Frequency-Assisted Mamba for Remote Sensing Image Super-Resolution

Yi Xiao, *Graduate Student Member, IEEE*, Qiangqiang Yuan, *Member, IEEE*, Kui Jiang, *Member, IEEE*, Yuzeng Chen, Qiang Zhang, and Chia-Wen Lin, *Fellow, IEEE*

***Abstract*—Recent progress in remote sensing image (RSI) super-resolution (SR) has exhibited remarkable performance using deep neural networks, e.g., Convolutional Neural Networks and Transformers. However, existing SR methods often suffer from either a limited receptive field or quadratic computational overhead, resulting in sub-optimal global representation and unacceptable computational costs in large-scale RSI. To alleviate these issues, we develop the first attempt to integrate the Vision State Space Model (Mamba) for RSI-SR, which specializes in processing large-scale RSI by capturing long-range dependency with linear complexity. To achieve better SR reconstruction, building upon Mamba, we devise a Frequency-assisted Mamba framework, dubbed FMSR, to explore the spatial and frequent correlations. In particular, our FMSR features a multi-level fusion architecture equipped with the Frequency Selection Module (FSM), Vision State Space Module (VSSM), and Hybrid Gate Module (HGM) to grasp their merits for effective spatial-frequency fusion. Considering that global and local dependencies are complementary and both beneficial for SR, we further recalibrate these multi-level features for accurate feature fusion via learnable scaling adaptors. Extensive experiments on AID, DOTA, and DIOR benchmarks demonstrate that our FMSR outperforms state-of-the-art Transformer-based methods HAT-L in terms of PSNR by 0.11 dB on average, while consuming only 28.05% and 19.08% of its memory consumption and complexity, respectively.**

***Index Terms*—Frequency selection, remote sensing image, state space model, super-resolution.**

I. Introduction

HIGH-RESOLUTION remote sensing imagery (RSI), which records high-quality earth observation details, provides promising prospects for large-scale and fine-grained applications [1], [2], [3], [4], [5], [6], [7], [8]. However, the complex imaging environment (e.g., scattering and tremor) often impedes high-resolution (HR) image acquisition [9], [10], [11], [12]. Moreover, RSI often undergoes severe compression and downsampling to tame the transmission instability between satellites and ground stations, resulting in suboptimal scene representation. Hence, reconstructing HR images from low-resolution (LR) observations is crucial for improving both human perception and subsequent applications.

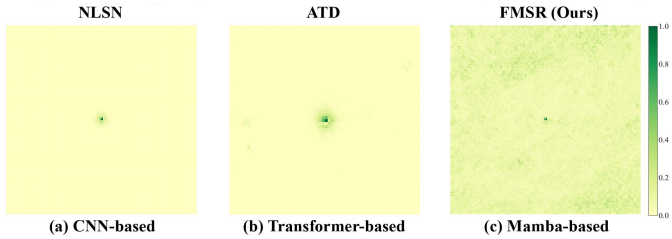


Fig. 1. The Effective Receptive Field (ERF) [21] comparison for (a) CNN-based method NLSN [22], (b) Transformer-based model ATD [23], and the proposed Mamba-based network FMSR. A wider distribution of dark areas demonstrates larger ERF. Our FMSR effectively obtains the largest ERF, indicating favorable global exploration capability.

In contrast to upgrading hardware maintenance, super-resolution (SR) techniques provide a flexible and cost-effective alternative by predicting latent HR images from their LR counterparts [12], [13], [14], [15]. Early efforts often relied on hand-crafted priors to tame the ill-posedness [16], [17], [18], [19], [20]. However, they struggled to produce accurate results and involved laborious optimization processes. Recently, deep neural networks have demonstrated remarkable progress in SR tasks and achieved superior performance over traditional approaches, such as Convolutional Neural Networks (CNNs) and Transformers. While CNN-based methods commonly invent elaborate attention mechanisms to grasp the informative features, restrained by the inherent nature of convolution units, they have limited respective fields and cannot capture long-range dependencies. As shown in Fig. 1, the effective receptive field (ERF) [21] of CNN-based model NLSN [22] is limited. This requirement is essential for SR tasks, as a predicted pixel needs prior knowledge from its surrounding region to be super-resolved, especially in wide-range RSI.

Transformer-based methods achieve increased respective fields by leveraging global interaction among all input data through a self-attention (SA) mechanism, demonstrating impressive performance across various domains [24], [25], [26],

Received 1 May 2024; revised 14 July 2024; accepted 28 August 2024. Date of publication 30 December 2024; date of current version 4 April 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 423B2104 and in part by the Fundamental Research Funds for the Central Universities under Grant 2042024kf0020 and Grant 2042023kfyq04. The associate editor coordinating the review of this article and approving it for publication was Dr Liang Lin. *(Corresponding authors: Qiangqiang Yuan; Kui Jiang.)*

Yi Xiao, Qiangqiang Yuan, and Yuzeng Chen are with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China (e-mail: xiao_yi@whu.edu.cn; yqiang86@gmail.com; yuzeng_chen@whu.edu.cn).

Kui Jiang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: jiangkui@hit.edu.cn).

Qiang Zhang is with Information Science and Technology College, Dalian Maritime University, Dalian 116026, China (e-mail: qzhang95@dlmu.edu.cn).

Chia-Wen Lin is with the Department of Electrical Engineering, Institute of Communications Engineering, National Tsing Hua University, Hsinchu 300, Taiwan (e-mail: cwlin@ee.nthu.edu.tw).

Code will be available at https://github.com/XY-boy/FreMamba

Digital Object Identifier 10.1109/TMM.2024.3521798

[27], [28], [29]. Despite achieving superior performance against CNN-based approaches, these methods exhibit quadratic complexity with respect to the token size. In this context, taming Transformer for high-resolution scenarios presents a significant challenge, particularly for large-scale remote sensing images. Although some approaches seek a lightweight SA for global modeling, such as recursive SA [30] and window-based SA [31], they usually come at the expense of global modeling accuracy and require stacking many blocks to establish a global dependency, thus increasing the computational budget. Moreover, the inherent issue of quadratic complexity remains unsolved. Therefore, a natural question arises: *can a more efficient yet effective solution be developed to grasp the long-range dependencies across large-scale RSI?*

The recent-popular State Space Model (SSM) could be a promising answer to this question. Originating from Kalman filtering [32], SSM primarily employs linear filtering and prediction methods to represent the evolution of the internal state of the system, thus naturally enjoying linear complexity. By integrating SSM with the MultiLayer Perceptron (MLP) block of Transformer, the simplified architecture Mamba [33] is achieved, which introduces a selection scanning mechanism in SSM to filter out irrelevant information for long-sequence modeling. Recently, Mamba demonstrated impressive results in various domains, making it a possible replacement for the Transformer model. Nevertheless, although Mamba exhibits favorable performance and can serve as an alternative to Transformer, some potential problems persist in large-scale earth observation scenarios, making introducing Mamba into RSI SR more challenging. **Firstly, images captured by satellite platforms often lose crucial frequency information for perception, which requires heterogeneous representation for accurate reconstruction.** The underlying reason lies in that the original Mamba processes each token equally in the spatial domain, limiting its ability to perceive informative frequency signals across entire images. **Secondly, there exists spatial diversity among observed objects in RSI, which is barely explored.** While Mamba is effective in exploring long-range dependencies, it lacks explicit consideration of spatially-varying contents, resulting in suboptimal pixel-wise representation during local modeling.

To mitigate the aforementioned problem, we first attempt to extend Mamba from the perspective of frequency analysis and propose a frequency-assisted Mamba for RSI SR, termed FMSR. Specifically, instead of solely employing VSSM for long-range modeling, we devise an effective Frequency Selection Module (FSM) to adaptively identify informative frequency cues vital for perception. By incorporating FSM with Mamba, the resulting frequency-assisted mamba block can better utilize the complementary strengths between Mamba and frequency analysis for accurate SR. Considering that the VSSM aggregates image features via patch-wise linear scaling, it inevitably overlooks some pixel-wise localities. To address this, we further develop a Hybrid Gate Module (HGM) to better introduce a local inductive bias. Unlike the commonly used channel attention, HGM allows for selective amplification or attenuation of local features on spatial position, effectively enhancing spatially-varying representation during channel-wise correction learning.

Furthermore, there is inherent misalignment between different level features (global and local). The direct fusion of multi-level features inevitably arises in confused and conflicted representation. To this end, we introduce a learnable adapter to rescale cross-level representation for improved integration. Overall, equipped with the above designs, our FMSR can capture both global and local dual-domain dependencies for RSI SR while maintaining moderate complexity.

Our main contribution can be summarized as follows:

1) We introduce the first state space model for remote sensing image super-resolution (FMSR), highlighting Mamba's capability for efficient and effective global modeling in large-scale remote sensing scenarios.
2) To integrate more high-frequency cues into Mamba, we develop a Frequency Selection Module (FSM), which adaptively identifies and selects the most informative frequency signals during the fast fourier transformation process.
3) We design a Hybrid Gate Module (HGM) that integrates the local bias of CNN operators with spatially-varying coordinates to enhance the locality of feature representation, leading to more accurate and faithful SR performance.

The remainder of this paper is organized as follows. Section II reviews the related knowledge pertinent to our FMSR. In Section III, we provide detailed descriptions of the implementation of our FMSR. Section IV contains extensive experiments conducted on widely used remote sensing benchmarks. Section V concludes our work.

II. Related Work

In this section, we first present a comprehensive review of remote sensing image super-resolution. Then, we introduce relevant background knowledge for this study, including state space models and frequency learning.

A. Remote Sensing Image Super-Resolution

RSI SR has witnessed significant advancements with the booming of deep learning [34], [35], [36], [37], [38]. The primary focus of this task lies in extracting prior knowledge from LR images, which can be broadly categorized into three categories: CNN-, Transformer-, and Mamba-based methods.

CNN-based: Drawing inspiration from SRCNN [39], early efforts usually elaborate CNNs with advanced modules, such as residual [40] and dense structure [41] and attention mechanisms [42], [43]. Mei et al. [22] introduced non-local sparse attention to capture global dependencies inherent in LR images. Similarly, Lei et al. [44] extended the non-local mechanism by exploring cross-scale similarity in RSI. While these methods improve the local receptive field nature of CNNs, they suffer from significant computational overhead during non-local exploration, making them less efficient in large-scale remote sensing scenes. Moreover, limited by the local bias of CNN, they cannot capture critical long-range dependencies and reach a plateau in performance.

