remote sensing

Article

ConvMambaSR: Leveraging State-Space Models and CNNs in a Dual-Branch Architecture for Remote Sensing Imagery Super-Resolution

Qiwei Zhu [1,2], Guojing Zhang [1,2,*], Xuechao Zou [3], Xiaoying Wang [1,2], Jianqiang Huang [1,2] and Xilai Li [4]

[1] School of Computer Technology and Application, Qinghai University, Xining 810016, China; qiweizhu@qhu.edu.cn (Q.Z.); wang_cta@qhu.edu.cn (X.W.); 2011990026@qhu.edu.cn (J.H.)
[2] Intelligent Computing and Application Laboratory of Qinghai Province, Qinghai University, Xining 810016, China
[3] School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China; xuechaozou@bjtu.edu.cn
[4] College of Agriculture and Animal Husbandry, Qinghai University, Xining 810016, China; 1985990024@qhu.edu.cn
* Correspondence: zhanggj@qhu.edu.cn

Abstract: Deep learning-based super-resolution (SR) techniques play a crucial role in enhancing the spatial resolution of images. However, remote sensing images present substantial challenges due to their diverse features, complex structures, and significant size variations in ground objects. Moreover, recovering lost details from low-resolution remote sensing images with complex and unknown degradations, such as downsampling, noise, and compression, remains a critical issue. To address these challenges, we propose ConvMambaSR, a novel super-resolution framework that integrates state-space models (SSMs) and Convolutional Neural Networks (CNNs). This framework is specifically designed to handle heterogeneous and complex ground features, as well as unknown degradations in remote sensing imagery. ConvMambaSR leverages SSMs to model global dependencies, activating more pixels in the super-resolution task. Concurrently, it employs CNNs to extract local detail features, enhancing the model's ability to capture image textures and edges. Furthermore, we have developed a global–detail reconstruction module (GDRM) to integrate diverse levels of global and local information efficiently. We rigorously validated the proposed method on two distinct datasets, RSSCN7 and RSSRD-KQ, and benchmarked its performance against state-of-the-art SR models. Experiments show that our method achieves SOTA PSNR values of 26.06 and 24.29 on these datasets, respectively, and is visually superior, effectively addressing a variety of scenarios and significantly outperforming existing methods.

Keywords: remote sensing; state-space models; convolutional neural networks; super-resolution

check for updates

Citation: Zhu, Q.; Zhang, G.; Zou, X.; Wang, X.; Huang, J.; Li, X. ConvMambaSR: Leveraging State-Space Models and CNNs in a Dual-Branch Architecture for Remote Sensing Imagery Super-Resolution. *Remote Sens.* **2024**, *16*, 3254. https://doi.org/10.3390/rs16173254

Academic Editor: Gong Cheng

Received: 22 July 2024
Revised: 15 August 2024
Accepted: 27 August 2024
Published: 2 September 2024

1. Introduction

High-resolution (HR) remote sensing imagery is of great importance in urban planning and management [1], agricultural resource optimization [2], biodiversity conservation [3], climate change research [4], environmental monitoring [5], and other fields. For instance, in environmental monitoring, HR imagery can identify changes in land cover [6], vegetation health [7], and water resources [8], which can assist in evaluating ecological trends and assessing the effectiveness of environmental protection efforts.

However, the acquisition of high-resolution remote sensing images is often constrained by technical and financial limitations. The image super-resolution technique employs an advanced algorithm to reconstruct the relatively low-resolution (LR) image from the HR image, thereby offering a cost-effective and efficient means of acquiring HR images.

In recent years, single-image super-resolution (SISR) techniques have emerged as a significant area of research [9,10]. The core objective of this technique is to recover an

HR image from a single LR image. Offering advantages such as cost-effectiveness and efficiency, SISR techniques have attracted extensive academic attention [11], especially deep learning-based methods [12], which have made significant progress in reconstructing images with detail and clarity. Deep learning-based SISR methods rely on training a large number of LR/HR image pairs to build mapping models. In addition, the performance of super-resolution networks can be significantly improved by introducing techniques such as residual connections [13], dense connections [14], generative adversarial networks [15], and attention mechanisms [16]. Despite the success of SISR techniques in many fields, their application in remote sensing images still faces many challenges. Ground objects in remote sensing images exhibit complex structural characteristics, significant size differences, and a substantial amount of noise, in addition to highly localized features [17]. These characteristics present significant challenges for the application of SISR to remote sensing.

Despite the advancements in SISR methods, several fundamental challenges remain unresolved, particularly in capturing long-range dependencies and handling complex and unknown degradations in remote sensing images. These challenges highlight the need for a more sophisticated approach that can overcome the limitations of both CNN-based and Transformer-based models.

Typical CNN-based SR methods include SRCNN [18], RRDBNet [14], and EDSR [19]. SRCNN is the first CNN-based image SR method that learns an end-to-end nonlinear mapping from LR to HR images through a three-layer convolutional network. RRDBNet combines the advantages of residue-in-residue and dense block residue-in-residue structures and is widely used in high-magnification super-resolution and generative adversarial training. EDSR improves super-resolution performance by removing the batch normalization layer and increasing the depth and width of the network. Although these CNN-based SR methods have significantly improved in performance, they are constrained by the receptive field due to the limitations of convolutional operations, which impede their ability to capture global contextual information [20]. Furthermore, while the superior efficiency of convolutional parallel operations makes them well-suited for deployment on resource-constrained devices, these methods face significant challenges in processing remotely sensed images with complex structures and diverse features. These factors present significant challenges to the super-resolution reconstruction of remote sensing images.

Transformer-based deep learning models have achieved state-of-the-art performance in a variety of computer vision applications [21–24], which have demonstrated an efficient ability to capture global background information by utilizing the self-attention mechanism [25], Transformer-based SR methods have also evolved significantly. Despite its global receptive field, Transformer [26] exhibits quadratic complexity in processing input sequences, which presents a challenge when dealing with common large-size image restoration tasks. SwinIR [20] represents a state-of-the-art approach to SISR tasks based on Swin Transformer [27]. In comparison with purely convolutional structures and ViT-based architectures, SwinIR is more efficient in public datasets such as DIV2K. Nevertheless, certain studies have demonstrated that the performance of a single Transformer may not be superior to that of a CNN due to the compression of image blocks into a 1D sequence, which may result in the loss of structural information [28]. Conversely, the incorporation of efficient attention techniques, such as the window-shift attention mechanism [27], often entails the compromise of a globally effective receptive field, indicating that there is a trade-off between the global receptive field and efficient computation.

The challenge in modeling long-range dependencies stems from the inherent limitations of convolutional operations, which are restricted by their local receptive fields. Transformer models, while capable of capturing global context, struggle with computational complexity and may lose structural information during the process of sequence transformation. These issues necessitate a re-evaluation of existing methodologies and drive the development of novel approaches.

Recently, state-space models derived from control systems have attracted attention for their linear complexity in dealing with input sequences. In particular, the enhanced version of Mamba has become an efficient and effective backbone for developing complex networks [29–33]. The discrete state-space equations in Mamba can be formalized into recursive form and, when equipped with a specifically designed structured parameterization [34], very long dependencies can be modeled. However, the standard Mamba [30] designed for 1D sequential data in NLP processes 1D image sequences recursively, which may result in spatially close pixels being very far away in the spread sequence. This can lead to localized pixel forgetting problems that are not suitable for image super-resolution scenarios.

To address the aforementioned challenges, this paper proposes a novel remote sensing image super-resolution framework (ConvMambaSR) for hybrid models, offering a new perspective on efficiently modeling long-range dependencies while maintaining the integrity of local details in remote sensing images. The framework is comprised of three principal steps: shallow feature extraction, deep feature extraction, and upsampling. In the deep feature extraction stage, ConvMambaSR employs an elaborate residual state-space group (RSSG) and residual convolution group (RCG) to capture high-dimensional features at different levels. Subsequently, GDRM is employed to facilitate the efficient integration of local details and global contextual information.

The principal contributions of this paper are summarized as follows:

1. ConvMambaSR is proposed as a hybrid model combining CNN and Mamba. It employs a dual-branch architecture: the CNN branch extracts local features and processes spatial information, while the Mamba branch captures global features and long-range dependencies.
2. A global–detail reconstruction module is introduced within ConvMambaSR, designed to integrate local details from the CNN with global contextual information from the Mamba. This module enhances the synergy between the branches by merging local features with global information, thereby improving model performance across various tasks.

2. Related Works

2.1. Advances in SISR and Applications to Remote Sensing

Recent advancements in SISR have been driven by both Convolutional Neural Networks and Vision Transformers.

CNN-based methods have achieved significant milestones, starting with the SRCNN model proposed by Dong et al. [18], which pioneered deep learning in SR. Shi et al. [35] introduced subpixel convolution, while Ledig et al. [36] incorporated ResNet and GANs into SISR, resulting in models like SRGAN [36] and ESRGAN [37]. Attention mechanisms were introduced by Zhang et al. [16], leading to advanced models such as RCAN [16] and HAN [38]. Despite their success, these CNN-based models often struggle with modeling long-range dependencies [20,39].

Vision Transformers, introduced by Dosovitskiy et al. [40], reshaped visual tasks by treating images as sequences of patches. The Swin Transformer [27] reduced computational complexity through window-based self-attention and became the backbone for many visual tasks, including SISR. Liang et al. [20] developed SwinIR, and subsequent improvements like HAT [41] and NGswin [42] achieved competitive reconstruction performance in capturing long-range dependencies and cross-window connections.

In the realm of remote sensing, SISR has gained traction due to the challenges of acquiring multiple images for Multi-Image SR (MISR) techniques [43]. Building on natural image SR techniques, models like SRCNN were adapted to remote sensing by Ducournau and Fablet [44], while other advancements include RDBPN [45], PMSRN [46], and hybrid models like SWCGAN [47], HSTNet [48], and ConvFormerSR [49], enhancing spatial resolution and spectral consistency in remote sensing images.

2.2. State-Space Models in Deep Learning

State-space models have recently gained prominence in deep learning, particularly for their ability to address long-range dependency challenges by drawing on continuous state-space modeling from control systems [32,50,51]. A notable example is the Structured State-Space Sequence model (S4) [50], which uses parameter normalization with diagonal structures as an alternative to CNNs and Transformers for modeling long-distance dependencies. Building on this, the S5 model [32] introduces multiple-input multiple-output (MIMO) SSM and efficient parallel scanning, while the gated state-space layer [31] enhances the expressive power of S4 by incorporating gating mechanisms.

