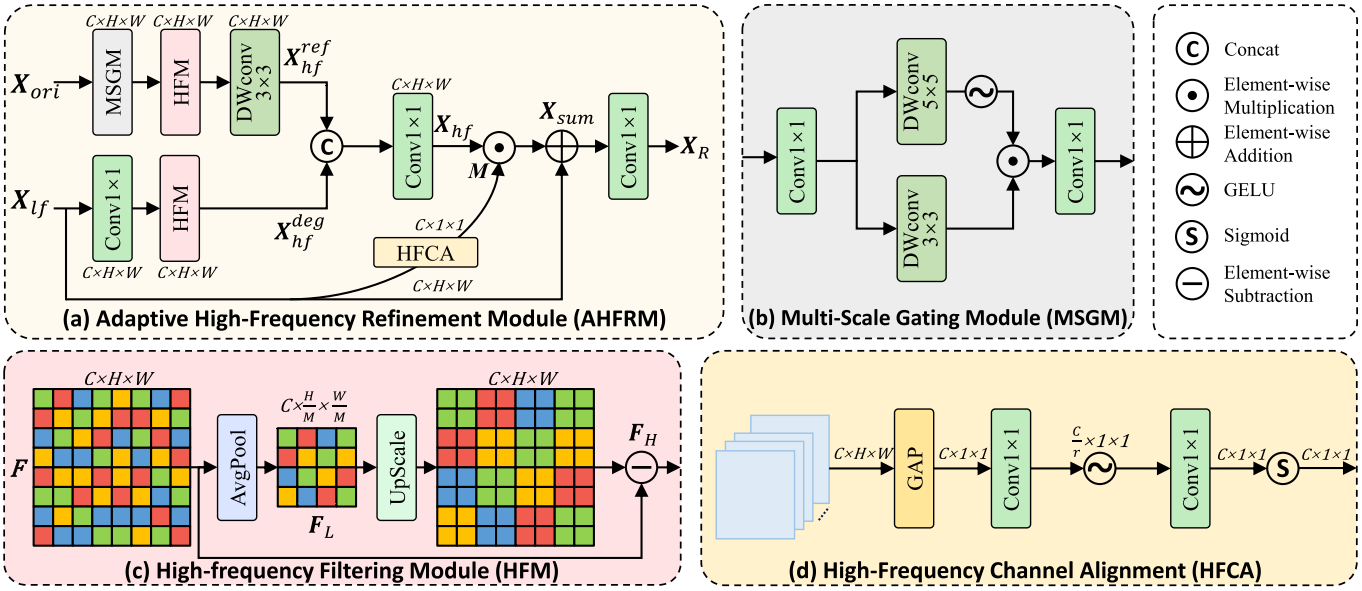


Fig. 4. The architecture of our (a) Adaptive High-Frequency Refinement Module (AHFRM), (b) Multi-Scale Gating Module (MSGM), (c) High-frequency Filtering Module (HFM), and (d) High-Frequency Channel Alignment (HFCA).

Moreover, both window attention and the standard 2D selective scan are ill-suited for long-range interactions between discrete, window-shaped feature regions of specific size, which is vital for maintaining coherence in the restoration of global features. To this end, we propose MWSS to address these challenges.

Differing from the traditional 2D selective scan paradigm of feature replication, multi-directional scanning, and summation, our MWSS operation enhances efficiency by splitting the feature tensor into two halves along the channel dimension. These two parts are first flattened separately by a Window Interaction Flatten (WIF) and a Window Fusion Flatten (WFF) operation. As depicted in Fig. 3, WIF flattens the input feature with its scanning window size $\boldsymbol{L}_{wi}$ set to 16, consistent with the window self-attention, to facilitate long-range interactions between windows. Meanwhile, WFF sets $\boldsymbol{L}_{wf}$ to 32 to serve as a smooth transition from local to global receptive fields. Within each MWSS, WIF and WFF flatten feature maps on the same axis (*e.g.*, both horizontal or vertical) but in opposite directions, while the flattening axes and directions alternate across different MWSS to ensure comprehensive coverage of all four raster-scan orders (*i.e.*, top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right). Subsequently, these two flattened feature sets are processed independently by the Selective Scan Module, after which their resulting outputs are concatenated:

$$\begin{aligned} \boldsymbol{X}_i, \boldsymbol{X}_f &= \text{Split}(\boldsymbol{X}_m^W), \\ \boldsymbol{X}_w &= \text{Concat}\begin{pmatrix} \text{SSM}(\text{WIF}(\boldsymbol{X}_i)), \\ \text{SSM}(\text{WFF}(\boldsymbol{X}_f)) \end{pmatrix}. \end{aligned} \tag{5}$$

This parallel design achieves efficient long-range window interaction while establishing a smooth multi-scale receptive field transition.

D. Multi-Head Global State Space Module

After a sequence of window self-attention and WISSM dedicated to fine-grained, multi-scale local modeling, as shown in Fig. 2, we introduce the Multi-head Global State Space Module (MGSSM) within our GSML, where our purpose is to further expand the receptive field to a global scale, ensuring that our feature representation captures both fine local details and long-range structural coherence.

As depicted in Fig. 3 (b), our MGSSM is structurally analogous to WISSM, differing only in its utilization of the Multi-head Global Selective Scan (MGSS). The process is:

$$\begin{aligned} \boldsymbol{X}_p^G &= \text{Linear}(\boldsymbol{X}_{in}^G), \\ \boldsymbol{X}_m^G &= \boldsymbol{X}_p^G \odot \sigma(\text{DWConv}(\boldsymbol{X}_p^G)), \\ \boldsymbol{X}_{out}^G &= \text{Linear}(\text{LN}(\text{MGSS}(\boldsymbol{X}_m^G))), \end{aligned} \tag{6}$$

where $\odot$ denotes the element-wise multiplication.

Multi-head Global Selective Scan (MGSS). Adopting a paradigm similar to MWSS, we process the input tensor by first splitting it into four parts along the channel dimension, each with C/4 channels. Each part is then flattened and scanned in parallel along one of the four raster-scan orders (*i.e.*, top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right). The four resulting output features are subsequently concatenated. This design reduces computational complexity while achieving performance that is nearly identical to that of a version using the traditional 2D selective scan, as validated in Table IV.

E. Adaptive High-Frequency Refinement Module

We propose an Adaptive High-Frequency Refinement Module (AHFRM) that extracts pristine high-frequency information from an initial, unprocessed feature as a reference. This reference guides the adaptive restoration of high-frequency information that has been attenuated within the feature after

undergoing low-frequency-biased modeling by a series of Transformer and Mamba operations.

Inspired by the design of extracting and enhancing high-frequency features in ESRT [26], as shown in Fig. 4 (c), for a given input $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$, we first apply a pooling layer to reduce its spatial resolution by half, utilizing the low-frequency extraction properties of pooling to obtain its low-frequency components, $\boldsymbol{F}_L \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$. This low-frequency feature is then upsampled back to the original resolution using bilinear interpolation. Finally, the upsampled low-frequency feature is subtracted from the original feature $\boldsymbol{F}$, and the resulting difference constitutes the desired high-frequency feature $\boldsymbol{F}_H \in \mathbb{R}^{C\times H\times W}$. This process can be formulated as:

$$\begin{aligned} \boldsymbol{F}_L &= \text{Pooling}(\boldsymbol{F}), \\ \boldsymbol{F}_H &= \boldsymbol{F} - \text{Upsample}(\boldsymbol{F}_L). \end{aligned} \tag{7}$$

