RWKVSR: Receptance Weighted Key-Value Network for Hyperspectral Image Super-Resolution

Xiaofei Yang, *Member, IEEE*, Sihuan Li, Weijia Cao, *Member, IEEE*, Dong Tang, *Member, IEEE*, Yifang Ban, *Senior Member, IEEE*, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*—Deep learning has achieved significant success in hyperspectral image super-resolution (HSISR) by leveraging advanced feature extraction techniques to reconstruct high-resolution images from low-resolution counterparts. However, existing methods predominantly utilize 2D/3D convolutions or Transformer architectures, which are often hindered by limited receptive fields, quadratic computational complexity, and inadequate fusion of spatial-spectral dependencies. To address these challenges, this paper proposes RWKVSR, a novel lightweight network that integrates a Receptance Weighted Key-Value (RWKV) architecture for efficient HSISR. The proposed RWKVSR comprises of three key components: 1) A linear-complexity RWKV module replacing quadratic self-attention, enabling efficient global spectral-spatial modeling; 2) A Spectral-Spatial Residual Module (SSRM) employing anisotropic, direction-separable 3D convolutions to hierarchically extract multi-scale features while enhancing local-global interactions; and 3) A Hyperspectral Frequency Loss (HFL) optimizing spectral consistency by prioritizing high-frequency structural alignment between reconstructed and ground-truth images in the frequency domain. Extensive experiments conducted on the CAVE and Harvard datasets demonstrate that RWKVSR outperforms the existing state-of-the-art methods, effectively balancing accuracy and efficiency, and providing a practical solution for high-quality HSI reconstruction. Our paper code is publicly available at https://github.com/backy-1/RWKVSR.git**

***Index Terms*—Hyperspectral image super-resolution, deep learning, Transformer, receptance weighted key-value network.**

I. Introduction

HYPERSPECTRAL images (HSIs) capture continuous spectral information across hundreds of narrow bands, enabling precise material identification and environmental monitoring in applications such as agriculture, mineral exploration, and military surveillance. Unlike RGB images, HSIs provide rich spectral-spatial signatures, but their spatial resolution is often limited due to hardware constraints [1]. High spatial resolution is critical for fine-grained analysis, such as detecting small objects or preserving structural details. Single-image hyperspectral super-resolution (HSISR) aims to reconstruct high-resolution (HR) HSIs from low-resolution (LR) inputs without auxiliary data, offering a cost-effective solution for enhancing spatial fidelity while retaining spectral integrity. However, HSISR faces significant challenges: (1) spectral-spatial coupling, jointly modeling correlations across bands and spatial dimensions; (2) global dependency modeling, capturing long-range interactions in high-dimensional HSIs; and (3) computational efficiency, balancing accuracy and resource consumption for practical deployment.

Traditional HSISR methods rely on handcrafted priors, such as sparse representation [2], [3], [4], non-negative matrix factorization [5], [6], [7], and self-similarity [8], [9]. These approaches encode shallow statistical properties of HSIs but struggle to generalize across diverse scenes due to their limited capacity to capture complex spectral-spatial relationships [2]. For instance, sparse coding-based methods decompose HSIs into dictionaries but fail to adapt to nonlinear variations in real-world data. Similarly, matrix decomposition techniques [5] often produce oversmoothed results, losing fine details. These limitations highlight the need for data-driven approaches to exploit deeper feature representations.

Attention-based methods, particularly Transformers [10], have emerged as powerful tools for global modeling. Transformer architectures usually leverage self-attention mechanisms to model global dependencies across spatial-spectral dimensions. For example, SwinIR [11] introduces shifted window self-attention, where images are divided into non-overlapping windows to reduce computational costs. DADI-GAN [12] uses dual attention blocks to iteratively remove clouds and shadows, effectively restoring high-resolution optical image details from degraded SAR and optical inputs. Within each window, self-attention dynamically weights inter-pixel relationships, enabling long-range feature interactions. However, the quadratic complexity of standard self-attention remains prohibitive for high-dimensional HSIs, which often contain hundreds of spectral bands and large spatial resolutions. To address this, ESSAformer [13] proposes a

Received 27 July 2025; revised 11 October 2025; accepted 27 October 2025. Date of publication 30 October 2025; date of current version 7 April 2026. This work was supported in part by the National Natural Science Foundation of China (NSFC) Fund under Grant 62301174 and in part by Guangzhou Basic and Applied Basic Research Topics under Grant 2024A04J2081. This article was recommended by Associate Editor J. Liu. *(Corresponding authors: Weijia Cao; Dong Tang.)*

Xiaofei Yang, Sihuan Li, and Dong Tang are with the School of Electronic and Communication Engineering, Guangzhou University, Guangzhou 510182, China (e-mail: xiaofeiyang@gzhu.edu.cn; 2112330082@e.gzhu.edu.cn; tangdong@gzhu.edu.cn).

Weijia Cao is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the Department of Computer and Information Science, University of Macau, Macau, China (e-mail: caowj@aircas.ac.cn).

Yifang Ban is with the Division of Geoinformatics, School of Architecture and the Built Environment, KTH Royal Institute of Technology, 11428 Stockholm, Sweden (e-mail: yifang@kth.se).

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Macau, China (e-mail: yicongzhou@um.edu.mo).

Digital Object Identifier 10.1109/TCSVT.2025.3626779

spectral-spatial decoupling strategy, applying separate attention modules to spectral and spatial dimensions. While this reduces redundancy, its rigid window partitioning limits adaptability to irregular spectral correlations. Recent variants like TC-HISRNet [14] integrate Transformers with 3D convolutions, using cross-attention to fuse multi-scale spectral features. Despite these advancements, Transformers still face challenges in balancing computational efficiency with spectral fidelity, particularly when scaling to large datasets or real-time applications.

Inspired by state-space sequence models, Mamba-based architectures [15] employ linear-time complexity mechanisms to address the scalability limitations of Transformers. Mamba utilizes selective state-space models (SSMs) to dynamically propagate contextual information across sequences, enabling efficient modeling of long-range dependencies. Unlike Transformers, which explicitly compute pairwise interactions, Mamba implicitly captures global correlations through hidden state transitions, significantly reducing memory overhead. Recent work explores Mamba for HSISR by treating spectral bands as sequential data [16], where SSMs learn cross-band dependencies while preserving spatial details. However, Mamba's inherent spectral insensitivity—stemming from its uniform state transitions—fails to adapt to the nonlinear spectral variations inherent in HSIs. Additionally, its reliance on 1D sequential processing neglects the 3D structural nature of hyperspectral data, leading to suboptimal spatial-spectral fusion.

Recently, RWKV [17] is an innovative network architecture proposed by Peng that combines the advantages of Transformer and RNN [18], aiming to achieve efficient global modeling and linear complexity computing capabilities at the same time. Its linear attention mechanism reduces computational costs to $\mathcal{O}(N)$ while maintaining parallelizability, offering a potential solution for HSISR. However, existing RWKV implementations lack tailored designs for hyperspectral data, such as multi-scale spectral-spatial fusion or frequency-domain optimization. Moreover, directly applying RWKV to HSISR fails to address the unique challenges of spectral distortion and spatial detail degradation.

To address these challenges, we propose RWKVSR, a novel framework that integrates RWKV for hyperspectral image super-resolution (HSISR). Specifically, RWKVSR comprises three key components: (1) a linear-complexity R-WKV module that facilitates efficient global dependency modeling; (2) a spectral-spatial residual module (SSRM) that employs direction-separable 3D convolutions combined with residual connections to hierarchically fuse multi-scale local and global features; and (3) a hyperspectral frequency loss (HFL) that aligns reconstructed and ground-truth hyperspectral images in the frequency domain, thereby enhancing spectral consistency and preserving high-frequency details. All the contributions of this paper are listed as follows:

1) We present the Receptance Weighted Key-Value (RWKV) architecture for Hyperspectral Image Super-Resolution (HSISR), which addresses the computational inefficiencies associated with traditional self-attention mechanisms. To our knowledge, this work is the first to introduce the RWKV architecture into the HSISR domain.
2) To mitigate the limitations of conventional three-dimensional convolutions in modeling anisotropic spectral-spatial relationships, we propose a spectral-spatial relationship module (SSRM) based on direction-separable convolutions.
3) Furthermore, we propose a High-Frequency Loss (HFL) function into HSISR, optimizing the alignment between reconstructed hyperspectral images and their corresponding ground-truth data in the Fourier domain.
4) Finally, extensive experiments are conducted on CAVE and Harvard datasets, demonstrating the superiority of RWKVSR.

II. Related Work

A. CNNs-Based Hyperspectral Image Super-Resolution Methods

Convolutional neural networks (CNNs) have dominated hyperspectral image super-resolution (HSISR) due to their ability to extract hierarchical spatial-spectral features. For example, numerous deeper CNN architectures have been developed to further enhance reconstruction performance by leveraging residual learning and dense connection strategies, including the very deep residual network VDSR [19], SRDenseNet [20], and the enhanced residual network EDSR [21]. To maintain high performance under constrained parameter budgets, recursive structures such as DRCN [22], DRRN [23], and EBRN [24] were introduced, which promote parameter sharing and deep feature reuse.

Further improvements in feature extraction and fusion were achieved by Jiang et al., who proposed the hierarchical dense recursive network (HDN) [25] with staggered diagonal connections for efficient multi-level feature integration. They subsequently introduced the Deep Extraction Recurrent Network (DDRN) [26], which incorporates a high-frequency detail compensation mechanism to mitigate information loss during propagation. Moreover, 3D-FCNN [27] introduced volumetric convolutions to jointly model spatial and spectral dimensions, achieving improved spectral fidelity. However, its isotropic 3D kernels blended spatial and spectral information indiscriminately, resulting in redundant computations and limited receptive fields. Subsequent efforts, such as MCNet [28], decomposed 3D convolutions into spatial and spectral branches, reducing parameters by 60% but still failing to capture anisotropic spatial-spectral dependencies. EUNet [29] integrated U-Net's encoder-decoder architecture with 3D convolutions, leveraging skip connections to preserve spatial details. However, its fixed-kernel convolutions constrained global dependency modeling, particularly for reconstructing large-scale textures. Recent methods like SRDNet [30] incorporated dense residual blocks to enhance feature reuse but remained bottlenecked by local operations, struggling to model nonlinear spectral transformations across bands.

Traditional CNN-based HSISR methods exhibit three significant limitations: (1) isotropic 3D convolutions insufficiently capture directional spectral–spatial correlations, (2) fixed

receptive fields restrict the ability to model long-range dependencies, and (3) high computational costs impede scalability to high-dimensional hyperspectral data. These challenges highlight the necessity for architectures that explicitly decouple spatial and spectral processing while preserving computational efficiency.

B. Attention-Based Hyperspectral Image Super-Resolution Methods

1) Transformers-Based Architectures: Transformers have become powerful tools in HSISR by leveraging self-attention mechanisms to capture global dependencies, effectively addressing the limited receptive fields of traditional convolutional models [31], [32], [33], [34], [35], [36]. For instance, SwinIR [11] employs shifted window self-attention, dividing hyperspectral tensors into non-overlapping spatial windows to reduce computational cost. However, this rigid partitioning compromises spectral continuity, leading to spectral inconsistencies. Global Context Networks (GCNet) [37] enhance feature representation and long-range modeling via a global context module, yet they fall short in capturing fine-grained structures and local details. To balance accuracy and efficiency, Lu et al. [38] proposed the Efficient Super-Resolution Transformer (ESRT) for natural images, which achieves strong performance with low computational overhead and high memory efficiency. ESSAformer [13] decoupled spectral and spatial attention, applying 1D self-attention along bands and 2D attention spatially. Although this improved spectral-spatial consistency, its quadratic complexity, remained prohibitive for high-resolution HSIs. TC-HISRNet [14] adopted a hybrid architecture, integrating Transformer encoders for spectral-spectral cross-attention with 3D convolutional decoders for spatial-spectral feature fusion. By introducing cross-attention between multi-scale spectral embeddings, this approach enhances inter-band feature interaction while leveraging convolutional inductive bias for spatial structure reconstruction. Cao et al. [39] propose an unsupervised hybrid Transformer-CNN network (uHNTC) for blind hyperspectral-multispectral image (HSI-MSI) fusion, which achieves high-accuracy HR-HSI recovery without predefined degradations and outperforms ten state-of-the-art methods on three public datasets. However, its static window designs biased attention toward local spectral contexts, failing to resolve subtle material variations in mixed pixels.

