Scholar Publishing Group
International Journal of Multimedia Computing
https://doi.org/10.38007/IJMC.2024.050105
ISSN 2789-7168 Vol. 5, Issue 1: 43-53

Super-Resolution Analysis of Remote Sensing Images Based on Cross-Attention

Yunhe Li[1,a], Mei Yang[1,b], Tao Bian[1,c], Wenhui Xia[1,d] and Haitao Wu[2,e*]

[1]*Zhaoqing University, Zhaoqing 526060, Guangdong, China*
[2]*Shenzhen Chenzhuo Technology Company, Shenzhen 518055, China*
[a]*liyunhe@zqu.edu.cn*, [b]*202308540517@stu.zqu.edu.cn*, [c]*202408540517@stu.zqu.edu.cn*,
[d]*202408540530@stu.zqu.edu.cn*, [e]*frank@cztek.cn*
[*]*corresponding author*

Keywords: Sentinel-2 Satellite; Remote Sensing Imagery; Super-Resolution Analysis; Attention Mechanism

Abstract: To achieve the goal of enhancing the resolution of Sentinel-2 satellite remote sensing images by a factor of four, this paper innovatively proposes a super-resolution model (S2SR) that combines VMamba and Transformer technologies. By skillfully introducing the Mixed Attention Block (MTB) and Cross Attention Block (CAMB), the model effectively integrates the channel attention mechanism with two-dimensional selective scanning technology. This design not only enhances the synergistic utilization of global and local features but also significantly improves the interaction capability of cross-window information through the overlapping cross-attention mechanism, effectively suppressing the common block effect issue in traditional super-resolution methods and thereby significantly enhancing the quality of reconstructed images. Experimental results demonstrate that on the SEN12MS standard dataset, the S2SR model exhibits superior performance compared to existing advanced methods in multiple no-reference image quality assessment metrics (such as NIQE, BRISQUE, PIQE). Especially when processing images with complex geographical features, the super-resolution images generated by the S2SR model exhibit clear edges and rich details, fully verifying the efficiency and practicality of the model.

1 Introduction

Satellite remote sensing imagery plays a crucial role in numerous fields, including agriculture, environmental protection, land use, urban planning, natural disaster monitoring, hydrology, and climate research. With continuous advancements in technologies such as optical instruments, the spatial resolution of satellite imagery has seen significant improvements. For instance, the WorldView-3/4 satellites are capable of providing 8-band multispectral data with a ground resolution as high as 1.2 meters. However, the utilization of such data requires payment, and when it comes to

large-area coverage or multi-temporal analysis, the cost of data becomes a significant constraint. Therefore, exploring the utilization of openly accessible data with acceptable spatial quality, such as those provided by satellite programs like Landsat or Sentinel, has emerged as a worthwhile research direction.

The Sentinel-2 project utilizes two satellites to achieve remote sensing coverage of equatorial regions globally every five days, providing multi-resolution layers composed of 13 spectral bands. Among these, the four bands of visible red (B4), green (B3), blue (B2), and near-infrared (B8) (RGBN) offer images with a 10m resolution, while the other bands provide images with 20m and 60m resolutions, respectively. The 10m and 20m resolution bands are frequently applied in fields such as land cover, hydrological mapping, agriculture, and forestry, while the 60m resolution bands are primarily used for monitoring water vapor and other purposes. Due to Sentinel-2's open data distribution policy, its 10m resolution RGBN images are increasingly becoming an important resource for many applications. However, such spatial resolution remains inadequate for many applications. On the other hand, high-resolution (up to 2m) multispectral images provided by commercial satellites like WorldView, due to their high cost, cannot be widely used in large-area or multi-temporal analyses [1].

To fully leverage the free availability of Sentinel-2 imagery while achieving a spatial resolution close to 2m, it is of practical significance to investigate methods for spatially enhancing low-resolution images through post-processing techniques, which can recover high-frequency details to produce high-resolution images. To improve the spatial resolution of Sentinel-2 imagery, some studies have attempted to fuse data from different spatial resolution bands of Sentinel-2 to obtain higher-resolution images. However, this paper focuses more on methods that directly utilize 10m resolution images to achieve super-resolution analysis (SR).