As depicted in Fig. 4 (a), each Adaptive High-Frequency Refinement Module (AHFRM) receives two inputs: the pristine, unprocessed feature $\boldsymbol{X}_{ori} \in \mathbb{R}^{C\times H\times W}$ with its high-frequency information intact, and the deeply processed, low-frequency-biased feature $\boldsymbol{X}_{lf} \in \mathbb{R}^{C\times H\times W}$. To obtain the reference high-frequency information $\boldsymbol{X}_{hf}^{ref} \in \mathbb{R}^{C\times H\times W}$, $\boldsymbol{X}_{ori}$ is first processed by our Multi-Scale Gating Module (MSGM) for selective multi-scale extraction, followed by an HFM and a 3×3 depth-wise convolutional layer. In parallel, $\boldsymbol{X}_{lf}$ is passed through a 1×1 convolutional layer and an HFM to extract the degraded high-frequency information $\boldsymbol{X}_{hf}^{deg} \in \mathbb{R}^{C\times H\times W}$ targeted for adaptive restoration:

$$\begin{aligned} \boldsymbol{X}_{hf}^{ref} &= \text{DWConv}_{3\times 3}(\text{HFM}(\text{MSGM}(\boldsymbol{X}_{ori}))), \\ \boldsymbol{X}_{hf}^{deg} &= \text{HFM}(\text{Conv}_{1\times 1}(\boldsymbol{X}_{lf})). \end{aligned} \tag{8}$$

Then, the two extracted high-frequency features are concatenated and fused by a 1×1 convolution, which projects them back to the original channel dimension to form a refined high-frequency feature $\boldsymbol{X}_{hf} \in \mathbb{R}^{C\times H\times W}$. In parallel, we introduce a High-Frequency Channel Alignment (HFCA). As shown in Fig. 4 (d), HFCA generates a channel attention map $\boldsymbol{M} \in \mathbb{R}^{C\times 1\times 1}$ from the high-frequency degraded feature $\boldsymbol{X}_{lf} \in \mathbb{R}^{C\times H\times W}$ and uses it to apply channel-wise weights to the refined high-frequency feature. Subsequently, this weighted high-frequency feature is fused with the low-frequency-biased feature via a residual connection and a 1×1 convolution to generate the refined feature $\boldsymbol{X}_R$. This process is:

$$\begin{aligned} \boldsymbol{X}_{hf} &= \text{Conv}_{1\times 1}(\text{Concat}(\boldsymbol{X}_{hf}^{ref}, \boldsymbol{X}_{hf}^{deg})), \\ \boldsymbol{X}_R &= \text{Conv}_{1\times 1}(\boldsymbol{X}_{hf} \odot \text{HFCA}(\boldsymbol{X}_{lf}) + \boldsymbol{X}_{lf}). \end{aligned} \tag{9}$$

Based on these designs, our AHFRM supplements the high-frequency information weakened by the Transformer or Mamba.

F. Computational Complexity Analysis

For an input feature map of size $H\times W\times C$, standard Window-MHSA (W-MHSA) with a window size of $M\times M$ exhibits a computational complexity of $\mathcal{O}(HWC^2+HWM^2C)$. Although this scales linearly with the overall spatial resolution $H\times W$, it remains quadratic with respect to the window size. In contrast, the selective scan component in our proposed Mamba-based modules (WISSM and MGSSM) processes features with a complexity of $\mathcal{O}(HWCN)$, where N denotes the dimension of the SSM state. By adopting different scanning paths alone, MWSS and MGSS achieve efficient receptive field expansion and long-range inter-window interactions without increasing the computational cost. This design preserves strict linearity and avoids the quadratic bottleneck caused by enlarging window sizes, thereby significantly reducing the computational burden.

IV. Experiments

A. Experimental Settings

1) Datasets and Metrics: Following prior works [10], [11], [12], we train our model on the DIV2K [44] dataset, which contains 800 pairs of images, and evaluate it on five benchmark datasets: Set5 [36], Set14 [13], BSDS100 [37], Urban100 [14], and Manga109 [38]. The LR images are obtained from their corresponding HR counterparts using bicubic degradation. We conduct experiments under three upscaling factors: $\times 2$, $\times 3$, and $\times 4$. To further validate our method, we train and test our model on the RealSRv3 [43] dataset. It contains real-world LR-HR image pairs obtained by adjusting the focal length of a camera on the same scene and subsequently aligning the images. To evaluate performance, we use PSNR and SSIM [45], both calculated on the Y channel in the YCbCr space. For real-world SR tasks, we also utilize the LPIPS [46] metric, as it aligns better with human subjective visual perception compared to traditional pixel-wise metrics.

2) Implementation Details: Consistent with previous methods, we perform data augmentation with random horizontal flips and rotations of 90°, 180°, and 270°. Our proposed model is constructed with 4 Transformer-Progressive Mamba Groups, and within each group, we set the number of Transformer-Window Scan Mamba Blocks to $N = 4$ and the channel width to 48. During training, we use a batch size of 32, and the patch size is set to 64×64. The model is trained for a total of 600K iterations using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. We optimize the network using L1 loss with a weight of 1. The initial learning rate is set to 2×10^{-4} and is halved at [300K, 450K, 525K, 575K] iterations. All experiments are implemented using the PyTorch framework and conducted on two NVIDIA GeForce RTX 4090 GPUs.

B. Comparisons With State-of-the-Art Methods

To evaluate the performance of our model, we compare it with several state-of-the-art lightweight SR approaches, including Transformer-based methods: SwinIR-light [8], ELAN-light [9], SRFormer-light [10], Omni-SR [39], HiT-SIR [11], and ESC-lt [40]; Mamba-based methods: CNMC [41], MambaIR-light [15] and MaIR-Small [42]; as well as the hybrid Transformer-Mamba approach MambaIRv2-light [16].

1) Quantitative Comparisons: Table I presents the quantitative comparison of our T-PMambaSR with existing methods at scale factors of $\times 2$, $\times 3$, and $\times 4$. Our model outperforms existing methods with architectures based solely on Transformer

TABLE I

QUANTITATIVE EVALUATION OF LIGHTWEIGHT SR METHODS, INCLUDING TRANSFORMER-BASED AND MAMBA-BASED METHODS. FLOPS ARE MEASURED ON UPSAMPLED IMAGES OF SIZE 1280 × 720. THE BEST AND SECOND-BEST ARE BOLD AND UNDERLINED