2) Mamba-Based Architectures: State Space Models (SSMs), exemplified by Mamba [40], offer a scalable alternative for sequence modeling with linear time complexity. For instance, SpectralMamba [41] treats spectral bands as sequences and applies selective SSMs for cross-band dynamic modeling, significantly reducing the computational overhead typically associated with quadratic-time Transformer methods. This enables efficient hyperspectral image (HSI) processing with minimal latency, even in high spectral dimensions.

MambaIR [42] further introduces a Selective State Space 2D (SS2D) mechanism, leveraging statistical SSMs with directional scanning strategies to capture long-range dependencies while maintaining linear complexity with respect to input size. This design alleviates the heavy computational cost of conventional attention mechanisms, enabling MambaIR to effectively model fine-grained image details in high-resolution image restoration tasks, achieving a favorable balance between performance and efficiency. However, its one-dimensional spectral processing disregards spatial adjacency and the inherent three-dimensional tensor structure of HSIs, which can lead to a misalignment between spatial textures and spectral signatures and consequently, to a degradation in the structural consistency of the reconstructed images.

3) Lightweight Image Super-Resolution Methods: While the aforementioned methods focus on general or hyperspectral-specific SR, a parallel research thrust aims to develop lightweight models for resource-constrained environments. These methods often employ strategies like parameter sharing, neural architecture search, or efficient operator design.

For example, SPH-Net [43] incorporates spectral-preserving hierarchical connections and weight sharing to reduce parameters but primarily relies on 2D convolutions, limiting its efficacy in capturing spectral correlations. Fluid Micelle Networks [44] leverage dynamic, self-organizing filters inspired by colloidal science to adaptively adjust to image content, offering a parameter-efficient solution. However, their computational overhead for generating dynamic filters can be non-trivial for high-dimensional HSI data. Heat Transfer-Inspired Networks [45] formulate image restoration as a heat diffusion process, employing a physics-informed prior to guide the reconstruction with minimal parameters. While innovative, such methods may struggle with the complex, non-linear degradations present in real-world HSISR tasks.

In contrast to these approaches, RWKVSR achieves lightweight efficiency through a fundamentally different strategy: a scalable architecture with linear complexity. Rather than simply reducing parameters or using dynamic filters, our model leverages the RWKV's efficient attention mechanism to enable global receptive fields without the quadratic cost. This design is inherently more suitable for HSI data, as it explicitly addresses the core challenge of long-range dependency modeling across spectra. Furthermore, the incorporation of the anisotropic SSRM and the HFL loss provides a targeted approach to spectral-spatial feature extraction and consistency enforcement, which are not jointly considered in the general lightweight SR methods mentioned above. Therefore, RWKVSR establishes a new state-of-the-art in efficient HSISR by integrating a scalable global model with specialized spectral-spatial processing modules.

III. Proposed Approach

A. Overview

The proposed RWKVSR is a novel hyperspectral image super-resolution (HSISR) framework designed to address the limitations of existing methods in global dependency modeling, spectral-spatial fusion, and computational efficiency. Speciafily, the architecture comprises three core components: 1) Receptance Weighted Key-Value (RWKV) Module: A linear-complexity attention mechanism for efficient global spectral-spatial modeling. 2) Spectral-Spatial Residual Module (SSRM): A hierarchical feature fusion module with direction-aware 3D convolutions. 3) Hyperspectral Frequency Loss

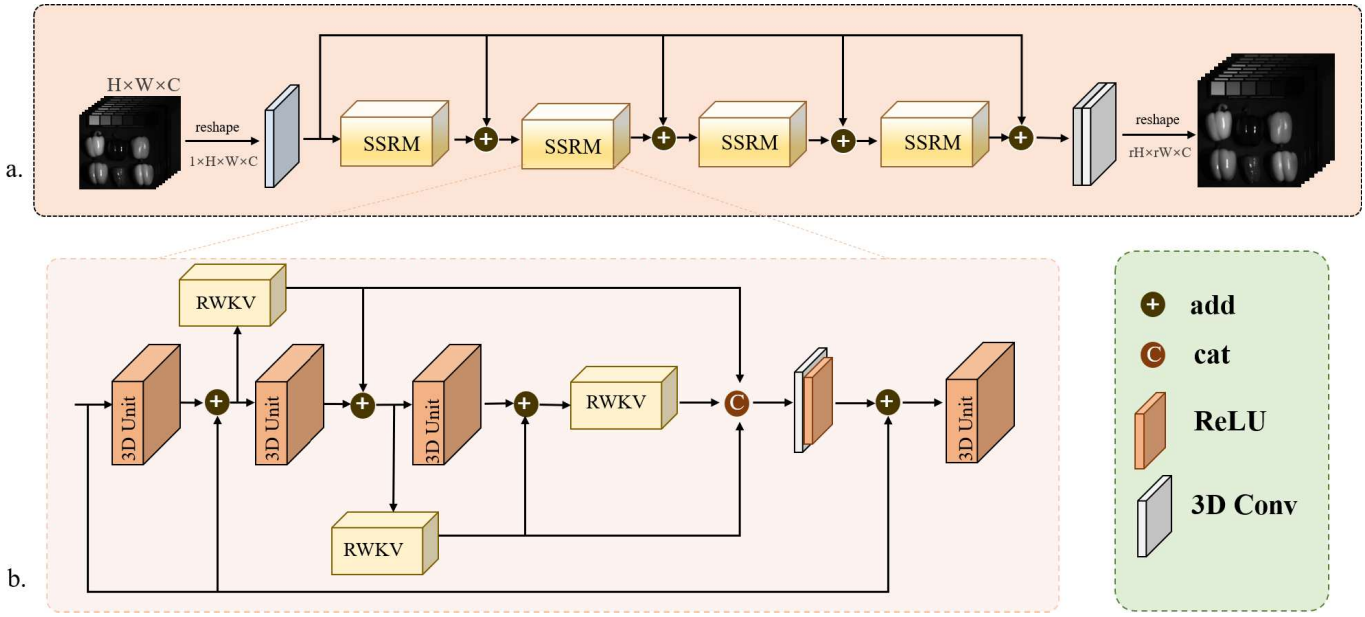

Fig. 1. The overall architecture of RWKVSR, depicted in Fig. a, comprises three main modules: a feature extraction module, a deep spatial-spectral channel fusion module, and an image reconstruction module. The feature extraction module is designed to extract shallow spatial and spectral features from low-resolution hyperspectral images, thereby providing fundamental information for subsequent modeling. Fig. b presents the detailed structure of the SSRM module, which enables efficient feature interaction and detail enhancement through the integration of parallel spatial and spectral R-WKV submodules.

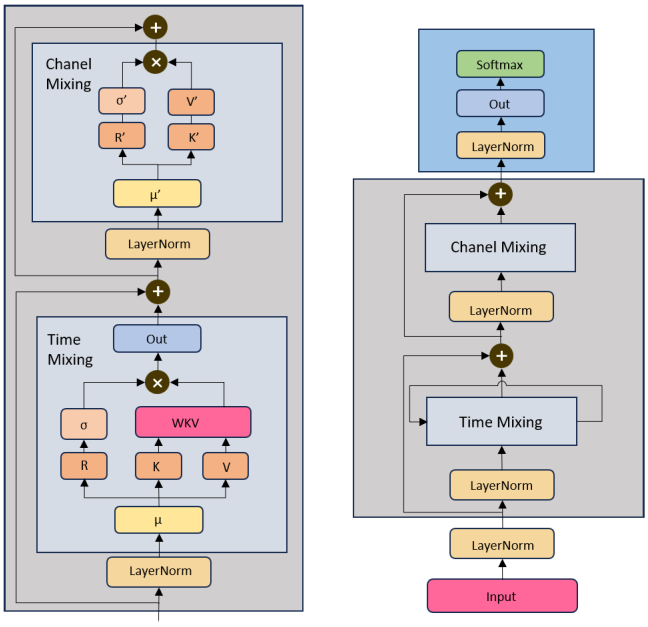

Fig. 2. Architecture of the RWKV.

(HFL): A frequency-domain optimization loss to enhance spectral consistency and structural details. These components form a unified framework that enables efficient global dependency modeling, hierarchical multi-scale feature extraction, and spectral fidelity enhancement, as illustrated in Fig. 1 (a).

B. Linear-Complexity RWKV Module

The Receptance Weighted Key-Value (RWKV) [17]model is a novel sequence modeling architecture that integrates the advantages of Transformers and recurrent neural networks (RNNs). Designed to efficiently capture long-range dependencies, RWKV [17]achieves linear computational complexity with respect to sequence length. Its core architecture comprises stacked residual blocks, each containing two sub-modules: Time-Mixing and Channel-Mixing. The Time-Mixing module employs a recurrent, attention-inspired mechanism to aggregate temporal information across steps, avoiding the quadratic cost of traditional self-attention. Meanwhile, the Channel-Mixing module functions analogously to the feed-forward network in Transformers, modeling cross-feature interactions to enhance representation learning. This hybrid design enables RWKV to process long sequences with high efficiency and scalability, while retaining strong modeling capacity. By mimicking attention with a recurrent structure, RWKV [17] significantly reduces memory usage and computation, making it particularly well-suited for resource-constrained and long-context tasks. The model diagram of RWKV is shown in Fig. 2.

C. SSRM Modules

The overall architecture of the SSRM module is shown in Fig. 1(b). The module consists of four 3D Units, three R-WKV modules, a feature concatenation operation, a standard 3D convolution, and a nonlinear activation function.

Three 3D units are connected in series to gradually model the association between spectrum and space in the horizontal and vertical directions to capture deeper feature information. In each 3D unit, the input features are not only locally modeled through 3D convolution, but also kept information flowing through jump connections to improve training stability. For each 3D unit, its output is not only added to its own input through jump connections, but also input to three R-WKV modules in parallel to ensure that R-WKV can extract feature

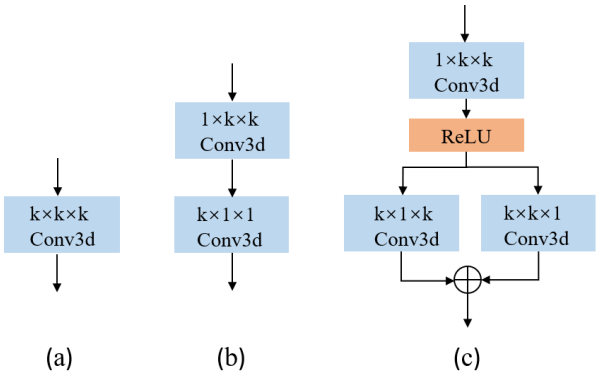

Fig. 3. Architecture of various convolutions. (a) Standard 3D convolution. (b) Separable Convolution. (c) 3D Unit.