Transformer-based: The core insight of Transformer lies in the Self-Attention (SA) mechanism [45], which has demonstrated superior long-range modeling capability and outperformed CNN-based methods. Lei et al. [46] devised a multi-stage Transformer-enhanced network. Recently, Chen et al. [31] proposed to activate more pixels in SR tasks by combining both CNN and self-attention. More Recently, an improved network [47] was proposed by integrating channel attention, improved SA, and anchored stripe attention. However, due to the quadratic complexity of SA regarding token size, they are less efficient in handling high-resolution images, large-scale RSI in particular. To alleviate the computational budget of SA, Chen et al. [30] proposed a recursive SA by recursively aggregating input features for enriched token representation. Nevertheless, efficient SA often sacrifices global modeling capability, and despite efforts to mitigate the quadratic complexity, the inherent problem of SA remains unsolved.

Mamba-based: In light of the success of Mamba, some scholars [48] attempt to introduce Mamba for efficient global modeling with linear complexity. However, to the best of our knowledge, the potential of Mamba in RSI SR remains unexplored. Since the imaging is wide-ranging, the content in RSI exhibits complex and diverse properties. Furthermore, compared to natural images, the texture information of RSI is less prominent, and the vital high-frequency information tends to vanish in deep models.

In summary, there is an urgent need for an efficient yet effective scheme to model the heterogeneous representation and to seek a practical solution to explore the critical high-frequency components. This paper pioneers exploring Mamba's potential in the RSI SR task and extends Mamba with frequency analysis, providing an effective and efficient paradigm for this challenging issue.

B. State Space Model

Recently, state space models (SSMs) [49] have emerged as a promising approach, demonstrating competitive performance in long-range modeling compared to transformers. The key advantage of SSMs lies in their linear scaling with sequence length, providing a global perspective with linear complexity. Gu et al. [50] pioneered the SSM to tackle long-sequence data modeling, illustrating promising linear scaling properties. Subsequently, they put forward a variant named Mamba [51], which adopts a selective mechanism and efficient network design. Mamba has shown superior performance to transformers in natural language processing tasks. In light of the success of Mamba, it has been introduced in computer vision tasks and demonstrated impressive performance, including object detection [52], image classification [53], and biomedical image segmentation [54]. However, research on Mamba in low-level vision tasks is still in its primary stage, and efforts for RSI SR remain unexplored.

This paper adapts Mamba for the SR task. Unlike previous works that solely replace self-attention with the Vision State Space Model (VSSM) for long-range modeling, we promote the perspective of global exploration in the spatial-frequency dual domain. Compared to spatial-wise modeling, our method is more effective by incorporating latent high-frequency cues for better global representation.

C. Fourier Transform

The Fast Fourier transform (FFT) can be viewed as a global statistical signal, making it suitable for global information analysis. In light of this, various visual tasks leverage FFT for frequency domain modeling, such as semantic segmentation [55] and image classification [56]. In low-level vision scenes, Mao et al. [57] introduced a residual FFT module capable of capturing comprehensive local high-frequency details for low-light enhancement. Li et al. [58] integrated the Fourier transform into a deep network to mitigate noise amplification during luminance enhancement. Guo et al. [59] proposed a window-based frequency channel attention mechanism. Wang et al. [60] conducted mutual learning of frequency and spatial domains to improve face image SR. Further, some approaches incorporate FFT into the loss function to enhance reconstructed sharp details [61].

However, these methods often prioritize elaborate networks for exploring frequency signals, neglecting frequency contribution analysis. This inevitably amplifies harmful frequencies and increases computational overhead. In this study, we dynamically adjust input frequency-domain features using lightweight activation weights and a convolutional layer to emphasize informative frequencies for accurate SR, which offers more flexibility to modulate selection thresholds.

III. Methodology

In this section, we introduce the implementation details of our FMSR. The FMSR comprises a straightforward backbone with convolution layers, Frequency-assisted Mamba Groups (FMG), and a pixel-shuffle layer. We start with an overview of FMSR, and then we dive into its components by explaining: the Frequency-assisted Mamba Block (FMB), Vision State Space Module (VSSM), Hybrid Gate Module (HGM), and Frequency Selection Module (FSM).

A. Overview of FMSR

As illustrated in Fig. 2, the proposed FMSR consists of three major stages: shallow feature extraction of LR, deep feature acquisition, and reconstruction of HR. Firstly, the given LR input I_{LR} is fed into a 3×3 convolution layer ϕ with learnable parameters θ to generate initial feature F_0, which can be written as:

$$F_0 = \phi(I_{LR}, \theta). \tag{1}$$

Subsequently, the F_0 undergoes deep feature extraction with multiple Frequency-assisted Mamba Groups (FMGs), followed by a 3×3 convolution for feature refinement. This process can be expressed as:

$$F_m = \mathrm{FMG}_m(\mathrm{FMG}_{m-1}(\cdots \mathrm{FMG}_1(F_0))), \tag{2}$$

where m represents the number of FMG and F_m is the output of m-th FMG. We incorporate a global skip connection

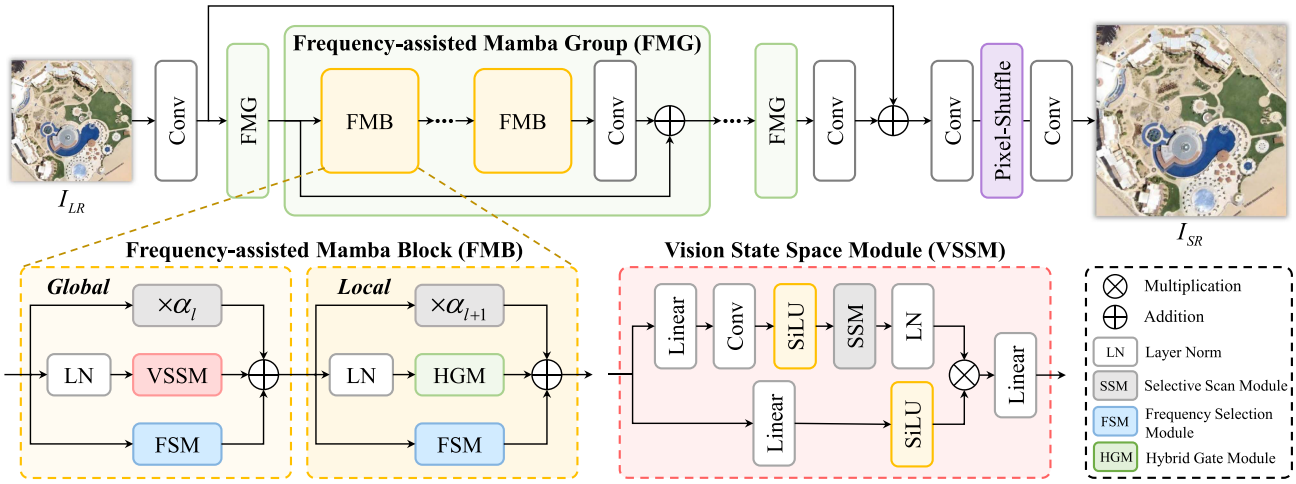


Fig. 2. Overview of the proposed FMSR. The Frequency-assisted Mamba Blocks (FMB) are arranged sequentially in Frequency-assisted Mamba Groups (FMG). In FMB, a Frequency Selection Module (FSM) is adopted to assist the learning process of the Vision State Space Module (VSSM) and Hybrid Gate Module (HGM). α_l is a learnable adaptor for hybrid adaptive integration in the l-th FMB.

to prepare high-quality features F_{rec} for reconstruction, i.e., $F_{rec} = \mathrm{Conv}(F_m) + F_0$. Finally, a 3×3 convolution, a pixel-shuffle layer PS, and a terminal convolution are employed to upscale and restore the super-resolved output I_{SR}:

$$I_{SR} = \mathrm{Conv}(\mathrm{PS}(\mathrm{Conv}(F_{rec}), s)), \tag{3}$$

where s means the upscale factor.

B. Frequency-Assisted Mamba Block

The structure of FMB is shown in Fig. 2, from which we can find that FMB serves as a primary component in FMG. In particular, each FMG contains cascaded FMB, a convolution layer, and a residual connection. The FMB is responsible for exploring global and local representations in a frequency-spatial dual domain. The function of i-th FMG can be summarized as follows:

$$F_i = \mathrm{Conv}(\Psi_n(\Psi_{n-1}(\cdots\Psi(F_{i-1})))) + F_{i-1}, \tag{4}$$

where Ψ_n denotes the n-th FMB in each FMG.

Specifically, FMB performs *global* and *local* modeling in the frequency-spatial dual domain. Given output of Ψ_{l-1}, termed x_{l-1}, the l-th FMB Ψ_l processes x_{l-1} with three parallel branches to grasp their merits of global representations: 1) A Layer Norm (LN) followed by the 2D Vision State Space Module (VSSM) is developed to capture the spatial-wise long-term information, 2) a Frequency Selection Module (FSM) is equipped to introduce more high-frequency cues in the frequency domain while promoting the capability of VSSM, 3) a learnable scaling factor α_l for dynamic feature aggregation. The feature y after global frequency-spatial exploration can be formulated as:

$$y = \alpha_l \cdot x_{l-1} + \mathrm{VSSM}(\mathrm{LN}(x_{l-1})) + \mathrm{FSM}(x_{l-1}). \tag{5}$$

After that, y undergoes further local modeling. Similarly, we adopt LN to normalize y and then use a Hybrid Gate Module (HGM) to grasp the spatial locality. Also, FSM is employed to assist the local modeling process with frequency learning. Finally, another scale factor α_{l+1} is used to adaptively integrate the output from local and global representations, which can be expressed as:

$$x_l = \alpha_{l+1} \cdot y + \mathrm{HGM}(\mathrm{LN}(y)) + \mathrm{FSM}(y). \tag{6}$$

C. Vision State Space Module

Previous efforts often rely on Transformers to explore global dependency, which calculate the long-range response with the self-attention mechanism. Despite achieving favorable performance, they suffer from high complexity, hindering the efficient modeling in large-scale remote sensing images. Inspired by the success of the vision state space module in long-term modeling and aggregation with linear complexity, we first introduce VSSM to the RSI SR task.

