FASwinNet: Frequency-Aware Swin Transformer for Remote Sensing Image Super-Resolution via Enhanced High-Similarity-Pass Attention and Octave Residual Blocks

Zhongyang Wang *, Shilong Liu, Keyan Cao, Xinlei Wang

Posted Date: 30 October 2025

doi: 10.20944/preprints202510.2414.v1

Keywords: remote sensing; super resolution; frequency-aware network

Article

FASwinNet: Frequency-Aware Swin Transformer for Remote Sensing Image Super-Resolution via Enhanced High-Similarity-Pass Attention and Octave Residual Blocks

Zhongyang Wang *, Shilong Liu, Keyan Cao, Xinlei Wang

School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang, Liaoning, China
* Correspondence: wangzhongyang@sjzu.edu.cn

Abstract

Remote Sensing Image Super-Resolution (RSISR) plays a vital role in enhancing the spatial details and interpretability of satellite imagery. However, existing methods often struggle to recover fine textures and high-frequency information effectively. In this paper, we propose a frequency-aware super-resolution network for remote sensing images, termed FASwinNet. The network introduces an Enhanced High-Similarity-Pass Attention (EHSPA) module, which improves high-frequency detail modeling through a similarity-aware mechanism guided by edge and positional information. Additionally, we design an Octave-based Residual Attention Block that explicitly separates and optimizes high and low-frequency features, further enhancing texture reconstruction. Experimental results demonstrate that FASwinNet outperforms state-of-the-art methods in both visual quality and quantitative metrics, achieving the best PSNR and SSIM performance on the AID and UCMerced datasets.

Keywords: remote sensing; super resolution; frequency-aware network

1. Introduction

Remote Sensing Image Super-Resolution (RSISR) [1] is a long-standing low-level vision problem that aims to reconstruct high-resolution (HR) remote sensing images from their low-resolution (LR) counterparts. Remote sensing images are widely used in applications such as land cover classification, environmental monitoring, and urban planning. However, the spatial resolution of raw satellite imagery is often limited due to hardware, bandwidth, and other environmental constraints, preventing satellite sensors from capturing images at the desired high spatial resolution. As an alternative solution, image super-resolution (SR) has attracted significant attention in the field of remote sensing. SR techniques effectively leverage one or multiple LR images to generate corresponding HR outputs, offering advantages such as cost efficiency and the ability to recover historical data. Consequently, SR has become a prominent research focus in the domain of remote sensing.

Recent advances in deep learning have led to significant progress in remote sensing image super-resolution (RSISR) [1]. Deep learning techniques are increasingly applied to image super-resolution tasks[2–9]. However, due to the limited receptive field, convolutional neural networks (CNNs) tend to focus only on local regions of an image. Recently, Transformers [10], which have achieved great success in natural language processing, have gained considerable attention in the vision community. With the rapid development of high-level vision tasks [11–13], Transformer-based methods have also been introduced to low-level vision problems [14–16] and super-resolution (SR) tasks [17,18]. In particular, the SwinIR [18] architecture demonstrates impressive capability in modeling long-range dependencies.

Nevertheless, most existing methods primarily emphasize global context modeling while paying insufficient attention to high-frequency textures, edge structures, and frequency-aware feature fusion.

 doi:10.20944/preprints202510.2414.v1

These limitations hinder their ability to reconstruct high-definition images with rich details and textures. To address the aforementioned challenges, we propose a Frequency-Aware Swin Transformer Network (FASwinNet), specifically designed for remote sensing image super-resolution (RSISR). This network aims to more effectively exploit frequency-domain features, edge information, and contextual modeling capabilities.

Built upon the SwinIR framework [18], we introduce the Enhanced High-Similarity-Pass Attention (EHSPA) module, which incorporates Sobel edge information and spatial positional encoding to guide the model's attention toward structurally critical regions. By leveraging spatial similarity matching and a soft-thresholding sparsity mechanism, EHSPA explicitly enhances the representation of high-frequency components, significantly improving the reconstruction quality of edges and textures in remote sensing images.

At the same time, we design the Octave-based Residual Attention Block (ORAB) based on the improved Octave Convolution concept [19], which explicitly decomposes feature representations into high-frequency and low-frequency channels for separate modeling and enhancement. In the high-frequency branch, High-Frequency Residual Enhancement and Efficient Localization Attention (ELA) [20] are incorporated to boost texture representation, while the low-frequency branch retains the global structural and semantic information of the image. This decoupled modeling enables effective fusion of frequency-specific features, facilitating more accurate and detailed image reconstruction.The main contributions of this paper are as follows:

- A frequency-aware super-resolution framework, FASwinNet, is proposed for remote sensing images, integrating frequency-domain modeling and attention mechanisms to enhance reconstruction quality.
- An Enhanced High-Similarity-Pass Attention (EHSPA) module is designed to guide the network toward structurally significant high-frequency regions, thereby enhancing the restoration of fine details in reconstructed images.
- An Octave-based Residual Attention Block (ORAB) is proposed to perform frequency-separated processing and fusion, effectively enhancing the network's feature representation capability.