Methods	Scale	Params↓	FLOPs↓	Set5 [36]		Set14 [13]		BSDS100 [37]		Urban100 [14]		Manga109 [38]	
				PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SwinIR-light [8]	×2	910K	244.4G	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
ELAN-light [9]	×2	621K	201.3G	38.17	0.9611	33.94	0.9207	32.30	0.9012	32.76	0.9340	39.11	0.9782
SRFormer-light [10]	×2	853K	236.2G	38.23	0.9613	33.94	0.9209	**32.36**	0.9019	32.91	0.9353	39.28	0.9785
Omni-SR [39]	×2	772K	194.5G	38.22	0.9613	33.98	0.9210	**32.36**	**0.9020**	33.05	0.9363	39.28	0.9784
HiT-SIR [11]	×2	772K	209.9G	38.22	0.9613	33.91	0.9213	32.35	0.9019	33.02	0.9365	**39.38**	0.9782
ESC-lt [40]	×2	603K	359.4G	38.24	0.9615	33.98	0.9211	32.35	**0.9020**	33.05	0.9363	39.33	**0.9786**
CNMC [41]	×2	769K	—	38.14	0.9610	33.84	0.9203	32.29	0.9011	32.74	0.9337	39.13	0.9775
MambaIR-light [15]	×2	905K	334.2G	38.13	0.9610	33.95	0.9208	32.31	0.9013	32.85	0.9349	39.20	0.9782
MaIR-Small [42]	×2	1,355K	542.0G	38.20	0.9611	33.91	0.9209	32.34	0.9016	32.97	0.9359	39.32	0.9779
MambaIRv2-light [16]	×2	774K	286.3G	**38.26**	0.9615	**34.09**	**0.9221**	**32.36**	0.9019	**33.26**	**0.9378**	39.35	0.9785
T-PMambaSR (ours)	×2	687K	238.1G	**38.26**	**0.9616**	**34.09**	0.9219	**32.36**	0.9019	33.17	0.9371	39.36	0.9782
SwinIR-light [8]	×3	918K	110.8G	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624	33.98	0.9478
ELAN-light [9]	×3	629K	89.5G	34.61	0.9288	30.55	0.8463	29.21	0.8081	28.69	0.8624	34.00	0.9478
SRFormer-light [10]	×3	861K	104.8G	34.67	0.9296	30.57	0.8469	29.26	0.8099	28.81	0.8655	34.19	0.9489
Omni-SR [39]	×3	780K	88.4G	34.70	0.9294	30.57	0.8469	29.28	0.8094	28.84	0.8656	34.22	0.9487
HiT-SIR [11]	×3	780K	94.2G	34.72	0.9298	30.62	0.8474	29.27	0.8101	28.93	0.8673	34.40	0.9496
ESC-lt [40]	×3	612K	162.8G	34.61	0.9295	30.52	0.8475	29.26	0.8102	28.93	0.8679	34.33	0.9495
CNMC [41]	×3	775K	—	34.57	0.9285	30.52	0.8456	29.19	0.8080	28.61	0.8614	34.02	0.9471
MambaIR-light [15]	×3	913K	148.5G	34.63	0.9288	30.54	0.8459	29.23	0.8084	28.70	0.8631	34.12	0.9479
MaIR-Small [42]	×3	1,363K	241.4G	34.75	0.9300	30.63	0.8479	**29.29**	0.8103	28.92	0.8676	**34.46**	0.9497
MambaIRv2-light [16]	×3	781K	126.7G	34.71	0.9298	**30.68**	0.8483	29.26	0.8098	29.01	0.8689	34.41	0.9497
T-PMambaSR (ours)	×3	694K	105.6G	**34.80**	**0.9304**	30.65	**0.8485**	**29.29**	**0.8104**	**29.06**	**0.8699**	34.45	**0.9500**
SwinIR-light [8]	×4	930K	63.6G	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
ELAN-light [9]	×4	640K	53.7G	32.43	0.8975	28.78	0.7858	27.69	0.7406	26.54	0.7982	30.92	0.9150
SRFormer-light [10]	×4	873K	62.8G	32.51	0.8988	28.82	0.7872	27.73	0.7422	26.67	0.8032	31.17	0.9165
Omni-SR [39]	×4	792K	50.9G	32.49	0.8988	28.78	0.7859	27.71	0.7415	26.64	0.8018	31.02	0.9151
HiT-SIR [11]	×4	792K	53.8G	32.51	0.8991	28.84	0.7873	27.73	0.7424	26.71	0.8045	31.23	0.9176
ESC-lt [40]	×4	624K	91.0G	32.52	0.8995	28.87	0.7878	27.72	0.7423	26.76	0.8058	31.26	0.9173
CNMC [41]	×4	782K	—	32.39	0.8975	28.73	0.7851	27.68	0.7402	26.47	0.7978	30.96	0.9135
MambaIR-light [15]	×4	924K	84.6G	32.42	0.8977	28.74	0.7847	27.68	0.7400	26.52	0.7983	30.94	0.9135
MaIR-Small [42]	×4	1,374K	136.6G	**32.62**	0.8998	**28.90**	0.7882	**27.77**	0.7431	26.73	0.8049	**31.34**	0.9183
MambaIRv2-light [16]	×4	790K	75.6G	32.51	0.8992	28.84	0.7878	27.75	0.7426	26.82	0.8079	31.24	0.9182
T-PMambaSR (ours)	×4	703K	63.0G	32.61	**0.9000**	**28.90**	**0.7887**	**27.77**	**0.7432**	**26.87**	**0.8091**	31.29	**0.9185**

or Mamba in terms of both PSNR and SSIM across most benchmarks. Compared to MambaIRv2-light [16], a powerful hybrid Transformer-Mamba model, our method achieves comparable performance on the ×2 task and surpasses it on the ×3 and ×4 tasks while utilizing fewer parameters and FLOPs. Specifically, our model outperforms SRFormer-light [10] on Urban100 [14] by 0.26 dB, 0.25 dB, and 0.20 dB in PSNR at three scale factors, respectively, while saving 20% of the parameters. Compared to the latest Mamba-based method, MaIR-Small [42], our method achieves an average PSNR gain of 0.10dB on the ×2 task and demonstrates consistent performance advantages across all benchmarks on the scale of ×3 and ×4, while using only approximately 51% of its parameters and 45% of its FLOPs, showing the lightweight nature of ours.

For the real-world scenario, we conduct experiments on RealSRv3 [43], which provides authentic LR-HR pairs captured by modulating camera focal lengths. We compare our T-PMambaSR with several lightweight models: SRFormer-light [10], HiT-SIR [11], MambaIR-light [15], and MambaIRv2-light [16]. For a fair comparison, all models are initialized with their pre-trained ×4 weights and subsequently fine-tuned for 200K iterations on the RealSRv3 training set. Evaluation is then performed on the official RealSRv3 test set. As presented in Table II, our model achieves the highest PSNR, ties for the best SSIM, and obtains the lowest LPIPS among the listed methods, while using the fewest parameters.

TABLE II

QUANTITATIVE EVALUATIONS OF OUR METHOD AND EXISTING LIGHTWEIGHT SR METHODS ON A REAL-WORLD TEST SET

Scale	Methods	Params↓	FLOPs↓	RealSRv3 [43]
				PSNR↑/SSIM↑/LPIPS↓
×4	SRFormer-light [10]	873K	62.8G	29.38/0.8314/0.2697
	HiT-SIR [11]	792K	**53.8G**	29.34/0.8312/0.2724
	MambaIR-light [15]	924K	84.6G	29.39/0.8308/0.2706
	MambaIRv2-light [16]	790K	75.6G	29.40/**0.8325**/0.2682
	T-PMambaSR (Ours)	**703K**	63.0G	**29.43/0.8325/0.2658**

2) Qualitative Comparisons: Fig. 5 presents the visual comparisons on synthesized datasets for the ×2, ×3, and ×4 tasks, while Fig. 6 shows the ×4 qualitative results on the real-world test set, including Nikon and Canon datasets.