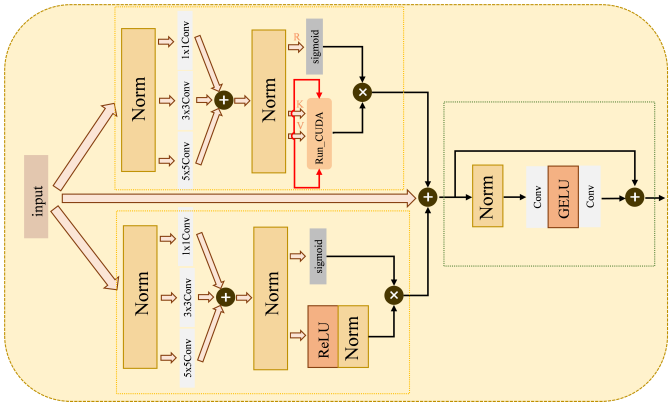

Fig. 4. Architecture of the proposed R-WKV.

representations at different levels. Next, we concatenate the outputs of the three R-WKV modules through the Concat operation and use 3D convolution to further fuse the concatenated features. Subsequently, an activation function is used to enhance the feature expression ability. Finally, we residually connect the activated features with the features initially input to the SSRM module and feed them into the final 3D unit to further refine and optimize the features. Through the above design, the SSRM module can fully explore the relationship between spectrum and space, improve the network's modeling ability for hyperspectral data, and enhance the feature expression ability, thereby improving the final reconstruction or classification effect. The output of the nth R-WKV module can be expressed as:

$$F_{0,R_c} = f_{R-WKV}(f_{3D}(F_{0,c-1}) + F_{0,c-1}). \tag{1}$$

where c is the cth R-WKV module, $f_3D(\cdot)$is the 3D Unit module, the first number indicates the number of SSRM modules, and the second number indicates the number of R-WKV modules.

1) 3D Unit: When processing hyperspectral data, traditional 3D convolution usually uses a standard $k \times k \times k$ convolution kernel, as shown in Fig. 3 (a), to perform feature modeling in both spectral and spatial dimensions. However, this approach has certain limitations in hyperspectral super-resolution tasks, and the amount of convolution calculation is large. SSRNet proposes separable convolution, as shown in Fig. 3 (b), which performs a structured decomposition of the standard 3D convolution kernel into two series-connected branches: $1 \times k \times k$ and $k \times 1 \times 1$ convolution kernels are used to extract the correlation of spectral information in the horizontal and vertical directions, respectively. This may destroy the ability to model joint features between channels when processing highly correlated multi-channel data such as hyperspectral images.

The 3D Unit we introduced first uses $1 \times k \times k$ 3D convolution to focus on extracting spatial features while keeping the spectral dimension unchanged and retaining its structural information; then the ReLU activation function is used to enhance the nonlinear expression ability of the model. Then, the features are further modeled through two parallel deep convolution branches: Branch 1 uses $k \times 1 \times k$ convolution to capture the correlation between the spectrum and the width direction; Branch 2 uses $k \times k \times 1$ convolution to focus on the interaction between the spectrum and the height direction. The two branches work together to extract joint features of different dimensions.

Finally, the outputs of the two branches are fused by element-by-element addition, which effectively improves the expression ability of the spectral information and the overall modeling performance of the network. This structure is called 3D Unit, and its schematic is shown in Fig. 3 (c).

2) R-WKV: Inspired by RWKV, we introduce the linear attention mechanism into it, which reduces the computational complexity and makes it more suitable for hyperspectral super-resolution tasks. At the same time, the linear attention is parallelized on the GPU to accelerate the calculation. Among them, f_{R-WKV} is the linear attention module. The structure of R-WKV is shown in the Fig. 4.

This method mainly consists of two core parts: multi-scale spatial mixing and multi-scale channel mixing. These two parts share the same first half, that is, using convolution kernels of different scales for feature extraction and then weighted fusion of information of different scales. However, in the subsequent information processing stage, they adopt different strategies to enhance the feature expression ability.

Spatial mixing: Transform the obtained features from (B, C, H, W) to (B, $H \times W$, C) to complete a spatial fusion. In order to more effectively capture spatial information of different scales, we designed a multi-scale spatial mixing method. This method uses three two-dimensional convolution kernels of different scales (1×1, 3×3 and 5×5) to extract features, so as to obtain spatial information under different receptive fields. The specific calculation process is as follows:

$$X_{1\times1} = Conv_{1\times1}(X). \tag{2}$$

$$X_{3\times3} = Conv_{3\times3}(X). \tag{3}$$

$$X_{5\times5} = Conv_{5\times5}(X). \tag{4}$$

After obtaining features of different scales, we introduce hyperparameters γ_1, γ_2, and γ_3 for weighted fusion to complete the integration of multi-scale features:

$$X_{mix} = \gamma_1 X_{1\times1} + \gamma_2 X_{3\times3} + \gamma_3 X_{5\times5}. \tag{5}$$

This fusion strategy allows the model to simultaneously focus on local details (1×1), medium-scale features (3×3), and larger-scale spatial relationships (5×5), thereby improving the expressiveness of features.

Similar to Transformer, we perform linear projection on the fused features and calculate Key (K), Value (V) and Receptance (R) respectively. However, unlike the standard Transformer, we do not use the full self-attention calculation, but instead adopt an efficient linear attention mechanism. The calculation process is as follows:

$$K = W_k X_{mix}, \quad V = W_v X_{mix}, \quad R = W_r X_{mix}. \tag{6}$$

Among them, W_k, W_v and W_r are learnable linear projection matrices. Then, we send K and V to the efficient attention calculation module Run_CUDA for linear attention calculation.

Traditional self-attention calculations require the construction of a complete $QK^\top$ correlation matrix, with a computational complexity of $\mathcal{O}(N^2)$ (where N is the number of pixels). In our method, linear attention does not need to calculate the complete correlation matrix, but instead uses element-wise weighted exponential decay summation to efficiently model long-range dependencies. The core idea is to use cumulative summation instead of matrix multiplication, thereby reducing the computational complexity to $\mathcal{O}(N)$.

Decay function $\Phi(x)$:

$$\Phi(x) = \sum_{t=1}^{T} e^{-\beta(t-x)} V_t \tag{7}$$

Among them, β controls the decay rate, and V_t is the Value of the corresponding time step. This operation can be efficiently calculated through Run_CUDA, avoiding the explicit construction of the QK^T matrix.

After the exponential decay accumulation operation is completed, the result will be element-wise multiplied with the modulation factor R activated by the sigmoid$(\cdot)$ function to achieve adaptive adjustment of the feature response:

$$X_1 = \sigma(R) \odot \Phi(K, V). \tag{8}$$

Among them, σ represents the Sigmoid function and $\odot$ denotes element-wise multiplication. The purpose of this operation is to let the Receptance control the attention weights, thereby dynamically adjusting the feature information at different spatial positions.

The first half of multi-scale channel mixing is the same as spatial mixing, that is, using 1×1, 3×3, and 5×5 convolution kernels to extract information of different scales and weighted fusion. However, after information fusion, channel mixing takes a different approach to more effectively adjust the information interaction between channels. In the channel mixing module, we divide the fused features into two information streams:

1) The first information flow: directly through the Sigmoid activation function, used to learn the importance weights:

$$X_{sig} = \sigma(W_1 X_{mix2} + b_1). \tag{9}$$

2) The second information flow: Enhance the expressive power of features through linear transformation and nonlinear activation:

$$X_{relu} = ReLU(W_2 X_{mix2} + b_2). \tag{10}$$

$$X_{linear2} = W_3 X_{relu} + b_3. \tag{11}$$

We then perform element-wise dot multiplication on these two information streams to fuse the information after channel attention and feature transformation:

$$X_2 = X_{sig} \odot X_{linear2}. \tag{12}$$

This operation enables the channel attention to work together with the transformed features, enhancing the model's adaptability.

Finally, the model performs element-wise addition of spatial and channel fusion outputs by elementing to fully integrate feature information from different dimensions. The fused features first pass through a linear transformation to adjust the channel dimension, and then pass through a convolutional layer, a GELU activation function, and another convolutional layer to further extract and enhance contextual features. In this process, nonlinear activation helps to improve the expressiveness of the model, while continuous convolution operations enhance the modeling ability of local features. The final output of the module is:

$$X_{\text{out}} = conv(GELU(conv(Linear(X_1 + X_2)))). \tag{13}$$

D. Loss Function

In super-resolution tasks, the$\mathcal{L}_1$ loss is a common choice because it provides stable gradients during training and helps the model converge quickly. In this study, we use $\mathcal{L}_1$ loss to measure the pixel reconstruction error in the spatial domain, ensuring that the reconstructed image remains consistent with the high-resolution (HR) image at the pixel level. In addition, to further improve the model's fidelity to spectral information, we introduce $\mathcal{L}_{\text{HFL}}$ (spectral frequency domain loss), which measures spectral discrepancies in the frequency domain and enhances the model's reconstruction performance on hyperspectral data. Finally, we combine the losses from both the spatial and frequency domains to form the total loss function $\mathcal{L}_{\text{total}}$, which jointly optimizes the dual-domain network to balance spatial detail preservation and spectral consistency. The total loss is defined as:

$$L_{\text{total}} = L_1 + \beta L_{\text{HFL}}. \tag{14}$$

The formula of spatial loss L_1 is as follows:

$$L_1 = \frac{1}{N} \sum_{n=1}^{N} \left\| I_{\text{GT}}^n - H_{\text{Net}}(I_{\text{LR}}^n) \right\|_1. \tag{15}$$

Let I_n^{GT} and HNet(I_{LR}^{n}) represent the ground truth and the reconstructed hyperspectral image (HSI) for the n-th sample in a training batch, respectively. Here, N denotes the total number of HSIs in the batch, and θ is the set of learnable parameters in our network. The loss function is computed over all training samples to optimize θ and minimize the reconstruction error between the predicted and ground truth HSIs.

1) Hyperspectral Frequency Loss: The conventional $\mathcal{L}_1$ loss, while providing stable gradients, operates solely in the spatial domain and often leads to spectral distortion and over-smoothed reconstructions by neglecting the structural properties of the data. Our proposed HFL is founded on the core principle that for hyperspectral data, fidelity must be enforced not just in the pixel-value space but also in the frequency domain to preserve both spatial high-frequency details and the intricate spectral signatures that are characteristic of materials.

Specifically, we first apply a two-dimensional discrete Fourier transform (2D DFT) to each band image to map it from the spatial domain to the frequency domain. The transformation formula for the kth band is:

$$F_k(u,v) = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1} f_k(x,y)\cdot e^{-i2\pi\left(\frac{ux}{H}+\frac{vy}{W}\right)}. \tag{16}$$

The image size is $H \times W$, $f_k(x,y)$ is the spatial image of the kth (0, 1, 2, 3, …, C-1) band, (x,y) is the pixel value, (u,v) represents the spectral space coordinates. $F_k(u,v)$ is the complex frequency value.

Considering that the amplitude and phase in the spectrum carry different structural information of the image, these structural features may be ignored if only the pixel differences in the spatial domain are compared. Therefore, we calculate the square of the Euclidean distance between the reconstructed image and the true image in the frequency domain as the basic measure of the frequency domain error:

$$d_c(u,v) = \| F_k^{GT}(u,v) - F_k^{SR}(u,v) \|^2 . \tag{17}$$

However, neural networks tend to learn low-frequency and other frequency components that are easy to fit during training, and ignore high-frequency components that are structurally strong but difficult to fit. In order to encourage the network to pay more attention to these difficult frequencies, we introduce a weight mechanism based on the error size, which is defined as follows:

$$w_k(u,v) = \| F_k^{GT}(u,v) - F_k^{SR}(u,v) \|^{\alpha} . \tag{18}$$

Here, α is an adjustable scaling factor, typically set to 1. Subsequently, we normalize the frequency-domain error weights $w_k(u,v)$ to the interval $[0,1]$. A value of 1 indicates the frequency component with the highest reconstruction error, which tends to be overlooked by the network and thus requires enhanced attention.