2. Related Work

2.1. Remote Sensing Image Super-Resolution

Recent remote sensing image super-resolution (RSISR) methods can be broadly categorized into CNN-based and Transformer-based approaches. Early methods such as SRCNN and EDSR leveraged convolutional neural networks (CNNs) to progressively improve reconstruction quality. In recent years, the remarkable success of Transformers in natural language processing (NLP) has led to their widespread adoption in computer vision tasks. For high-level vision applications such as image classification, object detection, and semantic segmentation, Transformer-based methods [11,12,21–30] have demonstrated outstanding performance.

Among them, the Vision Transformer (ViT) was the first pure Transformer model applied to visual tasks. By dividing images into fixed-size patches and performing sequential modeling, ViT enables end-to-end feature extraction and global context modeling, achieving state-of-the-art results on several benchmark datasets [11,13]. SwinIR [18] introduced a hierarchical feature extraction and restoration architecture based on the Swin Transformer, employing a sliding window mechanism. This design efficiently captures both local and non-local dependencies while maintaining computational efficiency, providing a strong backbone for visual restoration tasks.

2.2. Image Restoration Using Swin Transformer

Swin Transformer for Image Restoration (SwinIR) is the first model to successfully introduce the Swin Transformer architecture into the field of image restoration, marking a significant breakthrough for Transformer-based methods in low-level vision tasks [11]. SwinIR inherits the core design principles of the Swin Transformer, incorporating a hierarchical structure and a local window-based self-attention

 doi:10.20944/preprints202510.2414.v1

mechanism. This design enables efficient computation while effectively modeling both local and global dependencies. SwinIR has been widely applied to various tasks such as super-resolution, image denoising, and compressed sensing recovery, achieving state-of-the-art performance across the board.The architecture of SwinIR primarily consists of the following key components:

Shallow Feature Extraction: As shown in Equation (1), a single convolutional layer is applied to the input low-resolution image to extract shallow features, providing an effective initial representation for the subsequent Transformer layers.

Deep Feature Extraction: As shown in Equation (2), this part consists of multiple Residual Swin Transformer Blocks (RSTBs), each of which contains several Swin Transformer layers and introduces residual connections at the block level to enhance feature propagation and gradient flow. The stacked RSTB structure endows the model with stronger representational capacity.

Image Reconstruction: As shown in Equation (4), the deep features are reconstructed into a high-resolution image through an upsampling module (e.g., PixelShuffle).

Window-based Self-Attention: this mechanism performs multi-head self-attention (W-MSA) within fixed-size local windows and introduces a window-shifting strategy (SW-MSA) to capture cross-window contextual information. This design significantly reduces computational costs, enabling SwinIR to efficiently process large-scale images.

Although SwinIR has achieved state-of-the-art performance in various image restoration tasks, it still faces significant challenges when applied to high-resolution and structurally complex remote sensing images. Remote sensing imagery is characterized by rich high-frequency textures, distinct edge structures, and large-scale repetitive patterns, which impose stringent requirements on the model's capacity for effective feature representation [31,32]. While SwinIR improves computational efficiency via its local window-based self-attention mechanism, it exhibits inherent limitations in frequency modeling and edge reconstruction.

Firstly, SwinIR's ability to model high-frequency components remains constrained. Its attention mechanism primarily operates within localized windows and lacks explicit mechanisms to capture and enhance high-frequency details such as fine textures and edge signals, often leading to reconstructed images with oversmoothed textures and blurred edges. Secondly, the network architecture is predominantly designed in the spatial domain without explicit disentanglement or differential modeling of high- and low-frequency information, thereby limiting its capability to comprehensively learn frequency-specific features. Lastly, the model demonstrates insufficient sensitivity to edge regions and does not explicitly guide the attention towards structurally salient areas within remote sensing images (e.g., building outlines, road edges), which adversely affects the preservation of structural fidelity and overall reconstruction quality.

Therefore, despite the strong Transformer backbone provided by SwinIR, its full potential in remote sensing image super-resolution has yet to be fully realized. To further enhance the reconstruction of structural details and fine textures, it is imperative to integrate frequency-aware and edge-focused mechanisms into the SwinIR framework, thereby facilitating high-quality and high-fidelity image restoration.

3. Methods

To address the challenges of remote sensing image super-resolution (RSISR) [1], we aim to design an efficient Transformer-based framework with enhanced frequency modeling and detail reconstruction capabilities. To this end, we propose the Frequency-Aware Swin Transformer Network (FASwinNet), which integrates frequency-aware attention mechanisms and a high–low frequency feature decoupling strategy to improve reconstruction performance. In this section, we first present the overall architecture of the proposed FASwinNet. We then provide a detailed description of its two core components: the Frequency-Aware Attention Block (FAB), which facilitates the recovery of fine high-frequency details, and the Octave-based Residual Attention Block (ORAB), which enables

 doi:10.20944/preprints202510.2414.v1

effective frequency-domain feature modeling. Finally, we describe the loss functions and optimization strategies used during training.

