sensors

MDPI

Article

MSFANet: A Multi-Scale Feature Fusion Transformer with Hybrid Attention for Remote Sensing Image Super-Resolution

Jie Yu [1], Chengcheng Lin [1], Luyao Peng [1], Cheng Zhong [1] and Hui Li [2,*]

[1] Badong National Observation and Research Station of Geohazards, China University of Geosciences, Wuhan 430074, China; jieyu@cug.edu.cn (J.Y.); linchengcheng@cug.edu.cn (C.L.); luyaopeng@cug.edu.cn (L.P.); zhonglxm@cug.edu.cn (C.Z.)
[2] The School of Earth Sciences, China University of Geosciences, Wuhan 430074, China
* Correspondence: rslihui@cug.edu.cn

Abstract

To address the issue of insufficient resolution in remote sensing images due to limitations in sensors and transmission, this paper proposes a multi-scale feature fusion model, MSFANet, based on the Swin Transformer architecture for remote sensing image super-resolution reconstruction. The model comprises three main modules: shallow feature extraction, deep feature extraction, and high-quality image reconstruction. The deep feature extraction module innovatively introduces three core components: Feature Refinement Augmentation (FRA), Local Structure Optimization (LSO), and Residual Fusion Network (RFN), which effectively extract and adaptively aggregate multi-scale information from local to global levels. Experiments conducted on three public remote sensing datasets (RSSCN7, AID, and WHU-RS19) demonstrate that MSFANet outperforms state-of-the-art models (including HSENet and TransENet) across five evaluation metrics in $\times 2$, $\times 3$, and $\times 4$ super-resolution tasks. Furthermore, MSFANet achieves superior reconstruction quality with reduced computational overhead, striking an optimal balance between efficiency and performance. This positions MSFANet as an effective solution for remote sensing image super-resolution applications.

Keywords: remote sensing; super-resolution reconstruction; Swin Transformer; deep learning; attention mechanism

check for updates

Academic Editor: Chiman Kwan

Received: 21 September 2025
Revised: 20 October 2025
Accepted: 1 November 2025
Published: 3 November 2025

Citation: Yu, J.; Lin, C.; Peng, L.; Zhong, C.; Li, H. MSFANet: A Multi-Scale Feature Fusion Transformer with Hybrid Attention for Remote Sensing Image Super-Resolution. *Sensors* **2025**, *25*, 6729. https://doi.org/10.3390/s25216729

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Remote sensing images are widely used in fields such as urban planning, disaster assessment, and agriculture [1–4]. Spatial resolution—the smallest distinguishable ground feature—directly determines image detail. High-resolution imagery enables the identification of fine-grained details such as buildings and roads. In contrast, insufficient spatial resolution significantly compromises application accuracy [5]. For example, imagery with 10–30 m resolution cannot distinguish between residential and commercial zones [6] nor effectively monitor crop growth stages [7]. Similarly, features such as collapsed structures [8] and subtle geological formations like minor fault lines remain undetectable in low-resolution imagery [9]. Nevertheless, pursuing higher resolution faces two fundamental constraints: sensor hardware advancements demand expensive microfabrication processes [10], while increased data volumes strain transmission and storage infrastructures [10,11]. In this context, super-resolution technology emerges as a cost-effective alternative, enhancing image resolution algorithmically without hardware investment, mitigating data burdens, and ultimately improving data utility for fine-grained applications [9–12].

Initially, people primarily relied on methods based on interpolation and transform domains to enhance image resolution [13–15]. Common interpolation techniques include bilinear interpolation and bicubic interpolation. These methods offer the advantage of low computational complexity, enabling rapid image processing with strong real-time capabilities [16]. However, their ability to restore images is constrained by the quality of the original low-resolution image [17]. These methods often struggle to preserve the true structure and high-frequency details of an image, resulting in insufficient clarity and detail richness. Transform-domain methods, such as those based on wavelet domains and sparse representations, can enhance image details to some extent but also present limitations. These approaches may introduce blurring and excessive smoothing during image processing, posing significant challenges for tasks requiring fine details, such as object recognition [18,19]. While using multiple images can enhance overall resolution and capture more details, this method also introduces a series of challenges. For instance, issues like image alignment problems, increased computational complexity, data redundancy, and inconsistencies arising from temporal variations can all adversely affect the quality of image reconstruction [20–23].

Owing to its capacity to learn complex patterns, restore fine details, adapt to complex degradation processes, and enhance image quality, deep learning has been gaining increasing popularity in the field of image super-resolution reconstruction. Deep learning can effectively reduce common artifacts, blurring, or blocky effects that are often encountered in traditional methods. Since around 2014, convolutional neural networks (CNNs) have been widely applied in the field of image super-resolution. For instance, SRCNN [24], as an early representative work, achieved super-resolution reconstruction by learning the mapping relationship between low-resolution and high-resolution images. Subsequently, numerous improved CNN architectures have been proposed, such as VDSR [25], SRResNet [26], CFSRCNN [27], ESRGCNN [28], and LESRCNN [29], all of which have achieved significant performance improvements. CNNs are highly adaptable to various input data, including different types of remote sensing images, sensor conditions, and degradation processes. In addition, generative adversarial networks (GANs) have also achieved good results in super-resolution reconstruction. For example, SRGAN [30], by introducing adversarial training, is capable of generating more realistic high-resolution images, which not only shows improvement in objective metrics but also appears more natural in visual effects. Since then, improved methods such as ESRGAN [31], GAN-CIRCLE [32], SOUP-GAN [33], and CAL-GAN [34] have continued to emerge, further optimizing the texture details and visual quality of images and providing new ideas and methods for the development of the field of image super-resolution.