Recently, Gu et al. [30] introduced a data-dependent SSM layer and the Mamba language model backbone, which outperforms Transformers on large-scale real-world data with linear scalability for sequence lengths. Mamba's computational efficiency in image processing highlights the potential of SSMs in image restoration, offering novel insights and advantages over traditional deep learning models.

3. Methodology

In this section, we first introduce the overall structure of ConvMambaSR and then describe its three important modules, namely RSSG, RCG, and GDRM.

3.1. Overall Structure of ConvMambaSR

As illustrated in Figure 1, ConvMambaSR is comprised of three stages: shallow feature extraction, deep feature extraction, and high-quality reconstruction. In the shallow feature extraction stage, ConvMambaSR first extracts shallowfeatures $\mathbf{F}_S \in \mathbb{R}^{H\times W\times C}$ from the low-resolution input image $\mathbf{I}_{LQ} \in \mathbb{R}^{H\times W\times 3}$ through a 3 × 3 convolutional layer, where H and W denote the height and width of the input image, respectively, and C denotes the number of channels.

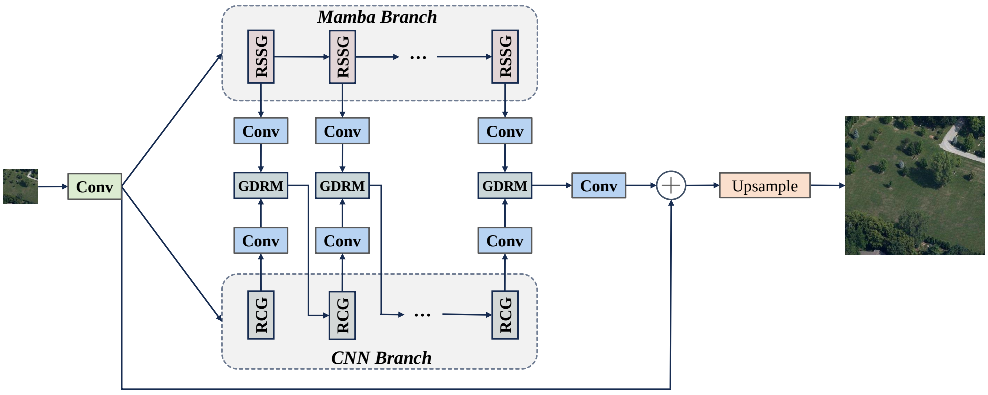


Figure 1. Architecture of the proposed ConvMambaSR.

In the deep feature extraction stage, the deep features $\mathbf{F}_{D\_m}$ and $\mathbf{F}_{D\_c}$ are obtained by parallel Mamba branching and CNN branching, respectively. These features are then fused by the GDRM module.

In the high-quality reconstruction stage, global residual concatenation is employed to integrate the low-level features with the deep-level features, thereby generating the input $\mathbf{F}_R$ for the high-quality reconstruction stage. Finally, the pixel rearrangement method [35] is utilized for upsampling, resulting in the SR result $\mathbf{I}^{SR}$.

The SR process of ConvMambaSR can be expressed mathematically as

$$\mathbf{F}_R = G(\mathbf{F}_{D\_m}, \mathbf{F}_{D\_c}) + \mathbf{F}_S \tag{1}$$

$$\mathbf{I}^{\mathrm{SR}} = \mathrm{Up}(\mathbf{F}_R) \tag{2}$$

where $\mathrm{Up}(\cdot)$ denotes the upsampling function and $\mathrm{G}(\cdot)$ denotes the global–detail reconstruction operation. $\mathbf{I}^{SR}$ is the reconstructed super-resolution image.

3.2. Residual State-Space Group

3.2.1. Vision State-Space Module

Transformer-based super-resolution networks typically partition the input into small patches [52] or employ shift-window attention [20] to ensure efficiency. However, this approach hinders interaction at the whole image level. In response, we introduce the Visual State-Space Module (VSSM) [53] to the image super-resolution task, thereby enabling the model to benefit from Mamba's success in long-range modeling of linear complexity.

VSSM is capable of capturing long-range dependencies through the use of state-space equations. As illustrated in Figure 2, which follows [53], the input feature $\mathbf{F}_{in} \in \mathbb{R}^{H \times W \times C}$ will undergo processing through two parallel branches. In the initial branch, the feature channel is expanded to βC through a linear layer, where β is a predefined channel extension factor. This is followed by a depth-wise convolution, a SiLU [54] activation function, and a 2D Selective Scanning Module (2D-SSM) layer [55] and LayerNorm (LN). In the subsequent branch, the feature channel is also extended to βC through a linear layer, followed by the SiLU activation function. Subsequently, the features from both branches are aggregated using the Hadamard product. Finally, the number of channels is projected back to C, generating an output of the same shape as the input X_{out}:

$$\mathbf{F}_1 = \text{LN}(\text{2DSSM}(\sigma(\text{DWConv}(\text{Linear}(\mathbf{F}_{in}))))) \tag{3}$$

$$\mathbf{F}_2 = \sigma(\text{Linear}(\mathbf{F}_{in})) \tag{4}$$

$$\mathbf{F}_{out} = \text{Linear}(\mathbf{F}_1 \odot \mathbf{F}_2) \tag{5}$$

where DWConv denotes the depth-wise convolution, σ denotes the SiLU activation function, and $\odot$ denotes the Hadamard product.

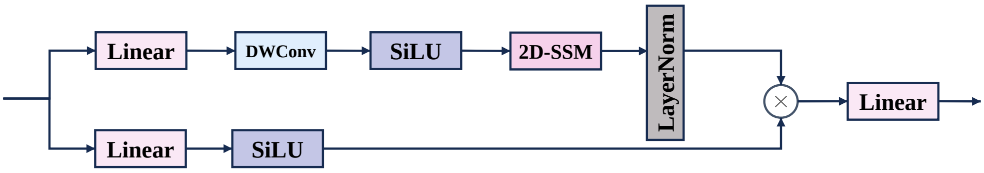


Figure 2. Structure of VSSM, which is a component of the RSSB.

3.2.2. Two-Dimensional Selective Scan Module

Standard Mamba [30] captures information in scanned data during causal processing, making it well-suited for sequential NLP tasks but less effective for noncausal data-like images. We introduce the 2D Selective Scanning Module, as shown in Figure 3. Two-dimensional image features are converted into 1D sequences and scanned in four directions, and long-range dependencies are captured using the discrete state-space equation. Finally, the sequences are merged and reshaped to restore the 2D structure.

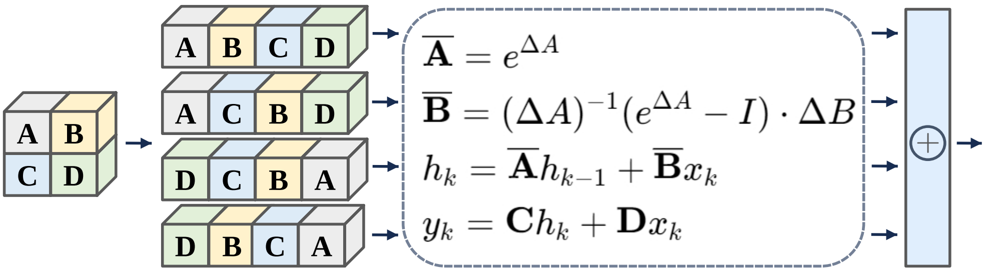

A B C D
A B C D
A C B D
D C B A
D B C A
$\overline{\mathbf{A}} = e^{\Delta A}$
$\overline{\mathbf{B}} = (\Delta A)^{-1}(e^{\Delta A} - I) \cdot \Delta B$
$h_k = \overline{\mathbf{A}} h_{k-1} + \overline{\mathbf{B}} x_k$
$y_k = \mathbf{C} h_k + \mathbf{D} x_k$

Figure 3. Structure of 2D-SSM, which is a component of the VSSM.

3.2.3. Residual State-Space Block

Since SSM deals with flattened feature maps as one-dimensional token sequences, it has been demonstrated that the number of neighboring pixels in the sequence can be significantly influenced by the flattening strategy [55]. When employing the four-direction unfolding strategy proposed by [53], only four nearest neighbors are accessible for anchor pixels. This discrepancy between spatial proximity in a two-dimensional feature map and temporal proximity in a one-dimensional token sequence can result in local pixel forgetting. Furthermore, SSMs typically incorporate a substantial number of hidden states to accommodate long-range dependencies, and there is a notable degree of channel redundancy in the activation results of different channels [55]. To address these issues, we propose the incorporation of a Channel Attention Block (CAB) within the residual state-space block (RSSB) framework. In this manner, SSMs can assist in alleviating the local pixel forgetting issue by extracting local features through convolution. Furthermore, it can be configured to prioritize the learning of diverse channel representations while avoiding channel redundancy by selecting key channels through subsequent Channel Attention.

As illustrated in Figure 4, given a feature map $\mathbf{F}_A \in \mathbb{R}^{C\times H\times W}$, CAB is first compressed by a compression factor γ_1 to obtain features with shape $\mathbb{R}^{\frac{c}{\gamma_1}\times H\times W}$. Thereafter, a channel expansion is performed to recover the original shape. By first compressing and then expanding the channel dimensions, the model can efficiently learn the different channels' nonlinear interactions while constraining the model complexity. Subsequently, Channel Attention (CA) [56] is introduced. CAB is calculated as follows:

$$\mathbf{F}_B = W_1\sigma(W_0(\mathbf{F}_A)) \tag{6}$$

$$\mathbf{F}_C = \mathbf{F}_B \otimes \mathcal{S}(W_3\sigma(W_2(\mathbf{F}^B_{\text{avg}}))) \tag{7}$$

where σ denotes the StarReLU [57] activation function, expressed as

$$\text{StarReLU}(x) = s\cdot(\text{ReLU}(x))^2 + b \tag{8}$$

where $s \in \mathbb{R}$ and $b \in \mathbb{R}$ are scalars of scale and bias, respectively. $W_0 \in \mathbb{R}^{\frac{c}{\gamma 1}\times C}$, $W_1 \in \mathbb{R}^{C\times\frac{C}{\gamma 1}}$, $W_2 \in \mathbb{R}^{\frac{c}{\gamma 2}\times C}$, $W_3 \in \mathbb{R}^{C\times\frac{C}{\gamma 2}}$, $\mathbf{F}^B_{\text{avg}}$ denotes global average pooling over the features $\mathbf{F}_B$, $\mathcal{S}(\cdot)$ denotes the sigmoid function, $\otimes$ denotes matrix multiplication, and $\gamma 1$, $\gamma 2$ denote two different compression factors.

