HYBRID ATTENTION AND FEATURE MODULATION TRANSFORMER FOR REMOTE SENSING IMAGE SUPER-RESOLUTION

WEN-LIANG YU[1], MIN ZHANG[1,*], JUN-KAI ZHAO[2], KAI-FANG LI[3,*], XIN YUAN[4], YUAN-YAN TANG[5], YUAN-HAO JIN[1], XIANG LI[1]

[1]College of Artificial Intelligence and Big Data, Hefei University, Hefei, 230601, China
[2]School of Software Engineering, Changchun University of Science and Technology, Changchun, 130022, China
[3]School of Advanced Engineering and Manufacturing, Hefei University, Hefei, 230601, China
[4]School of Electrical and Mechanical Engineering, Adelaide University, Adelaide, SA 5005, Australia
[5]Zhuhai UM Science and Technology Research Institute, FST University of Macau, 999078, Macau
*Corresponding Author's E-mail: hfuuzhangmin@hfuu.edu.cn

Abstract:

Recently, remote sensing image super-resolution (RSISR) has attracted considerable attention, and Transformer-based networks have achieved remarkable progress in this field. However, existing methods still struggle to sufficiently extract multi-scale texture features and effectively focus on critical channels and key spatial regions. To address these issues, we propose a novel Transformer-based network, named HATFM-SR, for remote sensing image super-resolution. Specifically, a lightweight channel attention (LCA) mechanism is embedded within Transformer blocks to emphasize informative channels. Furthermore, a feature importance modulation (FIM) module is introduced to enhance the extraction of critical local information in both spatial and channel domains. In addition, a multi-scale convolution module (MSConv) is employed to strengthen shallow feature representations for image reconstruction. Experimental results demonstrate that the proposed method achieves promising performance and shows significant potential for remote sensing image super-resolution reconstruction.

Keywords: Remote sensing image super-resolution; Transformer; Multi-scale feature extraction; Channel attention; Feature modulation

1 Introduction

The availability of high-resolution remote sensing images has significantly improved the accuracy of land-use classification, object detection, environmental assessment, and infrastructure monitoring. However, acquiring high resolution remote sensing data is often constrained by the physical limitations of imaging sensors and transmission bandwidth. Therefore, super-resolution (SR) techniques have attracted increasing attention due to their ability to generate high quality remote sensing images for practical applications.

Single image super-resolution (SISR) methods for remote sensing images have recently attracted increasing attention. Early deep learning-based SR methods, such as SRCNN [2], EDSR [3] and RCAN [4], employ convolutional neural networks (CNNs) to learn nonlinear mappings between low-resolution (LR) and high-resolution (HR) images.

Recently, Transformer-based models have shown superior performance in vision tasks due to their self-attention mechanisms. For example, IPT is based on a standard Transformer and addresses various image restoration tasks, including SR. The Swin Transformer [8] introduces shifted window attention to balance computational efficiency and global modeling capability. SwinIR [1] successfully adopts the Swin Transformer to image restoration and super-resolution tasks. Zhang et al. [14] proposed ELAN, which adopts a Grouped Multi-Scale Self-Attention (GMSA) mechanism to partition feature maps along the channel dimension and assign different window sizes to each group. However, remote sensing images contain rich details, including spectral and texture information, and exhibit obvious multi-scale characteristics. Existing Transformer-based methods mainly focus on global information modeling, which limits their ability to exploit multi-scale relationships in remote sensing images.

To address the limitations of previous methods, this paper proposes a novel network (HATFM-SR). The main contributions of this study are as follows.

(1) We design a Lightweight Channel Attention (LCA) module to emphasize critical channels. Moreover, a Multi-Scale Convolution module(MSConv) is introduced to enhance shallow features.

(2) We design a feature importance modulation module(FIM), which integrates the spatial feature and the channel feature through a dual-dimension modulation mechanism.

(3) Comprehensive evaluation on the AID dataset demonstrate the effectiveness and reliability of our method.