In the field of image super-resolution (SR) research, early studies such as those by Li et al. provided effective solutions based on dictionary learning and sparse coding techniques [2]. Subsequently, Lei et al. applied structural self-similarity and compressive sensing to SR tasks [3], while Shao et al. improved SR performance by adopting multiple different image representation spaces [4]. In recent years, deep learning has received increasing attention in the field of super-resolution analysis. Deep learning does not require direct modeling of the relationship between high-resolution (HR) and low-resolution (LR) bands; with sufficient training data, deep learning networks can, in principle, learn very complex nonlinear relationships. Among them, methods based on convolutional neural networks (CNNs) can effectively utilize high-order features of images for image super-resolution analysis, significantly enhancing SR performance. Dong et al. proposed the SRCNN network with high learning capabilities based on CNNs and optimized the network using pixel loss, but the results were overly smooth due to the lack of consideration for perceptual quality [5]. Since the successful application of deep learning to super-resolution tasks, CNN-based methods have emerged endlessly and have almost dominated this field in the past few years. Meanwhile, due to the success of Transformers in the field of natural language processing [6], they have attracted attention in the field of computer vision. A series of Transformer-based methods have been developed for advanced visual tasks, including image classification, object detection, segmentation, etc. Although vision Transformers have demonstrated superiority in modeling long-range dependencies, many works have shown that convolutions can help Transformers achieve better visual representations. Due to the outstanding performance of Transformers, they have also been introduced to low-level visual tasks. The Swin Transformer (SwinIR) has demonstrated excellent performance in image super-resolution (SR) [7]. However, due to the limited range of pixels utilized, SwinIR may restore incorrect textures. Furthermore, we can observe obvious blocking artifacts in the intermediate features of SwinIR, which are caused by the window partitioning mechanism, indicating that the shifted window mechanism is inefficient in establishing cross-window connections. Some works

targeting advanced visual tasks have also pointed out that enhancing the connections between windows can improve window-based self-attention methods.

To overcome the aforementioned limitations and further explore the potential of Transformers in super-resolution (SR) tasks, we have designed the MTOG module and introduced it into SR tasks, constructing the S2SR super-resolution analysis model. Notably, within our designed MTOG module, we concurrently utilize the VMamba-based SS2D module and the Transformer-based overlapping cross-attention block [8]. This is primarily due to the highly complementary nature of VMamba and Transformers, which allows this structure to simultaneously activate more pixels for reconstruction, enabling almost all pixels in the image to be visible and enabling the recovery of correct and clear textures.

2 Analysis of Sentinel-2 Image Super-Resolution (S2SR)

The task of super-resolution (SR) is inherently a pixel-intensive one, as its goal is to recover high-resolution (HR) details from low-resolution (LR) images. During this process, the model needs to perform dense computations at each pixel location to predict and generate new pixel points in the higher-resolution image. Therefore, modeling the contextual relationships among pixel points is particularly important in SR tasks. Based on this, the authors have designed the MTOG module and introduced it into SR tasks, constructing the S2SR super-resolution analysis model. Notably, within our designed MTOG module, we concurrently utilize the VMamba-based SS2D module and the Transformer-based Overlapping Cross-Attention (OCA) block. This is primarily due to the highly complementary nature of VMamba and Transformers. This structure can simultaneously activate more pixels for reconstruction, enabling almost all pixels in the image to be visible and allowing for the recovery of correct and clear textures. VMamba is characterized by its ability to model long-range dependencies in long sequences, likely benefiting from its parameterization method that enables VMamba to store information from long sequences. However, VMamba is an autoregressive model that typically exhibits unidirectionality, such as good temporal properties and causal sequence modeling. Compared to Transformers, it cannot model the relationships between sequence elements. Transformers have demonstrated powerful advantages across various tasks, but they struggle with processing long-sequence information.

2.1 Overall Structure

As shown in Figure 1, the overall network consists of three parts: shallow feature extraction, deep feature extraction, and image reconstruction. This architectural design has been widely applied in previous works. Specifically, for a given low-resolution (LR) input $I_{LR} \in R^{H \times W \times C_{in}}$, we first utilize a convolutional layer to extract shallow features $F_0 \in R^{H \times W \times C}$, where C_{in} and C denote the number of channels for the input and intermediate features, respectively. Subsequently, deep features are extracted using a series of Mixed-Attention Transformer Organic Groups (MTOGs) and a 3×3 convolutional layer $H_{Conv}(\cdot)$. After that, we add a global residual connection to fuse the shallow F_0 and deep features $F_D \in R^{H \times W \times C}$, and then reconstruct the high-resolution result through the reconstruction module. As illustrated in Figure 2, each MTOG contains several Mixed-Attention Blocks (MTBs), a Cross-Attention Mixed Block (CAMB), and a 3×3 convolutional layer with a residual connection. For the reconstruction module, a pixel shuffle method is employed to upsample the fused features. We optimize the network parameters using the loss function L_1 .