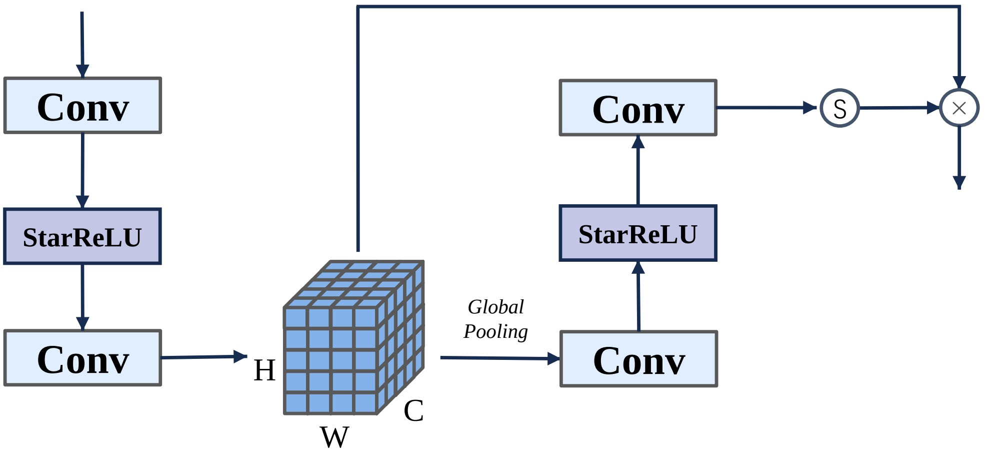


Figure 4. Structure of CAB, which is a component of the RSSB.

As illustrated in Figure 5, given the input deep features $\mathbf{F}^{l-1}_{D\_m} \in \mathbb{R}^{H\times W\times C}$, we first use LayerNorm, followed by the Visual State-Space Module [53] to capture spatial long-range dependencies, using a learnable scaling factor $s \in \mathbb{R}^C$ to control the jumps from the information obtained from the connections. At this point, the RSSB can be expressed as

$$\mathbf{F}^{l}_{D\_m} = \text{VSSM}(\text{LN}(\mathbf{F}^{l-1}_{D\_m})) + s\cdot\mathbf{F}^{l-1}_{D\_m} \tag{9}$$

$$\mathbf{F}^{l}_{D\_m} = \text{CAB}(\text{LN}(\mathbf{F}^{l}_{D\_m})) + s'\cdot\mathbf{F}^{l}_{D\_m} \tag{10}$$

where s and s' denote different learnable scaling factors.

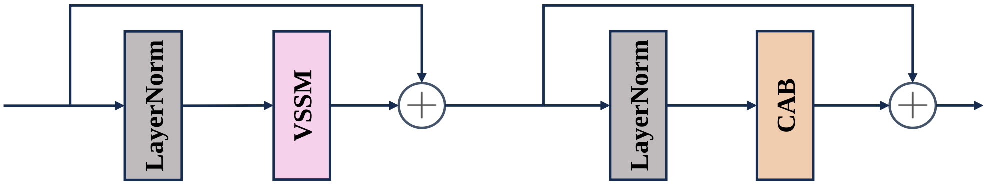


Figure 5. Structure of RSSB. A series of RSSB forms the RSSG in the Mamba branch depicted in Figure 1.

Ultimately, the RSSG can be expressed as

$$\text{RSSG}^k = \text{RSSG}^{k-1} + W\mathbf{F}^l_{D\_m}(\mathbf{F}^{l-1}_{D\_m}(\ldots\mathbf{F}^1_{D\_m}(\text{RSSG}^{k-1})\ldots)) \tag{11}$$

where RSSG^k and RSSG^{k-1} denote the kth RSSG and $k-1$th RSSG, respectively. $\mathbf{F}^l_{D\_m}$ is the lth RSSB. W is a convolutional layer that serves to enhance the translational equivariance of the Mamba layer.

3.3. Residual Convolution Group

It has been demonstrated that deep convolutional networks can enhance super-resolution performance [16,19]. Accordingly, as illustrated in Figure 6, we constructed RCGs with successive residual units to capture the details of LR images. A given RCB can be expressed as

$$\mathbf{F}^{l+1}_{D\_c} = \mathbf{F}^l_{D\_c} + W_1\sigma(W_0(\mathbf{F}^l_{D\_c})) \tag{12}$$

where σ denotes the leaky ReLU activation function.

The N RCBs are connected to form an RCG module expressed as

$$\text{RCG}^k = \mathbf{F}^{N-1}_{D\_c} + W_1\sigma(W_0(\mathbf{F}^{N-1}_{D\_c})), N \geq 1 \tag{13}$$

$$\mathbf{F}^0_{D\_c} = \text{RCG}^{k-1} \tag{14}$$

where RCG^k is the kth RCG feature map.

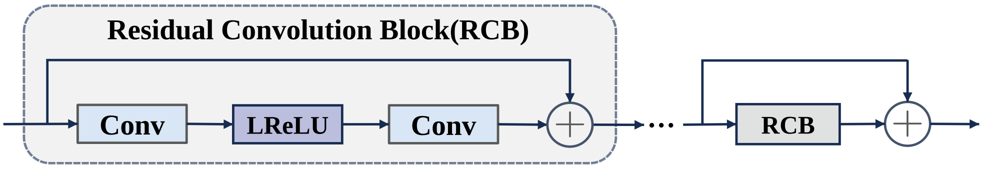


Figure 6. Structure of RCG, which corresponds to the CNN branch in Figure 1.

3.4. Global–Detail Reconstruction Module

CNN is capable of fully exploiting spatial and channel information in the receptive field, yet it is deficient in explicitly modeling inter-channel relationships [58]. In contrast, Mamba is capable of fully utilizing long-range and global information, although it is challenging to capture local details due to the absence of spatial generalization bias. Consequently, we devised GDRM intending to fuse the deep features of varying dimensions extracted by CNN and Mamba. The structure of GDRM and MSSA is depicted in Figure 7.

GDRM can be expressed as

$$\mathbf{F}_{\text{GDRM}} = W_1(\text{SimAM}(\mathbf{F}_{D\_m})) + W_2(\text{MSSA}(\mathbf{F}_{D\_c})) \tag{15}$$

where W_1 and W_2 denote the convolutional layers used to automatically learn the weights of different features extracted by the CNN and Mamba.

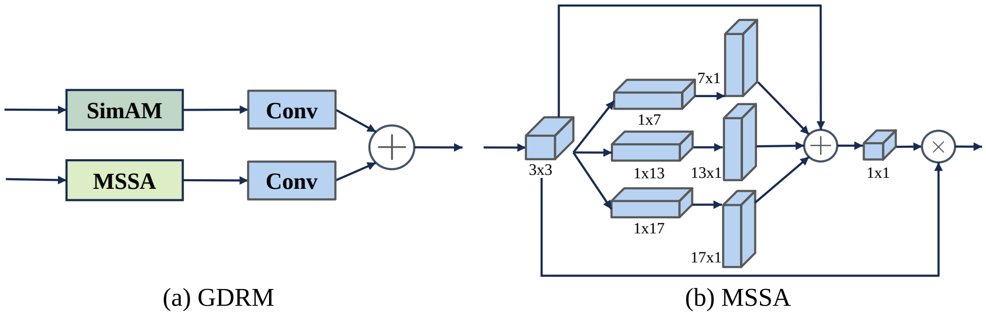


Figure 7. Structure of GDRM and MSSA. GDRM merges CNN and Mamba branch features.

MSSA and SimAM denote multiscale spatial attention and 3D attention based on energy functions, derived from SegNeXt [59] and SimAM [60].

We employed depth-wise strip convolutions to approximate standard depth-wise convolutions with large kernels. This approach is advantageous because strip convolution is lightweight. To simulate a standard 2D convolution with a 7 × 7 kernel size, we only require a pair of 7 × 1 and 1 × 7 convolutions. In contrast, remote sensing scenes frequently exhibit strip-like features, such as rivers and farmlands. Consequently, strip convolution can be employed as a supplement to grid convolutions [61–63], with the MSSA facilitating the extraction of strip-like features by the CNN branch.

The MSSA formula is expressed as follows:

$$\mathbf{F}_{D\_MSSA} = \mathbf{F}_{D\_c} \otimes \mathrm{Conv}_{1\times 1}\left(\sum_{i=0}^{3} \mathrm{Scale}_i(\mathrm{DWConv}(\mathbf{F}_{D\_c}))\right) \tag{16}$$

where $\mathbf{F}_{D\_c}$ and $\mathbf{F}_{D\_c}$ denote the CNN branch input features and Mamba branch input features received by the GDRM in the deep feature extraction stage. $\mathbf{F}_{D\_MSSA}$ denotes the output of the MSSA. DWConv denotes depth-wise convolution, and Scale_i, denotes the ith branch.

The SimAM formula is expressed as follows:

$$\mu = \frac{1}{M}\sum_{i=1}^{M}\mathbf{F}_i \tag{17}$$

$$\sigma = \frac{1}{M}\sum_{i=1}^{M}(\mathbf{F}_i - \mu)^2 \tag{18}$$

$$E_{\mathrm{inv}} = \frac{(\mathbf{F}-\mu)^2}{4(\sigma+\lambda)} + 0.5 \tag{19}$$

$$\mathbf{F}_e = \mathcal{S}(E_{\mathrm{inv}}) \cdot \mathbf{F} \tag{20}$$

where μ denotes the mean value of all neurons in the channel, x_i denotes the value of the ith neuron, M denotes the total number of neurons in the channel (i.e., $H{\times}W$, which denotes the spatial dimensions of the channel), σ denotes the variance of all the neurons in the channel, $\mathcal{S}(\cdot)$ denotes the sigmoid function, $\mathbf{F}$ denotes the input feature map, $\mathbf{F}_e$ denotes the output feature map, and λ is a set of hyperparameters.