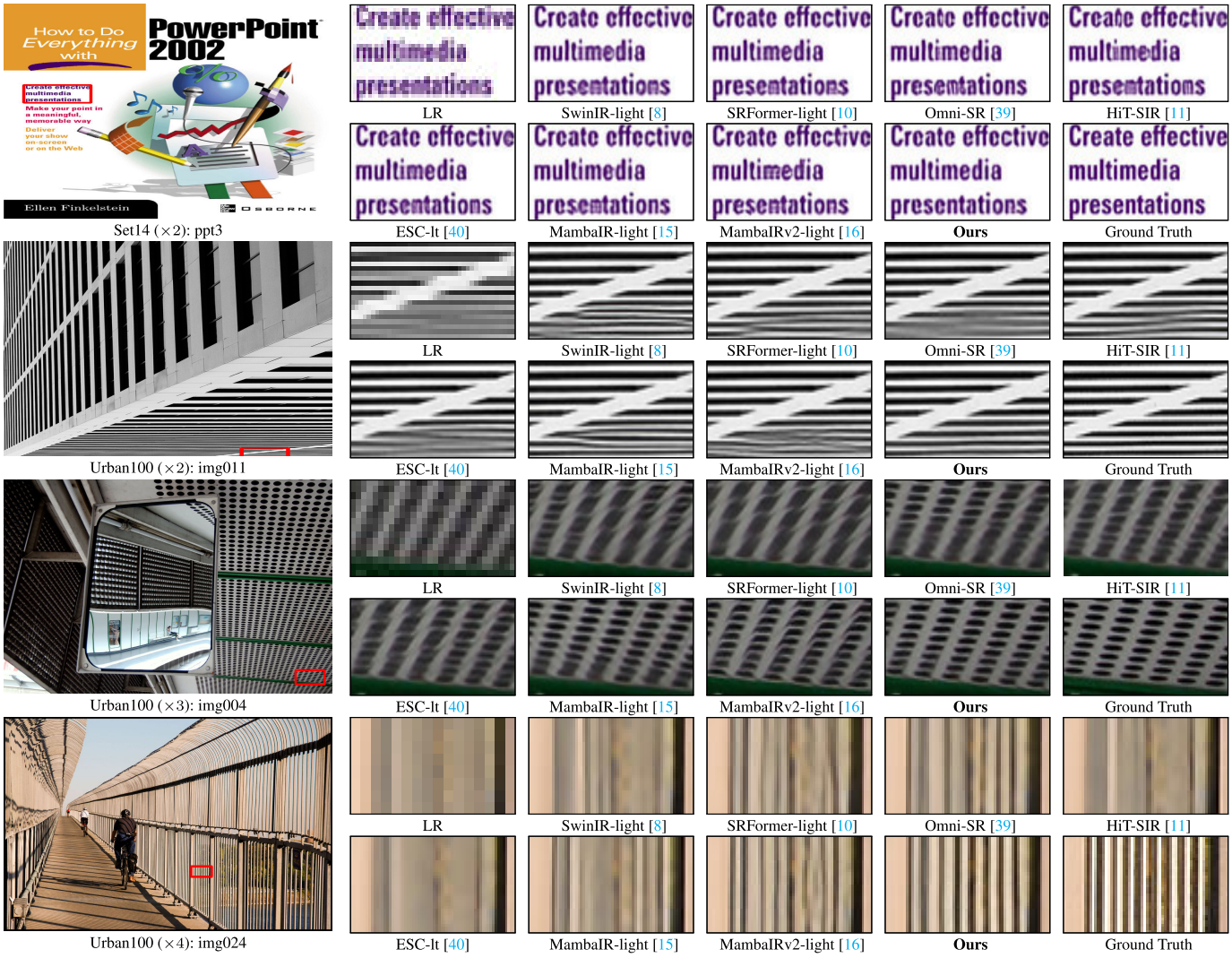


Fig. 5. Qualitative comparisons with existing methods in different scenes. Our method can restore clearer edges and structures.

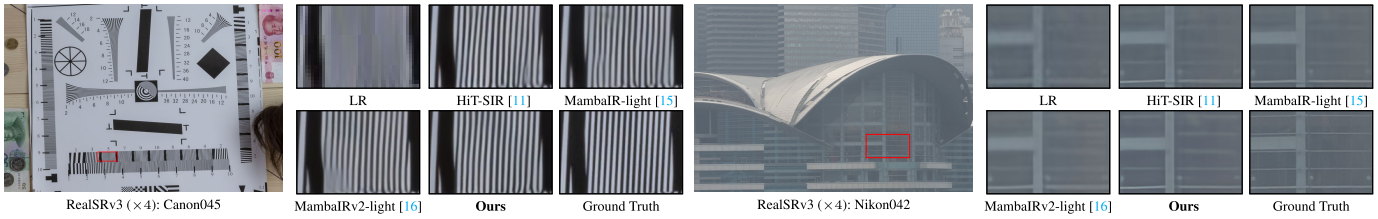


Fig. 6. Qualitative comparisons of our T-PMambaSR with existing methods on the real-world test set RealSRv3 [43].

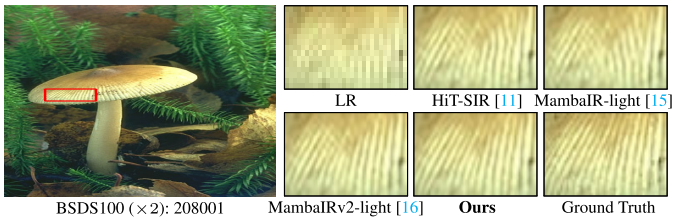


Fig. 7. Qualitative comparisons of our T-PMambaSR with existing methods on **208001** from the BSDS100 dataset (×2). Our method achieves sharper and more accurate restoration on natural textures.

Across all scales, our method demonstrates superior restoration clarity and accuracy. While other Transformer and Mamba-based models tend to blur or misalign high-frequency details, our adaptive refinement strategy excels at restoring these challenging textures, such as small letters in *ppt3* and fine textures in *img024* and *Canon045*. Moreover, other models lacking a fine-grained receptive field transition exhibit significant distortions or indistinctions on long-range structural details (*e.g.*, *img011*, *img004*, and *Nikon042*). In contrast, our progressive receptive field expansion enables a more accurate reconstruction of these structures, yielding results closest to the ground truth. Furthermore, to demonstrate the generality of our model beyond highly structured scenes, Fig. 7 visualizes its performance on natural textures using an image from the BSDS100 dataset (*image_208001*). Compared with existing models, our T-PMambaSR achieves sharper and more accurate restoration of complex natural patterns,

TABLE III

EFFICIENCY COMPARISON WITH EXISTING TRANSFORMER/MAMBA-BASED LIGHTWEIGHT SR METHODS [8], [10], [16], [42] ON THE ×4 SCALE, INCLUDING THE NUMBER OF PARAMETERS AND FLOPS, INFERENCE SPEED, AND AVERAGE PSNR/SSIM ON FIVE BENCHMARKS. FLOPS AND SPEED ARE MEASURED ON UPSAMPLED IMAGES WITH A SPATIAL SIZE OF 1280 × 720 USING A SINGLE NVIDIA RTX 4090 GPU

Methods	SwinIR-light [8]	SRFormer-light [10]	MaIR-Tiny [42]	MaIR-Small [42]	MambaIRv2-light [16]	**T-PMambaSR (Ours)**
Params↓	930K	873K	897K	1374K	790K	**703K**
FLOPs↓	63.6G	62.8G	**53.1G**	136.6G	75.6G	63.0G
Latency↓	**96ms**	108ms	242ms	425ms	138ms	124ms
PSNR↑/SSIM↑	29.26/0.8274	29.38/0.8296	29.35/0.8287	29.47/0.8309	29.43/0.8311	**29.49/0.8319**

TABLE IV

ABLATION STUDIES OF OUR CORE MODEL COMPONENTS. WE VALIDATE THE EFFECTIVENESS OF OUR **MWSS** AT THE WINDOW SCAN STAGE (REPLACED IN MODEL2 & 3), **MGSS** AT THE GLOBAL SCAN STAGE (REPLACED IN MODEL1 & 3), AND **HFM** AT THE HIGH-FREQUENCY REFINEMENT STAGE (REMOVED IN MODEL4). OUR FULL MODEL ACHIEVES THE BEST PERFORMANCE AND EFFICIENCY

Methods	Window Scan		Global Scan		High-Frequency Refinement	Params↓	FLOPs↓	PSNR↑	SSIM↑
	MWSS	2D-SS	MGSS	2D-SS	HFM				
Model1	✓			✓	✓	706.6K	241.2G	39.16	0.9781
Model2		✓	✓		✓	761.1K	249.8G	39.08	0.9780
Model3		✓		✓	✓	780.3K	253.0G	39.12	0.9779
Model4	✓		✓		✗	**687.0K**	**237.7G**	39.09	0.9781
Ours	✓		✓		✓	**687.0K**	238.1G	**39.20**	**0.9783**

confirming its robust generalization capability across diverse image types.

3) Efficiency: As shown in Table III, we evaluate the efficiency of our approach in comparison to existing Transformer- and Mamba-based methods, including SwinIR-light [8], SRFormer-light [10], MaIR-Tiny [42], MaIR-Small [42], and MambaIRv2-light [16]. Using ×4 scale as an example, we report the model's parameter count, FLOPs, inference speed, and average PSNR/SSIM across five benchmarks reported in Table I. These results demonstrate that with the fewest parameters, our model achieves the fastest inference speed among Mamba-based methods, reducing inference latency by 49% and 71% compared to MaIR-Tiny and MaIR-Small, respectively, while remaining competitive with Transformer-based models.

C. Ablation Studies

For ablation studies, all models are trained on the DIV2K dataset with 300K iterations and tested on the Manga109 (×2) test set. Our ablation studies focus on analyzing the impact of our proposed module components, the choice of scanning window size, and the strategy for receptive field transition.

1) Module Components: Table IV presents our ablation study validating that our proposed MWSS, MGSS, and HFM operations each contribute positively to performance. To verify the efficacy of our scan operations, we create three variants: Model1 replaces MGSS (in GSML) with the widely used 2D-selective scan (2D-SS) [15], [34], [47]; Model2 replaces MWSS (in WSML) with 2D-SS; and Model3 replaces all of our scan operations (MWSS and MGSS) with 2D-SS. Concurrently, to validate the HFM's role, Model4 removes it from the AHFRM. The results demonstrate that our model, with all components, achieves the best restoration metrics with the lowest parameter count and computational cost. Notably, replacing all our custom scans (Model3) increases parameters by approximately 100K and degrades PSNR by 0.08dB.