3.1. Network Architecture

As illustrated in Figure 1, the overall architecture of the proposed network consists of three main stages: shallow feature extraction, deep feature extraction, and image reconstruction. This architectural paradigm has been widely adopted in prior works [6,18], and our design follows the overall structure of SwinIR [18].

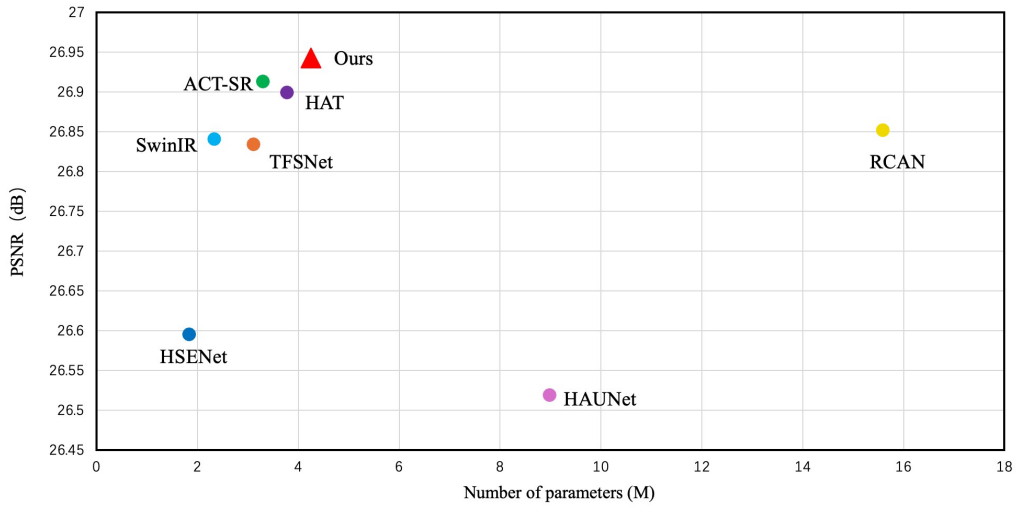

Figure 1. Quantitative comparison of parameter count and PSNR performance among different remote sensing image super-resolution methods on the AID test set.

Specifically, we denote the low-resolution input and the super-resolved output of the network as I_{LR} and I_{SR}, respectively.Given a low-resolution input image $I_{\mathrm{LR}} \in \mathbb{R}^{C\times H\times W}$, where C, H, and W denote the number of channels, height, and width of the image respectively, , we employ a 3×3 convolutional layer $H_{\mathrm{SF}}(\cdot)$ to extract shallow features. The resulting feature map $F_0 \in \mathbb{R}^{C\times H\times W}$ is computed as follows:

$$F_0 = H_{SF}(I_{LR}) \tag{1}$$

Convolutional layers are well-suited for early-stage visual processing and provide an efficient means of mapping input images from the spatial domain to a higher-dimensional feature space. Based on this observation, we feed the shallow feature map F_0 into a deep feature extraction module, denoted as $H_{\mathrm{DF}}(\cdot)$. This allows the network to extract more complex and abstract representations. The resulting deep feature map $F_{\mathrm{DF}} \in \mathbb{R}^{C\times H\times W}$ is computed as follows:

$$F_{\mathrm{DF}} = H_{\mathrm{DF}}(F_0) \tag{2}$$

The resulting deep feature map F_{DF} primarily captures high-frequency information. It is obtained through a sequence of K Residual Frequency-aware Attention Groups (RFAG) followed by a 3×3 convolutional layer. Specifically, the intermediate features $F_1, F_2, \ldots, F_K$ and the final deep feature F_{DF} are extracted in a block-wise manner as follows:

$$\begin{aligned} F_i &= H_{\mathrm{RFAG}}(F_{\text{i-1}}), \quad i = 1, 2, \ldots, K \\ F_{\mathrm{DF}} &= H_{\mathrm{conv}}(F_K) \end{aligned} \tag{3}$$

Here, $H_{\mathrm{RFAG}}(\cdot)$ denotes the i-th Residual Frequency-aware Attention Group, and $H_{\mathrm{conv}}(\cdot)$ represents the final convolutional layer. After deep feature extraction, a convolutional layer is used to reconstruct spatial features.To enhance information flow and facilitate the learning of residual

 doi:10.20944/preprints202510.2414.v1

components, we adopt a long-range skip connection. The final high-quality super-resolved image $I_{SR} \in \mathbb{R}^{C\times H\times W}$ is generated as:

$$I_{SR} = H_{RC}(F_0 + F_{DF}) \tag{4}$$

where $H_{RC}(\cdot)$ denotes the reconstruction module, which consists of a 3×3 convolutional layer followed by a pixel-shuffle operation [32] for upsampling. This design enables efficient and accurate reconstruction from the combined shallow and deep features.