3.5. Loss Function

We optimized ConvMambaSR for image super-resolution using L1 loss, which is the most common loss function in super-resolution tasks. Given a training set $\{\mathbf{I}^i_{LR}, \mathbf{I}^i_{HR}\}^N_{i=1}$, the loss can be expressed as

$$\frac{1}{N}\sum_{i=1}^{N}\left\| ConvMambaSR(\mathbf{I}^i_{LR}) - \mathbf{I}^i_{HR}\right\|_1 \tag{21}$$

4. Experiments

4.1. Datasets

In order to validate the effectiveness of our model, we conducted experiments on two datasets, the RSSCN7 public remote sensing dataset [64] and our own RSSRD-KQ remote sensing dataset.

The RSSCN7 dataset contains 2800 images of remotely sensed scenes from seven typical categories, namely grasslands, farmlands, industrial areas, rivers and lakes, forests, residential areas, and parking lots. Each category has 400 images from Google Earth, sampled at four scales of 100 images each. Each image is 400 × 400 pixels in size. The variety of scene images taken in different seasons and weather conditions and sampled at different scales made this dataset quite challenging. The LR images of RSSCN7 were obtained by bicubic interpolation. This dataset is divided into two equal parts, one is used as a training set with 1400 images, and the other is used as a test set, where 20% of the training set is used as the validation set.

The RSSRD-KQ dataset is situated within the Sanjiangyuan region of Qinghai Province, China. The HR images were captured by a DJI Phantom 4 RTK drone, equipped with a visible-light sensor, which served as the data collection platform for conducting aerial photography over two days, from 21 to 22 April 2024. To minimize the impact of solar radiation and atmospheric conditions on the imagery, the flights were conducted during the midday hours of 12:00 to 14:00 under clear sky conditions. The drone was flown at an altitude of 30 m, with a spatial resolution of 0.82 cm and a flight speed of 2.3 m per s. The total area of the test site that was imaged was 100 square meters. Its geographical coordinates are 33°44′36″N to 33°44′42″N and from 98°59′30″E to 98°59′38″E. The RSSRD-KQ LR images were obtained by unknown degradation of the blind super-resolution task [65,66], which performs a series of randomly sequenced degradation operations, including motion blurring, Gaussian blurring, random downsampling (nearest-neighbor interpolation, bilinear interpolation, bicubic interpolation), simple scaling of the image, addition of color, grayscale, or mixed Gaussian noise, addition of probabilistic JPEG compression noise, and finally, addition of final JPEG compression noise processing before random cropping. In this dataset, 90% of the 3162 images were randomly selected as the training set images, 10% as the validation set images, and another 672 images outside the region were used as the test set, making a total of 3834 images, each of which has a size of 480 × 480 pixels. We used such HR/LR paired images for further analysis and model training. Figure 8 shows sample examples of these two datasets.

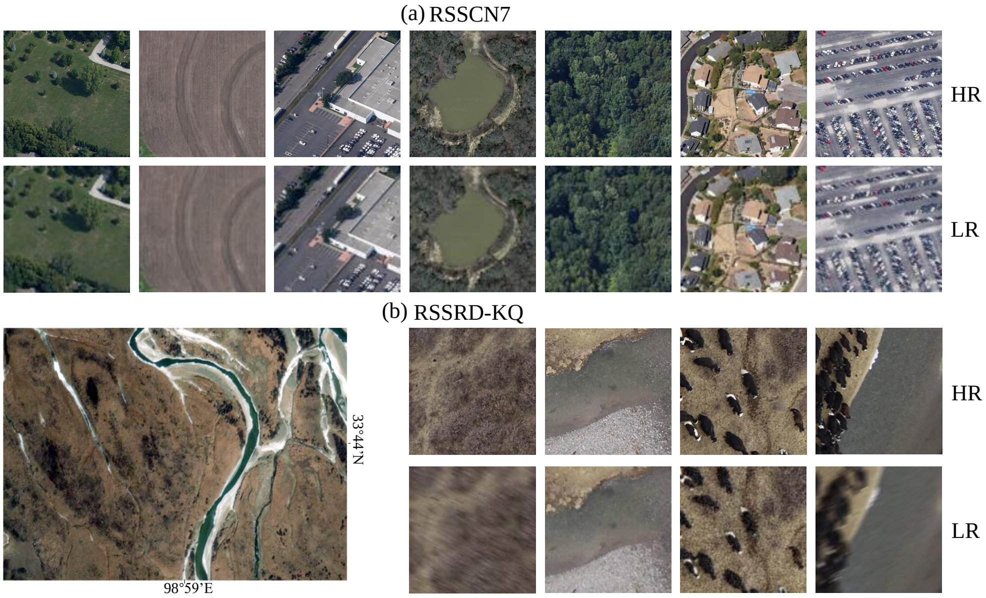


Figure 8. HR/LR sample examples of the two datasets.

4.2. Experiment Settings

In super-resolution tasks, the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [67] are commonly evaluated metrics. Assuming that the high-resolution image is x and the reconstructed super-resolution image is y, they are calculated as follows:

$$\mathrm{MSE}(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{22}$$

$$\mathrm{PSNR}(x,y) = 10\log_{10}\frac{1}{\mathrm{MSE}(x,y)} \tag{23}$$

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{24}$$

where μ_x and σ_x denote the mean and variance of x, respectively, σ_{xy} denotes the covariance between x and y, and c_1 and c_2 are constants.

In addition, root-mean-square error (RMSE), spectral angle mapper (SAM) [68], and relative dimensionless global error in synthesis (ERGAS) [69] are also used as the metrics that are widely used to evaluate the quality of reconstructed remote sensing images. They are calculated as follows:

$$\mathrm{SAM} = \cos^{-1}\frac{\sum xy}{\sqrt{\sum(x)^2\sum(y)^2}} \tag{25}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(x_i - y_i)^2} \tag{26}$$

$$\mathrm{ERGAS} = 100\frac{h}{l}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathrm{RMSE}_i}{\mu_i}\right)^2} \tag{27}$$

where h and l denote the spatial resolution of the super-resolution image and the original image, respectively, and N is the number of bands (channels).

We evaluate real-world images using the perception index (PI) [70], which combines Ma et al.'s reference-free image quality metric [71] and NIQE [72] and can be expressed as

$$\mathrm{PI} = \frac{1}{2}((10 - \mathrm{Ma}) + \mathrm{NIQE}) \tag{28}$$

Higher PSNR and SSIM values indicate a better SR quality, while lower RMSE, SAM, ERGAS, and PI values indicate a better reconstruction quality.

We implemented our SR model using PyTorch and performed all experiments on an HPC platform equipped with an Nvidia A100 80G GPU. The model in the RSSCN7 dataset takes as input a randomly cropped LR image with a size of 48 × 48 pixels, while the corresponding HR size is 192 × 192 pixels. The model in the RSSRD-KQ dataset takes as input a randomly cropped LR image with a size of 64 × 64 pixels, while the corresponding HR size is 192 × 192 pixels. During training, the batch size was set to 16 and the HR/LR images were randomly rotated for data enhancement. The initial learning rate was set to 1×10^{-4} and reduced by a factor of 0.1 after 80 epochs. A total of 200 epochs were trained using the Adam optimizer with β_1 = 0.9 and β_2 = 0.99.

Regarding the model parameter settings, we configured the number of RSSGs, RCGs, and GDRMs to be 4. The parameter settings of the RSSBs in the RSSGs were consistent with those of the MambaIR, including six VSSB modules and an SSM state expansion factor of 16. The value of λ in SimAM in GDRM was set to 1×10^{-4}, and the convolutional kernel sizes of MSSA were 3, 7, 13, and 17, with corresponding padding of 1, 3, 6, and 8.

4.3. Results

We performed comparative analyses with current state-of-the-art methods, including Bicubic, SRCNN [18], SRGAN [36], EDSR [19], RRDBNet [14], RCAN [16], SwinIR [20], HAT [41], SwinIR-NG [42], and MambaIR [55]. Visual comparison results on the RSSCN7 and RSSRD-KQ datasets are shown in Figures 9 and 10, respectively. Figure 9 shows the full results, while Figure 10 focuses on the local zoomed-in details of the results from the different models, where the subfigures within the red and green rectangles represent the zoomed-in view of the yellow rectangle.

In the RSSCN7 dataset, the agricultural scene contains a variety of complex geographic and artificial features, making detail and edge processing one of the key factors in evaluating model performance. As shown in Figure 9, the image displays a farm field, with the upper half showing plowed land and the lower half showing untreated land. ConvMambaSR shows significant advantages, especially in detail reconstruction and edge sharpening. The strip-like features commonly found in agricultural scenes require a model with high resolution and detail retention. In contrast, the results of other models are often blurry, failing to capture the boundaries and details of these features accurately, and they are particularly poor at reconstructing strip-like features.

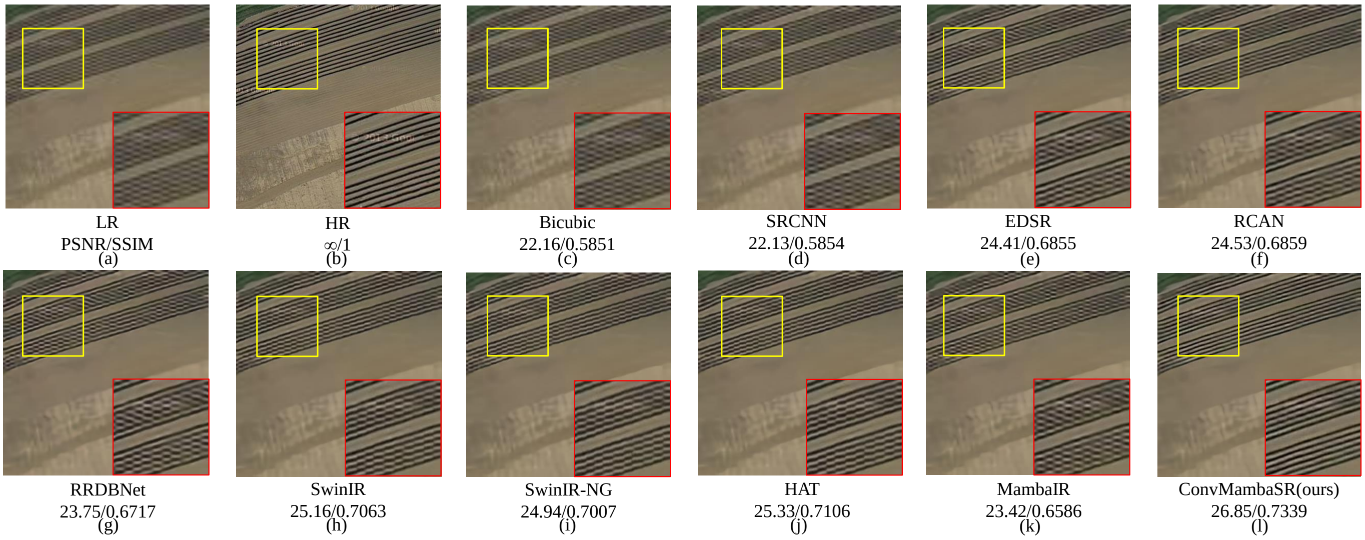


Figure 9. Qualitative comparison of our model with other works (**a**–**l**) on the RSSCN7 dataset.