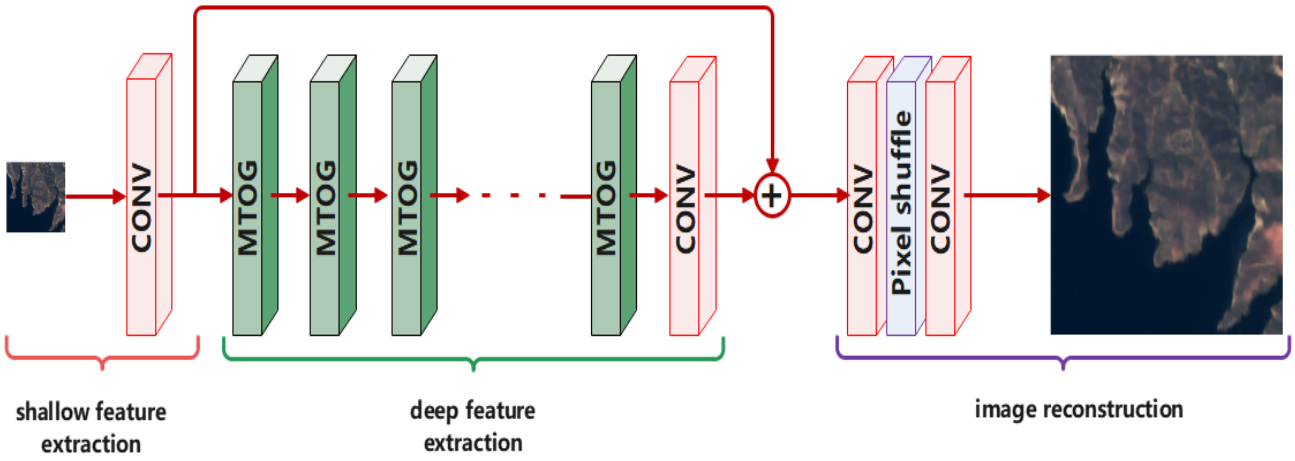

Figure 1. Architecture of the Super-Resolution Analysis Network (S2SR)

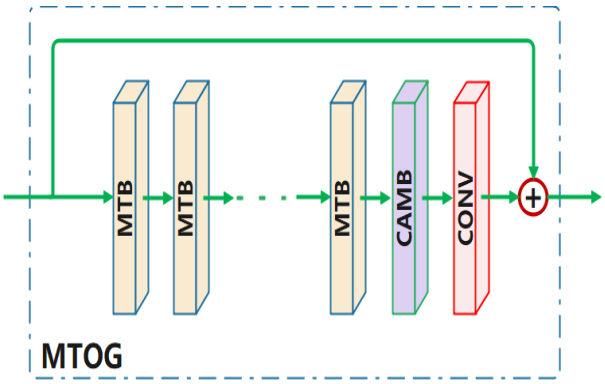

Figure 2. Structure of the Mixed Attention Residual Group (MTOG)

2.2 Mixed Attention Block (MTB)

When employing channel attention, more pixels are activated due to the incorporation of global information in computing the channel attention weights. Furthermore, numerous studies have demonstrated that convolutions can assist Transformers in obtaining superior visual representations or achieving more straightforward optimization. Therefore, we integrate convolutional blocks based on channel attention into standard Transformer blocks to enhance the representation capability of the network. As illustrated in Figure 3, within the standard Swin Transformer block, an Overlapping Cross-Attention block (OCA) is inserted in parallel after the first LayerNorm (LN) layer, alongside a 2D Selective Scanning (SS2D) module. In the SS2D module, the input block is traversed along four distinct scanning paths (cross-scanning), and each sequence is independently processed by different state-space model blocks (SSM + Selection) that integrate an input-dependent selection mechanism. Subsequently, the results are merged to construct a 2D feature map (cross-merging) as the final output. Notably, similar to previous practices, shifted window-based self-attention (adopted within SS2D) is intermittently applied across consecutive MTBs. To avoid potential conflicts between CA and SS2D in terms of optimization and visual representation, we multiply the output of CA by a small constant α. For a given input feature X, the entire computation process of the MTB is as follows:

$$X_N = \mathrm{LN}(X)$$
$$X_M = (S)W\text{-}\mathrm{MSA}(X_N) + \alpha \mathrm{CAB}(X_N) + X \quad (1)$$

$$Y = MLP(LN(X_M)) + X_M$$

Where X_N and X_M denote intermediate features, and Y represents the output of the Mixed Attention Block (MTB). Specifically, we treat each pixel as an embedded token (i.e., performing block embedding with a block size set to 1). MLP stands for Multi-Layer Perceptron. For the computation of the self-attention module, given an input feature of size $H \times W \times C$, self-attention $\frac{HW}{M^2}$ is first computed within each window $M \times M$. For local window features $X_W \in R^{M^2 \times C}$, the query, key, and value matrices, denoted as Q, K, and V, respectively, are computed through linear projections. Then, the window-based self-attention computation formula is:

$$Attention(Q, K, V) = SoftMax(QK^T/\sqrt{d} + B)V \quad (2)$$

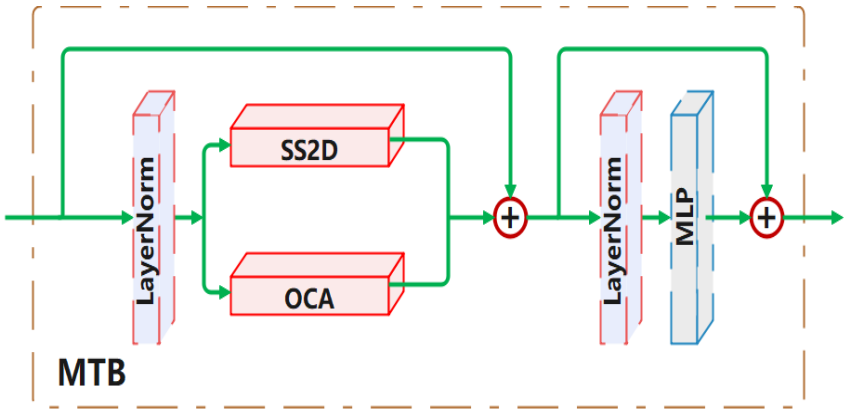

Figure 3. Mixed Attention Block (MTB)

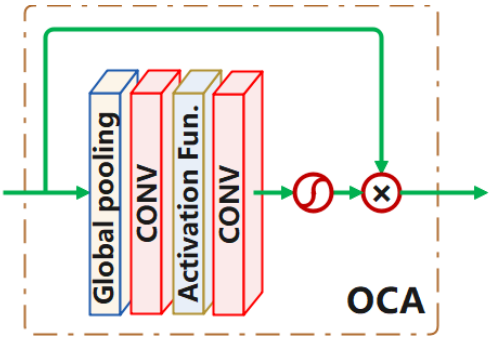

Figure 4. Overlapping Cross-Attention Block (OCA)

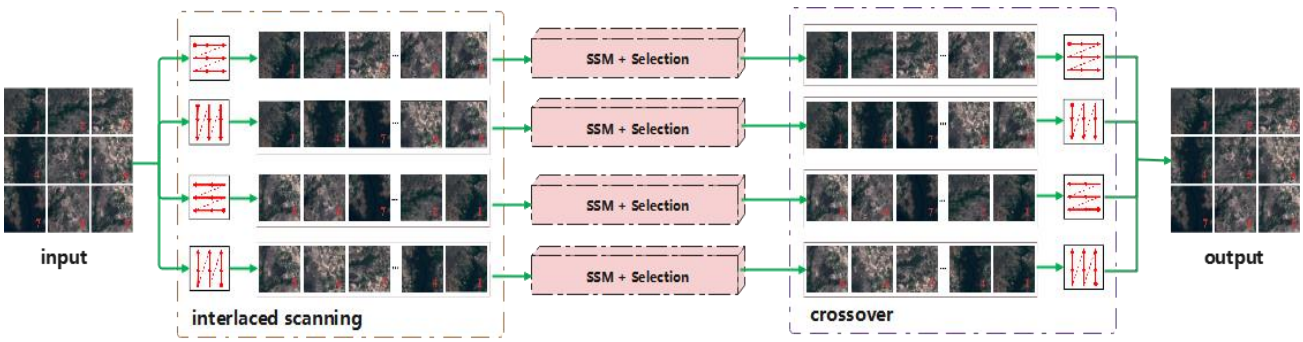

Figure 5. 2D Selective Scanning (SS2D)

2.3 Cross-Attention Module Block (CAMB)

We introduce the Cross-Attention Module Block (CAMB) to directly establish cross-window connections and enhance the representational capability of window-based self-attention. Our CAMB consists of a cross-attention (CA) layer and a multi-layer perceptron (MLP) layer, similar to the standard Swin Transformer block. However, for the CA layer, as shown in Figure 6, we employ different window sizes to partition the projected features. Specifically, for input features X, X_Q, X_K, $X_V \in R^{H\times W\times C}$ and, X_Q non-overlapping windows of size are used for $M\times M$, while X_K X_V non-overlapping windows of size $M_0 \times M_0$ are used for $\frac{HW}{M^2}$ and . The computation is performed as follows:

$$M_o = (1+\gamma)\times M \tag{3}$$

Where γ is a constant that controls the size of the overlap. To better understand this operation, the standard window partitioning can be viewed as a sliding partition with both the kernel size and stride equal to the window size M. In contrast, overlapping window partitioning can be seen as a sliding partition with a kernel size equal to M_o and a stride equal to $\frac{\gamma M}{2}$, ensuring consistency in the size of overlapping windows. The computation of the attention matrix is shown in Equation 2, and relative position biases $B\in R^{M\times M_o}$ are also adopted. Unlike window self-attention (WSA), where the queries, keys, and values are computed from the same window feature, CA computes keys/values from a larger context, allowing for more useful information to be queried. It is worth noting that, although the Multi-resolution Overlapping Attention (MOA) module in [reference] performs a similar overlapping window partitioning, our CA fundamentally differs from MOA because MOA uses window features as tokens to compute global attention, whereas CA computes cross-attention using pixel tokens within each window feature.

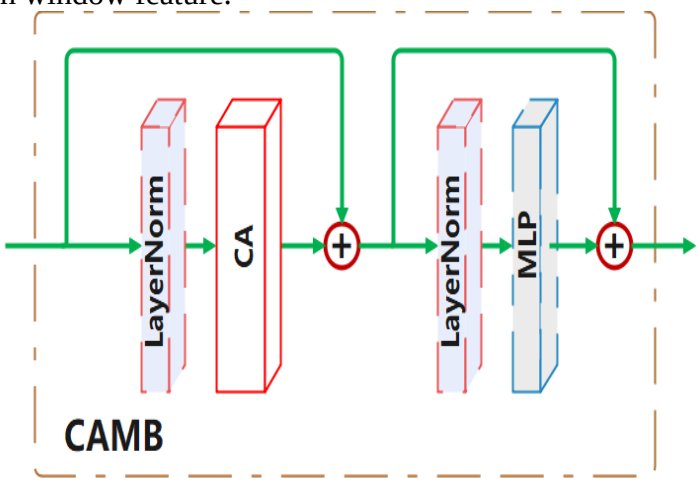

Figure 6. Cross-Attention Module Block (CAMB)

3 Experiments and Results

The proposed model, S2SR, along with other comparison models such as EDSR8-RGB[9], RCAN[10], RS-ESRGAN[11], and TS-SRGAN[12], were all run in the PyTorch environment, utilizing modules provided by the "sefibk/KernelGAN," "xinntao/BasicSR," and "Tencent/Real-SR" projects on GitHub.

Since the source images used are already the highest resolution (10m) images from the Sentinel-2 satellite, there are no actual high-resolution ground truth images (2.5m resolution) available for comparison with the generated images. Consequently, some commonly used image quality assessment metrics, such as PSNR and SSIM, are no longer applicable in this scenario. Therefore, this paper adopts no-reference image quality assessment (NR-IQA) metrics, including NIQE[13], BRISQUE[14], and PIQE[15]. The evaluation values for NIQE, BRISQUE, and PIQE can be calculated using the corresponding functions niqe, brisque, and piqe in Matlab, respectively. The output results of these three functions are all within the range of [0, 100], where a lower score indicates better perceived quality.

The 784 images in ROI_Te were processed using the EDSR8-RGB, RCAN, RS-ESRGAN, and S2SR models to generate x4 high-resolution images, and the NIQE, BRISQUE, and PIQE evaluation values for these images were calculated individually using Matlab. Table 1 provides the mean and extreme values based on the evaluation metrics. The proposed S2SR model outperforms the other models on various no-reference image quality assessment (NR-IQA) metrics. Figures 7-9 show the generated images of selected areas with strong geographical features in "ROIs1158_spring_106" for intuitive comparison of the differences between the different models. By comparing images of different terrains, it is evident that the images processed by the traditional BiCubic method are the most blurred and smooth due to the inherent limitations of interpolation algorithms. The EDSR8-RGB, RCAN, and RS-ESRGAN models fail to correctly distinguish noise with sharp edges, resulting in blurred outcomes where even houses and roads are indistinguishable. As shown in the results of the proposed S2SR, the boundaries between objects such as roads, bridges, and houses and the background are clearer, indicating that the noise estimated by our model is closer to real noise. Compared with the EDSR8-RGB, RCAN, and RS-ESRGAN models, the results of the proposed S2SR are clearer without any blurring.