3.2. *Frequency-Aware Attention Block(FAB)*

The Frequency-aware Attention Block (FAB) is a feature enhancement module designed for image reconstruction and enhancement tasks. It integrates the Swin Transformer with an Enhanced High-Similarity-Pass Attention mechanism to improve the model's capability in capturing local details and frequency-aware structures.As illustrated in Figure 2, the architecture of FAB is structured as follows:

Residual Swin Transformer Block. The Residual Swin Transformer Block (RSTB) is the core component of SwinIR, designed to enable efficient image restoration by combining the local window-based self-attention mechanism of the Swin Transformer with residual connections. This design allows RSTB to effectively extract both local and global features from the input while preserving important structural information through residual learning. The residual connection also mitigates the problem of gradient vanishing, thus enhancing the stability of feature propagation. Compared with conventional Transformers, the Swin Transformer adopts a non-overlapping window partitioning strategy, which significantly reduces computational complexity. This makes it particularly suitable for processing high-resolution images while maintaining strong performance and efficiency.

Enhanced High-Similarity-Pass Attention. As illustrated in Figure 2, to effectively recover high-frequency details such as edges and textures in remote sensing images, we propose the Enhanced High-Similarity-Pass Attention (EHSPA) based on the original High-Similarity-Pass Attention (HSPA) module [33].

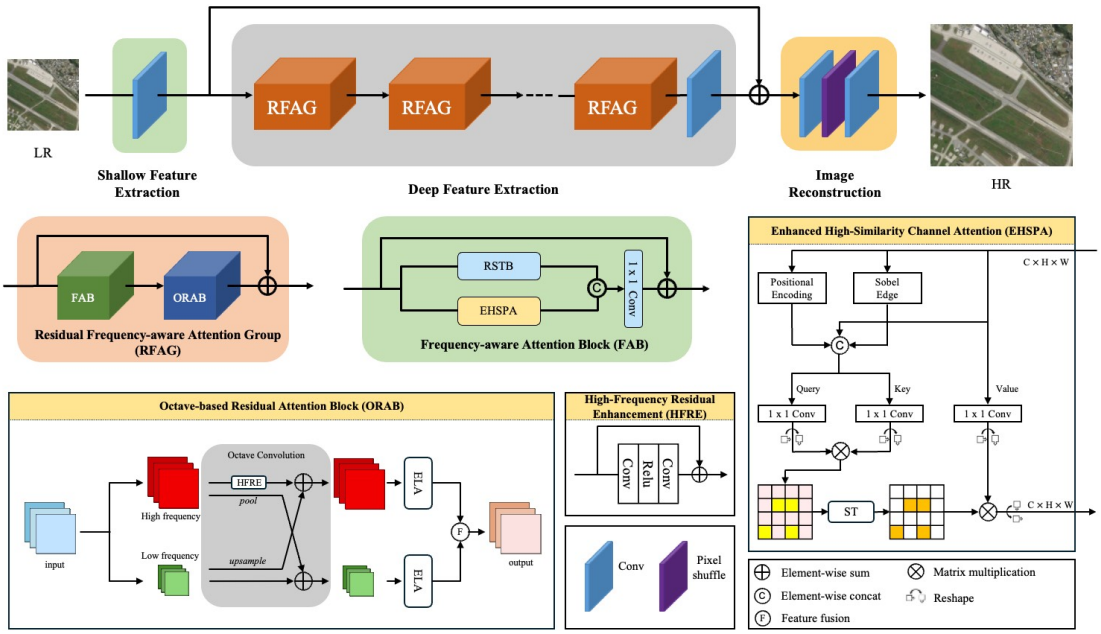

Figure 2. The overall architecture of FASwinNet and the structure of FAB and ORAB.

The original HSPA improves the modeling of structurally similar regions by constructing a global similarity map and selecting highly similar areas using a soft-thresholding operation for feature aggregation. While this approach enhances the representation of similar structures to some extent, it still suffers from several limitations. First, it lacks explicit modeling of high-frequency details, which are crucial for preserving fine-grained image content. Second, it offers limited spatial structural representation, making it less effective in capturing geometric layouts. Lastly, the matching-path

 doi:10.20944/preprints202510.2414.v1

representation capability is insufficient, leading to suboptimal performance in reconstructing edge- and texture-sensitive regions under the super-resolution task.

To address the aforementioned limitations, we introduce multi-modal auxiliary information to enhance the representational capacity of the EHSPA module. Specifically, considering that edge structures in remote sensing images often carry critical high-frequency information, we design an edge detection operator based on the Sobel filter to extract explicit edge features. These features serve as an additional signal to guide the construction of more accurate feature correspondences.For clarity, we denote the input feature map as $F_0 \in \mathbb{R}^{C\times H\times W}$, which is reshaped to $X \in \mathbb{R}^{C\times H\times W}$. We then apply the Sobel operator to extract an edge map $E \in \mathbb{R}^{1\times H\times W}$, defined as:

$$E = \sqrt{(X * K_x)^2 + (X * K_y)^2 + \varepsilon} \tag{5}$$

where $*$ denotes the convolution operation, K_x and K_y are the horizontal and vertical Sobel kernels, respectively, and ε is a small constant (e.g., 10^{-6}) added for numerical stability.