ConvMambaSR also shows significant advantages on the RSSRD-KQ dataset when processing images that have undergone complex quality reduction. As shown in Figure 10, the image displays four adult yaks and two young yaks. The LR image shows that after complex texture reduction processing, the image contains very little local information. Other models generally exhibit blurring phenomena when processing complex mass reduction images and lack sufficient detail extraction and reconstruction capabilities, resulting in blurred details in the reconstructed images, which cannot effectively recover the overall contours and shapes of the yak. In contrast, ConvMambaSR acquires the global receptive field by fusing Mamba, which enables the extraction of more globally dependent features, successfully reconstructs the overall contour and shape of the yak, and improves the recognizability of the image. It is worth noting that MambaIR also achieves good results.

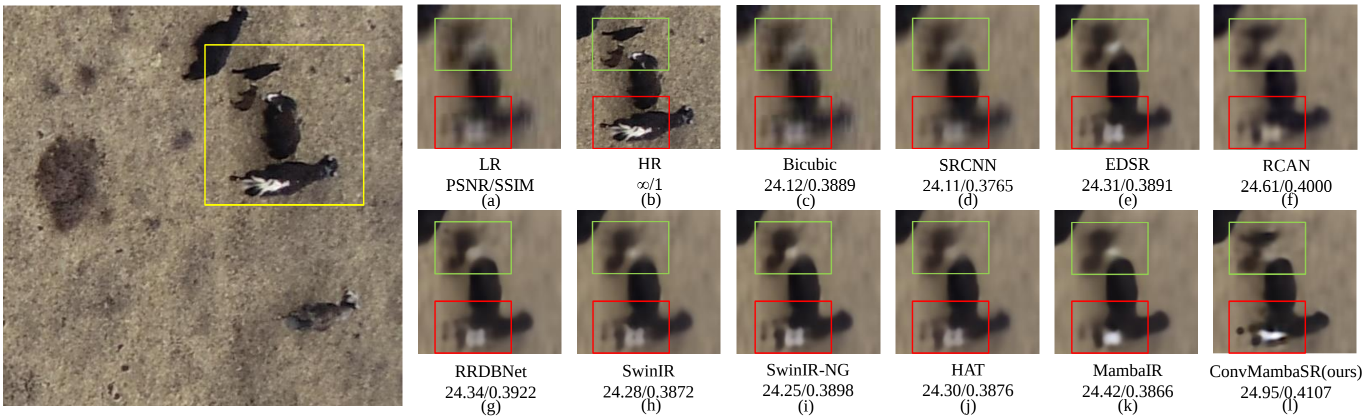


Figure 10. Qualitative comparison of our model with other works (**a**–**l**) on the RSSRD-KQ dataset.

In addition, a comprehensive quantitative evaluation of the model's performance on both datasets is shown in Table 1. The proposed ConvMambaSR model outperforms other models in terms of PSNR, SSIM, RMSE, SAM, and ERGAS. Transformer-based models outperformed CNN-based models in overall performance across both datasets. This superiority is especially evident in agricultural scenes with complex geographic and artificial features as shown in Figure 9. These results are consistent with previous research, such as the work on SwinIR [20], which demonstrated that Transformer architectures generally excel in capturing fine details and maintaining high-quality reconstructions compared with CNN-based approaches. The introduction of the attention mechanism [16,41,42] plays a key role in RSISR, and models based on the attention mechanism such as RCAN, HAT, SwinIR-NG, and ConvMambaSR show excellent performance. It is worth noting that MambaIR shows excellent performance under the RSSRD-KQ dataset, and the global receptive field brought by Mamba can better help the model to improve its performance in complex degradation scenarios compared with SwinIR, HAT, and SwinIR-NG. Furthermore, RCAN's Channel Attention mechanism allows it to adaptively adjust the features by taking into account the interdependencies between the feature channels, allowing it to achieve excellent results on the RSSRD-KQ dataset as well. Combining the excellent performance of MambaIR and RCAN and the achievements of other models, in the RSSRD-KQ dataset, after the complex degradation of the image, there is a lot of information between the global dependence and the channels, and the application of the global receptive field and the Channel Attention can greatly improve the performance of the model in complex degradation scenarios.

Figure 11 shows the performance of each model for the public dataset RSSCN7 in seven categories: grassland, farmland, industrial areas, rivers and lakes, forests, residential areas, and parking lots. The experiments show that grassland, rivers, and lakes have the highest SR reconstruction accuracy, while residential areas, forests, and parking lots have the lowest SR accuracy. The grassland and farmland scenes are mainly composed of low-frequency components, and the relatively smooth texture leads to higher PNSR accuracy, but in this case, the PSNR may not adequately reflect the real image quality improvement, and the SSIM metrics are more in line with human visual perception. On the other hand, residential areas, forests, and parking lots are full of complex details and contain a large amount of high-frequency information. The recovery and reconstruction of this high-frequency information is more challenging, resulting in a relatively low SR accuracy.

Overall, the SR performance of all deep learning models, except bicubic interpolation, is consistent across all land cover categories. ConvMambaSR and SwinIR-NG perform well across all categories.

Table 1. Quantitative comparison results for the RSSCN7 dataset and RSSRD-KQ dataset. Bold data indicate the best method.

Dataset	Method	PSNR (dB)↑	SSIM↑	RMSE↓	SAM↓	ERGAS↓
RSSCN7	Bicubic	25.50	0.6457	15.0848	0.1570	4.2513
	SRCNN [18]	25.70	0.6551	14.7016	0.1534	4.1536
	SRGAN [36]	25.96	0.6686	14.2594	0.1492	4.0402
	EDSR [19]	26.00	0.6706	14.2090	0.1487	4.0256
	RRDBNet [14]	25.98	0.6696	14.2405	0.1489	4.0346
	RCAN [16]	26.00	0.6694	14.2109	0.1487	4.0257
	SwinIR [20]	26.02	0.6727	14.1720	0.1483	4.0153
	HAT [41]	26.04	0.6734	14.1480	0.1481	4.0092
	SwinIR-NG [42]	26.04	0.6734	14.1453	0.1480	4.0084
	MambaIR [55]	25.98	0.6697	14.2493	0.1490	4.0363
	ConvMambaSR(ours)	**26.06**	**0.6751**	**14.1029**	**0.1477**	**3.9985**
RSSRD-KQ	Bicubic	23.74	0.3471	16.9950	0.1495	3.8872
	SRCNN [18]	24.06	0.3541	16.2812	0.1435	3.7173
	SRGAN [36]	24.06	0.3582	16.2846	0.1435	3.7155
	EDSR [19]	24.18	0.3648	16.0725	0.1415	3.6685
	RRDBNet [14]	24.19	0.3666	16.0470	0.1414	3.6614
	RCAN [16]	24.27	0.3740	15.9059	0.1404	3.6290
	SwinIR [20]	24.20	0.3672	16.0310	0.1412	3.6586
	HAT [41]	24.21	0.3684	16.0068	0.1411	3.6530
	SwinIR-NG [42]	24.20	0.3676	16.0305	0.1412	3.6587
	MambaIR [55]	24.23	0.3663	15.9726	0.1407	3.6450
	ConvMambaSR(ours)	**24.29**	**0.3752**	**15.8632**	**0.1398**	**3.6205**

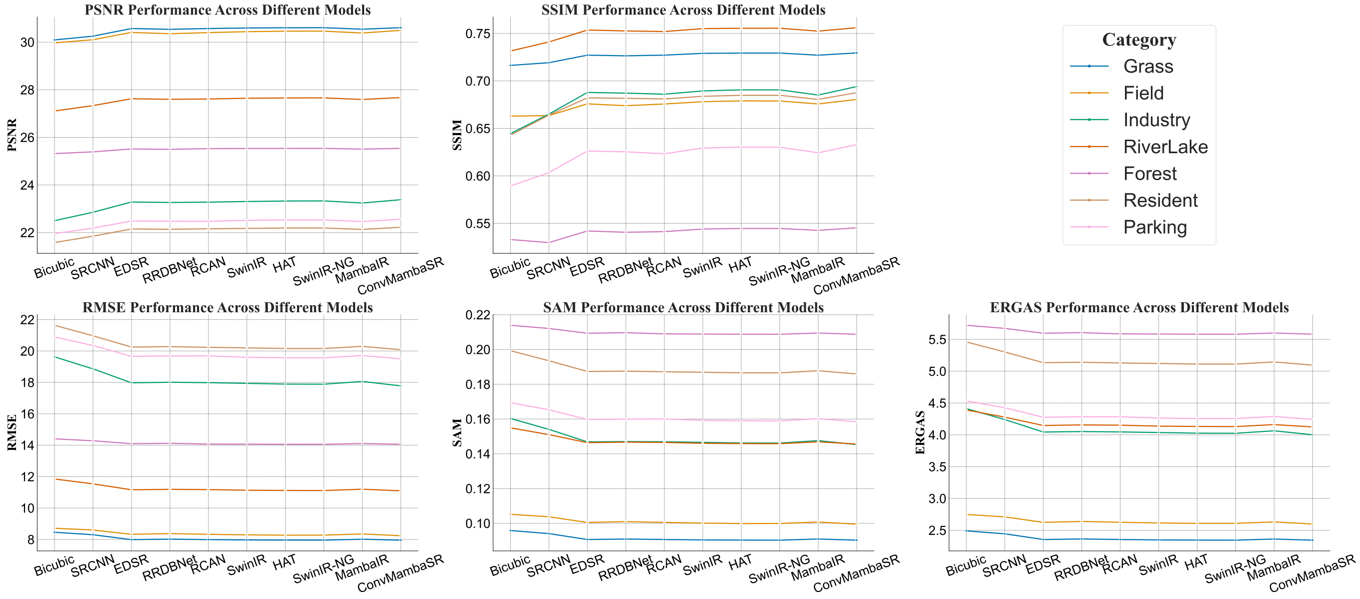


Figure 11. Performance of different land-cover categories on RSSCN7 dataset.