In recent years, the remarkable performance of large Transformer models in the field of natural language processing has spurred significant advancements in their application to computer vision. TransENet [35] leverages Transformer modules to capture long-range dependencies and effectively uncover correlations between high-dimensional and low-dimensional features. By integrating Transformer modules with convolutional layers, TransENet is capable of capturing fine-grained and comprehensive contextual information. Additionally, ESRT [36] and ESSAformer [37] have introduced Transformer architectures that have significantly enhanced the performance of single-image super-resolution, particularly excelling in handling complex textures and details. These contributions have provided innovative perspectives for research on Transformer-based image super-resolution.

In the realm of image super-resolution, non-Transformer architectures have also garnered significant achievements. HSENet [38], with its distinctive hierarchical structure, adeptly captures both global and local features across multiple scales. It excels not only in detail preservation but also ensures contextual consistency. OmniSR [39], through its

innovative architecture and training strategy, achieves efficient image processing across diverse content types and degradation scenarios, thereby substantially enhancing the quality and detail of reconstructed images. HAUNet [40], via optimized network design and efficient feature extraction mechanisms, significantly boosts processing speed and resource efficiency in super-resolution while maintaining image quality. This renders it particularly suitable for real-time or resource-constrained applications. BSRAW [41] proposes a blind super-resolution method for RAW domain images. It designs realistic degradation pipelines specifically for training models to process raw sensor data. This approach effectively addresses common issues such as sensor noise, defocus, and exposure, thereby significantly enhancing the enlargement and quality of real-world RAW images. ASID [42], a lightweight super-resolution network, reduces computational overhead by refining information distillation and introducing cross-block attention sharing.

Recent research has shifted its focus from designing individual convolutional structures toward a synergistic integration of generative priors, efficient global dependency modeling, and realistic physical degradation simulation [43–46]. On one hand, diffusion-based generative methods have demonstrated significant advantages in reconstructing high-frequency textures and complex structural details through iterative denoising processes [47,48]. These approaches mitigate the over-smoothing issues commonly found in deterministic reconstruction models, creating new possibilities for generating optically realistic imagery with enhanced detail [49,50]. On the other hand, visual Transformer architectures—exemplified by Swin Transformer—have shown powerful capabilities in capturing long-range spatial dependencies and global contextual information inherent in remote sensing imagery by leveraging window-based self-attention mechanisms and hierarchical designs [51–54]. In parallel, advances in self-supervised and weakly supervised learning are further reducing the reliance on large-scale, accurately paired datasets [55,56].

In this study, we aim to further enhance the quality of reconstructed images while maintaining efficiency and reliability by developing a novel model that leverages the latest advancements in deep learning. Specifically, we propose a new model called MSFANet, which is based on the Swin Transformer architecture. This model integrates three key components: Feature Refinement Augmentation (FRA), Local Structure Optimization (LSO), and Residual Fusion Network (RFN). These components work together to effectively extract and adaptively aggregate multi-scale information ranging from local to global, thereby achieving superior reconstruction quality and efficiency. Our model has demonstrated remarkable performance in terms of reconstruction quality and efficiency across multiple benchmark datasets.

The remainder of this paper is organized as follows: Section 2 provides a detailed description of the proposed MSFANet framework. Section 3 presents the experimental datasets and results. Section 4 offers a comprehensive discussion and analysis of the findings. This paper concludes with Section 5, which summarizes the main conclusions and implications of this study.

2. Methods

Inspired by SwinIR [55] and SwinFR [57], this paper proposes the MSFANet model based on the Swin Transformer architecture, whose structure is shown in Figure 1.

MSFANet comprises three key components: a shallow feature extraction module, a deep feature extraction module, and a high-quality image reconstruction module. The shallow feature extraction module and HQ image reconstruction module adopt the proven configuration from SwinFR to ensure model stability and effectiveness in both foundational feature extraction and final image reconstruction stages. Within the deep feature extraction module, MSFANet introduces a Feature Swin Transformer block (FSTB). Building upon

the Swin Transformer layers rom SwinIR, this module incorporates our independently designed FRA (Figure 1a,d). This design significantly enhances the model's hierarchical feature extraction and fusion capabilities. By emphasizing key feature weights, it enables the model to accurately capture both local details and global contextual information in remote sensing images, providing robust feature representation for high-quality super-resolution reconstruction. We employed a 5-layer FSTB architecture (5 STLs per layer) based on experimental findings. Additionally, MSFANet incorporates an LSO (Figure 1c) module. This module simulates a global self-attention mechanism through local convolutional operations, enhancing feature representation capabilities while effectively reducing computational complexity. Concurrently, MSFANet employs an RFN module that integrates dense residual learning with channel attention mechanisms. This module efficiently extracts, refines, and aggregates multi-level features from remote sensing images, providing rich and high-quality feature support for super-resolution reconstruction.