In particular, as illustrated in Fig. 2, the normalized feature $x_N = \mathrm{LN}(x_l)$ is expanded along the channel dimension by a linear projection operation ϕ_1 with an expansion factor λ. Then, a series of operations including a 1×1 Depth-Wise Convolution (DWConv) $f_{1\times1}$, a SiLU activation σ_1, as well as the 2D-selective scan module (SSM) and LN are sequentially stacked to generate the output of the first branch, denoted h_1. This branch can be defined as:

$$h_1 = \mathrm{LN}(\mathrm{SSM}(\sigma_1(f_{1\times1}(\phi_1(x_N, \lambda))))). \tag{7}$$

In the second branch, another linear layer ϕ_2 and SiLU function σ_2 are used. The output of this branch can be obtained by:

$$h_2 = \sigma_2(\phi_2(x_N, \lambda)). \tag{8}$$

Finally, to produce the final output h_{out}, the output h_1 and h_2 are incorporated via Hadamard product, followed by a linear layer ϕ_3. That is:

$$h_{out} = \phi_3(h_1 \otimes h_2). \tag{9}$$

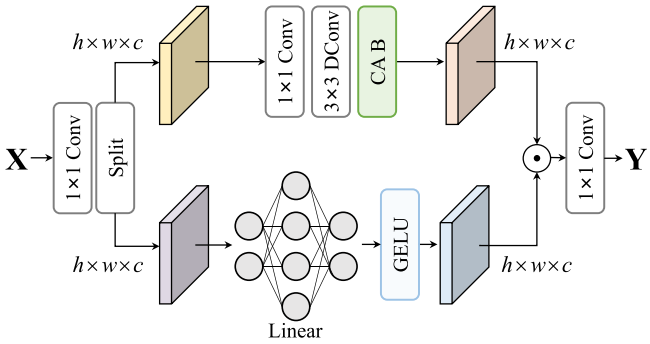


Fig. 3. The proposed Hybrid Gate Module (HGM) conceptual illustration. The input feature **X** is split in the channel dimension and fed through a Channel Attention Block (CAB) and a pixel-wise linear projection layer, respectively. After a Hadamard product operation, a 1×1 convolution generates the output tensor **Y**.

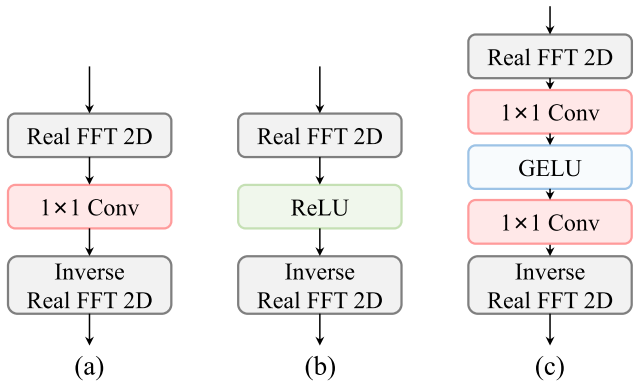


Fig. 4. Three variants of Frequency Selection Module (FSM). Here, we adopt 2D Fast Fourier Transformation (FFT) for frequency learning.

D. Hybrid Gate Module

Current methods often utilize either MLP layers [31] for feature propagation or incorporate convolution operations, such as attention mechanisms [48], after long-range exploration to introduce critical locality for improved performance. Our approach draws inspiration from these methods but condenses them into a unified hybrid module. To learn a comprehensive local context, the Hybrid Gate Module (HGM) incorporates the spatially-varying properties of remote sensing images (RSI) by selectively amplifying or attenuating local features in the pixel domain while preserving channel-specific features.

As shown in Fig. 3, HGM treats features captured by local convolution as coordinates and then multiplies them with a pixel-wise gating mask of the same size. Specifically, the input feature X is first processed through a 1×1 convolution to expand the channel dimension to $2c$. We then split X into two parts, X_1 and X_2, by halving the channel dimension. X_1 and X_2 are subsequently sent into the first and second branches, respectively. In the first branch, a 1×1 convolution, a 3×3 depth-wise convolution, and a channel attention block [42] are used to yield the coordinates.

$$\mathrm{X}_{coor} = \mathrm{CA}(\mathrm{Dconv}(\mathrm{Conv}(\mathrm{X}_1))). \tag{10}$$

In the second branch, we fed the feature through pixel-wise linear projection, somewhat similar to the MLP layer. In contrast to the MLP layer, we only activate the feature at the end of linear projection with the GELU activation function to generate the gate weights:

$$\mathrm{M} = \mathrm{GELU}(\mathrm{Linear}(\mathrm{X}_2)). \tag{11}$$

Finally, the output Y can be obtained by:

$$\mathrm{Y} = \mathrm{Conv}(\mathrm{M} \odot \mathrm{X}_{coor}). \tag{12}$$

E. Frequency Selection Module

To achieve frequency-spatial dual domain representation at global and local levels, we equip VSSM and HGM with frequency exploration. As illustrated in Fig. 4, we devise three variants of frequency selection operations using Fast Fourier transformation (FFT):

a) Using 2D real FFT and using a 1×1 convolution layer before inverse FFT, which means we do not perform frequency selection:

$$\mathrm{Z} = \mathcal{F}^{-1}(\mathrm{Conv}(\mathcal{F}(x))). \tag{13}$$

b) Inserting ReLU activation between FFT and Inverse FFT to dynamically select the frequency pattern:

$$\mathrm{Z} = \mathcal{F}^{-1}(\mathrm{ReLU}(\mathcal{F}(x))). \tag{14}$$

c) Applying two stacks of 1×1 convolution layer and GELU activation function:

$$\mathrm{Z} = \mathcal{F}^{-1}\mathrm{Conv}((\mathrm{GELU}(\mathrm{Conv}(\mathcal{F}(x))))). \tag{15}$$

We finally choose scheme (c) as our FSM as 1×1 convolution lets the network modulate flexible thresholds for frequency selection with lightweight design. Ablation experiments demonstrate the effectiveness of (c) compared to (a) frequency analysis without selection and (b) ReLU-based selection.

IV. Experiment

A. Datasets

In this paper, we report the results of the SR performance on three RSI benchmarks, including AID [62], DOTA [63], and DIOR [65]. In particular, AID is used to form the training and test simultaneously. We randomly select 3000 and 900 images from AID to form the training and test set, with an image size of 640×640. Note that the training and test parts of AID are non-overlapping. In DOTA and DIOR, 900 and 1000 images are randomly extracted for model evaluation, respectively. Both of them are with a size of 512×512.

B. Implementation Details

Model Details: This paper focuses on $\times 4$ SR. Our FMSR is constructed by 6 FMGs for deep feature exploration, with each FMG consisting of 6 FMBs, i.e., $m = n = 6$. Empirically, we set the internal channel dimension to $c = 96$. ll convolutional kernel sizes are set to 3×3, except for those in the Hybrid Gate

TABLE I
ABLATION STUDIES OF DIFFERENT COMPONENTS ON THE PROPOSED FMSR

Components	Model-1 (Base)	Model-2	Model-3	Model-4	Model-5	Model-6 (FMSR)
Window-based Self-Attention [31]	✓	✗	✗	✗	✗	✗
2D-VSSM (Ours)	✗	✓	✓	✓	✓	✓
MLP Layer [30]	✗	✗	✓	✗	✗	✗
Channel Attention [42]	✗	✗	✗	✓	✗	✗
Hybrid Gate Module (Ours)	✗	✗	✗	✗	✓	✓
Frequency Selection Module (Ours)	✗	✗	✗	✗	✗	✓
PSNR (dB)	27.751	27.846	28.104	28.088	<u>28.122</u>	**28.178**

The best and second-best PSNR performances are highlighted in **bold** and <u>underlined</u>. Note that we use residual blocks in model-1 and model-2 as feed-forward networks.

Module (HGM) and Frequency Selection Module (FSM), which utilize 1×1 kernels for increased efficiency. The expansion rate in the linear projection layer is set to $\lambda = 2$.

Training Details: During the training procedure, all the SR methods were retrained on the AID training set using L1 loss and ADAM algorithm with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. To train our FMSR, we randomly select 4 image patches with the size of 64×64 in each mini-batch. The learning rate is initialized to 1×10^{-4} and halved every 200 epochs until training stops at 500 epochs. All SR models were implemented in the PyTorch framework and trained on a single NVIDIA RTX 3090 GPU with 24 GB memory and a 3.40 GHz AMD Ryzen 5700X CPU.

C. Evaluation Metrics

Two classical full-reference indicators are used to evaluate the SR performance, i.e., Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [68]. Moreover, the LPIPS metric [69] is employed to analyze the perceptual quality of super-resolved results. Note that PSNR and SSIM are calculated on the luminance channel (Y) of YCbCr space.

D. Ablation Study

In this section, we discuss the proposed FMSR in depth by investigating the effect of its major components and their variants. All these models are trained on AID with scale factor $\times 4$. Following prior work [70], we employ a small-scale dataset, AID-tiny, consisting of 30 randomly selected images from AID, for efficient evaluation unless otherwise specified. The baseline model is derived by excluding FSM and substituting VVSM and HGM with the standard window-based self-attention and 5 residual blocks, respectively.

1) Effect of Key Modules: (a) Effect of VSSM: Table I reports that the Model-1 (baseline) obtains 27.751 dB. Model-2 exhibits a performance gain of 0.089 dB over the baseline model, which demonstrates the effectiveness of VSSM in global modeling. To demonstrate the linear complexity of VSSM, we conducted complexity experiments with inputs of different resolutions, as shown in Fig. 8. Specifically, we adopted the standard Multi-head Self-Attention (MSA) [71] with a dimension of 180 as the baseline and adjusted our model to have a dimension of 144. This adjustment ensured that the complexity in terms of parameters (0.1308 M vs. 0.1279 M) and FLOPs (0.1534 G vs. 0.1312 G) was roughly equivalent. Our method is more efficient

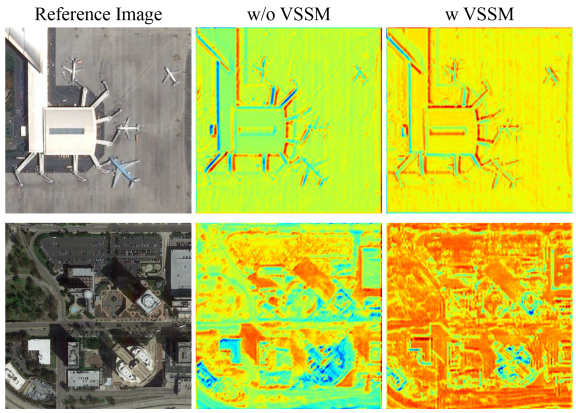


Fig. 5. Feature visualization comparisons. The feature maps corresponding to each reference image are the results of the 56-*th* channels in the final FMG.