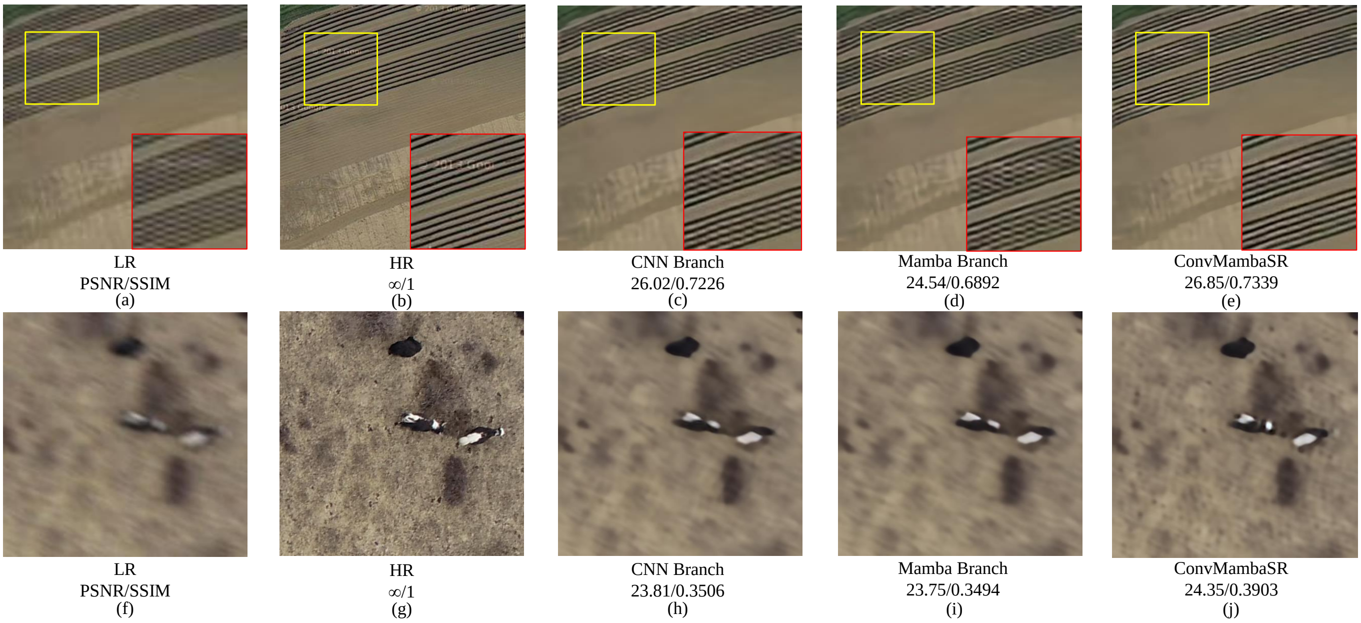
4.4. Effects of GDRM

To assess the efficacy of GDRM, we conducted ablation experiments utilizing CNN and Mamba branches, respectively. The experimental outcomes are presented in Table 2. The comparison demonstrates that GDRM exhibits superior performance compared with the single-branch model. ConvMambaSR exhibits higher PSNR and SSIM values, as well as lower RMSE, SAM, and ERGAS values. This indicates that ConvMambaSR effectively fuses the local details of CNN and the global context information of Mamba. Furthermore, we observe that the CNN branch outperforms the Mamba branch on the RSSCN7 dataset but is inferior to the Mamba branch on the RSSRD-KQ dataset. This suggests that deep residual-based CNNs produce favorable results for the pixel-intensive SR reconstruction task, despite the limited ability of CNNs to model long-range dependencies. However, for complex degraded scenes, the substantial reduction in local details in the image leads to a deteriorated CNN performance, whereas the global receptive field of Mamba enables the extraction of more information. This indicates that the fusion of local details and global features enables the model to achieve excellent performance in different SR reconstruction tasks.

Table 2. Comparison of the proposed model with different branch models on the RSSCN7 dataset and RSSRD-KQ dataset. Bold data indicate the best method.

Dataset	Method	PSNR(dB)↑	SSIM↑	RMSE↓	SAM↓	ERGAS↓
RSSCN7	CNN Branch	26.05	0.6740	14.1201	0.1478	4.0032
	Mamba Branch	26.00	0.6704	14.2115	0.1487	4.0269
	ConvMambaSR	**26.06**	**0.6751**	**14.1029**	**0.1477**	**3.9985**
RSSRD-KQ	CNN Branch	24.23	0.3698	15.9706	0.1408	3.6443
	Mamba Branch	24.25	0.3706	15.9296	0.1405	3.6353
	ConvMambaSR	**24.29**	**0.3752**	**15.8632**	**0.1398**	**3.6205**

Figure 12 presents a visual comparison of the RSSCN7 and RSSRD-KQ datasets across different branches; the image below displays three adult yaks. It can be observed that the performance is significantly enhanced following the fusion of the two branches using GDRM. ConvMambaSR is able to perfectly reconstruct the strip-like features in the farmland scene of the RSSCN7 dataset. Furthermore, ConvMambaSR is adept at effectively mitigating the adverse effects of motion blur degradation observed in the RSSRD-KQ dataset. Additionally, ConvMambaSR demonstrates satisfactory performance in terms of SSIM metrics. The high SSIM values validate the fused model's advantages in detailed and global feature reconstruction and demonstrate that it outperforms the CNN or Mamba branch alone in overall visual perception.



LR
PSNR/SSIM
(a)

HR
∞/1
(b)

CNN Branch
26.02/0.7226
(c)

Mamba Branch
24.54/0.6892
(d)

ConvMambaSR
26.85/0.7339
(e)

LR
PSNR/SSIM
(f)

HR
∞/1
(g)

CNN Branch
23.81/0.3506
(h)

Mamba Branch
23.75/0.3494
(i)

ConvMambaSR
24.35/0.3903
(j)

Figure 12. Qualitative comparison of our model with different branch models (**a**–**j**) on the RSSCN7 dataset and RSSRD-KQ dataset.

4.5. Ablation Study on RCG Count

To further explore the impact of the number of RCGs on model performance, we conducted an ablation study by varying the RCG count while monitoring the number of parameters, FLOPs, PSNR, and SSIM.

The results in Tables 2 and 3 indicate that using only the Mamba branch, without any RCGs, the model achieves a PSNR of 26.00 dB and an SSIM of just 1 RCG significantly improves the performance, raising the PSNR t0.6704. Introducingo 26.05 dB and the SSIM to 0.6742. This indicates that incorporating RCGs into the architecture enhances the model's ability to capture essential features, leading to better reconstruction quality.

Table 3. Performance metrics with varying RCG counts on the RSSCN7 dataset.

RCG Count	#Para ms (M)	FLOPs (G)	PSNR (dB)↑	SSIM↑
1	6.72	175	26.05	0.6742
4	8.71	226	26.05	0.6742
8	11.37	294	26.06	0.6751
12	14.02	362	26.06	0.6751
16	16.68	430	26.07	0.6756
20	19.34	498	26.07	0.6759

As the number of RCGs increases, further improvements in both PSNR and SSIM are observed. For instance, increasing the RCG count from 1 to 20 results in a PSNR increase from 26.05 dB to 26.07 dB and an SSIM increase from 0.6742 to 0.6759. However, these performance gains diminish with higher RCG counts, suggesting that while more RCGs contribute positively to the model's output, the marginal benefits decrease after a certain threshold.

Moreover, the increase in the number of RCGs leads to a corresponding rise in the model's computational complexity, as evidenced by the growing number of parameters and FLOPs. Therefore, it is crucial to balance the trade-off between the performance improvements and the associated computational costs when selecting the optimal RCG count, particularly for practical applications where efficiency is a critical consideration.

4.6. LAM Analysis and Feature Visualization

Local Attribution Map (LAM) [73] is a method based on Integrated Gradients [74] designed to analyze and visualize the contribution of individual input pixels to the output of deep SR networks. Additionally, LAM introduces a Diffusion Index (DI) to quantitatively measure the extent of pixel involvement in the reconstruction process.

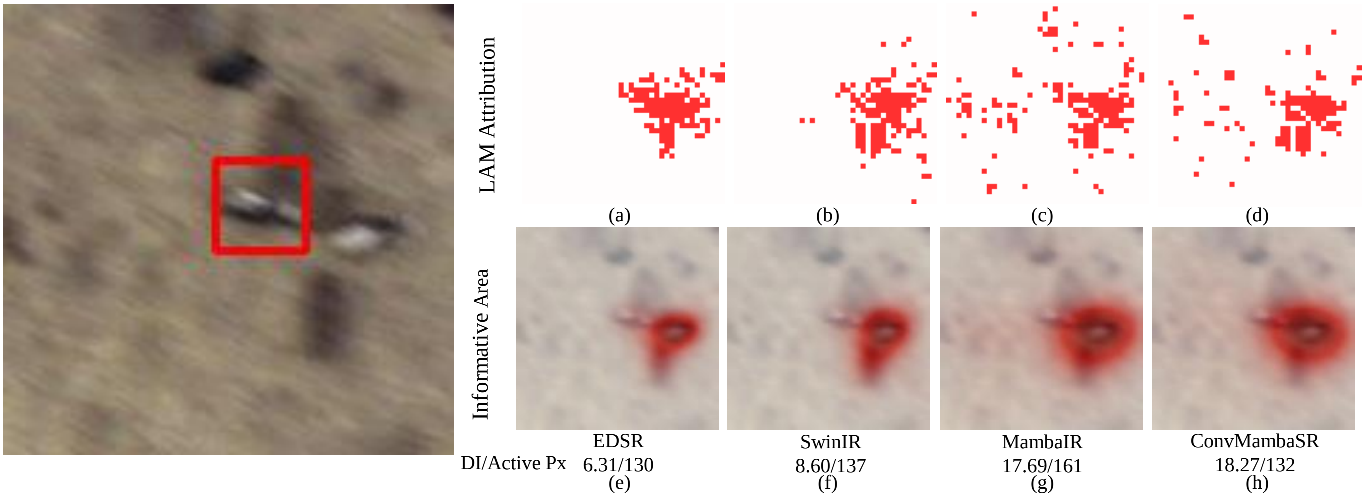


Figure 13. Visualization results for different networks (**a**–**h**). Active Px indicates the number of active pixels.