2 Related Work

2.1 CNN-Based Super-Resolution

Due to the powerful local feature extraction and nonlinear fitting capabilities of CNNs, CNN-based methods have been widely applied to SR tasks. Dong et al. proposed SRCNN [2], which learns an end-to-end nonlinear mapping from LR images to HR images. Kim et al. proposed VDSR [12], which increases the network depth to enlarge the receptive field, while EDSR [3] and RCAN [4] further improve reconstruction performance through residual learning and channel attention. In addition, ESRGAN [5] introduces adversarial learning to improve visual realism. However, convolution operations are inherently limited by local receptive fields, making it difficult for CNN-based methods to establish long-range dependencies.

2.2 Transformer-Based Super-Resolution

Transformer-based methods enhance long-range feature modeling through self-attention mechanisms. Swin Transformer [8] reduces the computational cost of standard self-attention by using window-based attention with shifted windows. SwinIR [1] applies this architecture to image restoration and super-resolution tasks and achieves competitive performance. Subsequently, ELAN [14] and TransENet [16] further explore Transformer-based structures for image super-resolution and remote sensing image super-resolution, respectively. However, existing Transformer-based methods still cannot explicitly address the multi-scale characteristics and spatial heterogeneity of remote sensing images.

2.3 Remote Sensing Super-Resolution

Lei et al. [11] proposed LGCNet for remote sensing image SR, which uses local and global representations to learn the residual between HR images and upsampled LR images. Li et al. [15] proposed a GAN-based remote sensing image SR network, namely SRAGAN, which adopts a local-global attention mechanism to capture both local details and global contextual information.

3 Proposed Method

In this section, we first introduce the overall architecture of the proposed HATFM-SR and then describe the Multi-Scale Convolution Module (MSConv). Finally, we present two key modules in HATFM-SR, namely LCA and FIM.

3.1 Overall Architecture

As illustrated in Fig. 1, our proposed HATFM-SR network mainly consists of three components: shallow feature extraction, deep feature extraction, and an image reconstruction module. Given a low-resolution (LR) remote sensing image I_{LR}, the model aims to reconstruct a high-resolution (HR) image I_{SR} through feature extraction and refinement.

The overall process can be formulated as:

$$I_{SR} = F\left(I_{LR};\theta\right) \tag{1}$$

where $F(\cdot)$ denotes the proposed super-resolution network and θ presents learnable parameters.

As defined in Eq. (1), the model aims to learn a nonlinear mapping from the low-resolution input to the corresponding high-resolution reconstruction through progressive feature extraction and refinement. HATFM-SR introduces multi-scale shallow feature enhancement and dual-dimension feature modulation mechanisms to extract complex features from remote sensing images.

3.2 Multi-Scale Convolution Module (MSConv)

Remote sensing images typically contain objects with varying spatial scales, including large structures, such as rivers and highways, and small targets, such as vehicles and small buildings. A single convolution kernel size may not effectively capture such scale diversity. To address this issue, we introduce the Multi-Scale Convolution (MSConv) module to extract the shallow feature.

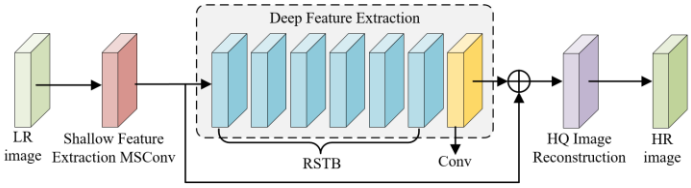

FIGURE 1. Overall architecture of the proposed framework

To extract complex texture distributions in remote sensing images, three parallel convolution branches with kernel sizes 3×3, 5×5 and 7×7 are employed to extract local features under different receptive fields. The features generated by the three branches are denoted as F_3, F_5 and F_7 respectively. By concatenating them along the channel dimension, a multi-scale shallow representation is obtained, which is defined as follows:

$$F_{cat} = [F_3, F_5, F_7] \tag{2}$$

The proposed design can expand the receptive field and capture fine-grained details. By introducing MSConv, the network can better capture multi-scale features in remote sensing images.

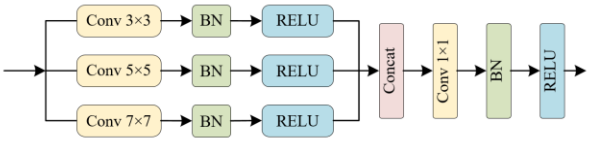

FIGURE 2. Structure of the Multi-Scale Convolution (MSConv) module.