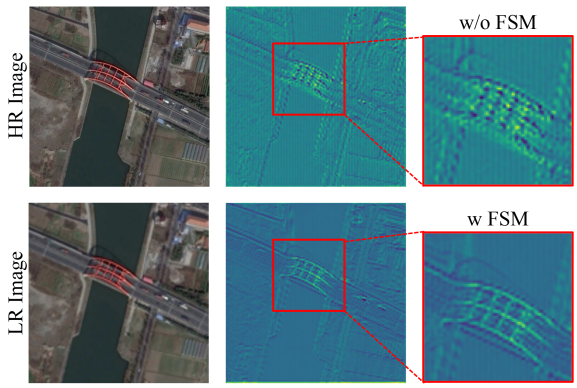


Fig. 6. Visualization of the feature maps. The proposed Frequency Selection Module (FSM) yields sharp and clear details for reconstruction.

than the widely used MSA and exhibits linear complexity with input resolution. In addition, in Fig. 5, we visualize the intermediate feature maps of the baseline model and our FMSR. The features are obtained by visualizing the 56th channels at the end of FMG. With VSSM, the results of our FMSR are more prominent across the entire feature maps, highlighting the favorable global modeling capability.

(b) Effect of HGM: Employing both VSSM and the MLP layer, Model-3 demonstrates a favorable performance improvement of 0.258 dB. However, when equipped with the channel attention

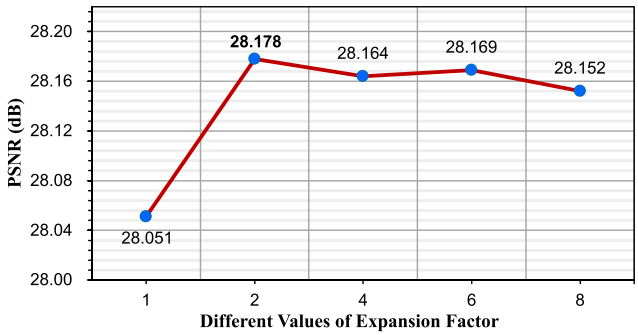


Fig. 7. Ablation studies on the effect of different expansion factors λ.

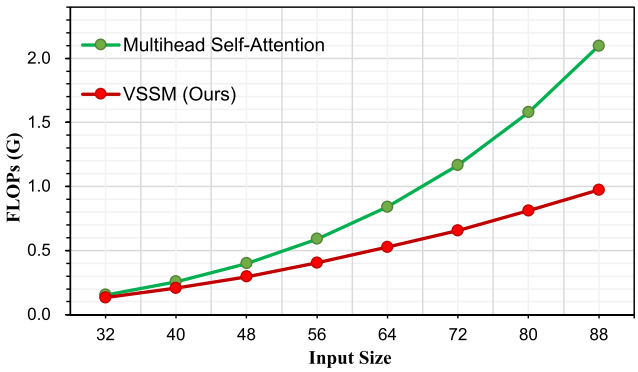


Fig. 8. Computational complexity comparison with inputs of different resolutions. We adopt the standard Multihead Self-Attention [1] as the baseline. Initially, we adjust the model to ensure that the consumption of parameters and FLOPs is roughly equivalent. Subsequently, we increase the input resolution from 32×32 to 88×88.

TABLE II
ABLATION STUDIES ON DIFFERENT VARIANTS OF FREQUENCY SELECTION MODULE (FSM) AS ILLUSTRATED IN FIG. 4

Variants	FSM (a)	FSM (a)	FSM (a)
PSNR (dB)	28.129	<u>28.147</u>	**28.178**

The best and second-best PSNR performances are highlighted in **bold** and <u>underline</u>, respectively.

mechanism to introduce more locality, Model-4 performs similarly to Model-2 (28.088 dB vs. 28.104 dB). This suggests that neither MLP nor channel attention is less effective in boosting the performance of Mamba. In this context, further improving the reconstruction performance becomes very challenging. Our HGM achieves a performance gain of 0.034 dB. Thus, to benefit from both global and local inductive bias, we combine VSSM and HGM in our FMSR. In addition, we have investigated the impact of different expansion factors λ used in the linear projection layer. The quantitative results are presented in Fig. 7. It is observed that the performance of FMSR does not fluctuate significantly with changes in λ. We ultimately selected $\lambda = 2$ as it delivers the best performance.

(c) Effect of FSM: By comparing the results of FMSR and Model-5, we can evaluate the performance of the proposed FSM. In this case, our FSM can bring an improvement of 0.056 dB. As reported in Table II, we discuss some variants of FSM shown in Fig. 4. If we do not perform frequency selection, FSM(a) produces the worst PSNR performance. By inserting a simple ReLU activation, FSM(b) could adaptively eliminate noisy frequencies and achieve a gain of 0.018 dB, demonstrating the effectiveness of the selection mechanism. If we adopt 1×1 convolution and GELU for selection, FMSR allows for a more flexible threshold for frequency selection, thus obtaining the best performance. Moreover, visual comparisons between FMSR and Model-5 are shown in Fig. 6. In Fig. 6, our FMSR generates superior textures in the high-frequency of the bridge. In contrast, without performing frequency selection, the intermediate features are blurred at the boundary and unclear at the edge. These results confirm that our FSM has the ability to explore more critical high-frequency cues for better SR performance.

TABLE III
ABLATION STUDIES ON HYBRID ADAPTIVE INTEGRATION (HAI) OF GLOBAL AND LOCAL REPRESENTATIONS

Method	Skip	Adaptive α	#Param.	PSNR (dB)	SSIM
w/o HAI			11.75M	<u>30.80</u>	<u>0.8121</u>
w/ Skip	✓		11.75M	30.57	0.7986
w/ HAI	✓	✓	11.76M	**30.93**	**0.8156**

The best and second best PSNR/SSIM performance are highlighted in **bold** and <u>underline</u>, respectively. We report these results on AID [62].

TABLE IV
ABLATION STUDIES FOR THE NUMBER OF FMB

Number of FMB	PSNR (dB)	#Param.	GFLOPs
2	27.744	4.55M	58.09G
4	28.026	8.15M	93.18G
6	<u>28.178</u>	11.76M	128.27G
8	**28.181**	15.37M	163.36G

The best and second-best PSNR performances are highlighted in **bold** and <u>underline</u>, respectively.

2) Effect of HAI: We show the influence of HAI in Table III, where we conduct three ablation analyses: 1) without HAI, 2) with residual connection (Skip), and with HAI. We observe that without HAI, the performance of FMSR drops by 0.13 dB. Additionally, comparing the model with Skip, our FMSR achieves a significant improvement of 0.36 dB. This may be because of the misalignment between global and local representations, which means simply adding them may not elaborate enough to integrate these different levels of knowledge, thus generating suboptimal performance. Benefiting from the adaptive scaling factor α, our FMSR could dynamically adjust the features at the global and local ranges, thus generating enriched feature integration.

3) Model Efficiency: The parameters, FLOPs, and SR performance of state-of-the-art (SOTA) methods are reported in Table VII. Intuitively, we plot the performance versus parameters in Fig. 9(c), where we observe that FMSR strikes a favorable trade-off between performance and parameters. Here, we further investigate the relationship between network structure and model complexity. Moreover, more intuitive metrics are involved to analyze the model efficiency, such as inference times and memory consumption.

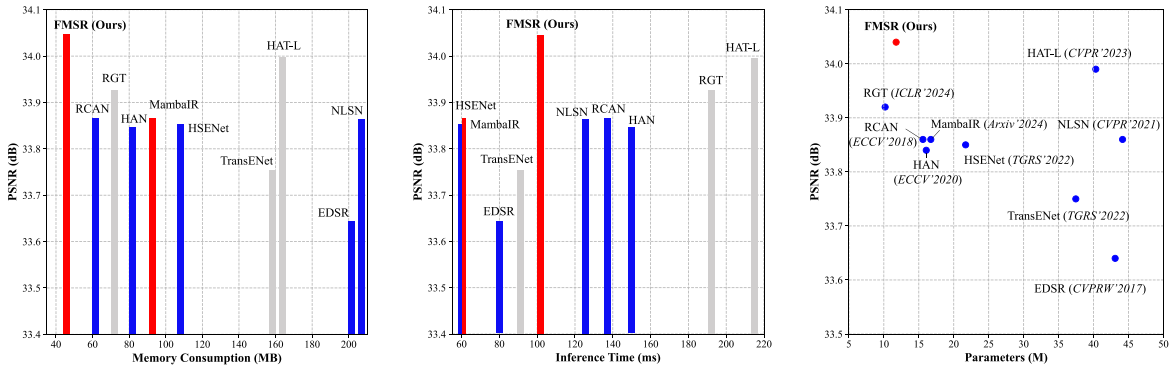


Fig. 9. Ablation studies of memory consumption, inference times, parameters, and PSNR performance on DOTA [63]. Note that the inference times are calculated on 100 images.