The final frequency domain loss is obtained by multiplying the error of each frequency point by its corresponding weight and summing them up:

$$d(F_{\mathrm{cr}}^k, F_{\mathrm{sR}}^k) = \frac{1}{HW}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1} w_k(u,v)\cdot \times \left\| F_k^{GT}(u,v) - F_k^{SR}(u,v) \right\|^2 , \tag{19}$$

$$L_{\mathrm{HFL}}(F_{\mathrm{GT}}, F_{\mathrm{SR}}) = \sum_{k=0}^{C-1} d(F_{\mathrm{GT}}^k, F_{\mathrm{sR}}^k). \tag{20}$$

2) Theoretical Advantage Over Common Losses: **(a) Vs. $\mathcal{L}_1$/$\mathcal{L}_2$:** Pixel-wise losses lack an explicit mechanism to prioritize structurally important information, often resulting in perceptually poor, blurry outputs. Unlike prior frequency losses that apply uniform weighting, HFL employs an *error-adaptive weighting scheme* that prioritizes spectral bands with higher reconstruction difficulty, which is particularly critical in HSIs due to non-uniform spectral correlations.

(b) Vs. Perceptual Loss: Perceptual losses use features from networks (e.g., VGG) pre-trained on RGB images. These features may not be optimal or directly relevant for capturing the unique spectral-spatial characteristics of hyperspectral data. HFL is *domain-agnostic* and directly minimizes the error in the space where spectral consistency is physically defined (frequency), making it uniquely suited for HSISR.

E. Theoretical Distinction From Competing Paradigms

The RWKV module is the first to be applied to HSISR. Its theoretical superiority stems from its unique formulation, which synergizes the parallelizable training of Transformers with the linear computational complexity of recurrent models.

(a) Complexity Analysis vs. Transformer. Standard self-attention in Transformers exhibits quadratic complexity $\mathcal{O}(N^2)$ with respect to token length N ($N = H \times W \times C$ for an HSI cube), making it prohibitively expensive for high-resolution HSIs. In contrast, the RWKV module achieves linear complexity $\mathcal{O}(N)$ through a novel recurrent formulation with a data-dependent decay mechanism:

$$\Phi(x) = \sum_{t=1}^{T} e^{-\beta(t-x)} V_t, \tag{21}$$

where $\Phi(x)$ denotes the output feature vector at the target position x, t is the summation index that iterates over all positions in the input sequence, from 1 to T. T is the total sequence length, which is equivalent to N. V_t is the value vector at the source position t, β is a learnable decay rate. And $e^{-\beta(t-x)}$ is the data-dependent decay weight that modulates the contribution of each V_t.

(b) Parallelizability vs. Mamba. Unlike Mamba and other State Space Models (SSMs) that rely on a hardware-inefficient sequential scanning process (limiting training parallelizability), RWKV's recurrence is designed to be fully parallelizable during training.

(c) Anisotropic Modeling vs. CNNs. Traditional 3D CNNs utilize isotropic kernels ($k \times k \times k$) that fail to account for the inherent *anisotropy* between spatial dimensions (shift-invariant) and the spectral dimension (sequential and highly correlated). The RWKV module, integrated with our direction-aware SSRM, provides a global receptive field from the first layer and enables adaptive, anisotropic interaction modeling across all dimensions, a capability beyond fixed-kernel CNNs.

IV. Experiments

In this section, we evaluate our network both qualitatively and quantitatively. First, we provide a benchmark dataset and implementation details. Finally, we analyze the effectiveness of our model. Finally, we compare it with other methods on the benchmark dataset.

A. Datasets

The CAVE dataset is a widely used benchmark in hyperspectral imaging research. It was acquired using a tunable filter and a cooled CCD camera, covering the spectral range of 400 – 700 nm with 10 nm intervals, resulting in 31 spectral bands. The dataset includes a variety of scenes, such as skin, beverages, and vegetables, making it suitable for evaluating hyperspectral imaging tasks. Each hyperspectral image has a spatial resolution of 512×512 pixels. In addition, the Harvard dataset consists of 70 hyperspectral images, each with a spatial resolution of 1392×1040 pixels and 31 spectral bands.

Unlike the CAVE dataset, which primarily consists of images captured under controlled conditions, the Harvard hyperspectral dataset emphasizes greater diversity and realism in natural scenes. Collected under natural illumination, it includes a wide range of indoor and outdoor environments with significant variations in lighting and scene complexity. The dataset contains 71 hyperspectral images, each with a spatial resolution of 1392×1040 pixels and 31 continuous spectral bands covering the 420 – 720 nm range.

The Pavia Center dataset contains 102 spectral bands with an original spatial size of 1096×1096 pixels. However, since the central region contains invalid data, it was removed, leaving a valid area of 1096×715 pixels. For testing, a $128 \times 715 \times 102$ subscene was extracted from the bottom part of the image. To avoid discontinuities along the left and right boundaries, the testing region was centrally cropped into four nonoverlapping patches, each measuring $128 \times 128 \times 102$.

B. Implementation Details

For our network, we randomly divide the data into 80% training set, 10% validation set, and 10% test set to ensure the fairness of the experiment and full use of the data. In order to improve the robustness of the model, we use random cropping and data augmentation methods during training. Specifically, we randomly select 24 patches for processing in each training iteration. To further enhance data diversity, these patches are randomly rotated (90°, 180°, 270°), horizontally flipped, and scaled ($1\times$, $0.5\times$, $0.75\times$). Subsequently, we perform bicubic interpolation on the processed patches to achieve downsampling with scale factor r, thereby obtaining a low spatial resolution image of size $32 \times 32 \times B$ (where B is the number of spectral channels). After the above processing, the CAVE dataset ultimately generates 8424 patches, providing sufficient data support for model training. During the testing phase, we crop a 512×512 pixel region from the top-left corner of the test image as the input for model evaluation, to ensure uniformity and comparability of the experimental settings.

Our RWKVSR model was implemented in PyTorch 1.11.0 and trained on an NVIDIA GeForce RTX 4090 GPU. The network was optimized using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with an initial learning rate of 1.0×10^{-4}, which was halved every 20 epochs. We trained the model for 100 epochs with a batch size of 6. The kernel size k in the 3D Unit was set to 3. The loss function was a weighted sum of $\mathcal{L}_1$ and the proposed Hyperspectral Frequency Loss ($\mathcal{L}_{\text{HFL}}$), with the weighting factor β set to 0.1 after empirical validation. Input patches of size $32 \times 32 \times B$ (where B is the number of spectral bands) were randomly cropped from the training images. Data augmentation techniques including random rotation (90°, 180°, 270°), horizontal flipping, and scaling ($1\times$, $0.5\times$, $0.75\times$) were applied to enhance the robustness and generalization of the model.

C. Evaluation Metrics

To evaluate the performance of our network, we use six popular image quality indicators, namely Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Spectral Angle Mapping (SAM), Correlation (COR), Root Mean Square Error (RMSE), and Erreur Relative Globale Adimensionnelle de Synthése (ERGAS). PSNR and SSIM are commonly used metrics in image restoration, and their valid value range is $(1, +\infty)$. The larger the value, the better the super-resolution reconstruction performance. The calculation formulas are defined as follows:

$$\text{PSNR} = \frac{1}{C}\sum_{c=1}^{C} 10\log_{10}\left(\frac{\text{MAX}_c^2}{\text{MSE}_c}\right). \tag{22}$$

$$\text{MSE}_c = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}(I_{\text{SR}}(w,h,c) - I_{\text{HR}}(w,h,c))^2. \tag{23}$$

where MAX_c is the maximum pixel value of the c_th band, I_HR is the HR hyperspectral image, W and H represent the width and height of the image respectively.

SSIM calculation formula:

$$\text{SSIM} = \frac{1}{C}\sum_{c=1}^{C}\frac{(2\mu_{I_{\text{SR}}}^c\mu_{I_{\text{HR}}}^c + b_1)(2\sigma_{I_{\text{SR}}I_{\text{HR}}}^c + b_2)}{m * n}. \tag{24}$$

$$m = (\mu_{I_{\text{SR}}}^c)^2 + (\mu_{I_{\text{HR}}}^c)^2 + b_1. \tag{25}$$

$$n = (\sigma_{I_{\text{SR}}}^c)^2 + (\sigma_{I_{\text{HR}}}^c)^2 + b_2. \tag{26}$$

where $\mu_{I_{\text{SR}}}^c$ and $\mu_{I_{\text{HR}}}^c$ represent the means of I_{SR} and I_{HR} of the c-th band, $\sigma_{I_{\text{SR}}}^c$ and $\sigma_{I_{\text{HR}}}^c$ represent the variances of I_{SR} and I_{HR} of the c-th band, and $\sigma_{I_{\text{SR}}I_{\text{HR}}}^c$ represents the covariance of I_{SR} and I_{HR} of the c-th band.

SAM measures the shape similarity between spectral curves. Spectral information of different bands is crucial. We should not only focus on the mean square error at the pixel level, but also take into account the fidelity of the spectrum. The value range is [0,90], and the smaller the value, the better.

$$\text{SAM} = \arccos\left(\frac{\langle I_{\text{SR}}, I_{\text{HR}}\rangle}{\|I_{\text{SR}}\|_2\|I_{\text{HR}}\|_2}\right). \tag{27}$$

Comparison with competing models: To ensure a fair and reproducible comparison, all baseline methods were re-implemented and carefully tuned on the same datasets using their officially released code and the recommended training strategies from their respective publications. The key hyper-parameters for each baseline are summarized in Table I.

- **Bicubic Interpolation**: Implemented using the OpenCV library (`cv2.resize` function) with the `INTER_CUBIC` flag.
- **EDSR** [21] and **SwinIR** [11]: Originally designed for RGB images, we adapted their first and last convolutional

TABLE I
SUMMARY OF KEY HYPERPARAMETERS AND TRAINING CONFIGURATIONS FOR THE COMPARED METHODS

Method	Primary Loss Function	Key Hyperparameters and Notes
Bicubic	-	Library: OpenCV, Interpolation method: `INTER_CUBIC`.
EDSR	Charbonnier	Scale-specific pre-training, number of blocks: 32, number of features: 256.
SwinIR	$\mathcal{L}_1$	Depth: 6, number of heads: 6, window size: 8, mlp ratio: 2.
GDRRN	$\mathcal{L}_2$	Recursive blocks: 6, number of features: 128.
3D-FCNN	$\mathcal{L}_2$	Kernel size: (3, 3, 3), number of layers: 4.
MCNet	$\mathcal{L}_1$	$\lambda_{\text{spectral}} = 0.1$, $\lambda_{\text{spatial}} = 1.0$.
EUNet	$\mathcal{L}_1$	Encoder/decoder channels: [64, 128, 256], number of epochs: 200.
TC-HISRNet	$\mathcal{L}_1$	Transformer layers: 4, CNN blocks: 6, number of features: 64.
ERCSR	$\mathcal{L}_1$	Ratio of 2D/3D convolutions: 0.5, feature channels: 64.
SFCSR	$\mathcal{L}_1$	Spectral attention bands: 8, spatial window size: 5.
ESSAformer	$\mathcal{L}_1$	Number of heads: 8, number of layers: 6, embedding dimension: 180.
SRDNet	$\mathcal{L}_1$	Dense blocks: 4, growth rate: 32.
RWKVSR (Ours)	$\mathcal{L}_1 + \beta\mathcal{L}_{\text{HFL}}$	$\beta = 0.1$, number of SSRM modules: 4, number of RWKV units per SSRM: 3.

layers to handle the hyperspectral input and output dimensions (B channels). They were trained with Charbonnier loss, a common choice in SISR, as per their original setup.