TABLE V

ABLATION STUDIES OF OUR MULTI-SCALE WINDOW SELECTIVE SCAN (MWSS) STRATEGY AND OUR CHOICE OF WINDOW SIZES IN WIF AND WFF, RESPECTIVELY. DISABLING MWSS (CASE1-3) OR USING WINDOW PAIRS WITH A LARGE SIZE DISPARITY (CASE4) LEADS TO PERFORMANCE DEGRADATION. THIS CONFIRMS THAT A SMOOTH MULTI-SCALE TRANSITION (OURS) IS CRUCIAL

Methods	MWSS	Window Sizes	PSNR↑	SSIM↑
Case1	✗	(16×16, 16×16)	39.15	0.9781
Case2	✗	(32×32, 32×32)	39.13	0.9780
Case3	✗	(64×64, 64×64)	39.13	0.9781
Case4	✓	(16×16, 64×64)	39.13	0.9779
Case5	✓	(32×32, 64×64)	39.16	0.9782
Ours	✓	(16×16, 32×32)	**39.20**	**0.9783**

Furthermore, removing the parameter-free and computationally negligible HFM (Model4) causes a 0.11dB PSNR drop, demonstrating its significant advantage for SR reconstruction.

2) Scanning Window Size: To validate the contribution of our Multi-scale Window Selective Scan (MWSS) strategy and demonstrate the optimality of the selected 16 × 16 and 32×32 scanning window sizes, we conduct a series of ablation experiments, with results presented in Table V. In Case1, Case2, and Case3, we deactivate the MWSS strategy and instead use two parallel scans with identical window sizes but different directions. In Case4 and Case5, the MWSS strategy is employed with different combinations of window sizes for the parallel scans. The results show that our proposed MWSS strategy and window size selection yield the best performance. Notably, simply adopting the MWSS strategy is insufficient if the disparity between window sizes is too large, as seen in the performance drop in Case4. We observe that a smooth receptive field transition (as in Ours) is essential for achieving optimal modeling performance.

3) Receptive Field Transition: Inspired by prior works [11], [16], we design a progressive modeling strategy that expands

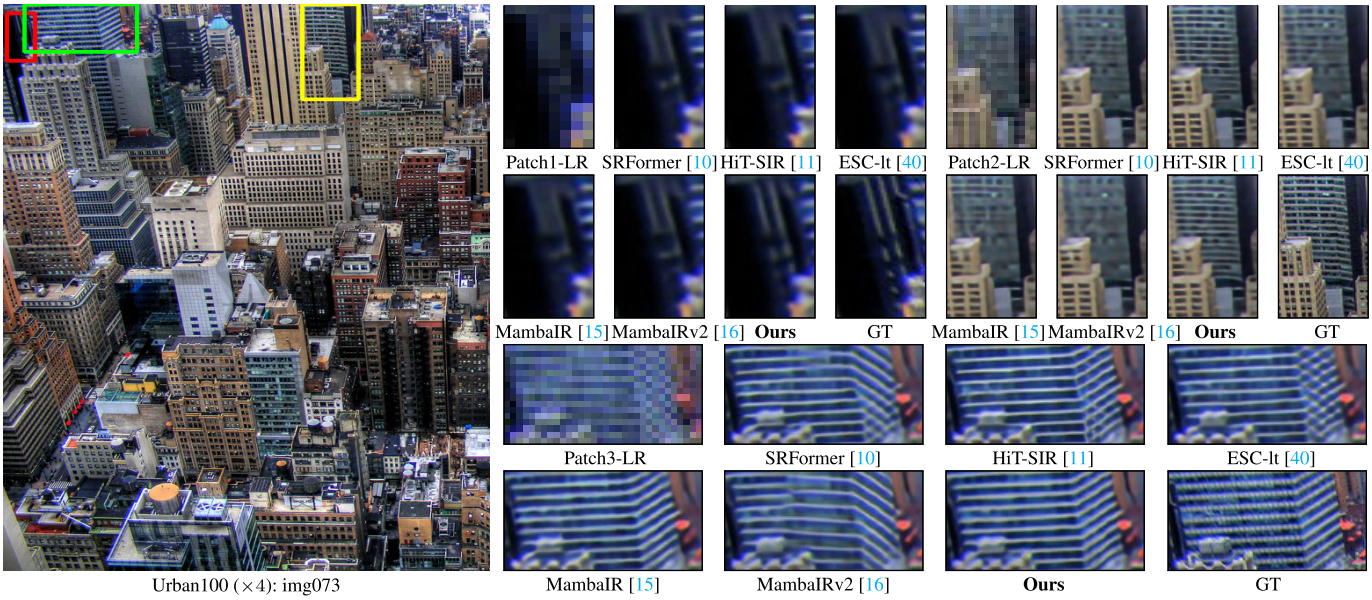


Fig. 8. Visual comparisons of different SR methods on multi-scale targets. The target in **Patch 1** represents local information, while the targets in **Patch 2** and **Patch 3** extend beyond a 16 × 16 W-MHSA window. Competing Transformer and Mamba architectures exhibit mixed performance across these varying scales. In contrast, our method's hybrid architecture and progressive receptive field expansion strategy consistently yields the best visual quality across all targets.

TABLE VI

ABLATION STUDIES OF THE RECEPTIVE FIELD TRANSITION SEQUENCE. OUR PROGRESSIVE SMALL-TO-LARGE STRATEGY YIELDS THE BEST PERFORMANCE. ANY DEVIATION FROM THIS DESIGNED SEQUENCE RESULTS IN PERFORMANCE DEGRADATION, CONFIRMING THE OPTIMALITY OF OUR APPROACH

Modeling Stage	Transition Sequence	PSNR↑	SSIM↑
GSML→TL→WSML	global→local→regional	39.16	0.9780
WSML→TL→GSML	regional→local→global	39.13	0.9781
TL→WSML→GSML	local→regional→global	**39.20**	**0.9783**

TABLE VII

QUANTITATIVE COMPARISON OF PERFORMANCE ON MULTI-SCALE PATCHES. "ARCH" DENOTES THE ARCHITECTURE ("T": TRANSFORMER, "M": MAMBA, "T-M": HYBRID), AND "RFT" INDICATES WHETHER A RECEPTIVE FIELD TRANSITION STRATEGY IS EMPLOYED. OUR HYBRID MODEL WITH A PROGRESSIVE RFT STRATEGY OBTAINS THE HIGHEST METRICS ACROSS ALL THREE PATCH SCALES

Methods	Arch	RFT	Patch1	Patch2	Patch3
			8×16 pixels	16×32 pixels	32×16 pixels
			PSNR↑/SSIM↑		
SRFormer [10]	T	✗	25.29/0.7997	19.32/0.5315	18.92/0.6958
HiT-SIR [11]	T	✓	25.25/0.7787	21.05/0.7151	21.02/0.8364
ESC-lt [40]	T	✗	25.40/0.7853	19.48/0.5455	20.60/0.8233
MambaIR [15]	M	✗	24.86/0.7489	19.36/0.5265	22.03/0.8663
MambaIRv2 [16]	T-M	✗	25.92/0.8206	18.93/0.4808	16.75/0.4775
Ours	T-M	✓	**26.81/0.8719**	**21.20/0.7261**	**22.99/0.8814**

the receptive field from small to large. We conduct an ablation study to validate the effectiveness of this sequence, with results presented in Table VI. Our modules—TL, WSML, and GSML—are designed to model features at local, regional, and global scales, respectively. The results confirm that our proposed small-to-large (local → regional → global) sequence achieves optimal performance. As shown, any deviation from this progressive order leads to degradation in performance.

V. FURTHER ANALYSIS

A. Progressive Expansion of Receptive Field

Our method proposes a progressive receptive field expansion strategy based on a hybrid Transformer-Mamba architecture. To demonstrate the superiority of this paradigm, we conduct a qualitative comparison on the *Urban100_img073*, which is rich in multi-scale features. As shown in Fig. 8, our model achieves the best visual results and the highest PSNR/SSIM (as shown in Table VII) on objects at various scales.