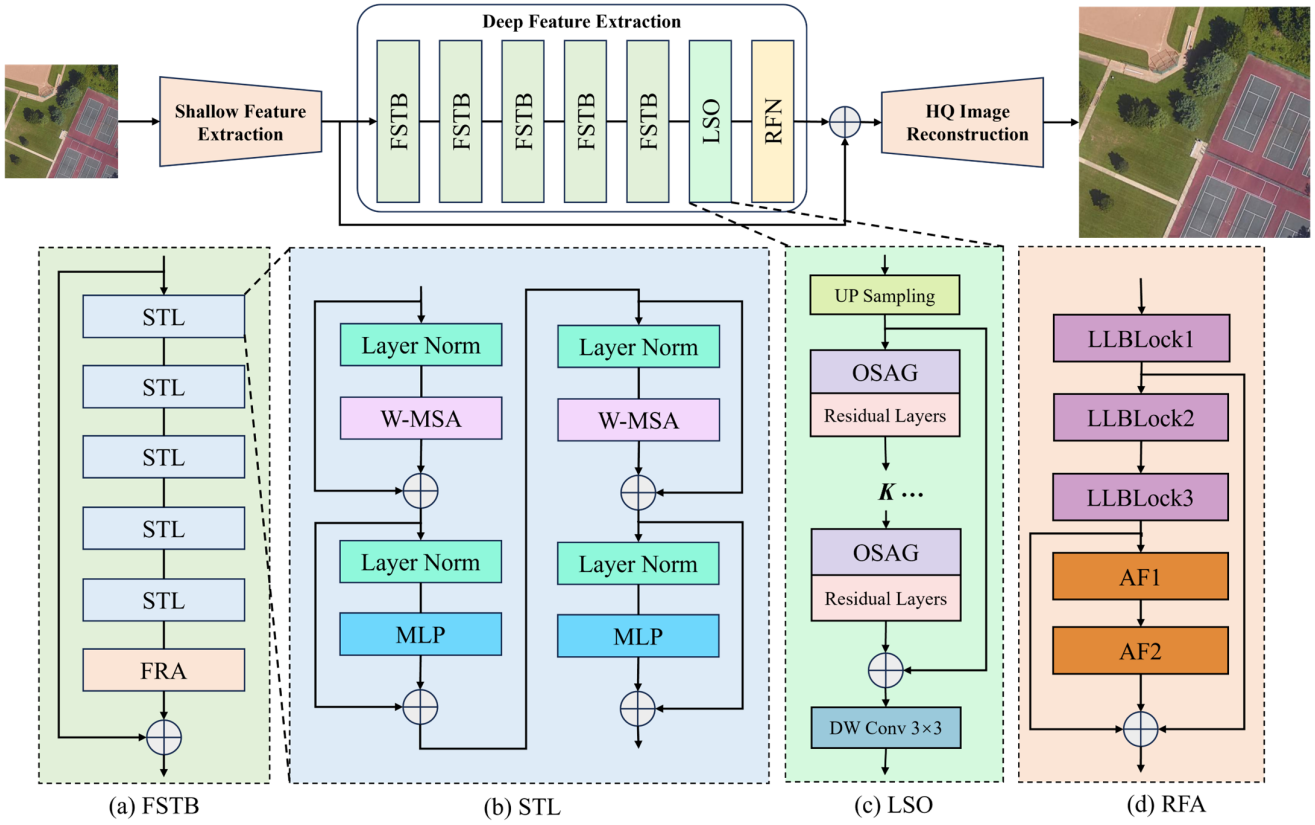

Figure 1. The network architecture of MSFANet. (**a**) FSTB: Feature Swin Transformer Block; (**b**) STL: Swin Transformer Layer; (**c**) LSO: Local Structure Optimization; (**d**) RFA: Feature Refinement Augmentation. In this and following figures, arrows indicate transmission medium and flow direction between components.

Remote sensing imagery exhibits multi-scale heterogeneity, where macroscopic scene structures (such as urban blocks, watershed basins) coexist with microscopic sparse details (such as individual buildings, narrow roads). Furthermore, the importance of different features (e.g., multispectral bands or deep semantic channels) varies significantly across channel dimension. Pure spatial attention mechanisms focus solely on pixel-level correlations and fail to capture the value differences among spectral channels—for instance, the unique value of the near-infrared band for vegetation analysis versus the role of visible light bands in structural identification. While pure self-attention mechanisms can capture global dependencies, their computational complexity increases quadratically with image

size, making them impractical for processing large-scale remote sensing data. To overcome these limitations, MSFANet introduces a tailored hybrid attention mechanism, with each component strategically designed to address specific bottlenecks in remote sensing image super-resolution.

2.1. Feature Refinement Augmentation

As depicted in Figure 2, the FRA is one of the core components of MSFANet. It is designed to enhance hierarchical feature extraction and fusion, thereby enabling the model to effectively capture both local details and global context in remote sensing images. The FRA module comprises three key structures: the Lightweight Lattice Block (LLBlock), the Channel Attention Layer (CALayer), and the Attention Fusion (AF) module.

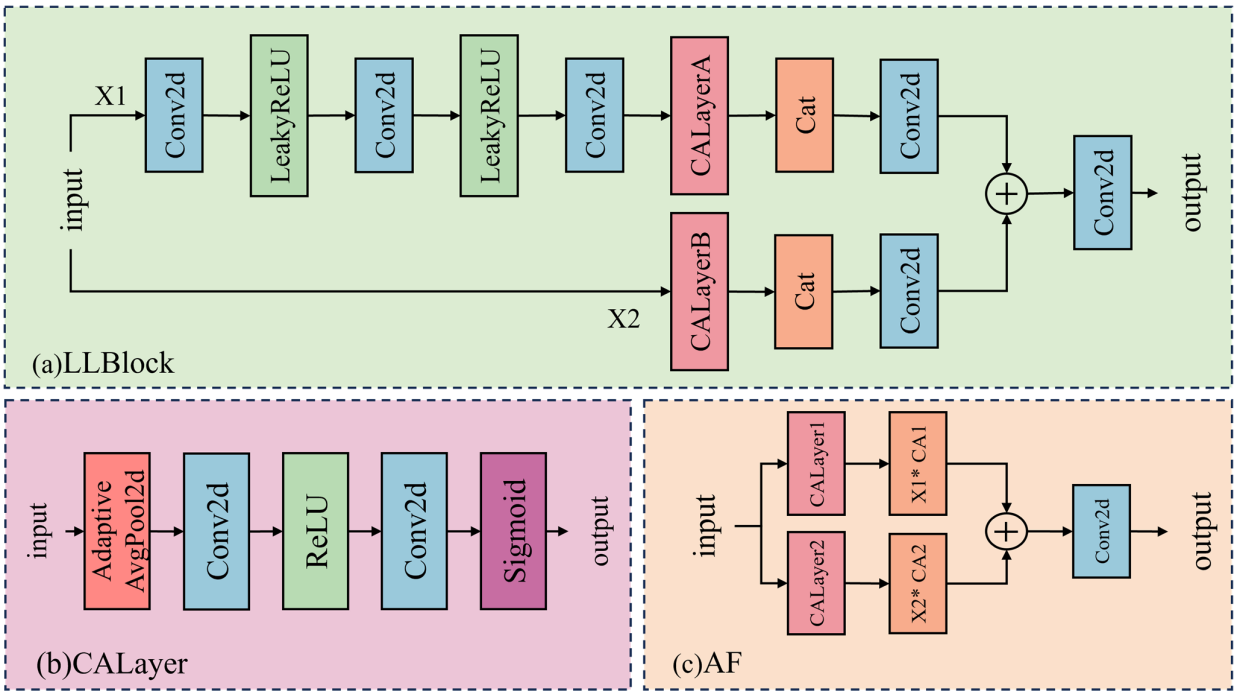

Figure 2. Detailed structure of the FRA block. (**a**) LLBLock: Lightweight Lattice Block; (**b**) CALayer: Channel Attention Layer; (**c**) AF: Attention Fusion module. In subfigure (**c**), the asterisk (*) symbolizes the element-wise multiplication operation performed between the feature tensor and its respective channel attention output.