3.3 Deep Feature Modeling with Swin Transformer and LCA

3.3.1 Lightweight Channel Attention (LCA)

The output features of the Transformer contain rich channel-wise information.

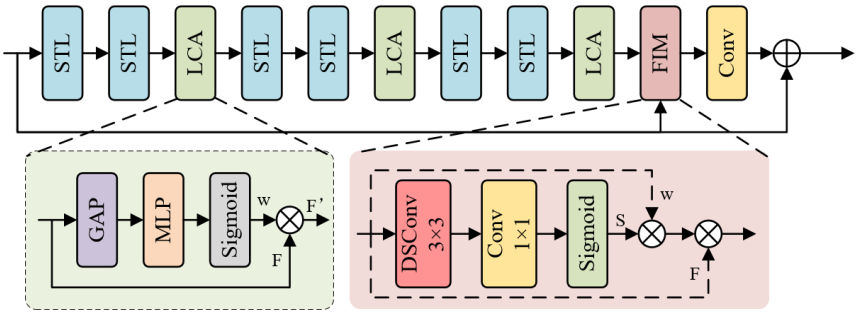

FIGURE 3. Structures of the Lightweight Channel Attention (LCA) and Feature Importance Modulation (FIM) modules.

However, different feature channels contribute unequally to the reconstruction of texture and structural details in remote sensing images. Channel attention mechanisms, such as the Squeeze-and-Excitation (SE) block [6], are effective for feature selection. To select important feature channels, a Lightweight Channel Attention (LCA) module is introduced after each Swin Transformer block, as illustrated in Fig. 3.

The LCA module first employs global average pooling (GAP) to obtain a compact channel descriptor:

$$z = 'GAP'(F) \in R\hat{}C \tag{3}$$

where $F \in R\hat{}(C\times H\times W)$ denotes the input feature map and C represents the number of channels.

The channel descriptor is then passed through a lightweight multi-layer perceptron (MLP) to generate channel attention weights:

$$w = \sigma(MLP(z)) \tag{4}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function.

Finally, the channel weights are applied to the feature map through channel-wise multiplication:

$$F' = w \mathrm{e}\ F \tag{5}$$

where e denotes element-wise multiplication for channel reweighting.

As shown in Eqs. (3)-(5), the LCA module applies channel-wise reweighting to emphasize informative feature channels while suppressing redundant responses. Moreover, the lightweight design reduces computational cost while introducing only a limited number of additional parameters.

3.3.2 Feature Importance Modulation (FIM)

We propose a Feature Importance Modulation (FIM) module, as illustrated in Fig. 3, which integrates spatial and channel importance through a dual-dimension modulation mechanism.

Spatial Importance Modeling. Given the channel-refined feature map $F' \in R\hat{}(C\times H\times W)$, spatial importance map is estimated using a depthwise separable convolution (DSConv), as defined below:

$$S = \sigma(DSConv(F')) \tag{6}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function and

$S \in R^{\wedge}(C \times H \times W)$ represents the spatial importance map.

By using depthwise separable convolution, the module captures spatial features without introducing excessive parameters.

*Dual-Dimension Importance Fusion***.** The channel attention weights $w \in R^{\wedge}C$ obtained from the LCA module are directly reused to avoid redundant computation.

To combine spatial and channel importance, the channel weights are first broadcast along the spatial dimensions and then fused with the spatial importance map through element-wise multiplication:

$$M = S \, e \, w \tag{7}$$

where the broadcasting operation expands w to match the spatial dimensions $H \times W$ and e denotes element-wise multiplication.

The final modulated feature is computed as:

$$F_m = M \otimes F \tag{8}$$

4 Experimental Results and Analysis

4.1 Experimental Setup

Experiments are conducted on the AID dataset [9] to evaluate the effectiveness of the proposed method for remote sensing image super-resolution. Following the common setting in image SR, the scale factor is set to ×4 throughout all experiments. The proposed network is built upon the SwinIR framework. In the deep feature extraction stage, the numbers of Residual Swin Transformer Blocks (RSTBs) and Swin Transformer Layers (STLs) in each block are set to 6 and 6, respectively. Following the original SwinIR setting, the window size, channel number, and attention head number are set to 8, 180, and 6, respectively. Based on this baseline architecture, a shallow multi-scale feature extraction module is introduced to enhance the representation of remote sensing textures, while a lightweight channel attention module and a feature importance modulation module are further incorporated to improve channel selection and enhance spatially salient features.