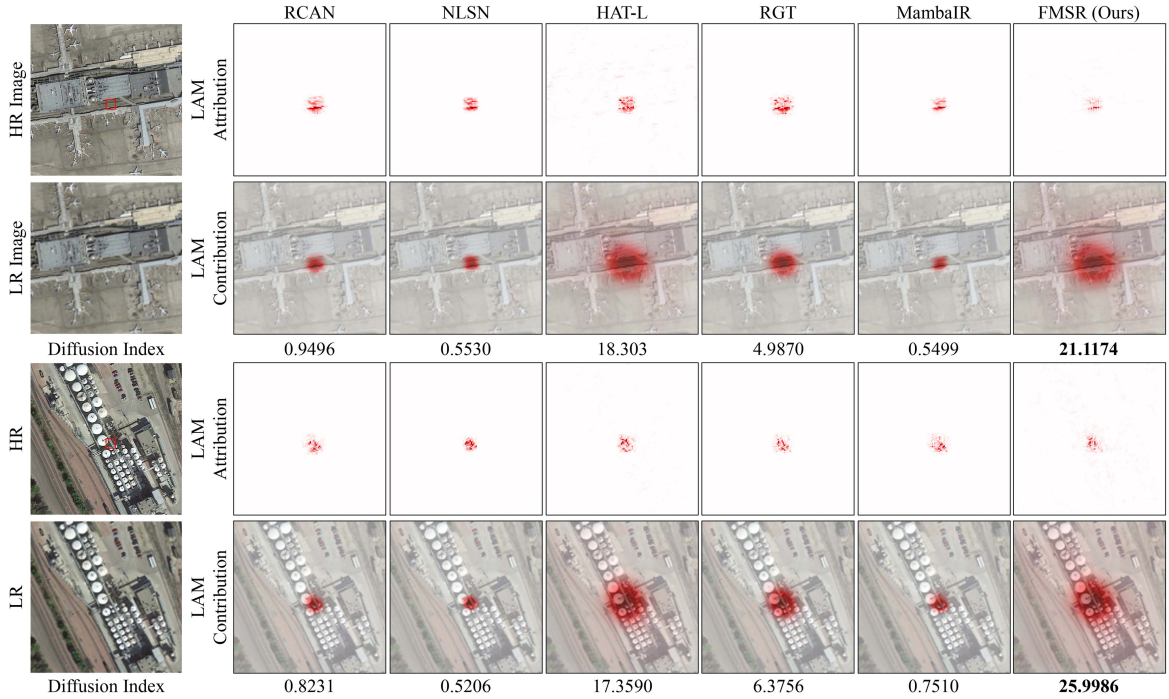


Fig. 10. The visualization of Local Attribution Maps (LAM) [67]. A wider range of LAM illustrates more pixels are involved in reconstruction. A higher diffusion index demonstrates better global activation capability.

TABLE V
ABLATION STUDIES FOR THE NUMBER OF FMG

Group	2	4	6	8
#Param.	4.21M	7.99M	11.76M	15.54M
GFLOPs	52.66G	90.46G	128.27G	166.07G
PSNR (dB)	27.691	27.985	**28.178**	<u>28.154</u>

The best and second-best PSNR performances are highlighted in **bold** and <u>underline</u>, respectively.

(a) Number of FMB and FMG: We study the inference of the number of FMB and FMG in Tables IV and V, respectively. As we can see in Table IV, using more FMB leads to the consistently increasing PSNR performance from 27.744 to 28.181 dB, demonstrating the effectiveness of FMB. Meanwhile, the parameters grow dramatically from 4.55 to 15.37 M. To strike a favorable balance between performance and model size, we finally picked 6 FMB in our FMSR. Regarding the number of FMG, as shown in Table V, the increasing number of groups leads to higher PSNR values. Nevertheless, the performance saturates at group 8, which may be caused by over-fitting. Ultimately, we insert 6 FMG in the FMSR for deep feature learning.

(b) Memory Consumption: The max CUDA memory consumption of our FMSR and SOTA models are shown in Fig. 9(a), from which we can see that our consumes the least memory consumption while outperforming other methods. Specifically, compared to competitive NLSN that adopts non-local attention for global modeling, we find the performance of FMSR is 0.18 dB higher, but also reduces the memory by 160 MB.

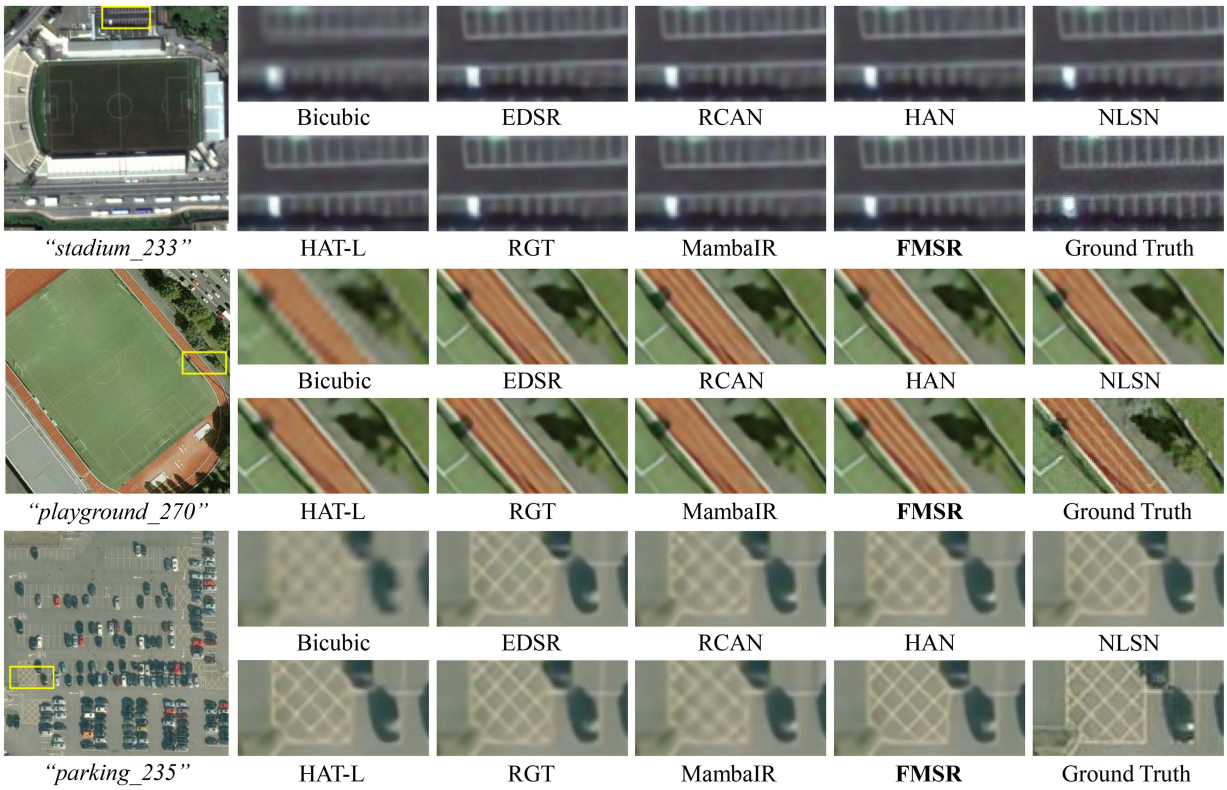


Fig. 11. Visual comparisons of our FMSR with CNN-, Transformer-, and Mamba-based methods on AID [62] with scale ×4. Zoom in for better observation.

TABLE VI
MODEL EFFICIENT COMPARISON

Metrics	EDSR [64]	RCAN [42]	HAN [43]	HSENet [44]	NLSN [22]	TransENet [46]	HAT-L [31]	RGT [30]	MambaIR [48]	FMSR
Parameters (M)	43.09	15.59	16.07	21.7	44.15	37.46	40.32	10.19	16.72	11.76
FLOPs (G)	823.34	261.01	268.89	306.31	840.79	87.85	672.15	40.12	280.82	128.27
Memory (MB)	202	62	82	108	206	158	164	72	92	46
Inference Times (ms)	79.188	136.71	149.24	59.95	125.42	91.09	214.56	191.48	60.99	100.88

The FLOPs results are calculated with an input tensor of size 1 × 3 × 128 × 128. The inference times are tested on 100 random images.

TABLE VII
QUANTITATIVE COMPARISON ON AID [62], DOTA [63], AND DIOR [65] TEST SET IN TERMS OF PSNR, SSIM, AND LPIPS

Methods	Venue	AID [62]			DOTA [63]			DIOR [65]			Average		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Bicubic	-	28.86	0.7382	0.4803	31.16	0.7947	0.4043	28.57	0.7432	0.4678	29.53	0.7587	0.4508
EDSR [64]	*CVPRW'2017*	30.65	0.8086	0.3068	33.64	0.8648	0.2616	30.63	0.8116	0.3020	31.64	0.8283	0.2901
RDN [41]	*CVPR'2018*	30.74	0.8112	0.3157	33.60	0.8670	0.2421	30.78	0.8147	0.3093	31.71	0.8310	0.2890
RCAN [42]	*ECCV'2018*	30.82	0.8121	0.3112	33.86	0.8680	0.2384	30.85	0.8159	0.3048	31.84	0.8320	0.2848
HAN [43]	*ECCV'2020*	30.80	0.8122	0.3079	33.84	0.8682	0.2363	30.84	0.8163	0.3026	31.83	0.8322	0.2823
HSENet [44]	*TGRS'2022*	30.72	0.8108	0.3124	33.85	0.8667	0.2412	30.77	0.8143	0.3070	31.78	0.8306	0.2869
NLSN [22]	*CVPR'2021*	30.81	0.8126	0.3044	33.86	0.8682	0.2349	30.82	0.8156	0.3004	31.83	0.8321	0.2799
HAUNet [66]	*TGRS'2023*	30.88	0.8132	0.3111	33.94	0.8687	0.2403	30.87	0.8160	0.3074	31.90	0.8326	0.2863
TransENet [46]	*TGRS'2022*	30.80	0.8109	0.3136	33.75	0.8675	0.2418	30.85	0.8148	0.3089	31.80	0.8311	0.2881
HAT-L [31]	*CVPR'2023*	30.81	0.8124	0.3078	**33.99**	0.8684	0.2392	30.87	0.8161	0.3062	31.89	0.8323	0.2844
GRL-L [47]	*CVPR'2023*	30.86	0.8127	0.3085	33.86	**0.8710**	0.2372	30.90	0.8177	0.3047	31.87	0.8338	0.2835
RGT [30]	*ICLR'2024*	**30.91**	**0.8159**	**0.3023**	33.92	**0.8709**	**0.2337**	**30.91**	**0.8182**	**0.2992**	**31.91**	**0.8350**	**0.2784**
MambaIR [48]	*Arxiv'2024*	30.85	0.8130	0.3098	33.86	0.8691	0.2388	30.89	0.8167	0.3060	31.87	0.8329	0.2849
FMSR (Ours)	*TMM'2024*	**30.93**	**0.8156**	**0.3001**	**34.04**	**0.8710**	**0.2325**	**30.97**	**0.8187**	**0.2983**	**31.98**	**0.8351**	**0.2770**
FMSR++ (Ours)	*TMM'2024*	**31.07**	**0.8185**	**0.3067**	**34.27**	**0.8735**	**0.2363**	**31.13**	**0.8219**	**0.3039**	**32.16**	**0.8380**	**0.2823**

FMSR++ means the self-embedding results of our FMSR.
Where the 1st, 2nd, and 3rd best performance are highlighted in red, blue, and green, respectively.