These components work in concert to optimize feature representation. The LLBlock is inspired by lattice filter banks. Upon receiving an input tensor, it first decomposes it into two sub-tensors $x_1 \in R^{C_1 \times H \times W}$ and $x_2 \in R^{C_2 \times H \times W}$. Subsequently, each sub-tensor is processed through three layers of 3×3 convolutional (Conv2d) operations, followed by LeakyReLU activation functions. This sequence of operations refines the deep features of each sub-tensor. The LLBlock employs a divide-and-conquer strategy to perform targeted feature extraction and enhancement on distinct sub-tensors, thereby laying the groundwork for subsequent feature fusion. The CALayer (as shown in Figure 2b) is designed to enhance key features. For the outputs x_1' and x_2' from the LLBlock, the CALayer calculates their respective channel attention weights. Specifically, it first applies global average pooling to compress the spatial dimensions, reducing the width and height dimensions of the feature map to 1 while preserving the global information in the channel dimension. Subsequently, two layers of 1×1 convolutions perform channel-wise feature extraction and transformation on the compressed features. Finally, a Sigmoid activation function generates a weight map. This weight map is then multiplied element-wise with the original feature map,

thereby amplifying features in important channels while suppressing redundant information. This process achieves precise enhancement of key features. The AF (Figure 2c) module is responsible for multi-scale feature integration. For the initially split sub-tensors x_1 and x_2, the AF module applies the CALayer for weighted processing. This ensures efficient fusion of features across different levels. By doing so, the AF module not only preserves the richness of multi-scale features but also further optimizes feature quality through the channel attention mechanism. This high-quality feature representation provides robust support for super-resolution reconstruction tasks, thereby significantly enhancing the model's performance in remote sensing image super-resolution reconstruction.

2.2. Local Structure Optimization

As depicted in Figures 1c and 3a, the LSO module is a pivotal component within the MSFANet architecture, tasked with enhancing the representation of local features in remote sensing images while balancing the precision of feature learning with computational efficiency. Functionally, the LSO module initiates the process by employing specialized convolutional layers to adjust the channel dimension of the input feature maps. This adjustment aligns the feature channels with the preset network processing dimensions, establishing a standardized feature input foundation for subsequent local structure optimization. By doing so, it prevents feature information loss or redundant computation due to channel dimension mismatch, ensuring that subsequent operations can focus efficiently on feature content optimization rather than dimension adaptation.

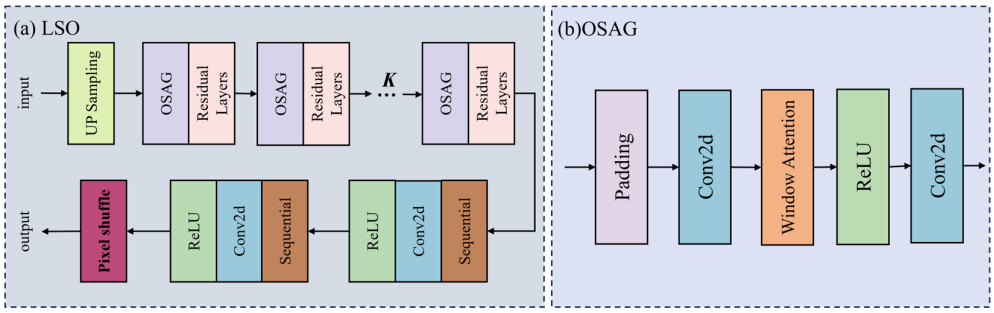

Figure 3. Detailed structure of the LSO block. (**a**) LSO: Local Structure Optimization; (**b**) OSAG: Overlapping Self-Attention Group.

At the core of the LSO module lies the Overlapping Self-Attention Group (OSAG) module (Figure 3b), which simulates the local feature correlation capture capability of global self-attention mechanisms. The OSAG module employs dynamic window partitioning and padding strategies, adaptively calculating the padding amount based on the spatial dimensions of the input feature maps. This approach ensures that edge regions of the feature maps are processed with the same window scale as the central regions, effectively avoiding the loss of edge details that can occur with traditional fixed-window processing. By stacking convolutions with non-linear activation operations, the OSAG module accurately models the dependencies between pixels within local windows, enhancing the learning of contour and detail features of fine targets in remote sensing images, such as small buildings, road edges, and vegetation textures. This capability compensates for the limitations of traditional convolutional networks in local fine feature extraction, providing richer local structural information support for super-resolution reconstruction. Moreover, the LSO module incorporates a residual connection mechanism that fuses the initially adjusted feature maps with those optimized by the OSAG module. This fusion retains the global structural information of the original features while integrating the refined local features, preventing feature degradation that can occur during deep network training. It enhances the stability of model training and ensures that the global scene structure is not

compromised by local optimizations, achieving a super-resolution effect where "global structure is coherent, and local details are clear".

Additionally, the LSO module utilizes a PixelShuffle layer to perform upsampling of the feature maps, mapping the optimized local features from the low-resolution space to the target high-resolution space. Precise dimension cropping ensures that the output feature maps match the spatial dimensions of the target resolution for the super-resolution task. This provides directly usable high-resolution local features for subsequent image reconstruction stages.