Secondly, to enhance the spatial awareness of the attention mechanism, we introduce a 2D normalized positional encoding. By concatenating positional information with the original features, the model is encouraged to establish spatially consistent similarity measurements among pixels. The normalized 2D coordinate map $P \in \mathbb{R}^{2\times H\times W}$ is defined as:

$$P_{i,j} = \left[\frac{2j}{W-1} - 1,\ \frac{2i}{H-1} - 1\right] \tag{6}$$

We concatenate the original feature, edge map, and positional encoding to form an augmented feature representation, which is then used to extract attention-related embeddings. Specifically, two separate matching paths are used to extract attention features F_1 and F_2, while a reconstruction path extracts input feature A. These features are reshaped accordingly for attention computation:

$$\begin{gathered} X_{\text{aug}} = \text{Concat}(X,\ E,\ P), \\ F_1 = f_1(X_{\text{aug}}), \quad F_2 = f_2(X_{\text{aug}}), \quad A = f_3(X), \\ Q = \text{reshape}(F_1), \quad K = \text{reshape}(F_2), \\ V = \text{reshape}(A) \end{gathered} \tag{7}$$

where $X_{\text{aug}} \in \mathbb{R}^{(C+3)\times H\times W}$, and f_1, f_2, f_3 are embedding functions implemented by 1×1 convolutional layers. The embeddings $F_1, F_2 \in \mathbb{R}^{C'\times H\times W}$, $A \in \mathbb{R}^{C\times H\times W}$, and the reshaped queries, keys, and values are $Q \in \mathbb{R}^{HW\times C'}$, $K \in \mathbb{R}^{C'\times HW}$, and $V \in \mathbb{R}^{HW\times C}$, respectively. Here, C' denotes the reduced channel dimension after compression. To suppress noise from irrelevant regions, we apply a soft-thresholding operation $\text{ST}(\cdot)$ that retains only the top-k most similar responses for each query:

$$S = \text{ST}(Q \cdot K) \tag{8}$$

where S is the sparsified attention map, and all non-top-k responses are zeroed out and re-normalized. The final output feature is obtained by aggregating the attention scores with the reconstruction features and applying residual learning:

$$Y = \alpha \cdot \text{reshape}(S \cdot V) + X \tag{9}$$

where α is a residual scaling factor that controls the strength of feature enhancement. This mechanism enables effective aggregation and preservation of features from highly similar regions, particularly in edge- and texture-rich areas.

Fusion and Residual connection. The output features from the RSTB and EHSPA modules are concatenated along the channel dimension. The fused features are then passed through a 1×1 convolutional layer to perform channel compression and feature integration. Finally, a residual

 doi:10.20944/preprints202510.2414.v1

connection is employed by adding the result to the input of the FAB, which helps preserve the original features and enhances both training stability and information flow. The Frequency-aware Attention Block (FAB) leverages a hybrid design combining RSTB and EHSPA, enabling joint modeling of global contextual information and local high-similarity patterns. Its multi-path fusion structure significantly enhances the model's representational capacity, making it particularly effective for remote sensing image super-resolution tasks.

3.3. Octave-Based Residual Attention Block(ORAB)

To more effectively decouple and exploit high-frequency and low-frequency components in image features, we design an Octave-based Residual Attention Block (ORAB) that integrates attention mechanisms and residual enhancement. As illustrated in Figure 2, ORAB consists of three main components: FirstOctaveConv, multiple stacked OctaveConv blocks, and a final LastOctaveConv.

In the FirstOctaveConv stage, the input feature map is decomposed into high-frequency and low-frequency components. The low-frequency features are obtained via average pooling, enabling frequency-aware feature separation at the early stage.

Each OctaveConv block is a stackable unit designed to process highlow and low-frequency features in parallel. In the high-frequency branch, we introduce High-Frequency Residual Enhancement (HFRE), implemented using two 3×3 convolutional layers followed by ReLU activation, which helps preserve and amplify fine-grained details. Additionally, an Efficient Localization Attention (ELA) mechanism [20] is integrated into both frequency branches. This attention design leverages 1D depth-wise convolutions within channels along with Group Normalization to model spatial relationships, thereby enhancing local feature sensitivity.

The LastOctaveConv module reunifies the high- and low-frequency branches into a single output stream. Specifically, the low-frequency features are upsampled to match the resolution of the high-frequency features and are then aggregated via element-wise addition. This design allows the module to retain global structural information from the low-frequency pathway while simultaneously improving high-frequency detail reconstruction.The entire ORAB can be formally expressed as:

$$Y = f_{\text{Last}}\left(\prod_{t=1}^{T} \text{OctaveConv}^{(t)}(f_{\text{First}}(X))\right) \tag{10}$$

where X is the input feature map, f_{First} and f_{Last} denote the FirstOctaveConv and LastOctaveConv modules respectively, and $\text{OctaveConv}^{(t)}$ represents the t-th OctaveConv block in a stack of T layers.By leveraging frequency separation, localized attention, and high-frequency residual enhancement, the proposed ORAB substantially improves the model's ability to capture fine-grained image structures and contributes to high-fidelity reconstruction in remote sensing image super-resolution.