Transformer-based methods [10], [11], [40], leveraging window-based multi-head self-attention (with a standard window size of 16×16 or smaller; here, we adopt 16×16 for our W-MHSA), excel at restoring regional features within small window sizes (*e.g.*, **Patch 1**), where Mamba-based methods [15] struggle. However, when the target object's size exceeds the size of windows (*e.g.*, **Patch 2** and **Patch 3**), these Transformer-based models show their limitations. Their fixed, small windows cannot capture the complete object, and they suffer from a lack of inter-window interaction. Meanwhile, Mamba-based models like MambaIR [15] and MambaIRv2 [16] also fail to achieve optimal results, as their scanning mechanisms (*e.g.*, 2D selective scan and semantic guided neighboring scan) disrupt the original spatial relationships of neighboring pixels and lack a smooth receptive field transition.

In contrast, our method performs well in these scenarios. For larger structures (*e.g.*, **Patch 2** and **Patch 3**), after the initial MHSA modeling, it employs the WSML as a transition. The WSML refines the preceding MHSA features via inter-window interaction and captures the complete object using larger-window modeling, before the GSML embeds the target

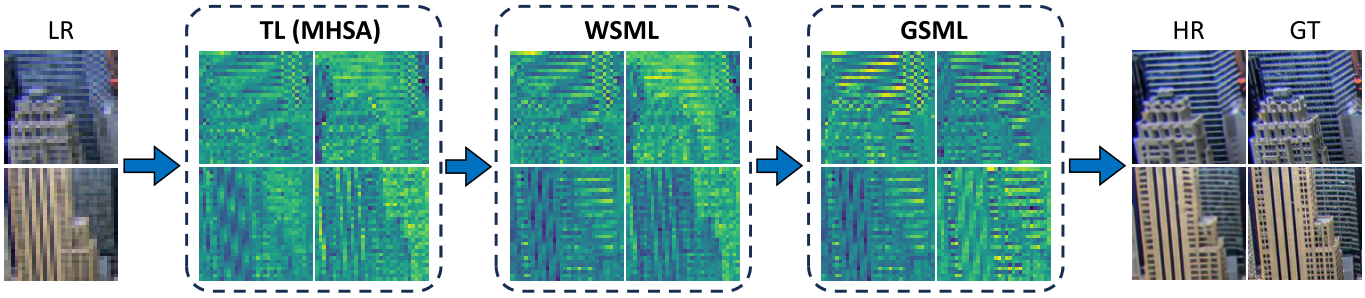


Fig. 9. Visualization of features from three stages. We show the outputs from our TL and WSML (from the final T-WSM Block of the last two T-PM Groups) and the subsequent GSML. The indistinct long-range features from the TL stage are progressively detailed by the subsequent WSML (regional) and GSML (global) stages.

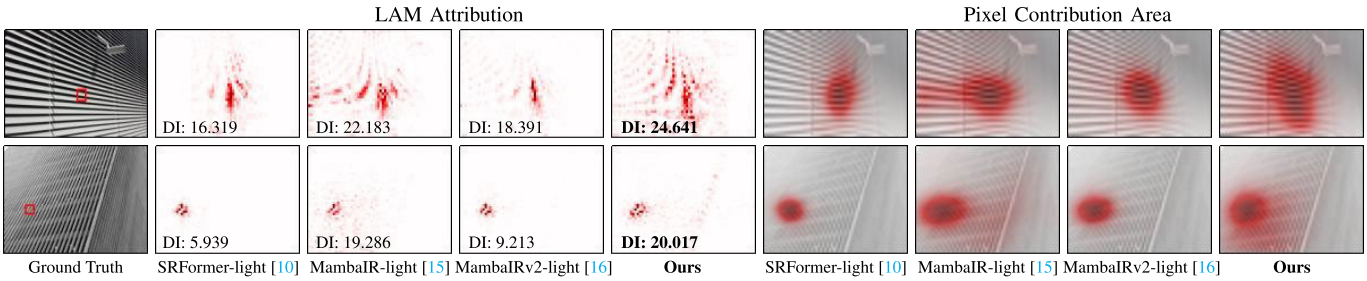


Fig. 10. LAM [48] comparison between our model and existing methods [10], [15], [16] on the $\times 4$ scale. Our method leads to the maximum pixel contribution areas, confirming its superior capacity for capturing broad contextual information.

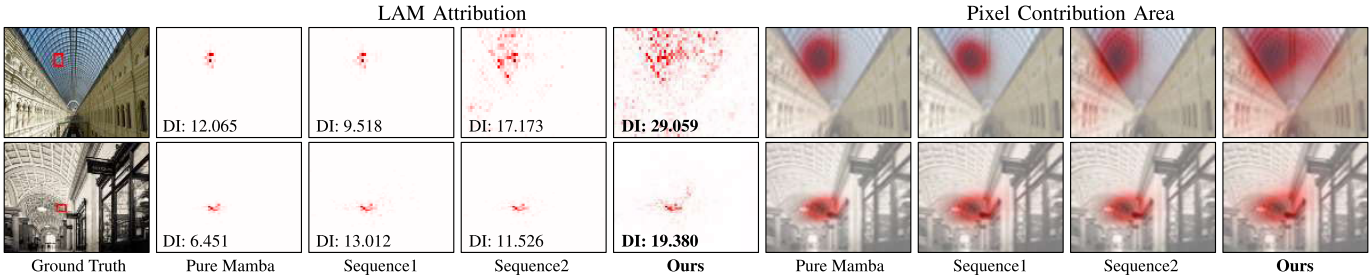


Fig. 11. LAM [48] comparisons at the $\times 4$ scale between our progressive sequence (TL $\rightarrow$ WSML $\rightarrow$ GSML) and its variants: (1) shuffled variants (Sequence 1: WSML $\rightarrow$ TL $\rightarrow$ GSML, Sequence 2: GSML $\rightarrow$ TL $\rightarrow$ WSML), and (2) a pure Mamba variant, which replaces all W-MHSA modules with MWSS. The proposed progressive sequence yields the highest DI value and the largest pixel contribution area, confirming that our local-to-global transition architecture more effectively enables broad contextual engagement than both the pure Mamba variant and the shuffled sequences.

in a global context for final refinement. This rich, progressive modeling process fully leverages the respective advantages of the hybrid Transformer-Mamba architecture. Notably, HiT-SIR [11] similarly recognizes the importance of progressive expansion, yet their gradually enlarging attention window only achieved suboptimal results. This is because HiT-SIR [11] is confined to information exchange within different windows, whereas our approach utilizes a scanning mechanism to concatenate information across different windows, thereby achieving superior reconstruction.

Fig. 9 visualizes the intermediate feature maps from our three modeling stages. Details that remain indistinct within the smaller receptive field of an earlier stage are progressively refined as they pass through subsequent stages with larger contextual windows, ultimately leading to a result that is clear at all scales. Furthermore, we employ Local Attribution Maps (LAM) [48] in Fig. 10 to analyze the contextual range of different architectures. LAM illustrates the model's contextual engagement by correlating all pixels with a target patch (red), where a higher Diffusion Index (DI) or larger contribution area signifies a broader receptive field. Our method achieves the highest DI value and the largest pixel contribution area, demonstrating its ability to leverage a wider context. Although MambaIR-light [15] shows a comparable contribution area, its lack of a fine-grained receptive field transition and rich modeling strategies results in significantly inferior performance. This stark contrast highlights the conceptual advancement of our architecture, which achieves a strictly fine-grained, progressive receptive field transition (local $\rightarrow$ regional $\rightarrow$ global) together with effective inter-window communication. To further validate the necessity of this design, we compare our modeling sequence (TL $\rightarrow$ WSML $\rightarrow$ GSML) with shuffled variants. As shown in Fig. 11, our proposed configuration achieves the highest DI value and the largest pixel contribution

TABLE VIII
ABLATION STUDIES OF OUR AHFRM COMPARED WITH EXISTING MODULES FROM ESRT [26] AND CRAFT [18]

Methods	Scale	Params↓	FLOPs↓	Manga109 [38]	
				PSNR↑	SSIM↑
ESRT-variant	×4	702K	62.6G	30.65	0.9109
CRAFT-variant	×4	670K	61.0G	30.66	0.9104
Ours	×4	703K	63.0G	**30.74**	**0.9110**

area, significantly outperforming both the shuffled variants (**Sequence 1**: WSML → TL → GSML, and **Sequence 2**: GSML → TL → WSML) and the **Pure Mamba variant**, which replaces all W-MHSA modules with MWSS. Notably, the clear superiority over the Pure Mamba variant further confirms that the observed gains do not stem merely from Mamba's inherent long-range modeling capability, but from the architectural synergy brought by our local-to-global transition design. These results demonstrate that our designed sequence effectively maximizes contextual engagement and validates the structural rationale behind our progressive architectural design.