Table 1. Statistical Data for NIQE, BRISQUE, and PIQE Evaluation Values

	EDSR8-RGB	**RCAN**	**RS-ESRGAN**	**T**	**S2SR**
NIQE mean	**5.851**	**5.041**	**4.108**	**3.743**	**2.286**
NIQE max	6.676	5.28	4.749	5.195	3.452
NIQE min	4.209	3.899	2.729	2.867	1.025
BRISQUE mean	**49.673**	**47.514**	**23.258**	**22.459**	**15.789**
BRISQUE max	60.312	58.572	33.774	44.027	42.839
BRISQUE min	42.859	36.064	8.648	3.943	3.108
PIQE mean	**80.299**	**60.502**	**15.044**	**14.709**	**13.122**
PIQE max	33.459	78.378	25.631	25.884	24.667
PIQE min	65.744	26.539	8.586	7.688	6.767

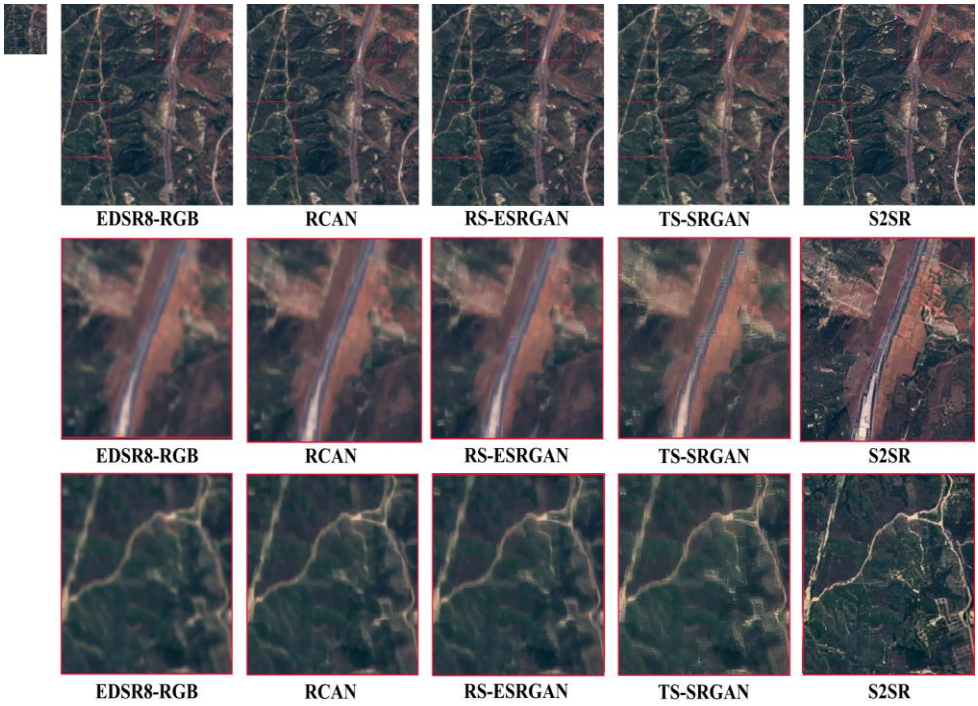

Figure 7. Comparison of Visual Effects of Images Generated in Areas with Mountain-Road Terrain. No-Reference Image Quality Assessment (NR-IQA) Values (NIQE, BRISQUE, PIQE) for Images: EDSR8-RGB (5.85, 50.25, 86.05), RCAN (4.64, 47.69, 62.20), RS-ESRGAN (3.31, 22.16, 8.42), TS-SRGAN (3.16, 15.75, 8.52), S2SR (2.43, 7.36, 7.48).