3.4. Learning Strategy

During model training, we adopt the widely used L1 loss as the objective function, and employ the Adam optimizer for parameter updates. Specifically, given a batch of N paired low- and high-resolution images denoted as $\{(I_i^{\text{LR}}, I_i^{\text{HR}})\}_{i=1}^{N}$, the model is trained to minimize the following L1 loss:

$$\mathcal{L}1(\theta) = \frac{1}{N}\sum i = 1^{N} \left\| f_{\text{FASwinNet}}(I_i^{\text{LR}}) - I_i^{\text{HR}} \right\|_1 \tag{11}$$

Here, $f_{\text{FASwinNet}}(\cdot)$ represents the proposed super-resolution network with trainable parameters θ, and $f_{\text{FASwinNet}}(I_i^{\text{LR}})$ denotes the predicted high-resolution image corresponding to the input I_i^{LR}.

4. Experiments

4.1. Datasets and Implementation

To evaluate the effectiveness of the proposed method on remote sensing image super-resolution (RSISR), we conduct experiments on the widely used AID dataset [34], which contains remote sensing

 doi:10.20944/preprints202510.2414.v1

images from 30 different categories. To ensure class balance and fairness, we randomly select 30 images from each category for training, resulting in 900 images in total, and 2 images per category for testing, totaling 60 test images. This setup enables a robust assessment of the model's reconstruction ability and generalization performance in small-sample RSISR scenarios.

Low-resolution (LR) images are generated from the high-resolution (HR) images using bicubic downsampling. All experiments are conducted with a scale factor of ×4. For evaluation, we adopt two standard metrics—Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) computed on the Y channel of YCbCr color space, which is commonly used in image restoration tasks.

The proposed network is optimized using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate is set to 2e-4, and the mini-batch size is set to 16. All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU.

4.2. Ablation Study

In this section, to thoroughly evaluate the effectiveness of the proposed modules, we conduct systematic ablation studies on the AID dataset, focusing on the impact of the Enhanced High-Similarity-Pass Attention (EHSPA) and the Octave-based Residual Attention Block (ORAB) on model performance. For all experiments, the same hyperparameters are adopted: the number of RSTBs and STLs is set to 2, the window size to 8, the channel dimension to 48, and the number of attention heads to 2.

Training is performed for 30,000 iterations, and the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) are reported on the test set.

To quantitatively assess the impact of the proposed ORAB module, we develop several comparative model variants, and the evaluation results are summarized in Table 1.

Table 1. Effectiveness analysis of ORAB on the ×4 AID test set.

Model	PSNR	SSIM
Baseline	26.5857	0.7177
Baseline + Octave	26.6381	0.7181
Baseline + ORAB	26.7015	0.7209

To verify the effectiveness of the EHSPA module, we construct a series of baseline variants, and the corresponding results are presented in Table 2.

Table 2. Effectiveness analysis of EHSPA on the ×4 AID test set.

Model	PSNR	SSIM
Baseline	26.5857	0.7177
Baseline + HSPA	26.6303	0.7186
Baseline + EHSPA	26.6542	0.7196

Finally, the overall ablation experiments consider four different combinations of EHSPA and ORAB. The results are reported in Table 3, where a ✓ denotes the inclusion of a module and a ✗ indicates its absence.

Table 3. Ablation study of EHSPA and ORAB on the ×4 AID test set.

EHSPA	ORAB	PSNR	SSIM
✗	✗	26.5857	0.7177
✓	✗	26.6500	0.7193
✗	✓	26.6902	0.7215
✓	✓	26.7348	0.7234

 doi:10.20944/preprints202510.2414.v1

4.3. Comparisons with State-of-the-Art Method

To comprehensively evaluate the performance of the proposed FASwinNet on remote sensing image super-resolution tasks, we conduct comparative experiments on the AID dataset against a range of mainstream and state-of-the-art super-resolution methods. These include classical approaches such as Bicubic and SRCNN [35], GAN-based methods like ESRGAN [2], as well as recent deep learning models with strong performance, including EDSR [8], RCAN[6], SwinIR [18], HAT [36], and several advanced architectures specifically designed for remote sensing images, such as CTNet [37], HSENet [38], HAUNet [39], TFSNet [40], and ACT-SR [41]. To achieve a better balance between performance and model complexity, we configure FASwinNet with 6 RSTBs, 6 STLs, a window size of 8, 90 channels, and 6 attention heads. Under this setting, the model is trained for 500,000 iterations.

As shown in Table 4, we compare the reconstruction performance of various methods at a ×4 upscaling factor using PSNR, SSIM, MSE, and LPIPS as evaluation metrics. The results demonstrate that our proposed method outperforms existing state-of-the-art approaches across all metrics on the AID dataset, achieving the best overall performance.

Table 4. Quantitative comparison of different methods on the AID dataset. The best results are marked in red, and the second-best in blue.