The proposed method is compared with several representative approaches, including Bicubic interpolation, SRCNN [2], EDSR [3], RCAN [4], and SwinIR [1]. These methods include both classical interpolation-based reconstruction methods and deep learning-based SR frameworks, providing a comprehensive set of baselines for remote sensing image super-resolution.

4.2 Quantitative Comparison

Fig. 4 shows the quantitative comparisons between the proposed method and several representative super-resolution approaches, including Bicubic interpolation, SRCNN, EDSR, RCAN, and SwinIR. The proposed method achieves the best performance on the AID dataset under the ×4 setting, with a PSNR of 31.24 dB and an SSIM of 0.8921.

Compared with Bicubic and CNN-based methods, the proposed method achieves superior reconstruction accuracy and structural similarity, indicating its capability in recovering complex textures and structural details in remote sensing images. Despite SwinIR achieving the best baseline performance among the compared methods, the proposed method still achieves further improvement. This demonstrates the effectiveness of the introduced shallow multi-scale feature extraction, lightweight channel attention, and feature importance modulation modules for remote sensing image super-resolution.

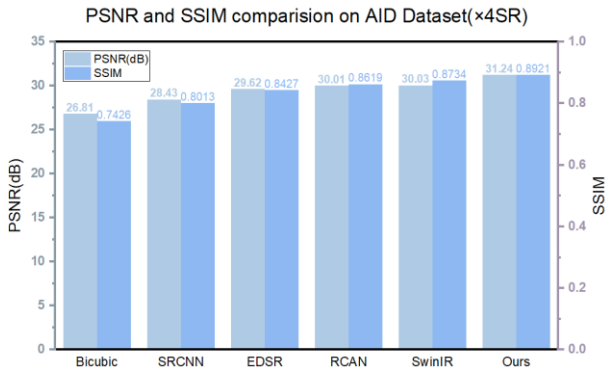

FIGURE 4. Comparison of PSNR and SSIM for Different Methods

4.3 Visual Comparison on the AID Dataset

To qualitatively evaluate the reconstruction performance of different methods on remote sensing images, three representative scene categories are selected from the AID dataset for visual comparison, as shown in Figs. 5-7.

(1) Urban Dense Regions

FIGURE 5. Visual comparison results in urban dense regions.

(2) Agricultural Texture Regions

FIGURE 6. Visual comparison results in agricultural texture regions.

(3) Complex Mixed Scenes

FIGURE 7. Visual comparison results in complex mixed scenes.

As shown in Fig. 5-7, Bicubic produces severely blurred results with significant texture loss, while SRCNN and EDSR recover only part of the structural information and still suffer from edge discontinuities and texture artifacts. RCAN improves contour reconstruction, and SwinIR better preserves the overall structure, but both remain limited in recovering fine-grained details in complex remote sensing scenes. In contrast, the proposed method generates clearer boundaries, more coherent textures, and better structural continuity, demonstrating stronger capability in restoring both global structure and local details.

4.4 Ablation Study

To verify the effectiveness of the proposed components, ablation experiments are conducted on the AID dataset with an upscaling factor of ×4. Based on the SwinIR baseline, the modules for shallow multi-scale feature extraction, lightweight channel attention, and feature importance modulation are progressively introduced. The quantitative results are reported in Table 1.

TABLE 1. Ablation experiments for the blocks on AID×4

Model Configuration	PSNR	SSIM
SwinIR	30.30	0.8734
SwinIR+MSConv	30.68	0.8802
SwinIR+MSConv+LCA	31.01	0.8867
SwinIR+MSConv+LCA+FIM	31.24	0.8921

Effect of shallow multi-scale extraction. After introducing the shallow multi-scale extraction module, the PSNR and SSIM improve by 0.38 dB and 0.0068, respectively, showing that multi-scale shallow representation is effective for capturing texture variations in remote sensing images.