With LAM, we can identify which input pixels contribute to the selected region. As shown in Figure 13, the points marked in red are the pixels that contribute to the reconstruction. It is evident that Transformer-based SwinIR [20] has a considerably larger receptive field and activates a greater number of pixels with a wider distribution than CNN-based EDSR [19]. However, SSM-based MambaIR [55] is capable of utilizing pixels across a variety of regions within the entire image, thereby facilitating reconstruction and activating the greatest number of pixels. The dual-branching structure of Mamba and CNN, along with the design of GDRM, enables ConvMambaSR to leverage the strengths of both SSMs and CNNs. Despite a reduction in the number of activated pixels, ConvMambaSR is capable of utilizing the correct pixels globally for reconstruction, as evidenced by its superior performance.

4.7. Complexity and Efficiency Evaluation

In order to quantitatively evaluate the complexity and computational efficiency of our proposed models, we calculated the number of training parameters (#Params), the number of floating-point operations (FLOPs), and the number of frames processed per second (FPS) for the different models. The FLOPs are measured with an input size of 160 × 160 × 3. The FPS is measured on an Nvidia A100 80G GPU. The scale factor was set to 4.

As demonstrated in Table 4, the CNN-based model exhibits the fastest inference speed among all the deep learning models, a result that can be attributed to the superior efficiency of the convolutional parallel operation. The inference speed of Mamba is considerably faster than that of Transformer due to its linear complexity and efficient inference. A comparison of Tables 1 and 4 reveals that the complexity and inference speed of our proposed model are comparable to those of most other models, while the model performance is markedly superior. This provides compelling evidence of the model's potential for practical applications.

Table 4. Comparison of model complexity and efficiency.

Model	#Params (M)	FLOPs (G)	FPS
Bicubic	-	-	22557.1
SRCNN [18]	0.02	8	620.7
EDSR [19]	1.51	50	268.0
RRDBNet [14]	16.69	459	27.5
RCAN [16]	15.59	408	20.6
MambaIR [55]	4.65	123	10.3
HAT [41]	26.03	703	4.4
SwinIR [20]	16.57	456	0.5
SwinIR-NG [42]	19.35	460	0.4
ConvMambaSR(ours)	14.02	362	8.9

4.8. Real-World Image Testing

The performance of each model was evaluated using real-world images in RSSRD-KQ, comprising a total of 48 images, each with a size of 480 × 480 pixels. The scale factor was set to 3.

As illustrated in Table 5, ConvMambaSR exhibits the lowest PI relative to other deep learning models. It is noteworthy that, similar to the outcomes of the preceding experiments, MambaIR continues to achieve highly favorable outcomes. The global receptive field introduced by Mamba is of significant benefit in addressing complex degradation cases, offering a substantial advantage over Transformer and CNN. This indicates that the presence of more information in the image is dependent on distance, and the enlargement of the model's receptive field will result in enhanced performance. The Mamba architecture is capable of significantly improving the blind SISR of remote sensing images.

Table 5. Quantitative comparison results for real-world images. Bold data indicate the best method.

Method	PI↓
SRCNN [18]	8.6022
EDSR [19]	8.2448
RRDBNet [14]	8.1700
RCAN [16]	8.2068
SwinIR [20]	8.2149
HAT [41]	8.1619
SwinIR-NG [42]	8.2149
MambaIR [55]	8.1346
ConvMambaSR(ours)	**8.0496**

Furthermore, the experiments demonstrate that the degradation method employed for RSSRD-KQ enables the model to learn certain degradation processes in the real world. As shown in Figure 14, the image displays a high-altitude wetland scene, with some snow covering the area, the phenomena such as motion blur and color fringes that are evident in the original image captured by the UAV are partially resolved.

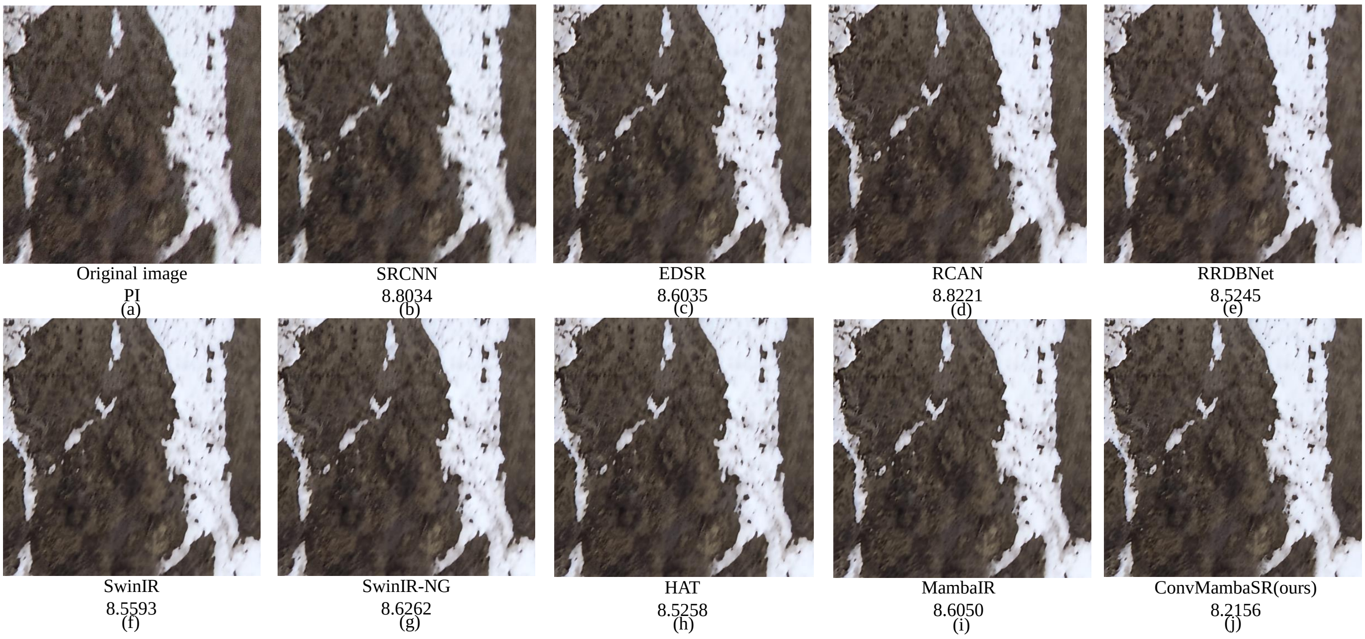


Figure 14. Examples of the proposed model with other works on real-world images (**a**–**j**).

5. Conclusions

This paper presents ConvMambaSR, a novel SISR method tailored for remote sensing applications. ConvMambaSR leverages the complementary strengths of CNNs and SSMs, effectively combining their abilities. Our experimental results reveal that a single model architecture is often inadequate for addressing the diverse challenges posed by remote sensing imagery, highlighting the necessity of integrating multiple models to achieve superior results. To comprehensively capture both local and global information from different branches, we introduce the global–detail reconstruction module, which fuses the locality-capturing capability of CNNs with the global dependency modeling power of SSMs. The experimental outcomes confirm the effectiveness of ConvMambaSR, demonstrating its significant improvements over existing methods. ConvMambaSR demonstrates state-of-the-art performance on the RSSCN7 and RSSRD-KQ datasets, outperforming existing methods in visual quality. Specifically, the model achieves PSNR and SSIM scores of 26.06 and 0.6751 on the RSSCN7 dataset and 24.29 and 0.3752 on the RSSRD-KQ dataset. Notably, across the seven representative categories within the RSSCN7 dataset, ConvMambaSR consistently exhibits superior performance. Furthermore, the model significantly exceeds the inference speed of Transformer-based approaches, achieving 8.9 FPS. This combination of efficiency and high performance underscores ConvMambaSR's potential for a wide range of super-resolution tasks, positioning it as a strong candidate for broader real-world applications.

Despite the promising results achieved with ConvMambaSR, there are several limitations in the current study. One notable limitation observed through LAM analysis and feature visualization is that, although the model inherits the global modeling capabilities of SSMs and activates more correct pixels for reconstruction, there is still an issue with the reduction in the number of activated pixels. Moreover, although the current approach is effective, it is computationally intensive, leading to significant resource consumption. In future research, we will explore further performance enhancements of hybrid models for blind remote sensing SISR and optimize our approach to improve model efficiency.