- **3D-FCNN** [46], **MCNet** [47], **EUNet** [29], **ERCSR** [28], **SFCSR** [48], **SRDNet** [30]: These are HSI-specific methods based on 2D/3D convolutions. We strictly followed the architectural details and training protocols (e.g., loss function, learning rate schedule) described in their original papers.
- **GDRRN** [49]: This recursive network was trained with $\mathcal{L}_2$ loss and a recursive depth of 6, as recommended by the authors.
- **TC-HISRNet** [14] and **ESSAformer** [13]: As hybrid Transformer-based models for HSISR, we used the authors' provided configurations, including their choice of optimizer (Adam), loss function ($\mathcal{L}_1$), and critical structural hyperparameters like the number of heads and layers.

The detailed quantitative analysis and visual comparisons are described as follows:

1) Analysis of experimental results of CAVE, Harvard and Pavia Center dataset: Table II shows the evaluation results of various methods on the Pavia Center dataset, Table III shows the quantitative evaluation results of various methods on two hyperspectral image datasets, CAVE and Harvard, with reconstruction scale factors of $\times 2$, $\times 3$ and $\times 4$ respectively. We use red to mark the best results under the current indicator and blue to mark the suboptimal results.

From the results in the table, it can be seen that our proposed method significantly outperforms the existing comparative methods in the PSNR indicator under each scale factor, showing strong reconstruction ability and good spectral structure preservation ability. It can be observed that the "Bicubic" (Traditional interpolation methods) performs poorly across all evaluation metrics, with a significant deficit in PSNR (> 5 dB) and SSIM (> 0.12) compared to deep learning-based methods like our RWKVSR. This is primarily because interpolation methods rely on fixed mathematical kernels that do not adapt to the content or structure of the image, leading to oversmoothing and loss of high-frequency details. In addition, although

TABLE II
COMPARISON OF DIFFERENT METHODS ON THE PAVIA CENTER DATASET

Method	Scale	PSNR ↑	SSIM ↑	SAM ↓
IFN	×2	30.89	0.9544	3.87
GDRRN	×2	33.13	0.9729	3.50
3DFCNN	×2	33.47	0.9749	3.43
SAGRDN	×2	33.40	0.9742	3.60
EUNet	×2	34.17	0.9781	3.26
ERCSR	×2	34.71	0.9802	3.09
RWKVSR	×2	**36.09**	0.9525	4.20
IFN	×3	27.82	0.9049	4.86
GDRRN	×3	28.97	0.9312	4.63
3DFCNN	×3	29.28	0.9362	4.58
SAGRDN	×3	29.44	0.9386	4.60
EUNet	×3	29.70	0.9420	4.53
ERCSR	×3	30.17	0.9485	4.14
RWKVSR	×3	29.16	0.8249	6.50
IFN	×4	26.19	0.8563	5.53
GDRRN	×4	27.00	0.8897	5.34
3DFCNN	×4	27.14	0.8931	5.35
SAGRDN	×4	27.35	0.8993	5.22
EUNet	×4	27.40	0.8996	5.36
ERCSR	×4	27.63	0.9062	4.96
RWKVSR	×4	**29.88**	0.8173	6.60

methods such as EDSR [21], SwinIR [11] and GDRRN [49] perform well in natural image super-resolution tasks, these methods are mainly based on two-dimensional convolution for spatial feature extraction, and fail to effectively model the inherent multi-channel spectral correlation in hyperspectral images, thus having obvious disadvantages in spectral fidelity. This further demonstrates that designing structures that can jointly model spatial and spectral features is crucial for high-quality hyperspectral image reconstruction.

In contrast, 3DFCNN [46] uses three-dimensional convolution to obtain spatial-spectral joint features. Although it has fewer parameters, it has high requirements for video memory and its reconstruction performance is still limited. The network that introduces 3D convolution can better capture the correlation between space and spectrum and retain the image structure information. Therefore, the network based on 3D convolution (such as MCNet [47], ERCSR [28], SFCSR

TABLE III
QUANTITATIVE COMPARISON ON THE CAVE AND HARVARD DATASETS UNDER SCALE ×2, ×3 AND ×4

Method	Scale	Param. (K)	Memory (M)	FLOPs (G)	Time (s)	CAVE			Harvard		
						PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓
Bicubic	×2	-	-	-	-	40.762	0.9622	2.647	43.650	0.9739	2.358
EDSR	×2	1402.143	1417	191	0.609	44.185	0.9728	2.522	45.434	0.9765	2.429
SWinIR	×2	985.864	2215	124	1.713	42.088	0.9678	3.139	43.968	0.9676	2.837
GDRRN	×2	218.880	1731	183	0.699	41.943	0.9653	3.808	45.070	0.9757	2.459
3DFCNN	×2	39.405	4319	120	0.625	42.899	0.9674	2.447	44.554	0.9753	2.335
MCNet	×2	1928.334	10177	7850	1.152	45.332	0.9739	2.211	46.361	0.9819	2.198
EUNet	×2	550.586	1902	349	0.573	43.550	0.9739	2.398	45.181	0.9796	3.057
HISRNet	×2	8230.364	7406	2030	5.559	45.569	0.9736	2.227	46.515	0.9825	2.182
SFCSR	×2	1848.882	2577	9052	1.568	45.324	0.9739	2.211	46.440	0.9822	2.195
ERCSR	×2	1348.742	10699	5920	0.955	45.320	0.9741	2.248	46.393	0.9819	2.199
ESSAformer	×2	8685.919	991	3365	0.199	44.323	0.9867	3.198	47.787	0.9889	3.008
SRDNet	×2	1531.359	962.4	2846	0.324	45.772	0.9910	2.385	49.247	0.9914	2.317
RWKVSR (Ours)	×2	1967.714	966.62	4659.68	1.8189	46.003	0.9906	2.877	49.077	0.9913	2.238
Bicubic	×3	-	-	-	-	37.535	0.9322	3.504	40.681	0.9535	2.731
EDSR	×3	1586.783	1909	100	0.429	40.725	0.9514	3.090	42.226	0.9563	2.916
SWinIR	×3	1069.719	2037	64	1.224	38.921	0.9423	3.795	41.526	0.9535	3.077
GDRRN	×3	218.880	2251	81	0.587	39.853	0.9473	3.142	42.022	0.9604	2.743
3DFCNN	×3	39.405	4815	52	0.679	39.328	0.9421	2.891	41.257	0.9540	2.702
MCNet	×3	2038.926	6569	3670	1.166	41.236	0.9526	2.799	43.385	0.9662	2.627
EUNet	×3	632.506	1832	215	0.340	40.362	0.9502	3.041	42.939	0.9638	2.685
HISRNet	×3	8415.324	3396	782	2.070	41.422	0.9506	2.812	43.532	0.9674	2.626
SFCSR	×3	1269.842	2177	4340	1.238	41.309	0.9527	2.814	43.357	0.9663	2.612
ERCSR	×3	1459.334	8793	3140	0.787	41.171	0.9517	2.784	43.417	0.9657	2.634
ESSAformer	×3	11193.759	953.51	3033	0.1881	41.450	0.9757	4.122	43.513	0.9578	2.418
SRDNet	×3	1679.396	954.91	3276	0.224	41.755	0.9783	2.999	43.359	0.9692	2.362
RWKVSR (Ours)	×3	2029.922	959.36	3057	0.8768	41.909	0.9778	3.020	43.929	0.9739	3.469
Bicubic	×4	-	-	-	-	35.749	0.9073	3.927	38.776	0.9356	2.949
EDSR	×4	1549.855	1895	74	0.548	38.647	0.9290	3.603	40.280	0.9434	3.071
SWinIR	×4	1187.116	1963	38	1.269	37.311	0.9204	4.301	39.736	0.9392	3.262
GDRRN	×4	218.880	2249	46	0.446	37.478	0.9209	4.489	40.107	0.9416	3.062
3DFCNN	×4	39.405	4815	28	0.496	37.251	0.9177	3.435	39.290	0.9358	2.884
MCNet	×4	2174.094	5339	2230	1.065	39.069	0.9320	3.235	41.334	0.9511	2.875
EUNet	×4	616.250	1942	179	0.385	38.521	0.9092	3.467	40.898	0.9481	2.923
HISRNet	×4	8378.332	136	449	1.466	39.523	0.9323	3.202	41.400	0.9520	2.861
SFCSR	×4	1232.850	2063	2852	0.7255	39.235	0.9322	3.202	41.169	0.9512	2.862
ERCSR	×4	1594.502	7473	2180	0.859	39.222	0.9323	3.180	41.333	0.9517	2.859
ESSAformer	×4	11193.759	991.00	3151	0.1948	39.378	0.9646	4.415	42.963	0.9702	3.511
SRDNet	×4	1560.031	962.40	738.55	0.289	39.161	0.9669	3.436	44.339	0.9787	2.811
RWKVSR (Ours)	×4	2105.954	967.14	2970.26	0.5738	39.673	0.9671	3.516	42.416	0.9687	3.151

[48] and SRDNet [30]) is generally better than the model based on 2D convolution in reconstruction quality. However, whether it is based on 2D or 3D convolution, its receptive field is still limited, and it is difficult to model global feature changes.

ESSAformer [13] introduced the Transformer structure to expand the receptive field, but its performance was not significantly better than the existing methods. Overall, the PSNR index of our proposed method was improved by 0.213 dB, 0.154 dB and 0.150 dB respectively compared with the suboptimal algorithm under all scale factors. In terms of the SSIM index, our method also ranks among the top two. In addition, the number of parameters of our model is significantly smaller than that of SFCSR [48] and ESSAformer [13], reflecting a strong performance-complexity balance ability.

2) Qualitative comparison: To more intuitively illustrate the reconstruction performance of the proposed method across different datasets, we selected the 30th spectral band of an image from the CAVE test set and the spectral band from the Harvard test set as representative visualization samples. Given the large number of comparison methods, we divided the visualization results—along with the Ground Truth (GT)—into two groups, each containing seven images. Specifically, the first and third rows of the visualizations display the reconstructed results of each method for the selected spectral bands. The visualization outcomes for the CAVE and Harvard datasets are presented in Fig. 5 and Fig. 7, respectively. At the same time, the waveform of the Pavia Center dataset is visualized as shown in Fig. 6. To further assess the reconstruction quality, we also provide absolute error maps between each reconstructed image and its corresponding ground truth. These error maps are rendered using a pseudo-color scheme, where colors closer to blue indicate lower reconstruction errors and thus higher fidelity to the ground truth. As shown in the figures, the proposed method consistently exhibits a wider distribution of blue regions in the error maps, suggesting superior reconstruction performance in both spatial and spectral dimensions, with notably lower reconstruction errors and enhanced overall visual quality. It can be observed that the "Bicubic" reconstructed image and the ground truth is particularly high along edges and textured regions and the spectral distortion introduced by bicubic interpolation is evident.