B. High-Frequency Restoration

Our AHFRM aims to refine the high-frequency information that is often lost after being processed by the Transformer and Mamba blocks. While existing frequency-aware methods like ESRT [26] and CRAFT [18] attempt to enhance high-frequency details either prior to deep spatial modeling or directly from features that have already been degraded by low-pass-like spatial blocks, our AHFRM overcomes this limitation by employing a **reference-guided restoration mechanism**. Specifically, its dual-branch design extracts *pristine, undecayed* high-frequency priors from early shallow features to serve as explicit guidance. This enables highly targeted recovery by using pristine priors to adaptively restore damaged high-frequency features, rather than merely refining already degraded information. To empirically validate the effectiveness of this design, we replace our AHFRM with equivalent baseline modules—namely, the HPB from ESRT and the HFERB from CRAFT (denoted as **ESRT-variant** and **CRAFT-variant** in Table VIII, respectively). As shown, our AHFRM consistently outperforms both variants. These results provide strong empirical evidence that using pristine high-frequency priors to explicitly guide the restoration process is fundamentally more effective for accurate texture recovery than simply refining already degraded features.

Fig. 12 visualizes the effectiveness of our AHFRM using its intermediate features. Fig. 12 (a) shows the features before and after the AHFRM module; the attenuated high-frequency textures become significantly clearer after modeling because we use undecayed high-frequency information as a supplement to guide the restoration of the degraded HF content. Fig. 12 (b) further demonstrates the superiority of this method. The restored high-frequency feature $\boldsymbol{X}_{hf}$ in Fig. 4 possesses visibly clearer textures than the degraded feature $\boldsymbol{X}_{lf}$. After an adaptive and weighted fusion, the resulting feature $\boldsymbol{X}_{sum}$ has shown superior, clearer edges prior to further convolutional processing, which confirms that our high-frequency targeted refinement is significantly effective.

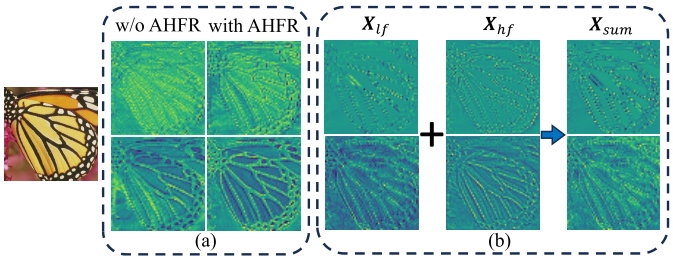


Fig. 12. Visualization of our AHFRM's high-frequency refinement mechanism. (a) Demonstrates the effectiveness of AHFRM in enhancing high-frequency features. (b) It details the internal mechanism: input features are enriched by the fusion of supplementary high-frequency features, effectively restoring details missing from the input.

TABLE IX
ABLATION STUDIES COMPARING OUR HYBRID ARCHITECTURE WITH PURE-TRANSFORMER AND PURE-MAMBA VARIANTS

Methods	Scale	Params↓	FLOPs↓	Manga109 [38]	
				PSNR↑	SSIM↑
Pure-Transformer	×4	737K	53.5G	30.57	0.9087
Pure-Mamba	×4	690K	51.4G	30.36	0.9061
Ours	×4	703K	63.0G	**30.74**	**0.9110**

C. Superiority of the Hybrid Architecture

Our T-PMambaSR deliberately integrates Transformer and Mamba modules to leverage their complementary strengths. While W-MHSA excels at extracting precise local features, expanding its receptive field (window size) incurs quadratic complexity, forcing pure-attention methods to rely on lossy operations that compromise spatial integrity. In contrast, purely Mamba-based models efficiently capture global contexts but often struggle to accurately reconstruct fine-grained local structures. Our hybrid architecture addresses these limitations. By transitioning from W-MHSA to our Mamba-based modules (MWSS and MGSS), we expand the receptive field and achieve effective inter-window communication solely by altering the selective scan paths. This ensures a smooth, fine-grained scale transition without additional costs or spatial feature loss.

To quantitatively validate our hybrid design, we construct two single-architecture baselines under a controlled parameter count: a **Pure Transformer** variant (using only hierarchical transformer blocks from HiT-SIR [11]) and a **Pure Mamba** variant (replacing all W-MHSA with MWSS). In these comparisons, AHFRM is kept unchanged to strictly isolate the effect of the spatial backbone. As shown in Table IX, our hybrid architecture significantly outperforms both variants. These results demonstrate that integrating Transformer and Mamba modules achieves a superior receptive field transition and better overall performance than either single paradigm.

VI. LIMITATIONS AND FUTURE WORKS

While our T-PMambaSR achieves competitive performance with high efficiency, we acknowledge several limitations that

offer avenues for future research. First, although our hybrid design is lightweight, the Window-MHSA component still accounts for a notable portion of the model's parameters and computational load. Exploring lighter-weight attention variants or dynamic attention mechanisms could further reduce the computational overhead without compromising their local modeling capabilities. Second, our progressive scanning strategy relies on pre-defined, fixed window sizes (*e.g.*, 16×16 and 32×32) for the WSML. While effective, this fixed-scale approach may not be optimal for all image content. A future direction would be to develop a more flexible or content-adaptive scanning mechanism, where the window sizes and scanning paths are dynamically determined based on the input image's structural properties. Finally, our model is specifically designed and optimized for the task of single-image SR. Its effectiveness and generalizability to other low-level vision tasks [49], such as image denoising and deblurring, have not yet been validated. Adapting and evaluating our progressive hybrid modeling paradigm for these related tasks remains a promising area for future investigation.

VII. Conclusion

In this paper, we proposed the T-PMambaSR, an efficient hybrid Transformer-Mamba architecture for lightweight SR. Our core contribution is a progressive receptive field expansion strategy that synergizes the local modeling strength of Transformers with Mamba's long-range dependency capabilities. This is achieved through two novel modules: the WSML for regional modeling and the GSML for global context. Furthermore, we design an AHFRM to effectively restore high-frequency texture details that may be lost during global interaction. Experiments demonstrate that our T-PMambaSR achieves competitive performance compared with existing lightweight Transformer- and Mamba-based SR methods on both synthetic and real-world benchmarks, while incurring lower computational cost and faster inference speed. Overall, we can conclude that our work validates the superiority of a smooth, progressive local-to-global paradigm for hybrid SR models.

References

[1] W. Li et al., "Efficient image super-resolution with feature interaction weighted hybrid network," *IEEE Trans. Multimedia*, vol. 27, pp. 2256–2267, 2025.

[2] L. Peng et al., "Towards realistic data generation for real-world super-resolution," in *Proc. ICLR*, 2024, pp. 1–24.

[3] Z. Feng et al., "PMQ-VE: Progressive multi-frame quantization for video enhancement," in *Proc. NeurIPS*, 2025, pp. 1–30.

[4] W. Li, X. Wang, H. Guo, G. Gao, and Z. Ma, "Self-supervised selective-guided diffusion model for old-photo face restoration," in *Proc. NeurIPS*, 2025, pp. 1–26.

[5] G. Gao, W. Li, J. Li, F. Wu, H. Lu, and Y. Yu, "Feature distillation interaction weighting network for lightweight image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 661–669.