Author Contributions: Conceptualization, Q.Z. and G.Z.; methodology, Q.Z. and X.Z.; software, Q.Z.; the investigation, Q.Z. and X.W.; writing—original draft preparation, Q.Z. and G.Z.; writing—review and editing, Q.Z., J.H. and X.L.; project administration, G.Z.; funding acquisition, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Qinghai Province (No. 2023-ZJ-906M), and this research was funded by the National Natural Science Foundation of China (No. 62162053, No. U23A20159, No. U21A20191). This research was supported by the High-performance Computing Center of Qinghai University.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mathieu, R.; Freeman, C.; Aryal, J. Mapping private gardens in urban areas using object-oriented techniques and very high-resolution satellite imagery. *Landsc. Urban Plan.* **2007**, *81*, 179–192. [CrossRef]
2. Kumar, S.; Meena, R.S.; Sheoran, S.; Jangir, C.K.; Jhariya, M.K.; Banerjee, A.; Raj, A. Remote sensing for agriculture and resource management. In *Natural Resources Conservation and Advances for Sustainability*; Elsevier: Amsterdam, The Netherlands, 2022; pp. 91–135.
3. Turner, W.; Spector, S.; Gardiner, N.; Fladeland, M.; Sterling, E.; Steininger, M. Remote sensing for biodiversity science and conservation. *Trends Ecol. Evol.* **2003**, *18*, 306–314. [CrossRef]
4. Yang, J.; Gong, P.; Fu, R.; Zhang, M.; Chen, J.; Liang, S.; Xu, B.; Shi, J.; Dickinson, R. The role of satellite remote sensing in climate change studies. *Nat. Clim. Chang.* **2013**, *3*, 875–883. [CrossRef]
5. Li, J.; Pei, Y.; Zhao, S.; Xiao, R.; Sang, X.; Zhang, C. A review of remote sensing for environmental monitoring in China. *Remote Sens.* **2020**, *12*, 1130. [CrossRef]
6. Singh, S.; Bhardwaj, A.; Verma, V. Remote sensing and GIS based analysis of temporal land use/land cover and water quality changes in Harike wetland ecosystem, Punjab, India. *J. Environ. Manag.* **2020**, *262*, 110355. [CrossRef]
7. Soubry, I.; Doan, T.; Chu, T.; Guo, X. A systematic review on the integration of remote sensing and GIS to forest and grassland ecosystem health attributes, indicators, and measures. *Remote Sens.* **2021**, *13*, 3262. [CrossRef]
8. Bhaga, T.D.; Dube, T.; Shekede, M.D.; Shoko, C. Impacts of climate variability and drought on surface water resources in Sub-Saharan Africa using remote sensing: A review. *Remote Sens.* **2020**, *12*, 4184. [CrossRef]
9. Wang, P.; Bayram, B.; Sertel, E. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Sci. Rev.* **2022**, *232*, 104110. [CrossRef]
10. Li, K.; Yang, S.; Dong, R.; Wang, X.; Huang, J. Survey of single image super-resolution reconstruction. *IET Image Process.* **2020**, *14*, 2273–2290. [CrossRef]
11. Wang, Y.; Bashir, S.M.A.; Khan, M.; Ullah, Q.; Wang, R.; Song, Y.; Guo, Z.; Niu, Y. Remote sensing image super-resolution and object detection: Benchmark and state of the art. *Expert Syst. Appl.* **2022**, *197*, 116793. [CrossRef]
12. Yang, W.; Zhang, X.; Tian, Y.; Wang, W.; Xue, J.H.; Liao, Q. Deep learning for single image super-resolution: A brief review. *IEEE Trans. Multimed.* **2019**, *21*, 3106–3121. [CrossRef]
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
14. Tong, T.; Li, G.; Liu, X.; Gao, Q. Image super-resolution using dense skip connections. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4799–4807.
15. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]

16. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
17. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
18. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [CrossRef]
19. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
20. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF international Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1833–1844.
21. Chen, B.; Zou, X.; Zhang, Y.; Li, J.; Li, K.; Xing, J.; Tao, P. LEFormer: A Hybrid CNN-Transformer Architecture for Accurate Lake Extraction from Remote Sensing Imagery. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 5710–5714.
22. Zou, X.; Li, K.; Xing, J.; Zhang, Y.; Wang, S.; Jin, L.; Tao, P. DiffCR: A Fast Conditional Diffusion Framework for Cloud Removal From Optical Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
23. Wang, S.; Zou, X.; Li, K.; Xing, J.; Cao, T.; Tao, P. Towards robust pansharpening: A large-scale high-resolution multi-scene dataset and novel approach. *Remote Sens.* **2024**, *16*, 62899. [CrossRef]
24. Li, K.; Xie, F.; Chen, H.; Yuan, K.; Hu, X. An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, 1–15. [CrossRef] [PubMed]
25. Zou, X.; Li, K.; Xing, J.; Tao, P.; Cui, Y. PMAA: A Progressive Multi-scale Attention Autoencoder Model for High-Performance Cloud Removal from Multi-temporal Satellite Imagery. In Proceedings of the European Conference on Artificial Intelligence (ECAI), Kraków, Poland, 30 September–4 October 2023; pp. 3165–3172.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
27. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
28. Gao, L.; Liu, H.; Yang, M.; Chen, L.; Wan, Y.; Xiao, Z.; Qian, Y. STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10990–11003. [CrossRef]
29. Fu, D.Y.; Dao, T.; Saab, K.K.; Thomas, A.W.; Rudra, A.; Ré, C. Hungry hungry hippos: Towards language modeling with state space models. *arXiv* **2022**, arXiv:2212.14052.
30. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
31. Mehta, H.; Gupta, A.; Cutkosky, A.; Neyshabur, B. Long range language modeling via gated state spaces. *arXiv* **2022**, arXiv:2206.13947.
32. Smith, J.T.; Warrington, A.; Linderman, S.W. Simplified state space layers for sequence modeling. *arXiv* **2022**, arXiv:2208.04933.
33. Li, K.; Chen, G. Spmamba: State-space model is all you need in speech separation. *arXiv* **2024**, arXiv:2404.02063.
34. Gu, A.; Dao, T.; Ermon, S.; Rudra, A.; Ré, C. Hippo: Recurrent memory with optimal polynomial projections. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1474–1487.
35. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
36. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
37. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
38. Niu, B.; Wen, W.; Ren, W.; Zhang, X.; Yang, L.; Wang, S.; Zhang, K.; Cao, X.; Shen, H. Single image super-resolution via a holistic attention network. In Proceedings of the Computer Vision–ECCV 2020: 16th Europea Conference, Proceedings Part XII 16, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 191–207.
39. Huang, J.; Li, K.; Wang, X. Single image super-resolution reconstruction of enhanced loss function with multi-gpu training. In Proceedings of the 2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Xiamen, China, 16–18 December 2019; pp. 559–565.
40. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
41. Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; Dong, C. Activating more pixels in image super-resolution transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 22367–22377.

42. Choi, H.; Lee, J.; Yang, J. N-gram in swin transformers for efficient lightweight image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2071–2081.
43. Fernandez-Beltran, R.; Latorre-Carmona, P.; Pla, F. Single-frame super-resolution in remote sensing: A practical overview. *Int. J. Remote Sens.* **2017**, *38*, 314–354. [CrossRef]
44. Ducournau, A.; Fablet, R. Deep learning for ocean remote sensing: An application of convolutional neural networks for super-resolution on satellite-derived SST data. In Proceedings of the IEEE 2016 9th IAPR Workshop on Pattern Recogniton in Remote Sensing (PRRS), Cancun, Mexico, 4 December 2016; pp. 1–6.
45. Pan, Z.; Ma, W.; Guo, J.; Lei, B. Super-resolution of single remote sensing image based on residual dense backprojection networks. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7918–7933. [CrossRef]
46. Huan, H.; Li, P.; Zou, N.; Wang, C.; Xie, Y.; Xie, Y.; Xu, D. End-to-end super-resolution for remote-sensing images using an improved multi-scale residual network. *Remote Sens.* **2021**, *13*, 666. [CrossRef]
47. Tu, J.; Mei, G.; Ma, Z.; Piccialli, F. SWCGAN: Generative adversarial network combining swin transformer and CNN for remote sensing image super-resolution. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 5662–5673. [CrossRef]
48. Shang, J.; Gao, M.; Li, Q.; Pan, J.; Zou, G.; Jeon, G. Hybrid-Scale Hierarchical Transformer for Remote Sensing Image Super-Resolution. *Remote Sens.* **2023**, *15*, 3442. [CrossRef]
49. Li, J.; Meng, Y.; Tao, C.; Zhang, Z.; Yang, X.; Wang, Z.; Wang, X.; Li, L.; Zhang, W. ConvFormerSR: Fusing Transformers and Convolutional Neural Networks for Cross-sensor Remote Sensing Imagery Super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 5601115. [CrossRef]
50. Gu, A.; Goel, K.; Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv* **2021**, arXiv:2111.00396.
51. Gu, A.; Johnson, I.; Goel, K.; Saab, K.; Dao, T.; Rudra, A.; Ré, C. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 572–585.
52. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12299–12310.
53. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Liu, Y. Vmamba: Visual state space model. *arXiv* **2024**, arXiv:2401.10166.
54. Shazeer, N. Glu variants improve transformer. *arXiv* **2020**, arXiv:2002.05202.
55. Guo, H.; Li, J.; Dai, T.; Ouyang, Z.; Ren, X.; Xia, S.T. MambaIR: A Simple Baseline for Image Restoration with State-Space Model. *arXiv* **2024**, arXiv:2402.15648.
56. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
57. Yu, W.; Si, C.; Zhou, P.; Luo, M.; Zhou, Y.; Feng, J.; Yan, S.; Wang, X. Metaformer baselines for vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 896–912. [CrossRef]
58. Li, K.; Yang, R.; Sun, F.; Hu, X. IIANet: An Intra-and Inter-Modality Attention Network for Audio-Visual Speech Separation. In Proceedings of the Forty-First International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024.
59. Guo, M.H.; Lu, C.Z.; Hou, Q.; Liu, Z.; Cheng, M.M.; Hu, S.M. Segnext: Rethinking convolutional attention design for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 1140–1156.
60. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 11863–11874.
61. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large kernel matters–improve semantic segmentation by global convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4353–4361.
62. Hou, Q.; Zhang, L.; Cheng, M.M.; Feng, J. Strip pooling: Rethinking spatial pooling for scene parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4003–4012.
63. Li, K.; Luo, Y. On the design and training strategies for rnn-based online neural speech separation systems. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes, Greece, 4–10 June 2023; pp. 1–5.
64. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep learning based feature selection for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325. [CrossRef]
65. Zhang, K.; Liang, J.; Van Gool, L.; Timofte, R. Designing a practical degradation model for deep blind image super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4791–4800.
66. Wang, X.; Xie, L.; Dong, C.; Shan, Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1905–1914.
67. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
68. Yuhas, R.H.; Goetz, A.F.; Boardman, J.W. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In Proceedings of the JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop, Volume 1: AVIRIS Workshop, Pasadena, CA, USA, 1–5 June 1992.

69. Ranchin, T.; Wald, L. Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation. *Photogramm. Eng. Remote Sens.* **2000**, *66*, 49–61.
70. Blau, Y.; Mechrez, R.; Timofte, R.; Michaeli, T.; Zelnik-Manor, L. The 2018 PIRM challenge on perceptual image super-resolution. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September l2018.
71. Ma, C.; Yang, C.Y.; Yang, X.; Yang, M.H. Learning a no-reference quality metric for single-image super-resolution. *Comput. Vis. Image Underst.* **2017**, *158*, 1–16. [CrossRef]
72. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.* **2012**, *20*, 209–212. [CrossRef]
73. Gu, J.; Dong, C. Interpreting super-resolution networks with local attribution maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9199–9208.
74. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 3319–3328.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.