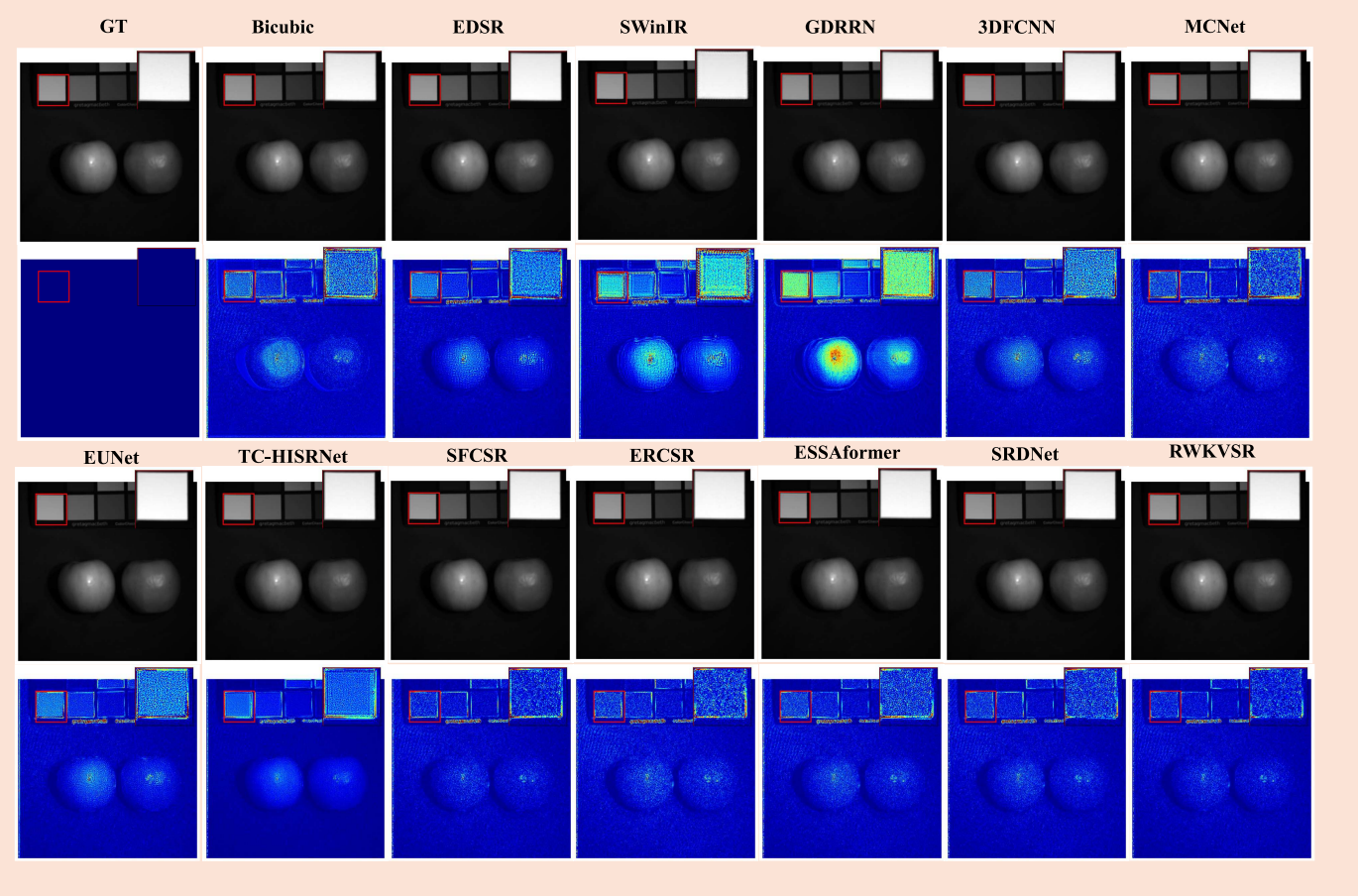

Fig. 5. Reconstruction results and corresponding absolute error maps generated by different methods, `fake_and_real_apples_ms` of CAVE dataset. The bluer the area in the absolute error map, the smaller the absolute error of GT.

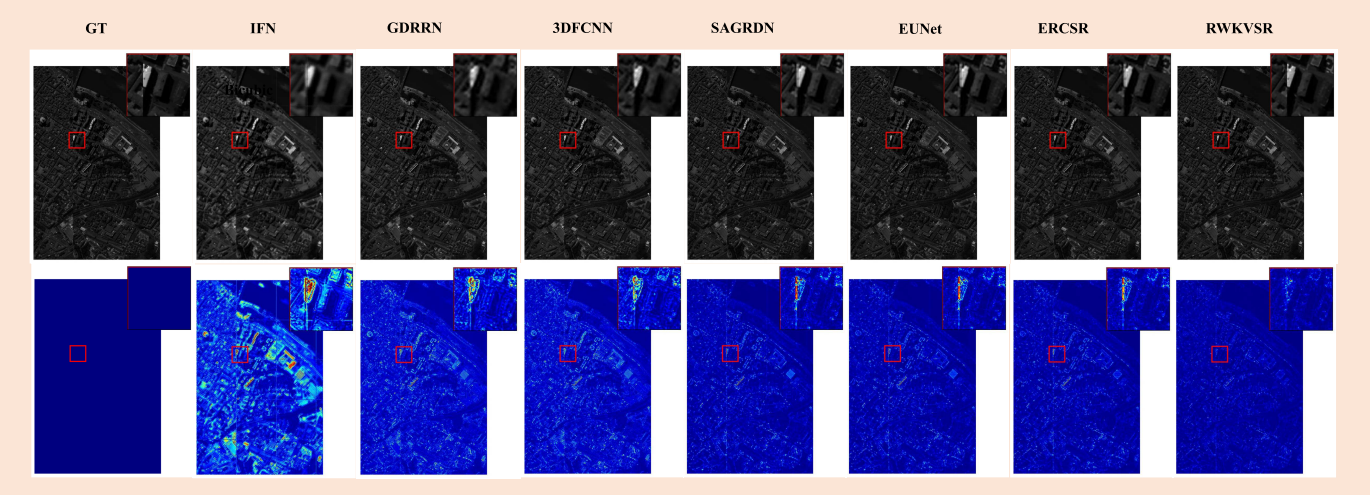

Fig. 6. Reconstruction results and corresponding absolute error maps obtained by different methods on the Pavia Center dataset. Blue regions in the absolute error maps indicate smaller deviations from the ground truth.

To provide a more intuitive comparison of the reconstruction performance of different methods in the spectral domain, several representative pixels were randomly selected, and their corresponding spectral curves reconstructed by 13 methods, including the proposed RWKVSR, were plotted. The spectral reconstruction results on the CAVE and Harvard datasets are shown in Fig. 8 and Fig. 9, respectively, while those on the Pavia Center dataset are presented in Fig. 10.

As shown in the figures, the spectral curves reconstructed by our method consistently align more closely with the ground truth across multiple pixels. This observation further confirms that the proposed method not only demonstrates superior

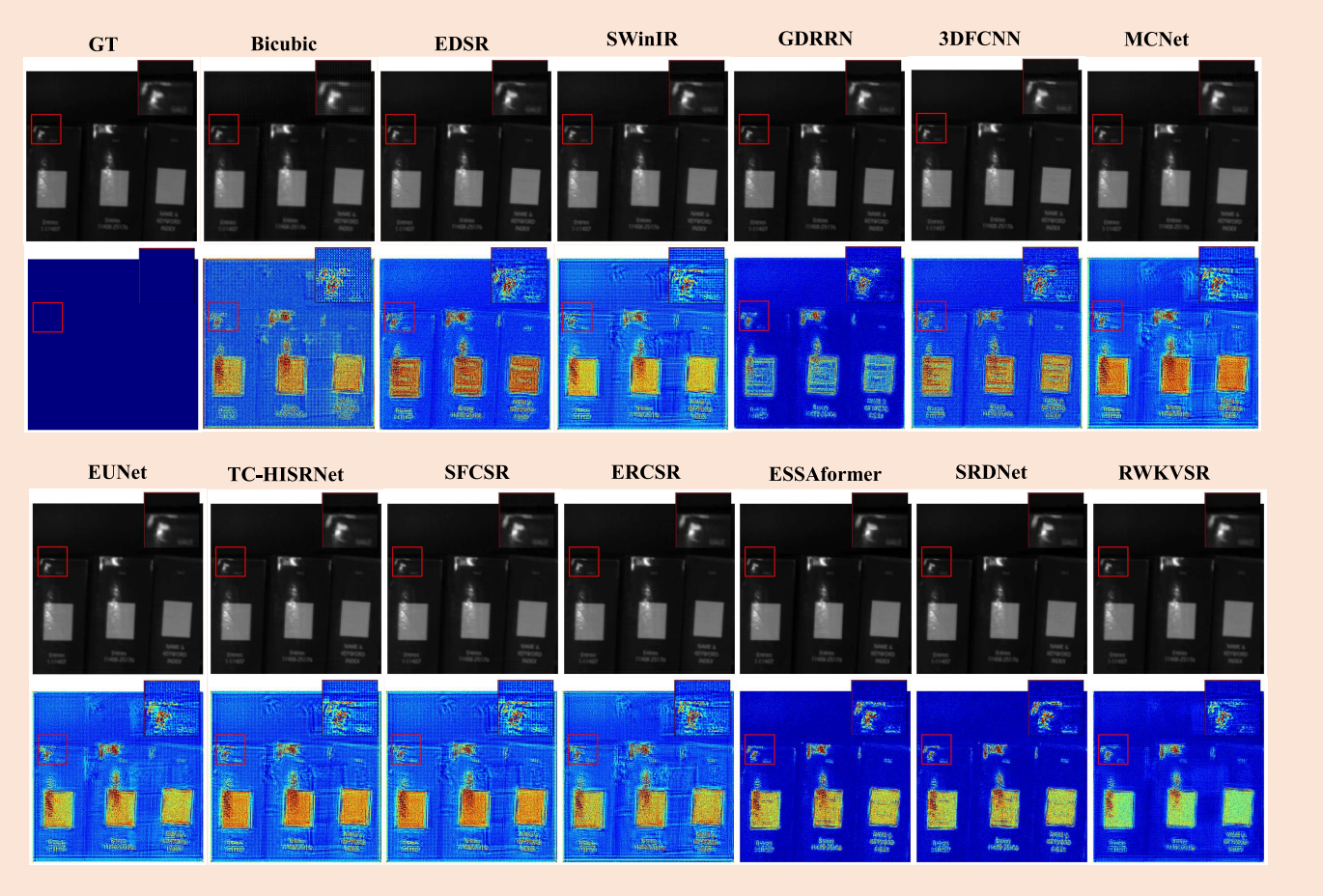

Fig. 7. Reconstruction results and corresponding absolute error maps generated by different methods, imgh2 of Harvard dataset. The bluer the area in the absolute error map, the smaller the absolute error of GT.

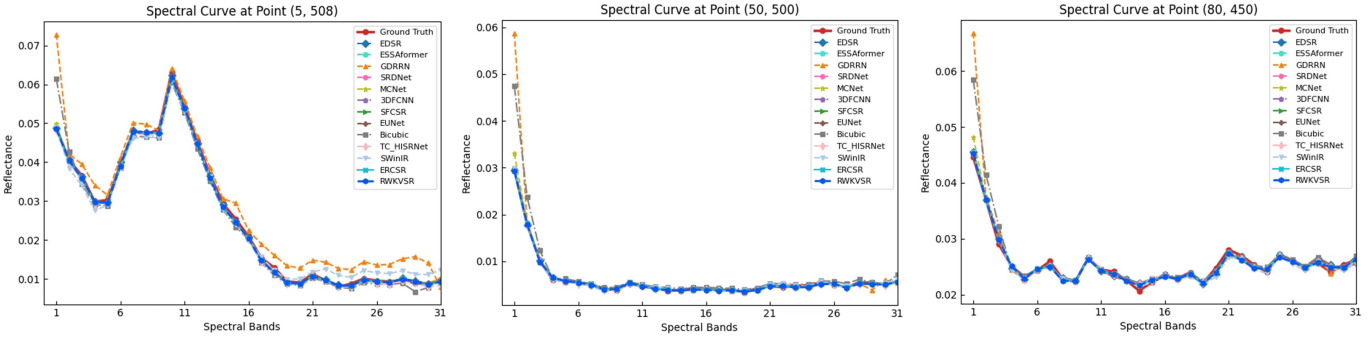

Fig. 8. Visual comparison of the `fake_and_real_apples_ms` image spectral curves at pixel locations (8, 508), (50, 500), and (80, 450) on the CAVE dataset.

performance in recovering spatial structures, but also excels in preserving spectral consistency—a critical factor in high-quality hyperspectral image reconstruction.