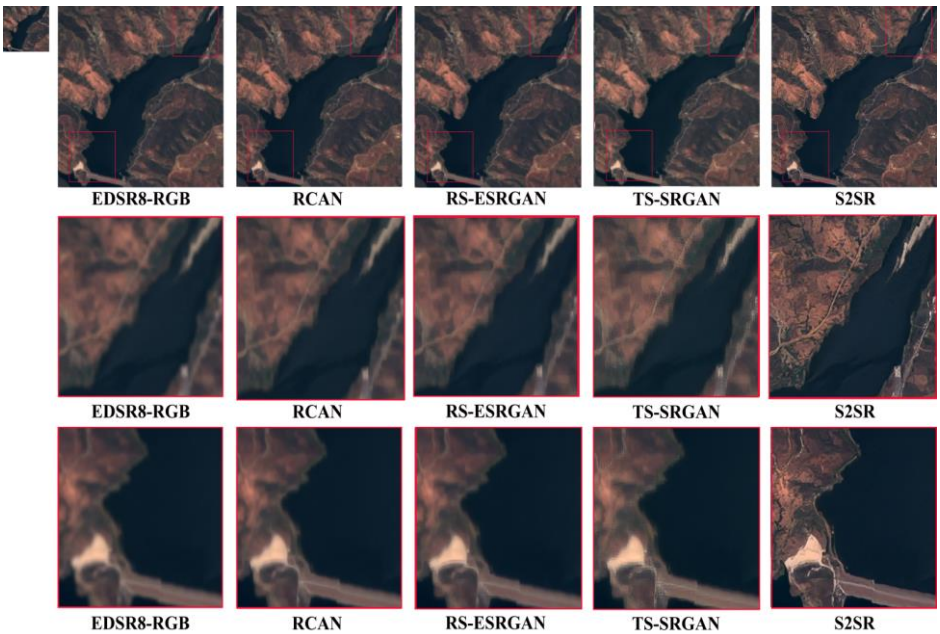

Figure 8. Comparison of Visual Effects of Images Generated in Areas with Surface Water Terrain. No-Reference Image Quality Assessment (NR-IQA) Values (NIQE, BRISQUE, PIQE) for Images: EDSR8-RGB (5.29, 46.00, 76.57), RCAN (4.67, 46.28, 74.49), RS-ESRGAN (3.53, 27.75, 12.08), TS-SRGAN (2.56, 13.68, 9.84), S2SR (2.42, 13.97, 10.03).

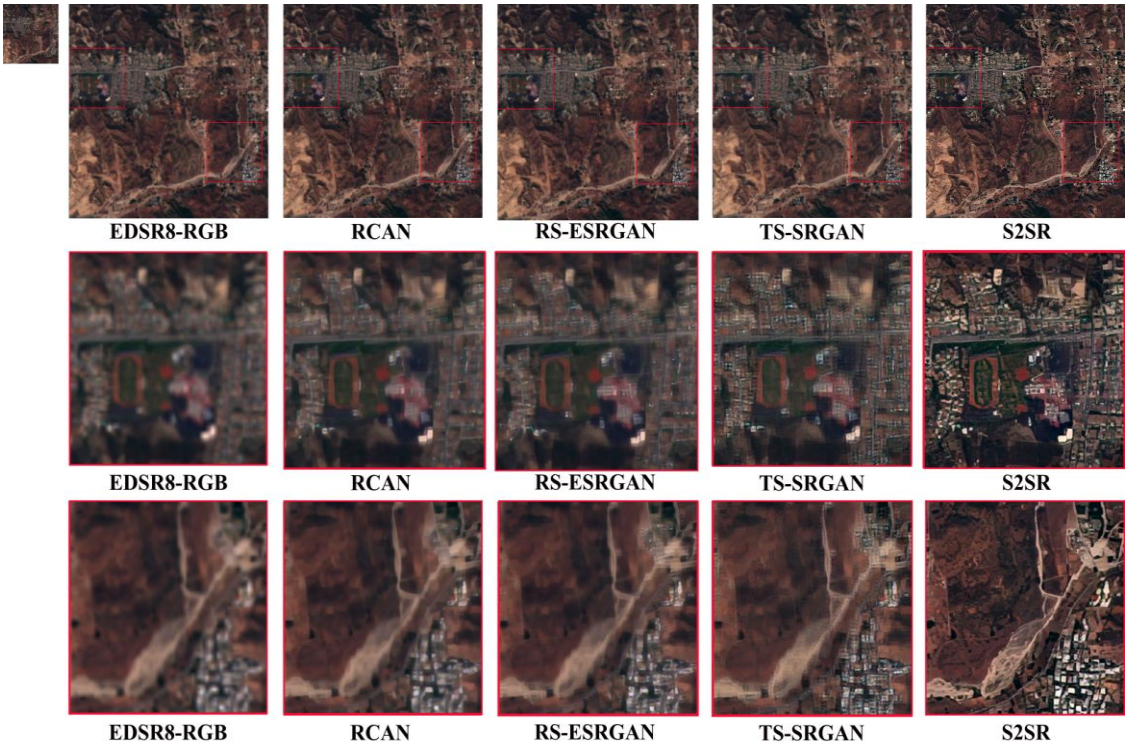

Figure 9. Comparison of Visual Effects of Images Generated in Residential Areas. No-Reference Image Quality Assessment (NR-IQA) Values (NIQE, BRISQUE, PIQE) for Images: EDSR8-RGB (4.82, 45.85, 74.28), RCAN (3.96, 45.61, 71.41), RS-ESRGAN (3.08, 22.47, 20.85), TS-SRGAN (3.12, 27.19, 15.42), S2SR (2.32, 18.27, 13.99).