2.3. Residual Feature Network

Figure 4 illustrates the detailed structure of the Residual Fusion Network (RFN) block in MSFANet.

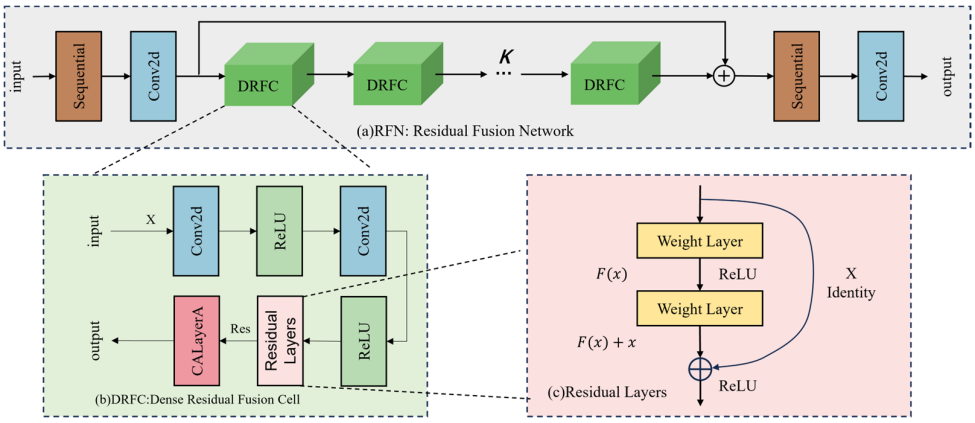

Figure 4. Detailed structure of the RFN block. (**a**) RFN: Residual Fusion Network; (**b**) DRFC: Dense Residual Fusion Cell; (**c**) Residual Layers.

As illustrated in Figure 4a, the RFN module is a central component within the MSFANet architecture, tasked with the efficient extraction, refinement, and aggregation of multi-level features from remote sensing images. Functionally, the RFN module begins by employing specialized convolutional layers to perform initial encoding of the input features. It transforms the diverse features originating from the FSTB and the LSO module into a unified dimensional initial feature representation. This standardized input foundation is crucial for subsequent deep feature learning, as it prevents information discontinuities or redundancies that might arise from feature source disparities, ensuring that multi-level features can be effectively integrated within the same processing framework.

The RFN module's innovative strength lies in the introduction of the Dense Residual Fusion Cell (DRFC), as shown in Figure 4b, and the dynamic configuration of its quantity. Each DRFC integrates convolutional operations, non-linear activation, and channel attention mechanisms. The convolutional layers are designed to capture feature details at various scales, such as architectural textures, vegetation distribution, and water body boundaries in remote sensing images. Non-linear activation functions enhance the model's ability to fit complex feature patterns, while the channel attention mechanism quantifies the importance of each feature channel, adaptively reinforcing channels that contain high-frequency details and structural information, and suppressing interference from noise and background channels. Furthermore, the DRFC employs a dense connection design, allowing each cell to directly receive the output features of all preceding cells. This not only maximizes the utilization of complementary information across different feature levels but also prevents the common issue of feature attenuation in deep networks. The residual propagation within the DRFC helps mitigate the vanishing gradient problem, enhancing

the stability and convergence efficiency of model training, thus providing ample deep feature support for super-resolution tasks. In the feature output phase, the RFN module uses a final convolutional layer to map the deep features processed by the DRFC to the target super-resolution output space, facilitating the transformation from abstract features to high-resolution image features. During this process, the aggregation effect of multi-level features is fully leveraged—shallow features provide basic texture and contour information, while deep features supplement fine structures and high-frequency details. The synergistic action of these features ensures that the super-resolution results maintain the structural coherence of the global scene while accurately restoring the details of local fine targets, such as small infrastructure and sparse vegetation in remote sensing images.

3. Experiment and Results

3.1. Dataset

In this study, the performance of the proposed method was evaluated using three publicly available remote sensing datasets: RSSCN7 [58], AID [59], and WHU-RS19 [60]. These datasets serve as critical benchmarks for training and evaluating deep learning models in various remote sensing applications, including classification, segmentation, scene recognition, object detection, and super-resolution tasks.

The RSSCN7 dataset, developed by the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, contains 2800 high-resolution images across seven classes: agriculture, commercial, forest, industrial, residential, river, and wasteland. Each class comprises 400 images (400 $\times$ 400 pixels), captured under varying seasonal, lighting, and platform conditions, ensuring high diversity and complexity. The AID dataset, created by Wuhan University, includes ~3000 high-resolution images spanning 30 land-use categories, such as airports, bridges, commercial areas, forests, and farmland. Each category contains 220–420 images (600 $\times$ 600 pixels), providing broad coverage of urban and natural scenes. The WHU-RS19 dataset, also from Wuhan University, consists of 950 high-resolution images divided into 19 categories, including airports, ports, bridges, industrial zones, forests, and sports fields. Each category has 50 images (600 $\times$ 600 pixels), sourced primarily from Google Earth, with diverse scenes and complex backgrounds.

All three datasets were partitioned into training, validation, and test sets using a consistent 70%/15%/15% ratio. Throughout this process, we strictly adhered to a category-balanced stratified random sampling strategy: samples were first grouped by category, followed by proportional random sampling within each group. This approach ensures that the category distributions across training, validation, and test sets remain fully consistent with the complete dataset, effectively mitigating experimental bias caused by data skew. Low-resolution (LR) images were synthesized from the original high-resolution (HR) images through bicubic interpolation downsampling. The downsampling factors strictly followed the specifications of the three super-resolution reconstruction tasks ($\times$2, $\times$3, and $\times$4). This synthesis strategy effectively mitigates confounding factors inherent in real remote sensing scenarios—such as sensor noise, atmospheric scattering, and spatio-temporal misalignment—enables strict control over experimental variables and helps validate the core innovations [38–55].