D. Model Analysis

In this section, we conduct an in-depth analysis of the performance of the proposed model on the CAVE dataset from four aspects:

1) **Number of SSRM Modules:** The Spectral-Spatial Residual Module (SSRM) is responsible for multi-scale feature extraction and fusion. Varying the number of SSRM modules allows the authors to identify the optimal network depth that maximizes representational power without inducing overfitting or excessive computational burden.
2) **Type of 3D Convolution:** Hyperspectral data exhibit inherent anisotropy between spatial and spectral dimensions. The proposed 3D Unit uses direction-aware convolutions to decouple these dimensions. The analysis compares this design against traditional isotropic 3D convolutions to validate its efficacy in capturing directional dependencies while reducing redundancy.
3) **Number of RWKV Units:** The RWKV module provides linear-complexity global attention. The number of units per SSRM affects the model's ability to capture

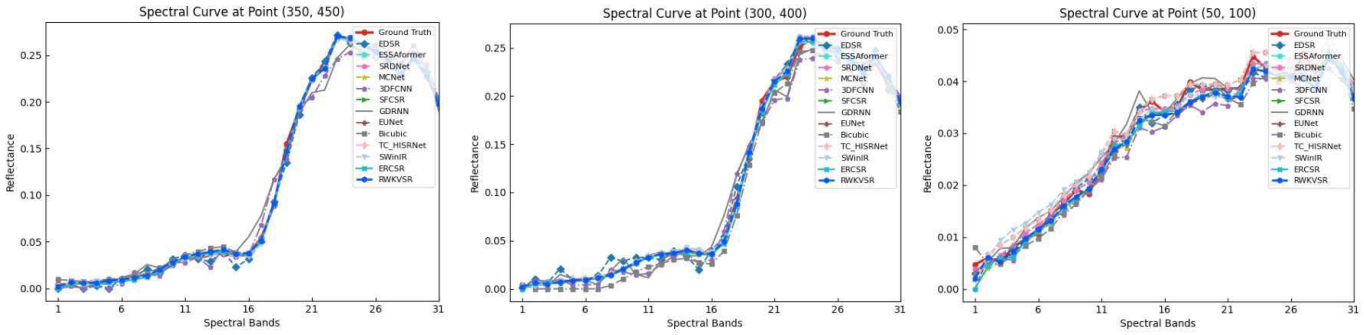

Fig. 9. Visual comparison of the imgh2 image spectral curves at pixel locations (350, 450), (300, 400), and (50, 100) on the Harvard dataset.

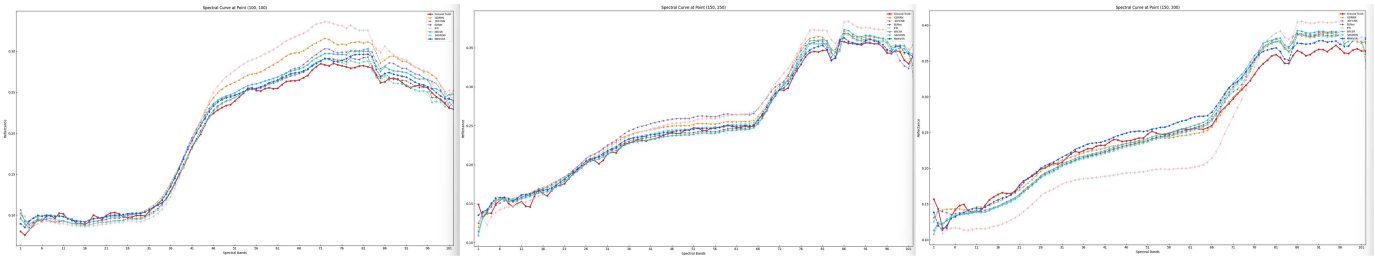

Fig. 10. Visual comparison spectral curves at pixel locations (100, 100), (150, 250), and (150, 300) on the Pavia Center dataset.

TABLE IV
ABLATION RESULTS ON MODEL COMPONENTS, ATTENTION MODULES, AND LOSS FUNCTIONS

Module	Setting	Param (K)	Memory (M)	Flops (G)	Time (s)	PSNR↑	SSIM↑	SAM↓
SSRM_num	×5	2036.974	983.78	5596.63	2.2920	45.986	0.9903	2.366
	×4 (ours)	1967.714	966.62	4659.68	1.8189	**46.003**	0.9906	**2.877**
	×3	1504.130	965.32	3722.73	1.1426	45.619	0.9890	2.468
	×2	1040.546	964.02	2785.78	1.0117	45.331	0.9886	2.510
	×0	1217.762	963.77	3143.09	0.4640	44.802	0.9878	2.596
3D conv	Separable 3D	1413.218	964.51	3536.280	2.0941	45.436	0.9890	2.683
	3D Unit (ours)	1967.714	966.62	4659.68	1.8189	**46.003**	**0.9906**	**2.877**
RWKV_num	×4	1976.930	966.65	6018.687	1.9518	45.953	**0.9908**	2.392
	×3 (ours)	1967.714	966.62	4659.68	1.8189	**46.003**	0.9906	**2.877**
	×2	1545.341	965.94	3806.21	0.9238	43.565	0.9851	3.035
	×0	1141.346	965.33	2990.18	0.4640	44.496	0.9872	2.632
RWKV	only spatial	1614.434	966.54	3940.392	1.0740	44.384	0.9873	2.661
	only spectral	1798.754	744.01	4314.860	1.4350	45.888	0.9905	2.422
	N/A	1445.474	966.46	3539.570	0.8244	45.614	0.9901	2.436
	All (ours)	1967.714	966.62	4659.68	1.8189	**46.003**	**0.9906**	**2.877**
Loss	L_1	1967.714	966.62	4659.68	1.4015	44.758	0.9877	2.680
	$L_1 + 0.05L_{\text{HFL}}$	1967.714	965.33	4659.68	1.7704	44.726	0.9878	3.069
	$L_1 + 0.10L_{\text{HFL}}$ (ours)	1967.714	966.62	4659.68	1.8189	**46.003**	**0.9906**	**2.877**
	$L_1 + 0.15L_{\text{HFL}}$	1967.714	965.33	4659.68	1.7704	44.684	0.9875	2.942
	$L_1 + 0.20L_{\text{HFL}}$	1967.714	965.33	4659.68	1.7704	45.020	0.9884	2.618

long-range dependencies. This analysis determines the optimal count that balances global contextual modeling with computational efficiency.

4) **Fusion via Loss Function:** The fusion of spatial and spectral information is further guided by the loss function. The comparison between $\mathcal{L}_1$ and $\mathcal{L}_1 + \mathcal{L}_{\text{HFL}}$ evaluates whether frequency-domain alignment enhances spectral consistency and preserves high-frequency details.

These aspects collectively address the core challenges in HSISR: efficient global modeling, anisotropic feature extraction, optimal network depth, and spectral-spatial consistency preservation. Taking the CAVE dataset with a magnification of 2 as an example, the relevant experimental results are shown in Table IV, showing the specific impact of each factor on the model performance.

1) Study on the number of SSRM modules: To determine the optimal number of SSRM modules, we set it to 2 to 5 and conducted comparative experiments. The results show that when the number increases from 2 to 4, the overall performance of indicators such as PSNR, SSIM and SAM improves, especially PSNR. However, when the number of modules increases to 5, the performance decreases, indicating that the model may be saturated or even overfitting. Therefore,

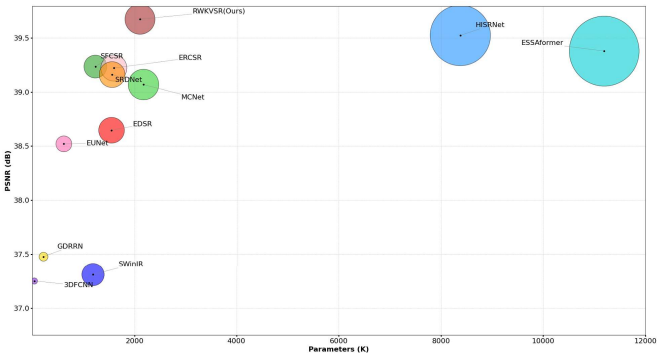

Fig. 11. Efficiency and performance comparison on the CAVE dataset with a scale factor of 3. Parameters are distributed on the horizontal axis, which represents accuracy.

4 SSRM modules are used as the default configuration in subsequent experiments.

2) Research on 3D convolution structure: This study introduces the 3D Unit module to jointly model the relationship between the spectrum and other dimensions in parallel. To verify its effectiveness, a comparative experiment with serial separable convolution was designed. Although the 3D Unit has a slight increase in the number of parameters and FLOPs, it performs better in PSNR and SSIM, and the detection time is also reduced. This shows that parallel convolution can more effectively mine multi-dimensional features, and SAEC was finally selected as the core structure.

3) Research on RWKV modules: Each 3D Unit module is followed by a RWKV module, forming 3 RWKV units. This module includes multi-scale spatial mixing, channel mixing and fusion operations, and uses multi-scale two-dimensional convolution to extract information of different receptive fields. In terms of quantity, increasing RWKV from 2 to 3 can significantly improve performance, but when it continues to increase to 4, the index decreases, so 3 modules are the best. In terms of structure, we tested different configurations to remove channel mixing. The results show that channel mixing is particularly critical to performance improvement, and the combination of spatial and channel mixing can achieve the best results, verifying the effectiveness of multi-dimensional collaborative modeling.

4) Study of loss function: L_1 loss is a widely used loss function, which is often used in image reconstruction tasks. In order to verify the effectiveness of the L_{HFL} loss we introduced, we conducted a set of comparative experiments, removing theL_{HFL} loss and only including the L_1 loss. The test results are shown in the last group of Table IV. When the L_{HFL} loss is added, the model has significantly improved in both PSNR and SSIM compared with the L_1 loss alone. It can be seen that the L_{HFL} loss plays an important role in improving the quality of hyperspectral image reconstruction.

5) Efficiency and performance comparison: To comprehensively evaluate the performance and resource efficiency of the methods, we conducted a comparative analysis of the models in terms of accuracy, memory usage, and number of parameters. As shown in Fig. 11 and Table III, the proposed method significantly outperforms the existing mainstream methods in terms of reconstruction accuracy. Although EDSR [21] and SwinIR are better in terms of memory usage, their accuracy is low and it is difficult to meet the reconstruction requirements of hyperspectral images. In contrast, thanks to the 3D convolution and global attention mechanism, although our method has a large number of parameters, it has obvious advantages in reconstruction quality and has the best overall performance.

6) Comprehensive Architectural Analysis Summary: The ablation study with "×0" configurations confirms that RWKVSR's design is balanced and synergistic. Each component significantly boosts performance without one dominating. The Receptance Weighted Key-Value mechanism provides 32.8% improvement through efficient global dependency modeling, followed by the Hyperspectral Frequency Loss (27.1%) for spectral consistency, the Spectral-Spatial Residual Module (26.1%) for multi-scale feature fusion, and the anisotropic 3D Unit (12.3%) for directional decoupling. A synergy gain of 1.7% further shows that the modules complement each other, while removing any component leads to a 1.201–1.507 dB PSNR drop, confirming the design's non-redundancy in addressing hyperspectral image super-resolution.