TABLE VIII
QUANTITATIVE COMPARISON WITH SOTA CNN-, TRANSFORMER-, AND MAMBA-BASED SR METHODS ACROSS 30 SCENE CATEGORIES ON AID [62]

Categories	Bicubic		EDSR [64]		HSENet [44]		NLSN [22]		HAT-L [31]		MambaIR [48]		RGT [30]		**FMSR (Ours)**	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Airport	27.83	0.7554	29.93	0.8282	30.08	0.8303	30.16	0.8322	30.15	0.8319	30.21	0.8327	**30.29**	**0.8352**	**30.26**	**0.8346**
Bare Land	35.60	0.8564	36.94	0.8837	36.79	0.8841	**37.00**	0.8845	36.88	0.8841	36.95	0.8843	37.00	**0.8852**	**37.02**	**0.8847**
Baseball Field	31.00	0.8305	33.05	0.8765	33.15	0.8774	33.24	0.8787	33.25	0.8789	33.32	0.8792	**33.35**	**0.8805**	**33.43**	**0.8803**
Beach	32.90	0.8446	34.18	0.8727	34.14	0.8746	34.31	0.8749	34.34	0.8756	34.36	0.8754	**34.41**	**0.8766**	**34.38**	**0.8759**
Bridge	30.22	0.8283	32.93	0.8800	33.06	0.8809	33.12	0.8818	33.04	0.8809	33.14	0.8821	**33.22**	**0.8832**	**33.28**	**0.8836**
Center	26.51	0.6944	28.77	0.7921	28.83	0.7937	28.95	0.7971	28.92	0.7956	28.99	0.7966	**29.09**	**0.8015**	**29.09**	**0.8010**
Church	24.29	0.6333	26.30	0.7469	26.47	0.7507	26.51	0.7528	26.56	0.7532	26.61	0.7543	**26.66**	**0.7580**	**26.64**	**0.7567**
Commercial	27.33	0.7174	29.01	0.7940	29.21	0.7989	29.21	0.7996	29.21	0.8007	29.29	0.8012	**29.30**	**0.8048**	**29.34**	**0.8039**
D-Residential	22.93	0.5671	24.38	0.6839	24.60	0.6912	24.60	0.6936	24.67	0.6936	24.74	0.6965	**24.78**	**0.7019**	**24.78**	**0.7019**
Desert	39.26	0.9100	40.20	0.9268	39.57	0.9271	40.27	0.9278	**40.37**	**0.9278**	40.09	0.9270	40.15	**0.9283**	**40.29**	0.9275
Farmland	33.10	0.8226	35.00	0.8683	35.02	0.8692	35.10	0.8699	35.03	0.8691	35.09	0.8696	**35.14**	**0.8715**	**35.18**	**0.8711**
Forest	28.79	0.6605	29.85	0.7315	30.00	0.7363	29.98	0.7369	**30.01**	0.7363	30.00	0.7370	30.01	**0.7400**	**30.04**	**0.7391**
Industrial	26.77	0.6952	28.88	0.7931	28.98	0.7956	29.04	0.7982	29.04	0.7980	29.09	0.7988	**29.20**	**0.8034**	**29.17**	**0.8027**
Meadow	33.86	0.7483	34.63	0.7804	34.55	0.7804	34.69	0.7821	**34.70**	0.7815	34.64	0.7811	34.69	**0.7831**	**34.74**	**0.7827**
M-Residential	26.36	0.6335	28.34	0.7365	28.45	0.7390	28.49	0.7418	28.46	0.7408	28.59	0.7435	**28.66**	**0.7479**	**28.65**	**0.7472**
Mountain	29.51	0.7349	30.63	0.7885	30.72	0.7907	30.74	0.7916	**30.78**	0.7923	30.76	0.7918	30.71	**0.7941**	**30.78**	**0.7928**
Park	29.06	0.7530	30.54	0.8130	30.71	0.8167	30.71	0.8177	30.71	0.8189	30.76	0.8185	**30.81**	**0.8211**	**30.80**	**0.8204**
Parking	24.24	0.7060	27.25	0.8317	27.32	0.8341	27.57	0.8408	27.56	0.8405	27.65	0.8414	**27.79**	**0.8472**	**28.02**	**0.8524**
Playground	32.64	0.8450	35.37	0.8943	35.46	0.8952	35.58	0.8967	35.49	0.8959	35.54	0.8961	**35.65**	**0.8978**	**35.72**	**0.8978**
Pond	30.70	0.8167	32.11	0.8542	32.17	0.8549	32.22	0.8559	32.18	0.8555	**32.24**	0.8559	**32.24**	**0.8567**	32.23	**0.8560**
Port	26.67	0.7986	28.50	0.8596	28.71	0.8623	28.71	0.8631	28.81	0.8638	28.83	0.8644	**28.90**	**0.8667**	**28.92**	**0.8668**
Railway Station	26.78	0.6793	28.72	0.7738	28.84	0.7762	28.89	0.7783	28.88	0.7780	28.99	0.7802	**29.03**	**0.7833**	**29.04**	**0.7830**
Resort	26.79	0.7029	28.52	0.7799	28.64	0.7825	28.68	0.7845	28.71	0.7849	28.76	0.7857	**28.80**	**0.7889**	**28.82**	**0.7886**
River	30.37	0.7402	31.55	0.7891	31.61	0.7904	31.64	0.7914	31.63	0.7909	31.64	0.7912	**31.67**	**0.7927**	**31.66**	**0.7918**
School	27.41	0.7237	29.36	0.8044	29.51	0.8074	29.55	0.8097	29.54	0.8104	29.62	0.8113	**29.71**	**0.8150**	**29.70**	**0.8143**
S-Residential	26.66	0.6006	27.71	0.6728	27.84	0.6754	27.84	0.6767	27.88	0.6759	27.88	0.6768	**27.91**	**0.6796**	**27.91**	**0.6789**
Square	28.55	0.7391	30.84	0.8200	30.94	0.8223	31.04	0.8244	31.00	0.8251	31.05	0.8247	**31.16**	**0.8284**	**31.16**	**0.8276**
Stadium	27.16	0.7547	29.63	0.8387	29.68	0.8391	29.79	0.8422	29.77	0.8422	29.83	0.8421	**29.94**	**0.8457**	**29.97**	**0.8460**
Storage Tanks	25.65	0.6793	27.44	0.7664	27.58	0.7688	27.61	0.7709	27.60	0.7698	27.64	0.7709	**27.71**	**0.7744**	**27.71**	**0.7739**
Viaduct	26.97	0.6755	28.99	0.7757	29.08	0.7772	29.17	0.7813	29.11	0.7794	29.19	0.7810	**29.26**	**0.7851**	**29.26**	**0.7840**
Average	28.86	0.7382	30.65	0.8086	30.72	0.8108	30.81	0.8126	30.81	0.8124	30.85	0.8130	**30.91**	**0.8159**	**30.93**	**0.8156**

Where the 1st, 2nd, and 3rd best performance are highlighted in red, blue, and green, respectively.

Furthermore, FMSR achieves better performance against impressive HAT-L with only 28% of its memory. This indicates that FMSR has stronger global modeling capability and greatly surpasses Transformer-based models in model efficiency.

(c) Inference Times: Regarding the inference times, as shown in Fig. 9(b), our FMSR shows a trade-off with other methods. For instance, FMSR achieves the best PSNR performance compared to the advanced Transformer-based model RGT by 0.12 dB, but at the cost of increased inference time usage of about 90.6 ms, measured by the test times of 100 images. These results provide intuitive evidence that our FMSR is very efficient in large-scale remote sensing image SR tasks.

E. Comparisons With State-of-the-Art

1) Comparative Methods: To evaluate the SR performance of our FMSR against SOTA methods on remote sensing images, advanced CNN-, Transformer-, and Mamba-based models are involved for comprehensive comparison, including EDSR [64], RDN [41], RCAN [42], HAN [43], NLSN [22], HSENet [44], TransENet [46], HAT-L [31], GRL-L [47], RGT [30], and MambaIR [48]. We also report the self-ensemble results of our FMSR, dubbed FMSR++.

2) Quantitative Evaluations: The quantitative results on the AID, DOTA, and DIOR datasets are shown in Table VII. Our FMSR++ and FMSR achieve the highest and second-highest average performance on all metrics, demonstrating their superior SR performance across various remote sensing benchmarks. Particularly on the AID dataset, FMSR generates a substantial gain of 0.12 dB PSNR over HAT-L with only 29% parameters and 19% FLOPs. Compared to GRL-L, FMSR obtains 0.18 dB higher PSNR on the DOTA dataset with 60% lower complexity. Our FMSR exhibits 0.3001 LPIPS on AID, which is much higher than other comparative approaches. Moreover, recent Mamba-based methods, like MambaIR, achieved marginal improvements against CNN-based models. For instance, compared to NLSN, which explores global dependency with non-local attention, MambaIR only leads NLSN by 0.04 dB on AID. This underscores the effectiveness of VSSM in exploring non-local dependency. However, it also indicates that spatial-wise global modeling reaches a plateau in complex remote sensing images. In contrast to simply employing VSSM for global modeling, our FMSR seeks a more practical solution in the frequency domain, thus achieving significant improvement over MambaIR.

In addition, to validate the generalization capability of SR models on diverse remote sensing senses, we further report the PSNR and SSIM results on AID across 30 scene categories. The results are shown in Table VIII. As we can see, FMSR demonstrates stronger generalization capability against SOTA methods and obtains the best performance on almost all remote sensing scenarios, receiving 0.18 dB PSNR and 0.039 SSIM over MambaIR on the "Industrial" scene. Furthermore, in the "Stadium" category, our FMSR significantly outperforms HAT-L by a large margin. These notable improvements obtained by FMSR align with our motivation, which aims to introduce Mamba for efficient yet favorable global modeling in large-scale remote sensing images.