3.2. Performance Assessment

We compared our proposed method with nine SOTA models, including VDSR, SRDD [61], and OmniSR for natural images, as well as HSENet, TransENet, FENet [62], LGCNet [63], BSRAW and ASID which were specifically designed for RSIS tasks. All comparison methods were retrained using open-source code and tested under the same conditions. We generated low-resolution images from the high-resolution images in the

datasets using bicubic interpolation to train the models, while the original high-resolution images were used for model validation. We employed five metrics to comprehensively evaluate the quality of the reconstructed images, including the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Spatial Correlation Coefficient (SCC), Spectral Angle Mapper (SAM), and Universal Quality Index (UQI).

PSNR is a widely used objective metric to evaluate the quality of an image after it has undergone some form of transformation, such as compression, noise reduction, or fusion. It quantifies the level of distortion by comparing the original (reference) image to the reconstructed image. A higher PSNR value indicates better quality, as it signifies a closer match between the reconstructed image and the original.

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_i^2}{MSE}\right) \tag{1}$$

$$MSE = \frac{1}{N_W N_H} \sum_{i=0}^{N_W-1} \sum_{j=0}^{N_H-1} [Y(i,j) - X(i,j)]^2 \tag{2}$$

SSIM is a measure of the similarity of the two images, to evaluate image fidelity after transformations. Unlike PSNR, which focuses solely on pixel-level error, SSIM considers perceptual factors, making it more aligned with human visual perception.

$$SSIM(X,Y) = \frac{(2\mu_X\mu_Y + C_1)(\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)} \tag{3}$$

SCC focuses on the spatial correlation between images, making it sensitive to spatial relationships rather than just pixel intensity differences. It is particularly useful for assessing the alignment or similarity of spatial structures, which is important in image processing applications such as image fusion, registration, and super-resolution.

$$\mathrm{SCC} = \frac{\sum_{i,j}\left(I_{ij} - \overline{I}\right)\left(R_{ij} - \overline{R}\right)}{\sqrt{\sum_{i,j}\left(I_{ij} - \overline{I}\right)^2 \sum_{i,j}\left(R_{ij} - \overline{R}\right)^2}} \tag{4}$$

SAM is frequently used in multispectral and hyperspectral image analysis to measure the spectral similarity between each pixel in a reference image and the corresponding pixel in a reconstructed image. SAM is effective for assessing how well spectral features are preserved during processes such as image fusion, compression, and enhancement.

$$SAM(X,Y) = cos^{-1}\left(\frac{X \cdot Y}{\|X\|\|Y\|}\right) \tag{5}$$

UQI is an image quality assessment metric that measures the similarity between two images by considering structural, luminance, and contrast aspects. UQI focuses on structural similarity rather than just pixel-level differences.

$$UQI(X,Y) = \frac{4 \cdot \mu_X\mu_Y \cdot \sigma_{XY}}{(\mu_X^2 + \mu_Y^2)(\sigma_X^2 + \sigma_Y^2)} \tag{6}$$

where Y denotes the reconstructed image, X denotes the reference image, σ_X and σ_Y represent the standard deviation of X and Y, σ_{XY} represents the covariance in X and Y, while $\|X\|$ and $\|Y\|$ denote the magnitudes (or norms) of the spectral vectors X and Y, σ_{XY} denotes the covariance between X and Y.

3.3. Experimental Setting

In this research, we conduct a thorough assessment of the proposed model's performance across ×2, ×3, and ×4 super-resolution tasks, benchmarking it against a selection of cutting-edge (SOTA) efficient models. All comparison methods were retrained using open-source code and tested under the same conditions. The training batch size for all tasks is uniformly set to 32. The model was optimized using the Adam optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\varepsilon = 1 \times 10^8$. The initial learning rate was set to 1×10^4 and gradually reduced to 1×10^6 over 300 training epochs using a cosine annealing scheme (Cosine Annealing LR). The model was implemented in PyTorch 2.1.0 and all experiments were conducted on a workstation equipped with two NVIDIA GeForce GTX 3090 GPUs (NVIDIA Corporation, Santa Clara, CA, USA).

3.4. Comparison with Other Models

We evaluated the performance of our proposed model across ×2, ×3, and ×4 super-resolution tasks, comparing it with VDSR [25], SRDD [61], HSENet [38], TransENet [35], FENet [62], LGCNet [63], OmniSR [39], BSRAW [41], and ASID [42] on the RSSCN7 [58], AID [59], and WHU-RS19 [60] datasets. The results of these comparisons are detailed in Tables 1–3 and Tables S1–S3.

Table 1. The performance of the proposed model on the RSSCN7 dataset. The best results are highlighted in bold. The units for PSNR and SAM are dB and rad, respectively.

Scale	Metric	VDSR	SRDD	HSENet	TransENet	FENet	LGCNet	OmniSR	BSRAW	ASID	MSFANet
×2	**PSNR**	28.05	27.98	28.22	28.16	28.20	28.09	28.17	28.21	28.23	**28.29**
	SSIM	0.7583	0.7552	0.7655	0.7587	0.7656	0.7605	0.7637	0.7658	0.7667	**0.7683**
	SCC	0.2625	0.2529	0.2761	0.2700	0.2738	0.2632	0.2751	0.2760	0.2769	**0.2793**
	SAM	0.1038	0.1046	0.1018	0.1038	0.1028	0.1031	0.1013	0.1018	0.1015	**0.1004**
	UQI	0.9913	0.9912	0.9917	0.9914	0.9917	0.9914	0.9917	0.9917	0.9917	**0.9921**
×3	**PSNR**	26.13	26.31	26.30	26.23	26.35	26.21	26.30	26.31	26.38	**26.52**
	SSIM	0.6528	0.6634	0.6625	0.6622	0.6650	0.6558	0.6615	0.6655	0.6669	**0.6714**
	SCC	0.1341	0.1487	0.1491	0.1484	0.1511	0.1368	0.1463	0.1535	0.1539	**0.1552**
	SAM	0.1301	0.1276	0.1278	0.1302	0.1271	0.1290	0.1278	0.1276	0.1266	**0.1230**
	UQI	0.9865	0.9870	0.9871	0.9864	0.9871	0.9866	0.9871	0.9869	0.9872	**0.9883**
×4	**PSNR**	25.12	25.16	25.26	25.13	25.32	25.24	25.26	25.30	25.35	**25.48**
	SSIM	0.5912	0.5902	0.5988	0.5898	0.6028	0.5966	0.6029	0.6013	0.6040	**0.6093**
	SCC	0.0735	0.0717	0.0841	0.0669	0.0862	0.0782	0.0881	0.0846	0.0885	**0.0891**
	SAM	0.1461	0.1455	0.1440	0.1461	0.1429	0.1441	0.1438	0.1433	0.1425	**0.1403**
	UQI	0.9829	0.9831	0.9835	0.9829	0.9837	0.9834	0.9832	0.9836	0.9838	**0.9856**