[6] W. Li et al., "Survey on deep face restoration: From non-blind to blind and beyond," *ACM Comput. Surveys*, vol. 58, no. 6, pp. 1–35, Apr. 2026.

[7] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–22.

[8] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[9] X. Zhang, H. Zeng, S. Guo, and L. Zhang, "Efficient long-range attention network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Tel Aviv, Israel: Springer, Oct. 2022, pp. 649–667.

[10] Y. Zhou, Z. Li, C.-L. Guo, S. Bai, M.-M. Cheng, and Q. Hou, "SRFormer: Permuted self-attention for single image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12780–12791.

[11] X. Zhang, Y. Zhang, and F. Yu, "Hit-Sr: Hierarchical transformer for efficient image super-resolution," in *Proc. ECCV*, 2024, pp. 483–500.

[12] W. Li, H. Guo, Y. Hou, G. Gao, and Z. Ma, "Dual-domain modulation network for lightweight image super-resolution," *IEEE Trans. Multimedia*, vol. 28, pp. 4679–4689, 2026.

[13] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. ICCS*, 2010, pp. 711–730.

[14] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.

[15] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S.-T. Xia, "Mambair: A simple baseline for image restoration with state-space model," in *Proc. ECCV*, 2024, pp. 222–241.

[16] H. Guo et al., "MambaIRv2: Attentive state space restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 28124–28133.

[17] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.

[18] A. Li, L. Zhang, Y. Liu, and C. Zhu, "Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12514–12524.

[19] S. Xu, W. Li, G. Gao, J. Yang, G.-J. Qi, and C.-W. Lin, "FADPNet: Frequency-aware dual-path network for face super-resolution," 2025, *arXiv:2506.14121*.

[20] X. Zhao et al., "Efficient single image super-resolution with entropy attention and receptive field augmentation," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 1302–1310.

[21] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis. Cham*, Switzerland: Springer, 2016, pp. 391–407.

[22] G. Gao, Z. Wang, J. Li, W. Li, Y. Yu, and T. Zeng, "Lightweight bimodal network for single-image super-resolution via symmetric CNN and recursive transformer," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 913–919.

[23] W. Li et al., "Cross-receptive focused inference network for lightweight image super-resolution," *IEEE Trans. Multimedia*, vol. 26, pp. 864–877, 2024.

[24] C. Liu, D. Zhang, G. Lu, W. Yin, J. Wang, and G. Luo, "SRMamba-T: Exploring the hybrid mamba-transformer network for single image super-resolution," *Neurocomputing*, vol. 624, Apr. 2025, Art. no. 129488.

[25] J. Qiao et al., "Hi-mamba: Hierarchical mamba for efficient image super-resolution," 2024, *arXiv:2410.10140*.

[26] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 457–466.

[27] W. Li et al., "Efficient face super-resolution via wavelet-based feature enhancement network," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 4515–4523.

[28] Z. Huang, Z. Zhang, C. Lan, Z.-J. Zha, Y. Lu, and B. Guo, "Adaptive frequency filters as efficient global token mixers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6026–6036.

[29] W. Li, H. Guo, Y. Hou, and Z. Ma, "FourierSR: A Fourier token-based plugin for efficient image super-resolution," *IEEE Trans. Image Process.*, vol. 35, pp. 732–742, 2026.

[30] P. Xu, Q. Liu, H. Bao, R. Zhang, L. Gu, and G. Wang, "FDSR: An interpretable frequency division stepwise process based single-image super-resolution network," *IEEE Trans. Image Process.*, vol. 33, pp. 1710–1725, 2024.

[31] G. Wu, J. Jiang, J. Jiang, and X. Liu, "Transforming image super-resolution: A ConvFormer-based efficient approach," *IEEE Trans. Image Process.*, vol. 33, pp. 6071–6082, 2024.

[32] Y. Cui, W. Ren, and A. Knoll, "Exploring the potential of pooling techniques for universal image restoration," *IEEE Trans. Image Process.*, vol. 34, pp. 3403–3416, 2025.

[33] Y. Yan, S. Yao, W. Ren, R. Zhang, Q. Guo, and X. Cao, "CAN: Cascade augmentations against noise for image restoration," *IEEE Trans. Image Process.*, vol. 34, pp. 5131–5146, 2025.

[34] J. Weng, Z. Yan, Y. Tai, J. Qian, J. Yang, and J. Li, "MambaLLIE: Implicit retinex-aware low light enhancement with global-then-local state space," in *Proc. Adv. Neural Inf. Process. Syst.* 37, 2024, pp. 27440–27462.
[35] W. Zou, H. Gao, W. Yang, and T. Liu, "Wave-Mamba: Wavelet state space model for ultra-high-definition low-light image enhancement," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 1534–1543.
[36] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L.-A. Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 135.1–135.10.
[37] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, Jul. 2001, pp. 416–423.
[38] Y. Matsui et al., "Sketch-based Manga retrieval using Manga109 dataset," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21811–21838, Oct. 2017.
[39] H. Wang, X. Chen, B. Ni, Y. Liu, and J. Liu, "Omni aggregation networks for lightweight image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22378–22387.
[40] D. Lee, S. Yun, and Y. Ro, "Emulating self-attention with convolution for efficient image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2025, pp. 24467–24477.
[41] X. Wang, J. Li, J. Li, S. Wang, L. Yan, and Y. Xu, "A collaborative network of mamba and CNN for lightweight image super-resolution," *IEEE Trans. Consum. Electron.*, vol. 71, no. 2, pp. 3591–3604, May 2025.
[42] B. Li, H. Zhao, W. Wang, P. Hu, Y. Gou, and X. Peng, "MaIR: A locality-and continuity-preserving mamba for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 7491–7501.
[43] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3086–3095.
[44] R. Timofte et al., "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 126–135.
[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[46] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[47] Y. Liu et al., "VMamba: Visual state space model," in *Proc. Adv. Neural Inf. Process. Syst.* 37, 2024, pp. 103031–103063.
[48] J. Gu and C. Dong, "Interpreting super-resolution networks with local attribution maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9199–9208.
[49] L. Peng et al., "Pixel to Gaussian: Ultra-fast continuous super-resolution with 2D Gaussian modeling," in *Proc. ICLR*, 2026, pp. 1–18.

Sichen Guo received the B.E. degree in computer science from the Bell Honors School, Nanjing University of Posts and Telecommunications, Nanjing, China. He is currently pursuing the M.S. degree in computer vision with the Robotics Institute, School of Computer Science, Carnegie Mellon University. His research interests include image restoration and multimodal learning.

Wenjie Li received the M.S. degree in control science and engineering from the College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, in 2023. He is currently pursuing the Ph.D. degree in artificial intelligence with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. His research interests include image restoration.

Yuanyang Liu received the B.E. degree in communication engineering from the Bell Honors School, Nanjing University of Posts and Telecommunications, Nanjing, China. He is currently pursuing the M.S. degree with the School of Artificial Intelligence, Nanjing University of Posts and Telecommunications. His research interests include 3D point cloud semantic segmentation and image restoration.

Guangwei Gao (Senior Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from Nanjing University of Science and Technology (NJUST), Nanjing, China, in 2014. He is currently a Professor with the School of Computer Science and Engineering, NJUST. His research interests include pattern recognition and computer vision. He has served as an Associate Editor for *Pattern Recognition* and IEEE TRANSACTIONS ON IMAGE PROCESSING. For more information https://guangweigao.github.io

Jian Yang (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from Nanjing University of Science and Technology (NJUST), Nanjing, China, in 2002. He is currently a Professor with the School of Computer Science and Engineering, NJUST. He is the author of more than 400 scientific articles in pattern recognition and computer vision. His research interests include pattern recognition, computer vision, and machine learning. He is a fellow of IAPR. He is/was an Associate Editor of *Pattern Recognition* and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Chia-Wen Lin (Fellow, IEEE) received the Ph.D. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000. He is currently a Distinguished Professor with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU. His research interests include image and video processing, computer vision, and video networking. He is currently serving as an Associate Editor-in-Chief for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.