V. Conclusion

This paper addresses the critical challenges in hyperspectral image super-resolution (HSISR), specifically, computational inefficiency, inadequate spectral-spatial dependency modeling, and spectral distortion by proposing RWKVSR, a novel framework integrating three key innovations. First, the Receptance Weighted Key-Value (RWKV) module replaces quadratic self-attention with a linear-complexity decay mechanism, enabling efficient global spectral-spatial interaction at $\mathcal{O}(N)$ cost, a pioneering application of RWKV to HSISR. Second, the Spectral-Spatial Residual Module (SSRM) employs anisotropic 3D convolutions to decouple directional spectral-spatial correlations, resolving feature homogenization in traditional 3D CNNs while preserving fine-grained details. Third, the Hyperspectral Frequency Loss (HFL) optimizes spectral consistency by prioritizing high-frequency structural alignment in the Fourier domain, effectively mitigating distortion from pixel-level losses. Extensive experiments on the CAVE dataset demonstrate state-of-the-art performance. Visual and spectral analyses further validate superior spatial detail recovery and spectral fidelity. These advancements establish RWKVSR as a benchmark for balancing efficiency, accuracy, and practicality in HSISR. Future work will explore lightweight variants for real-time deployment and integrate physics-based priors to address atmospheric distortions, extending applicability to multi-modal remote sensing tasks.

References

[1] D. A. Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 17–28, Jan. 2002.

[2] R. C. Patel and M. V. Joshi, "Super-resolution of hyperspectral images: Use of optimum wavelet filter coefficients and sparsity regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1728–1736, Apr. 2015.

[3] C. Wang, W. Pedrycz, M. Zhou, and Z. Li, "Sparse regularization-based fuzzy C-means clustering incorporating morphological grayscale reconstruction and wavelet frames," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 7, pp. 1826–1840, Jul. 2021.

[4] C. Wang, W. Pedrycz, Z. Li, and M. Zhou, "Residual-driven fuzzy C-means clustering for image segmentation," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 4, pp. 876–889, Apr. 2021.

[5] W. Xie, X. Jia, Y. Li, and J. Lei, "Hyperspectral image super-resolution using deep feature matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 6055–6067, Aug. 2019.

[6] C. Wang, Z. Yan, W. Pedrycz, M. Zhou, and Z. Li, "A weighted fidelity and regularization-based method for mixed or unknown noise removal from images on graphs," *IEEE Trans. Image Process.*, vol. 29, pp. 5229–5243, 2020.

[7] C. Wang, M. Zhou, W. Pedrycz, and Z. Li, "Comparative study on noise-estimation-based fuzzy C-means clustering for image segmentation," *IEEE Trans. Cybern.*, vol. 54, no. 1, pp. 241–253, Jan. 2024.

[8] C. Wang, W. Pedrycz, Z. Li, M. Zhou, and S. S. Ge, "G-image segmentation: Similarity-preserving fuzzy C-means with spatial information constraint in wavelet space," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 12, pp. 3887–3898, Dec. 2021.

[9] Z. Pan et al., "Super-resolution based on compressive sensing and structural self-similarity for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4864–4876, Sep. 2013.

[10] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[11] 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.

[12] Z. He et al., "DADIGAN: A dual attention blocks-based disentangled iterative generative adversarial network for cloud and shadow removal on SAR and optical images," *Inf. Fusion*, vol. 125, Jan. 2026, Art. no. 103487.

[13] M. Zhang, C. Zhang, Q. Zhang, J. Guo, X. Gao, and J. Zhang, "ESSAformer: Efficient transformer for hyperspectral image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 23073–23084.

[14] L. Yang et al., "TC-HISRNet: Hyperspectral image super-resolution network based on contextual band joint transformer and CNN," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 9632–9645, 2023.

[15] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Proc. 1st Conf. Lang. Modeling*, 2023, pp. 1–6.

[16] B. Li, X. Wang, and H. Xu, "SSRMamba: Efficient visual state space model for spectral super-resolution," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5.

[17] B. Peng et al., "RWKV: Reinventing RNNs for the transformer era," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2023, pp. 1–9.

[18] K. Wang, X. Liao, J. Li, D. Meng, and Y. Wang, "Hyperspectral image super-resolution via knowledge-driven deep unrolling and transformer embedded convolutional recurrent neural network," *IEEE Trans. Image Process.*, vol. 32, pp. 4581–4594, 2023.

[19] 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. (CVPR)*, Jun. 2016, pp. 1646–1654.

[20] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4809–4817.

[21] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.

[22] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.

[23] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2790–2798.

[24] Y. Qiu, R. Wang, D. Tao, and J. Cheng, "Embedded block residual network: A recursive restoration model for single-image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4179–4188.

[25] V. V. Duong, T. N. Huu, J. Yim, and B. Jeon, "A fast and efficient super-resolution network using hierarchical dense residual learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1809–1813.

[26] K. Jiang, Z. Wang, P. Yi, J. Jiang, J. Xiao, and Y. Yao, "Deep distillation recursive network for remote sensing imagery super-resolution," *Remote Sens.*, vol. 10, no. 11, p. 1700, Oct. 2018.

[27] S. Mei, X. Yuan, J. Ji, Y. Zhang, S. Wan, and Q. Du, "Hyperspectral image spatial super-resolution via 3D full convolutional neural network," *Remote Sens.*, vol. 9, no. 11, p. 1139, Nov. 2017.

[28] Q. Li, Q. Wang, and X. Li, "Exploring the relationship between 2D/3D convolution for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8693–8703, Oct. 2021.

[29] D. Liu et al., "An efficient unfolding network with disentangled spatial–spectral representation for hyperspectral image super-resolution," *Inf. Fusion*, vol. 94, pp. 92–111, Jun. 2023.

[30] T. Liu, Y. Liu, C. Zhang, L. Yuan, X. Sui, and Q. Chen, "Hyperspectral image super-resolution via dual-domain network based on hybrid convolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–18, 2024.

[31] H. Wang, C. Wang, and Y. Yuan, "Asymmetric dual-direction quasi-recursive network for single hyperspectral image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6331–6346, Nov. 2023.

[32] X. Wang, J. Chen, Q. Wei, and C. Richard, "Hyperspectral image super-resolution via deep prior regularization with parameter estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1708–1723, Apr. 2022.

[33] M. Zhang et al., "Spatial–spectral aggregation transformer with diffusion prior for hyperspectral image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 4, pp. 3557–3572, Apr. 2025.

[34] S. Jia, Z. Min, and X. Fu, "Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 96, pp. 117–129, Aug. 2023.

[35] X. Cao, X. Wang, X. Dun, Y. Lian, X. Cheng, and X. Hao, "Cross-domain-aware deep unfolding transformer for hyperspectral image super-resolution," *Pattern Recognit.*, vol. 172, Apr. 2026, Art. no. 112374.

[36] P. Wang et al., "MT_GAN: A SAR-to-optical image translation method for cloud removal," *ISPRS J. Photogramm. Remote Sens.*, vol. 225, pp. 180–195, Jul. 2025.

[37] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Global context networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6881–6895, Jun. 2020.

[38] 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.

[39] X. Cao, Y. Lian, K. Wang, C. Ma, and X. Xu, "Unsupervised hybrid network of transformer and CNN for blind hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

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

[41] J. Yao, D. Hong, C. Li, and J. Chanussot, "SpectralMamba: Efficient Mamba for hyperspectral image classification," 2024, *arXiv:2404.08489*.

[42] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S. Xia, "MambaIR: A simple baseline for image restoration with state-space model," in *Proc. Eur. Conf. Comput. Vis.*Cham, Switzerland: Springer, 2024, pp. 222–241.

[43] M. Zhang, J. Xu, J. Zhang, H. Zhao, W. Shang, and X. Gao, "SPH-Net: Hyperspectral image super-resolution via smoothed particle hydrodynamics modeling," *IEEE Trans. Cybern.*, vol. 54, no. 7, pp. 4150–4163, Jul. 2024.

[44] M. Zhang, Q. Wu, J. Zhang, X. Gao, J. Guo, and D. Tao, "Fluid micelle network for image super-resolution reconstruction," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 578–591, Jan. 2023.

[45] M. Zhang, Q. Wu, J. Guo, Y. Li, and X. Gao, "Heat transfer-inspired network for image super-resolution reconstruction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 1810–1820, Feb. 2024.

[46] A. Sanchez-Caballero et al., "3DFCNN: Real-time action recognition using 3D deep neural networks with raw depth information," 2020, *arXiv:2006.07743*.

[47] Q. Li, Q. Wang, and X. Li, "Mixed 2D/3D convolutional network for hyperspectral image super-resolution," *Remote Sens.*, vol. 12, no. 10, p. 1660, May 2020.

[48] Q. Wang, Q. Li, and X. Li, "Hyperspectral image superresolution using spectrum and feature context," *IEEE Trans. Ind. Electron.*, vol. 68, no. 11, pp. 11276–11285, Nov. 2021.

[49] Y. Li, L. Zhang, C. Dingl, W. Wei, and Y. Zhang, "Single hyperspectral image super-resolution with grouped deep recursive residual network," in *Proc. IEEE 4th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2018, pp. 1–4.

Xiaofei Yang (Member, IEEE) received the B.S. degree from Suihua University, Suihua, China, in 2011, and the M.S. and Ph.D. degrees from Harbin Institute of Technology, China, in 2014 and 2019, respectively.

He was a Post-Doctoral Researcher with the Department of Computer and Information Science, University of Macau, China, from 2020 to 2023. He is currently a Lecturer with the School of Electronics and Communications Engineering, Guangzhou University. His research interests are in the areas of semi-supervised learning, deep learning, remote sensing, transfer learning, and graph mining.

Sihuan Li is currently pursuing the master's degree with the School of Electronic and Information Engineering, Guangzhou University.

Her research interests include hyperspectral image super-resolution and remote sensing technology based on deep learning, focuses on building an efficient spectral-spatial joint reconstruction model to improve the resolution and quality of hyperspectral images.

Weijia Cao (Member, IEEE) received the master's and Ph.D. degrees in computer science from the University of Macau, Macau, China, in 2013 and 2017, respectively.

She is currently an Associate Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. Her main research interests revolve around machine learning and remote sensing image processing.

Dong Tang (Member, IEEE) received the B.S. degree from Nanhua University, Hengyang, China, in 1989, the M.S. degree from Hunan University, Changsha, China, in 1999, and the Ph.D. degree in communications and information systems from Sun Yat-sen University, Guangzhou, in 2006.

From 2014 to 2015, he was a Research Fellow with the University of California at Irvine, CA, USA. Currently, he is a Professor with the School of Electronics and Communications Engineering, Guangzhou University, Guangzhou. His main research interests include signal processing, deep learning, intelligent network systems, and wireless communications.

Yifang Ban (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from Nanjing University, Nanjing, China, in 1984 and 1987, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 1996.

Before joining KTH Royal Institute of Technology (KTH), Stockholm, Sweden, in 2004, she was a tenured Associate Professor with York University, Toronto, ON, Canada. She is currently the Chair Professor and the Director of the Geoinformatics Division and the Associate Director of Digital Futures in Stockholm. Her research interests include Earth observation big data analytics, machine learning/deep learning and their applications in mapping urban and land cover, monitoring urbanization, wildfires, other environmental changes, and assessing environmental impact. She has published extensively on these topics. She is the Co-Chair of the ICA Commission on Sensor-Driven Mapping and a Co-Lead of the Group on Earth Observations (GEO) Initiative "Global Urban Observation and Information" (2012–2022). Since 2016, she has been an Invited Expert of the UN Habitat Technical Committee on Human Settlements Indicators for UN Sustainable Development Goals (SDGs). She has been an associate editor and a guest editor for major remote sensing journals and an invited expert for EU and national grant application evaluations.

Yicong Zhou (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Tufts University, Medford, MA, USA.

He is a Professor with the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security. He is a fellow of the Society of Photo-Optical Instrumentation Engineers (SPIE). He was recognized as one of "Highly Cited Researchers" in 2020, 2021, 2023, and 2024. He serves as a Senior Area Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY; and an Associate Editor for IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.