The MSFANet model outperforms all SOTA models on the RSSCN7 dataset (Table 1). Specifically, under ×2 magnification processing, MSFANet achieves the highest PSNR (28.29 dB), SSIM (0.7683), and UQI (0.9921). In experiments on the RSSCN7 dataset, our proposed MSFANet demonstrated outstanding performance at ×2, ×3, and ×4 magnification levels. At ×2 magnification, MSFANet achieved a PSNR value of 28.29 dB, surpassing ASID (28.23 dB) and BSRAW (28.21 dB) by 0.21% and 0.28%, respectively. Simultaneously, its SSIM value reached 0.7683, surpassing ASID (0.21%) and BSRAW (0.33%). At ×3 magnification, MSFANet's PSNR was 26.52 dB, outperforming ASID (26.38 dB) and BSRAW (26.31 dB). Its SSIM value of 0.6714 was 0.68% and 1.49% higher than ASID and BSRAW, respectively. At ×4 magnification, MSFANet achieved a PSNR of 25.48 dB, surpassing ASID (25.35 dB) and BSRAW (25.30 dB). Its SSIM value reached 0.6093, representing improvements of 0.88% and 1.33% over ASID and BSRAW, respectively. Furthermore, MSFANet

consistently outperformed ASID and BSRAW on the SCC metric across all magnification levels. Specifically, at ×2, ×3, and ×4 magnification, MSFANet's SAM values improved by 1.09%, 2.84%, and 1.54% over ASID, and by 1.38%, 3.76%, and 2.10% over BSRAW. Compared to traditional algorithms like VDSR and SRDD, MSFANet demonstrates more significant improvements. At ×2 magnification, MSFANet achieves PSNR values 0.86% and 1.09% higher than VDSR and SRDD, respectively.

Table 2. The performance of the proposed model on the AID dataset. The best results are highlighted in bold. The units for PSNR and SAM are dB and rad, respectively.

Scale	Metric	VDSR	SRDD	HSENet	TransENet	FENet	LGCNet	OmniSR	BSRAW	ASID	MSFANet
×2	**PSNR**	34.16	35.17	34.88	35.24	35.19	35.03	35.21	35.47	35.56	**35.84**
	SSIM	0.9247	0.9354	0.9319	0.9344	0.9355	0.9340	0.9357	0.9364	0.9364	**0.9371**
	SCC	0.5941	0.6412	0.6250	0.6398	0.6404	0.6346	0.6418	0.0642	0.6432	**0.6451**
	SAM	0.0609	0.0548	0.0567	0.0573	0.0547	0.0557	0.0546	0.0542	0.0541	**0.0536**
	UQI	0.9959	0.9968	0.9966	0.9966	0.9968	0.9967	0.9968	0.9969	0.9969	**0.9971**
×3	**PSNR**	30.18	31.04	30.95	30.89	31.12	30.69	31.02	31.16	31.30	**31.37**
	SSIM	0.8311	0.8521	0.8492	0.8523	0.8542	0.8449	0.8517	0.8555	0.8581	**0.8594**
	SCC	0.3408	0.3939	0.3886	0.3789	0.3972	0.3738	0.3911	0.4013	0.4086	**0.4131**
	SAM	0.0942	0.0856	0.0865	0.0898	0.0848	0.0887	0.0859	0.0843	0.0830	**0.0813**
	UQI	0.9916	0.9929	0.9928	0.9924	0.9930	0.9923	0.9929	0.9930	0.9933	**0.9934**
×4	**PSNR**	28.13	28.89	28.74	28.41	28.94	28.89	28.89	28.94	29.05	**29.10**
	SSIM	0.7521	0.7779	0.7724	0.7603	0.7798	0.7788	0.7781	0.0787	0.7841	**0.7848**
	SCC	0.2086	0.2643	0.2517	0.2578	0.2665	0.2646	0.2628	0.2683	0.2739	**0.2745**
	SAM	0.1173	0.1080	0.1097	0.1103	0.1069	0.1077	0.1080	0.1072	0.1058	**0.1055**
	UQI	0.9873	0.9892	0.9890	0.9881	0.9894	0.9893	0.9893	0.9893	0.9896	**0.9898**

Table 3. The performance of the proposed model on the WHU-RS19 dataset. The best results are highlighted in bold. The units for PSNR and SAM are dB and rad, respectively.