3) Qualitative Results: Visual comparisons on AID, DOTA, and DIOR with a scale factor of $\times 4$ are shown in Figs. 11, 12,

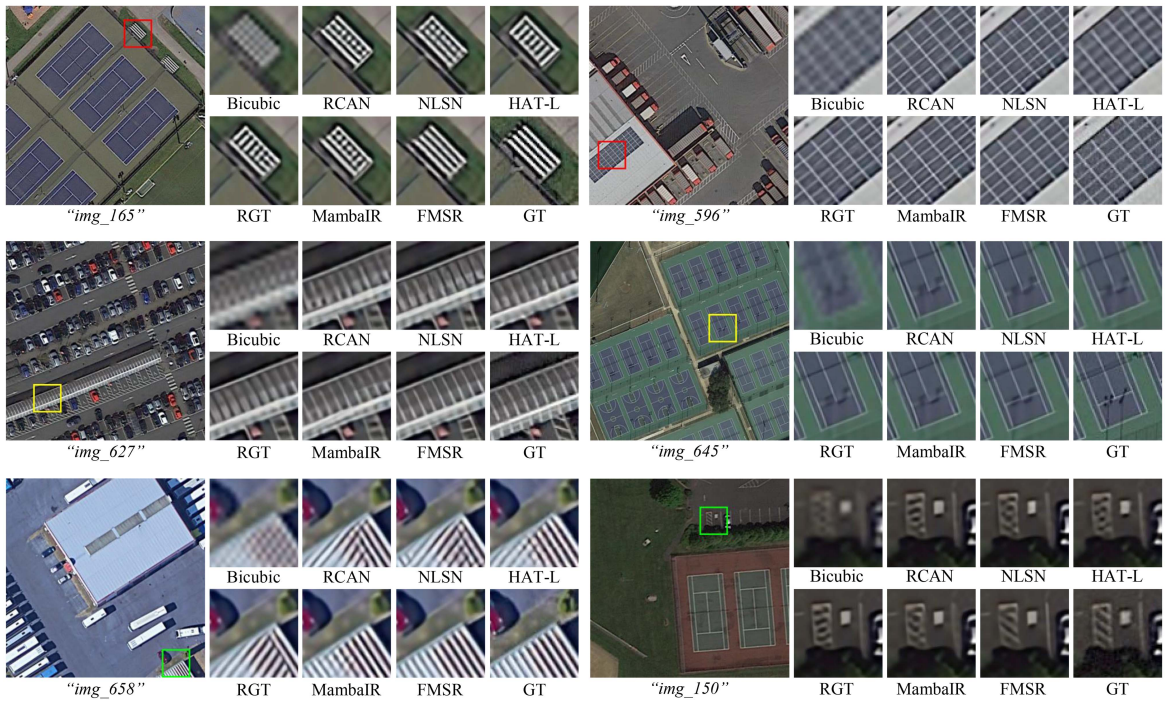


Fig. 12. Visual comparisons of our FMSR with CNN-, Transformer-, and Mamba-based methods on DOTA [63] with scale ×4. Zoom in for better observation.

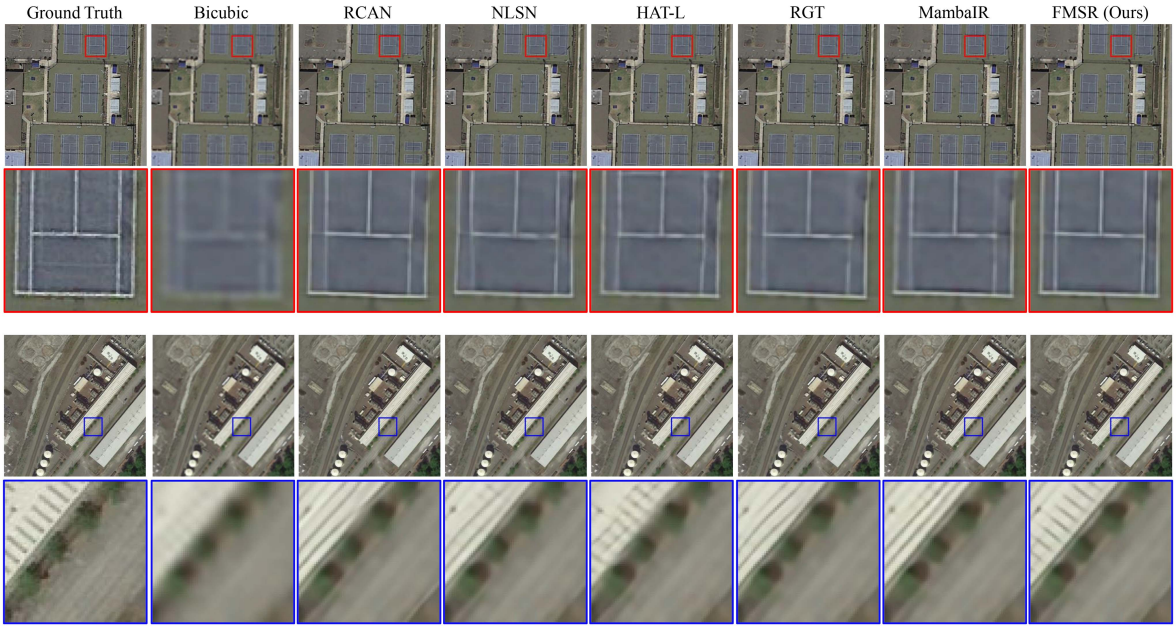


Fig. 13. Visual comparisons of our FMSR with CNN-, Transformer-, and Mamba-based methods on DIOR [65] with scale ×4. Zoom in for better observation.

and 13. From these visualizations, we observe that our FMSR recovers sharp edges with richer textures, especially capturing critical high-frequency details in these remote sensing images. For example, comparing the restored "playground_270" in Fig. 11, we can see that the recent competitive Transformer-based method HAT-L and Mamba-based model MambaIR struggle to reconstruct the HR lines on the runway. Moreover, as illustrated in "img_165" and "img_658" of Fig. 12, only the proposed FMSR can recover severely damaged lines on the ground, while other SR models fail to produce accurate distribution on the details. These visual comparisons further demonstrate the effectiveness of our FMSR in capturing and reconstructing fine details in RSI.

When comparing the visual results of the top image from the DIOR dataset in Fig. 13, where large-scale and global information exists, we observe that all CNN-, Transformer-, and

Mamba-based SR networks struggle to handle this issue, resulting in suboptimal results and losing some high-frequency textures. In contrast, benefiting from the global modeling capability of VSSM as well as the spatial-frequency dual-domain exploration, our FMSR produces results that are visually close to the ground truth and successfully generate multi-frequency lines. For example, MambaIR without frequency selection cannot recover accurate details, while HAT-L produces severe distortion. Our FMSR still maintains favorable visual quality with more high-frequency contextual information. Similarly, the visual comparisons in the bottom image of Fig. 13 provide additional evidence of the superiority of our FMSR.

In addition, the LAM comparisons shown in Fig. 10 demonstrate that FMSR can exploit more pixels during the SR process thanks to the favorable long-range modeling capability of VSSM. By comparing FMSR with HAT-L, FMSR surpasses HAT-L by 8.637 in terms of the diffusion index, indicating the strong wide-range pixel utilization capability due to our spatial-frequency dual-domain representation.

V. Conclusion

In this study, we first introduce the state space model for remote sensing image super-resolution. FMSR effectively models global dependencies in large-scale remote sensing images while enjoying the linear complexity. Specifically, we develop an efficient yet effective frequency selection module (FSM) to incorporate more relevant frequencies during spatial-frequency dual-domain learning. Meanwhile, channel-wise attention metrics are linearly scaled to enrich the spatially varying representation. Furthermore, to combine global and local representations, we employ a learnable adaptor to adaptively adjust features across different levels. Extensive quantitative and qualitative experiments across AID, DOTA, and DIOR benchmarks demonstrate the superior performance of our FMSR in remote sensing image super-resolution tasks compared to state-of-the-art CNN-, Transformer-, and Mamba-based SR models.

As prior research demonstrated that achieving optimal performance on the $\times 4$ SR task typically translates to favorable performance on lower scaling factors, this study focused solely on the $\times 4$ SR. While saving computational resources, this specificity lacks flexibility in exploring SR at different scales. In future work, we plan to extend our FMSR to include more scaling factors, further demonstrating its robustness and effectiveness.

References

[1] J. Zhang et al., "Frequency-aware multi-modal fine-tuning for few-shot open-set remote sensing scene classification," *IEEE Trans. Multimedia*, vol. 26, pp. 7823–7837, 2024.

[2] J. Nie et al., "MIGN: Multiscale image generation network for remote sensing image semantic segmentation," *IEEE Trans. Multimedia*, vol. 25, pp. 5601–5613, 2023.

[3] H. Chen et al., "ChangeMamba: Remote sensing change detection with spatio-temporal state space model," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4409720.

[4] H. Chen, J. Song, C. Wu, B. Du, and N. Yokoya, "Exchange means change: An unsupervised single-temporal change detection framework based on intra-and inter-image patch exchange," *ISPRS J. Photogrammetry Remote Sens.*, vol. 206, pp. 87–105, 2023.

[5] J. Wang et al., "EarthVQANet: Multi-task visual question answering for remote sensing image understanding," *ISPRS J. Photogrammetry Remote Sens.*, vol. 212, pp. 422–439, 2024.

[6] P. Liu et al., "Spectrum-driven mixed-frequency network for hyperspectral salient object detection," *IEEE Trans. Multimedia*, vol. 26, pp. 5296–5310, 2024.

[7] H. Guo, X. Su, C. Wu, B. Du, and L. Zhang, "SAAN: Similarity-aware attention flow network for change detection with VHR remote sensing images," *IEEE Trans. Image Process.*, vol. 33, pp. 2599–2613, 2024.

[8] H. Du et al., "Global mapping of urban thermal anisotropy reveals substantial potential biases for remotely sensed urban climates," *Sci. Bull.*, vol. 68, no. 16, pp. 1809–1818, 2023.

[9] X. Liu et al., "Rethinking pan-sharpening via spectral-band modulation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2023, Art. no. 5400716.