Effect of lightweight channel attention. Further introducing the lightweight channel attention module achieves additional performance gains, indicating that channel-wise reweighting is beneficial in enhancing informative features and suppressing redundant responses.

Effect of feature importance modulation. After adding the feature importance modulation module, the complete model achieves the best performance, with a PSNR of 31.24 dB and an SSIM of 0.8921. This confirms the effectiveness of feature importance modulation for improving detail recovery and structural consistency.

Discussion of the ablation results. As shown in Table 1, Each proposed component contributes positively to the overall performance of the model. The results indicate that the proposed modules complement each other, and their progressive integration consistently improves performance over the SwinIR baseline.

4.5 Computational Complexity Analysis

To further evaluate the efficiency of the proposed method, the number of parameters and FLOPs are compared with those of the SwinIR baseline. FLOPs are calculated under the ×4 upscaling setting with an LR input size of 256 × 256. The results are reported in Table 2.

TABLE 2. Computational complexity comparison on AID ×4

Method	Params (M)	FLOPs (G)	PSNR	SSIM
SwinIR	11.90	788.60	30.30	0.8734
HATFM-SR	12.27	806.80	31.24	0.8921

As shown in Table 2, compared with SwinIR, HATFM-SR increases the number of parameters from 11.90M to 12.27M and the FLOPs from 788.60G to 806.80G. The additional computational cost is limited, with only a 3.11% increase in parameters and a 2.31% increase in FLOPs. Meanwhile, HATFM-SR improves the PSNR by 0.94 dB and the SSIM by 0.0187. These results indicate that the proposed method improves reconstruction performance with low computational cost, demonstrating the lightweight and efficient design of HATFM-SR.

5 Conclusions

This paper proposes a new SR framework for remote sensing image super-resolution, named HATFM-SR. We de-

sign a lightweight channel attention (LCA) mechanism to emphasize informative channels, thereby enhancing the reconstruction of important texture details in remote sensing images. Moreover, a feature importance modulation (FIM) module is developed to integrate spatial and channel features through a dual-dimension modulation mechanism. In addition, a multi-scale convolution module (MSConv) is introduced to enhance shallow feature representations. Extensive experiments on the AID dataset show that HATFM-SR achieves promising performance, demonstrating the effectiveness of the proposed framework for remote sensing image super resolution.

Acknowledgements

This paper is supported by the Hefei University Talent Research Fund Project (21-22RC12) and the Anhui Provincial University Research Project (2025AHGXZK20110).

References

[1] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image Restoration Using Swin Transformer," *in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 1833-1844.

[2] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295-307, 2016.

[3] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 136-144.

[4] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image Super-Resolution Using Very Deep Residual Channel Attention Networks," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2940-2955, 2018.

[5] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," *in Proceedings of the European Conference on Computer Vision Workshops (ECCV)*, 2018, pp. 11-25.

[6] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132-7141.

[7] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," *in Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3-19.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," *in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10012-10022.

[9] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965-3981, 2017.

[10] W. Yang, X. Zhang, Y. Tian, W. Wang, and J.-H. Xue, "Deep Learning for Single Image Super-Resolution: A Brief Review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106-3121, 2019.

[11] S. Lei, Z. Shi and Z. Zou, "Super-Resolution for Remote Sensing Images via Local-Global Combined Network," *in IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 8, pp. 1243-1247, 2017.

[12] J. Kim, J. K. Lee, and K. M. Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1646-1654.

[13] S. Jia, Z. Wang, Q. Li, X. Jia, and M. Xu, "Multiattention Generative Adversarial Network for Remote Sensing Image Super-Resolution," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 8006-8018, 2022.

[14] X. Zhang, H. Zeng, S. Guo, et al. "Efficient Long-Range Attention Network for Image Super-resolution," *European Conference on Computer Vision (ECCV)*, 2022: 649-667.

[15] Y. Li, S. Mavromatis, F. Zhang, Z. Du, J. Sequeira, Z. Wang, X. Zhao, and R. Liu, "Single-image super-resolution for remote sensing images using a deep generative adversarial network with local and global attention mechanisms," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-24, 2021.

[16] S. Lei, Z. Shi, and W. Mo, "Transformer-based multi-stage enhancement for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens*, vol. 60, Art. no. 5622724, pp. 1-14, 2021