Method	Parameters	Flops	PSNR	SSIM	MSE	LPIPS
Bicubic	–	–	25.3463	0.6752	287.3178	0.4893
SRCNN	0.02M	0.57	25.8984	0.6929	249.1492	0.3859
ESRGAN	16.69M	645.25	25.4840	0.6708	264.7530	0.3359
EDSR	1.52M	16.25	26.7145	0.7282	205.2616	0.3145
RCAN	15.59M	106.28	26.8516	0.7340	198.7015	0.3056
SwinIR	2.33M	3.48	26.8410	0.7346	199.3198	0.3025
HAT	3.78M	4.78	26.8993	0.7360	195.7945	0.3038
CTNet	0.53M	0.57	26.3965	0.7157	221.3819	0.3424
HSENet	1.84M	38.44	26.5954	0.7250	210.2266	0.3170
HAUNet	8.99M	37.79	26.5190	0.7212	215.6408	0.3242
TFSNet	3.13M	111.17	26.8338	0.7318	199.2198	0.3168
ACT-SR	3.30M	131.5	26.9131	0.7376	195.9505	0.2984
Ours	4.26M	13.92	26.9423	0.7380	194.5582	0.3022

To further evaluate the generalization capability of the proposed model, we conduct cross-dataset experiments. The model is trained on the AID dataset, which serves as the source domain, and tested on the UCMerced dataset as the target domain. For evaluation, two images are randomly selected from each class of the UCMerced dataset, resulting in a representative subset for testing the model's performance on unseen data. The reconstruction quality is quantitatively assessed using two widely adopted metrics, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). The results of these cross-dataset tests are summarized in Table 5, which demonstrates the model's ability to generalize beyond the training dataset and maintain high-quality reconstruction across different remote sensing image distributions.

As shown in Table 5, even when tested on images from an unseen dataset, the proposed model consistently outperforms the baseline and other comparative methods in terms of PSNR and SSIM. This indicates that the model effectively learns features with strong generalization capability, rather than being limited to the training samples from the AID dataset. Notably, the improvement in SSIM further suggests that the model can well preserve structural and textural information across different remote sensing scenes. Overall, these results validate the strong generalization ability of the proposed approach and highlight its potential advantages in practical scenarios where the test data distribution differs from that of the training data.

 doi:10.20944/preprints202510.2414.v1

Table 5. Quantitative comparison of different methods on the UCMerced dataset. The best results are marked in red, and the second-best in blue.

Method	PSNR	SSIM	MSE	LPIPS
Bicubic	24.2996	0.5956	312.1537	0.5705
SRCNN	24.3741	0.6004	306.7432	0.5604
ESRGAN	24.3296	0.5963	309.8543	0.5452
EDSR	24.5884	0.6067	291.9466	0.5327
RCAN	24.5951	0.6129	291.4762	0.5301
SwinIR	24.5896	0.6137	292.3339	0.5221
HAT	24.6389	0.6141	288.8521	0.5212
CTNet	24.5673	0.6055	293.3512	0.5231
HSENet	24.5894	0.6121	292.1432	0.5323
HAUNet	24.5847	0.6104	292.3324	0.5384
TFSNet	24.6102	0.6112	290.5234	0.5312
ACT-SR	24.6345	0.6133	288.2546	0.5203
Ours	24.6411	0.6156	287.9915	0.5180

As shown in Figure 3, our method successfully reconstructs visually rich and structurally clear image content. In contrast, other methods exhibit varying degrees of blurring and detail loss. Our approach more accurately restores the textures of rooftop structures and road edges, with particularly sharp reconstruction observed in parking areas and along building boundaries. Notably, our method also preserves the circular line structures in the center of sports fields more effectively, exhibiting stronger continuity and clearer boundaries compared to other approaches. Visually, the reconstructed field textures appear more natural, with finer details better preserved. These visual results further demonstrate the superiority of our method in maintaining structural consistency and enhancing detail recovery in remote sensing image super-resolution.

Figure 3. *Cont.*

 doi:10.20944/preprints202510.2414.v1

Figure 3. Reconstruction results of different methods on test images from the AID dataset under ×4 upscaling. The red boxes indicate enlarged regions, showing that FASwinNet achieves clearer texture and edge restoration.

5. Conclusions

In this paper, we propose a novel frequency-aware Swin Transformer network, FASwinNet, for remote sensing image super-resolution, aiming to better capture frequency domain features and structural details. We design the Enhanced High-Similarity-Pass Attention (EHSPA) module, which integrates edge guidance and spatial similarity matching to effectively enhance the modeling of high-frequency textures. Additionally, we introduce the Octave-based Residual Attention Block (ORAB), which explicitly separates features into high- and low-frequency subspaces. By incorporating high-frequency residual enhancement and localized attention mechanisms, ORAB further improves reconstruction quality. Experimental results demonstrate that FASwinNet consistently outperforms existing mainstream methods on multiple remote sensing super-resolution datasets, especially excelling in edge preservation and detail recovery. This makes it suitable for a variety of real-world applications. In future work, we plan to further explore the joint modeling of frequency-domain and spatial transformations, and extend this framework to other low-level vision tasks such as image denoising and deblurring.