[10] Q. Zhang et al., "Hyperspectral image denoising: From model-driven, data-driven, to model-data-driven," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 13143–13163, Oct. 2024.

[11] X. Chen, Y. Li, L. Dai, and C. Kong, "Hybrid high-resolution learning for single remote sensing satellite image dehazing," *IEEE Geosci. remote Sens. Lett.*, vol. 19, 2021, Art. no. 6002805.

[12] Y. Zhao et al., "Activating more information in arbitrary-scale image super-resolution," *IEEE Trans. Multimedia*, vol. 26, pp. 7946–7961, 2024.

[13] J.-S. Yoo, D.-W. Kim, Y. Lu, and S.-W. Jung, "RZSR: Reference-based zero-shot super-resolution with depth guided self-exemplars," *IEEE Trans. Multimedia*, vol. 25, pp. 5972–5983, 2023.

[14] Z. Xiao, Z. Xiong, X. Fu, D. Liu, and Z.-J. Zha, "Space-time video super-resolution using temporal profiles," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 664–672.

[15] Z. Xiao, X. Fu, J. Huang, Z. Cheng, and Z. Xiong, "Space-time distillation for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2113–2122.

[16] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1127–1133, Jun. 2010.

[17] J. Jiang et al., "Single image super-resolution via locally regularized anchored neighborhood regression and nonlocal means," *IEEE Trans. Multimedia*, vol. 19, pp. 15–26, 2017.

[18] J. Zhou, S. Wang, Z. Lin, Q. Jiang, and F. Sohel, "A pixel distribution remapping and multi-prior retinex variational model for underwater image enhancement," *IEEE Trans. Multimedia*, vol. 26, pp. 7838–7849, 2024.

[19] J. Zhou et al., "Underwater camera: Improving visual perception via adaptive dark pixel prior and color correction," *Int. J. Comput. Vis.*, pp. 1–19, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s11263-023-01853-3

[20] Z. Xiao, Y. Liu, R. Gao, and Z. Xiong, "CutMIB: Boosting light field super-resolution via multi-view image blending," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1672–1682.

[21] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4905–4913.

[22] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 3517–3526.

[23] L. Zhang, Y. Li, X. Zhou, X. Zhao, and S. Gu, "Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 2856–2865.

[24] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[25] T. Zheng, K. Jiang, and H. Yao, "Dynamic policy-driven adaptive multi-instance learning for whole slide image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 8028–8037.

[26] J.-N. Su, M. Gan, G.-Y. Chen, W. Guo, and C. P. Chen, "High-similarity-pass attention for single image super-resolution," *IEEE Trans. Image Process.*, vol. 33, pp. 610–624, 2024.

[27] Q. Liu, P. Gao, K. Han, N. Liu, and W. Xiang, "Degradation-aware self-attention based transformer for blind image super-resolution," *IEEE Trans. Multimedia*, vol. 26, pp. 7516–7528, 2024.

[28] X. Chen, H. Li, M. Li, and J. Pan, "Learning a sparse transformer network for effective image deraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 5896–5905.

[29] J. Hou et al., "Linearly-evolved transformer for pan-sharpening," in *Proc. 32nd ACM Int. Conf. Multimedia*, 2024, pp. 1486–1494.

[30] Z. Chen et al., "Recursive generalization transformer for image super-resolution," in *Proc. Int. Conf. Learn. Representations*, 2024, pp. 1–12.
[31] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 22367–22377.
[32] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, 1960.
[33] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *Comput. Vis. Media*, vol. 9, no. 4, pp. 733–752, 2023.
[34] J.-N. Su, M. Gan, G.-Y. Chen, J.-L. Yin, and C. P. Chen, "Global learnable attention for single image super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8453–8465, Jul. 2023.
[35] W. Huang, M. Ye, Z. Shi, and B. Du, "Generalizable heterogeneous federated cross-correlation and instance similarity learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 712–728, Feb. 2024.
[36] W. Huang, M. Ye, and B. Du, "Learn from others and be yourself in heterogeneous federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10143–10153.
[37] Y. Xiao et al., "EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5601514.
[38] S. Chen, L. Zhang, and L. Zhang, "Cross-scope spatial-spectral information aggregation for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 33, pp. 5878–5891, 2024, doi: 10.1109/TIP.2024.3468905.
[39] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.
[40] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.
[41] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2472–2481.
[42] Y. Zhang et al., "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.
[43] B. Niu et al., "Single image super-resolution via a holistic attention network," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., Springer, 2020, pp. 191–207.
[44] S. Lei and Z. Shi, "Hybrid-scale self-similarity exploitation for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5401410.
[45] S. Chen, L. Zhang, and L. Zhang, "MSDformer: Multiscale deformable transformer for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5525614.
[46] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5615611.
[47] Y. Li et al., "Efficient and explicit modelling of image hierarchies for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 18278–18289.
[48] H. Guo et al., "MambaIR: A simple baseline for image restoration with state-space model," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 222–241.
[49] J. T. Smith, A. Warrington, and S. Linderman, "Simplified state space layers for sequence modeling," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–13.
[50] A. Gu, K. Goel, and C. Re, "Efficiently modeling long sequences with structured state spaces," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–12.
[51] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.
[52] W. Dong et al., "Fusion-mamba for cross-modality object detection," 2024, *arXiv:2404.09146*.
[53] Y. Liu et al., "VMamba: Visual state space model," 2024, *arXiv:2401.10166*.
[54] J. Ma, F. Li, and B. Wang, "U-Mamba: Enhancing long-range dependency for biomedical image segmentation," 2024, *arXiv:2401.04722*.
[55] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4085–4095.
[56] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A fourier-based framework for domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14383–14392.
[57] X. Mao, Y. Liu, W. Shen, Q. Li, and Y. Wang, "Deep residual fourier transformation for single image deblurring," 2021, *arXiv:2111.11745*.
[58] C. Li et al., "Embedding Fourier for ultra-high-definition low-light image enhancement," in *Proc. 11th Int. Conf. Learn. Representations*, 2023, pp. 1–12.
[59] S. Guo, H. Yong, X. Zhang, J. Ma, and L. Zhang, "Spatial-frequency attention for image denoising," 2023, *arXiv:2302.13598*.
[60] C. Wang, J. Jiang, Z. Zhong, and X. Liu, "Spatial-frequency mutual learning for face super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22356–22366.
[61] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4641–4650.
[62] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[63] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.
[64] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 136–144.
[65] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.
[66] J. Wang, B. Wang, X. Wang, Y. Zhao, and T. Long, "Hybrid attention-based u-shaped network for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612515.
[67] J. Gu and C. Dong, "Interpreting super-resolution networks with local attribution maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9199–9208.
[68] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[69] 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 Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[70] Y. Xiao et al., "TTST: A top-k token selective transformer for remote sensing image super-resolution," *IEEE Trans. Image Process.*, vol. 33, pp. 738–752, 2024.
[71] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

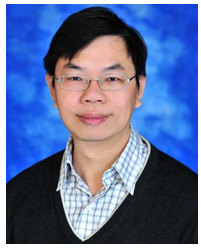
Yi Xiao (Graduate Student Member, IEEE) received the B.S. degree from the School of Mathematics and Physics, China University of Geosciences, Wuhan, China, in 2020. He is currently working toward the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan. His research interests include remote sensing image/video processing and computer vision. More details can be found at https://xy-boy.github.io/.

Qiangqiang Yuan (Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively. In 2012, he joined the School of Geodesy and Geomatics, Wuhan University, where he is currently a Professor. He has authored or coauthored more than 90 research papers, including more than 70 peer-reviewed articles in international journals, such as *Remote Sensing of Environment, ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE TRANSACTION ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. His research interests include image reconstruction, remote sensing image processing and application, and data fusion. Dr. Yuan was the recipient of the Youth Talent Support Program of China in 2019, Top-Ten Academic Star of Wuhan University in 2011, and recognition of Best Reviewers of the IEEE GRSL in 2019, and in 2014, he was the recipient of the Hong Kong Scholar Award from the Society of Hong Kong Scholars and the China National Postdoctoral Council. He is also an Associate Editor for five International Journals and has frequently served as a referee for more than 40 international top journals, such as *Nature Climate Change* and *Nature Communications*.

Kui Jiang (Member, IEEE) received the M.E. and Ph.D. degrees from the School of Computer Science, Wuhan University, Wuhan, China, in 2019 and 2022, respectively. He was a Research Scientist with Cloud BU, Huawei. He is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology. His research interests include image/video processing and computer vision. He was the recipient of the 2022 ACM Wuhan Doctoral Dissertation Award.

Yuzeng Chen received the B.S. degree in geographic information science from the Southwest University of Science and Technology, Mianyang, China, in 2020, and the M.S. degree in surveying and mapping engineering from Central South University in Changsha, China, in 2023. He is currently working toward the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan. His research interests include remote-sensing video processing and computer vision.

Qiang Zhang received the B.E. degree in surveying and mapping engineering, the M.E. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2017, 2019, and 2022, respectively. He is currently an Associate Professor with the Center of Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China. He has authored or coauthored more than ten journal papers on IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, ESSD, and ISPRS P&RS. His research interests include remote sensing information processing, computer vision, and machine learning. More details could be found at https://qzhang95.github.io.



Chia-Wen Lin (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000. He is currently a Professor with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU. He is also Deputy Director with AI Research Center,f NTHU. During 2000–2007, he was with the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan. During 1992–2000, he was with Information and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan. His research interests include image and video processing, computer vision, and video networking. From 2018 to 2019, he was a Distinguished Lecturer of IEEE Circuits and Systems Society, Steering Committee Member of IEEE TRANSACTIONS ON MULTIMEDIA from 2014 to 2015, and Chair of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society from 2013 to 2015. His articles were the recipient of the Best Paper Award of IEEE VCIP 2015 and the Young Investigator Award of VCIP 2005, and he was the recipient of the Outstanding Electrical Professor Award presented by the Chinese Institute of Electrical Engineering in 2019, and the Young Investigator Award presented by the Ministry of Science and Technology, Taiwan, in 2006. He is also the Chair of the Steering Committee of IEEE ICME. He was a Technical Program Co-Chair for IEEE ICME 2010, General Co-Chair for IEEE VCIP 2018, and Technical Program Co-Chair for IEEE ICIP 2019. He was also an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE MULTIMEDIA, and *Journal of Visual Communication and Image Representation*.