4 Conclusions

In this paper, we address the super-resolution problem of Sentinel-2 satellite remote sensing images by proposing a novel Hybrid Attention Transformer model, termed S2SR, aimed at enhancing images from a 10-meter resolution to near-2.5-meter high resolution. By integrating channel attention and self-attention mechanisms, and leveraging the complementary characteristics of VMamba and Transformer, the model significantly boosts its representation capability and reconstruction quality. Firstly, we devise an image degradation model based on Generative Adversarial Networks (GANs), utilizing KernelGAN to explicitly estimate the degradation kernels of images and incorporating noise injection to construct natural low-resolution to high-resolution image pairs. This approach overcomes the issue of lacking real high-resolution images in training data for traditional methods, enabling the model to train under degradation conditions that are closer to reality. Secondly, we innovatively propose Mixed Attention Blocks (MTB) and Cross Attention Mixed Blocks (CAMB), combining the advantages of VMamba and Transformer. MTB, by introducing channel attention blocks and a 2D Selective Scanning (SS2D) module, effectively utilizes global information and local features. CAMB, through an overlapping cross-attention mechanism, enhances cross-window information interaction and reduces blocking artifacts. These designs enable the S2SR model to activate more pixels for reconstruction, almost perceiving all pixels in the image and restoring correct and sharp textures. In experiments, we train and test the model using the SEN12MS dataset and evaluate its performance through No-Reference Image Quality Assessment (NR-IQA) metrics, including NIQE, BRISQUE, and PIQE. The results demonstrate that the S2SR model outperforms existing state-of-the-art methods such as EDSR8-RGB, RCAN, RS-ESRGAN, and TS-SRGAN on various NR-IQA metrics. Especially when processing areas with complex geographical features, the S2SR model generates

clearer images with sharper edges, showcasing its powerful super-resolution reconstruction capability. In summary, the proposed S2SR model in this paper achieves efficient super-resolution reconstruction of Sentinel-2 satellite remote sensing images through innovative hybrid attention mechanisms and degradation model designs. The model not only performs excellently on quantitative metrics but also demonstrates significant improvements in visual effects. Future work can further explore more types of remote sensing images and higher-resolution super-resolution reconstruction to expand the application scope of the S2SR model.

Funding

This research was funded by the General University Key Field Special Project of Guangdong Province, China (Grant No. 2024ZDZX1004); the General University Key Field Special Project of Guangdong Province, China (Grant No. 2022ZDZX1035)

References

[1] Phiri D, Simwanda M, Salekin S, et al. Sentinel-2 data for land cover/use mapping: A review[J]. Remote Sensing, 2020, 12(14): 2291.

[2] Li X, Chen J, Cui Z, et al. Single image super-resolution based on sparse representation with adaptive dictionary selection[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2016, 30(07): 1654006.

[3] Lei S, Shi Z. Hybrid-scale self-similarity exploitation for remote sensing image super-resolution[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-10.

[4] Shao Z, Wang L, Wang Z, et al. Remote sensing image super-resolution using sparse representation and coupled sparse autoencoder[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(8): 2663-2674.

[5] Dong C, Loy C C, Tang X. Accelerating the super-resolution convolutional neural network[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer International Publishing, 2016: 391-407.

[6] Han K, Xiao A, Wu E, et al. Transformer in transformer[J]. Advances in neural information processing systems, 2021, 34: 15908-15919.

[7] Liu Z, Hu H, Lin Y, et al. Swin transformer v2: Scaling up capacity and resolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 12009-12019.

[8] Xu R, Yang S, Wang Y, et al. A survey on vision mamba: Models, applications and challenges[J]. arXiv preprint arXiv:2404.18861, 2024.

[9] Galar M, Sesma R, Ayala C, et al. Super-resolution of sentinel-2 images using convolutional neural networks and real ground truth data[J]. Remote Sensing, 2020, 12(18): 2941.

[10] Lin Z, Garg P, Banerjee A, et al. Revisiting rcan: Improved training for image super-resolution[J]. arXiv preprint arXiv:2201.11279, 2022.

[11] Salgueiro L, Marcello J, Vilaplana V. SEG-ESRGAN: A multi-task network for super-resolution and semantic segmentation of remote sensing images[J]. Remote Sensing, 2022, 14(22): 5862.

[12] Li Y, Wang Y, Li B, et al. Super-resolution of remote sensing images for × 4 resolution without reference images[J]. Electronics, 2022, 11(21): 3474.

[13] Wu L, Zhang X, Chen H, et al. Unsupervised quaternion model for blind colour image quality assessment[J]. Signal Processing, 2020, 176: 107708.

[14] Li H, Cao W, Li S, et al. Blind Image Quality Assessment Based on Natural Scene Statistics[J]. Journal of System Simulation, 2020, 28(12): 2903-2911.

[15] Ganesan P, Sathish B S, Vasanth K, et al. Color Image Quality Assessment Based on Full Reference and Blind Image Quality Measures[M]//Innovations in Electronics and Communication Engineering: Proceedings of the 8th ICIECE 2019. Singapore: Springer Singapore, 2020: 449-457.