Author Contributions: Conceptualization, Z.W. and K.C.; methodology, K.C. and X.W.; validation, X.W.; formal analysis, S.L.; investigation, S.L.; resources, K.C.; data curation, X.W.; writing—original draft preparation, Z.W.; writing—review and editing, Z.W., S.L., K.C. and X,W.; visualization, Z.W.; supervision, S.L.; software, Z.W.,S.L. ; project administration, K.C.; funding acquisition, Z.W.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Foundation of Liaoning Educational Committee (Grant Nos. LJ212410153001, LJ212410153013) and the Liaoning Applied Basic Research Program (Grant No. 2025JH2/101300003).

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

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

 doi:10.20944/preprints202510.2414.v1

References

1. Yang, D.; Li, Z.; Xia, Y.; Chen, Z. Remote sensing image super-resolution: Challenges and approaches. In Proceedings of the 2015 IEEE international conference on digital signal processing (DSP). IEEE, 2015, pp. 196–200.
2. 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 Proceedings of the European conference on computer vision (ECCV) workshops, 2018, pp. 0–0.
3. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence* **2015**, *38*, 295–307.
4. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer, 2016, pp. 391–407.
5. Kong, X.; Zhao, H.; Qiao, Y.; Dong, C. Classsr: A general framework to accelerate super-resolution networks by data characteristic. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12016–12025.
6. 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 Proceedings of the European conference on computer vision (ECCV), 2018, pp. 286–301.
7. Li, Z.; Liu, Y.; Chen, X.; Cai, H.; Gu, J.; Qiao, Y.; Dong, C. Blueprint separable residual network for efficient image super-resolution. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 833–843.
8. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 136–144.
9. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2472–2481.
10. Ashish, V. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*, I.
11. 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 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
12. 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 Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
13. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems* **2021**, *34*, 12116–12128.
14. 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12299–12310.
15. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5728–5739.
16. Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; Li, H. Uformer: A general u-shaped transformer for image restoration. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 17683–17693.
17. Li, W.; Lu, X.; Qian, S.; Lu, J.; Zhang, X.; Jia, J. On efficient transformer-based image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175* **2021**.
18. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 1833–1844.
19. Chen, Y.; Fan, H.; Xu, B.; Yan, Z.; Kalantidis, Y.; Rohrbach, M.; Yan, S.; Feng, J. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 3435–3444.
20. Xu, W.; Wan, Y. ELA: Efficient Local Attention for Deep Convolutional Neural Networks, 2024, [arXiv:cs.CV/2403.01123].

 doi:10.20944/preprints202510.2414.v1

21. Yuan, K.; Guo, S.; Liu, Z.; Zhou, A.; Yu, F.; Wu, W. Incorporating convolution designs into visual transformers. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 579–588.
22. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in neural information processing systems* **2021**, *34*, 9355–9366.
23. Wu, S.; Wu, T.; Tan, H.; Guo, G. Pale transformer: A general vision transformer backbone with pale-shaped attention. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 2731–2739.
24. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 12124–12134.
25. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 22–31.
26. Huang, Z.; Ben, Y.; Luo, G.; Cheng, P.; Yu, G.; Fu, B. Shuffle transformer: Rethinking spatial shuffle for vision transformer. *arXiv preprint arXiv:2106.03650* **2021**.
27. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 568–578.
28. Li, K.; Wang, Y.; Zhang, J.; Gao, P.; Song, G.; Liu, Y.; Li, H.; Qiao, Y. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**, *45*, 12581–12600.
29. Li, Y.; Zhang, K.; Cao, J.; Timofte, R.; Van Gool, L. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707* **2021**.
30. Patel, K.; Bur, A.M.; Li, F.; Wang, G. Aggregating global features into local vision transformer. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR). IEEE, 2022, pp. 1141–1147.
31. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote sensing of Environment* **2020**, *241*, 111716.
32. 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 Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1874–1883.
33. Su, J.N.; Gan, M.; Chen, G.Y.; Guo, W.; Chen, C.P. High-similarity-pass attention for single image super-resolution. *IEEE Transactions on Image Processing* **2024**, *33*, 610–624.
34. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *55*, 3965–3981.
35. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13. Springer, 2014, pp. 184–199.
36. Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; Dong, C. Activating more pixels in image super-resolution transformer. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 22367–22377.
37. Peng, G.; Xie, M.; Fang, L. Context-aware lightweight remote-sensing image super-resolution network. *Frontiers in Neurorobotics* **2023**, *17*, 1220166.
38. Lei, S.; Shi, Z. Hybrid-scale self-similarity exploitation for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing* **2021**, *60*, 1–10.
39. Wang, J.; Wang, B.; Wang, X.; Zhao, Y.; Long, T. Hybrid attention-based U-shaped network for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing* **2023**, *61*, 1–15.
40. Wang, J.; Lu, Y.; Wang, S.; Wang, B.; Wang, X.; Long, T. Two-stage spatial-frequency joint learning for large-factor remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–13.
41. Kang, Y.; Wang, X.; Zhang, X.; Wang, S.; Jin, G. ACT-SR: Aggregation Connection Transformer for Remote Sensing Image Super-Resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2024**.

 doi:10.20944/preprints202510.2414.v1