Scale	Metric	VDSR	SRDD	HSENet	TransENet	FENet	LGCNet	OmniSR	BSRAW	ASID	MSFANet
×2	**PSNR**	30.36	30.87	30.03	30.91	30.95	30.58	30.92	30.67	30.97	**31.04**
	SSIM	0.8369	0.8488	0.8302	0.8514	0.8515	0.8421	0.8505	0.8440	0.8519	**0.8527**
	SCC	0.3279	0.3544	0.2934	0.3628	0.3610	0.3373	0.3578	0.3427	0.3614	**0.3657**
	SAM	0.0827	0.0786	0.0866	0.0781	0.0778	0.0806	0.0782	0.0804	0.0776	**0.0772**
	UQI	0.9891	0.9901	0.9887	0.9896	0.9893	0.9893	0.9902	0.9897	0.9904	**0.9905**
×3	**PSNR**	27.76	28.19	27.96	28.22	28.15	27.99	28.23	28.11	28.28	**28.34**
	SSIM	0.7400	0.7543	0.7445	0.7565	0.7530	0.7451	0.7541	0.7543	0.7565	**0.7590**
	SCC	0.1837	0.2077	0.1953	0.2150	0.2061	0.1937	0.2091	0.2093	0.2119	**0.2169**
	SAM	0.1127	0.1079	0.1106	0.1072	0.1080	0.1095	0.1075	0.1109	0.1065	**0.1062**
	UQI	0.9827	0.9840	0.9833	0.9834	0.9836	0.9833	0.9842	0.9832	0.9846	**0.9847**
×4	**PSNR**	25.93	26.59	26.51	26.63	26.47	26.46	26.62	26.69	26.70	**26.75**
	SSIM	0.6505	0.6812	0.6780	0.6860	0.6819	0.6770	0.6818	0.6855	0.6872	**0.6873**
	SCC	0.0943	0.1326	0.1296	0.1398	0.1392	0.1257	0.1341	0.1393	0.1398	**0.1416**
	SAM	0.1402	0.1302	0.1313	0.1291	0.1307	0.1308	0.1298	0.1285	0.1283	**0.1280**
	UQI	0.9751	0.9783	0.9779	0.9770	0.9779	0.9779	0.9784	0.9788	0.9791	**0.9793**

As illustrated in Table S1, the MSFANet model demonstrates exceptional performance across diverse categories within the RSSCN7 dataset for the ×4 upscaling task, consistently achieving the highest PSNR and SSIM values in the majority of these categories. Specifically, MSFANet attains a PSNR of 28.34 dB in the "Grasslands" category, 30.63 dB in the "Forests"

category, and 26.94 dB in the "Parking Lots" category. Additionally, it secures the top SSIM values with 0.6756 in "Grasslands," 0.6337 in "Forests," and 0.7362 in "Parking Lots". In the "Village" and "Factory" categories, MSFANet also shows strong competitive performance, recording PSNR values of 24.77 dB and 21.88 dB, and SSIM values of 0.4736 and 0.5787, respectively. On average, MSFANet achieves an impressive PSNR of 25.48 dB and an SSIM of 0.61, underscoring its superior capabilities in image super-resolution tasks when compared to other models.

Tables 2 and 3 showcase the superior performance of the MSFANet model across various metrics and scales on both the AID and WHU-RS19 datasets. The model consistently achieves the highest scores in PSNR, SSIM, SCC, SAM, and UQI, highlighting its exceptional capabilities in terms of image quality, structural coherence, and detail preservation. Tables S2 and S3 provide a detailed analysis of various scenes within the AID and WHU-RS19 datasets, further confirming that the MSFANet model delivers the most balanced performance. This well-rounded efficacy lays a solid foundation for remote sensing image super-resolution reconstruction.

Figure 5 demonstrates that the MSFANet algorithm shows outstanding performance in super-resolution reconstruction in various scenarios and surpasses other algorithms, especially in accurately capturing and restoring complex image details. In the river image (Figure 5a), MSFANet creates a clearer and sharper portrayal of the riverbank textures and contours, leading to a higher-resolution output. For the port scene (Figure 5b), the algorithm effectively distinguishes the port structure and its boundary with the surrounding water, presenting more intricate details. In the parking lot image (Figure 5c), MSFANet achieves more precise reconstruction of the vehicle shapes, parking space markings, and surrounding roads. Figures S1 and S2 also illustrate the excellent super-resolution effects of MSFANet.

3.5. Ablation Study

This ablation study systematically assesses the influence of various parameter configurations on model efficacy, with a focus on Memory, PSNR, and SSIM as principal metrics of evaluation. The analysis was executed across three distinct remote sensing datasets: RSSCN7, AID, and WHU-RS19, encompassing a total of eight experimental setups denoted as L0 through L7.

Table 4 summarizes the performance comparison of various model configurations on image processing tasks in terms of parameter count, memory consumption, and PSNR/SSIM metrics. The baseline model (L0) features a low parameter count (1.37 M) and memory footprint (5.26 MB), but delivers relatively poor PSNR and SSIM results across all three datasets. For instance, on the RSSCN7 dataset, it achieves a PSNR of 24.84 and an SSIM of 0.5738. As additional modules are incorporated, both performance and resource requirements evolve. The FRA (L1) and LSO (L2) modules introduce only marginal increases in parameters and memory, while providing limited performance gains. In contrast, the RFN module (L3) substantially increases the parameter count (8.93 M) and memory consumption (12.10 MB), accompanied by noticeable performance improvements. Combined configurations—FRA+LSO (L4), FRA+RFN (L5), and LSO+RFN (L6)—achieve varying trade-offs between performance and resource cost. The full model, MSFANet (L7), which integrates all three modules, attains the highest parameter count (11.98 M) and memory usage (16.3 MB), but also yields the best PSNR and SSIM results across all datasets: on RSSCN7, PSNR = 25.48, SSIM = 0.6093; on AID, PSNR = 29.1, SSIM = 0.7848; and on WHU-RS19, PSNR = 26.75, SSIM = 0.6873. Specifically, compared to the baseline, MSFANet achieves approximate improvements of 2.64 in PSNR and 0.0355 in SSIM on RSSCN7; 1.46 in PSNR and 0.054 in SSIM on AID; and 0.8 in PSNR and 0.0379 in SSIM on WHU-RS19. These results demonstrate that MSFANet significantly enhances reconstruction quality

through the integration of multiple modules, despite the increased computational cost. The fusion architecture effectively captures fine image details and structural information, enabling high-quality outputs and underscoring its strong suitability for complex image processing tasks.

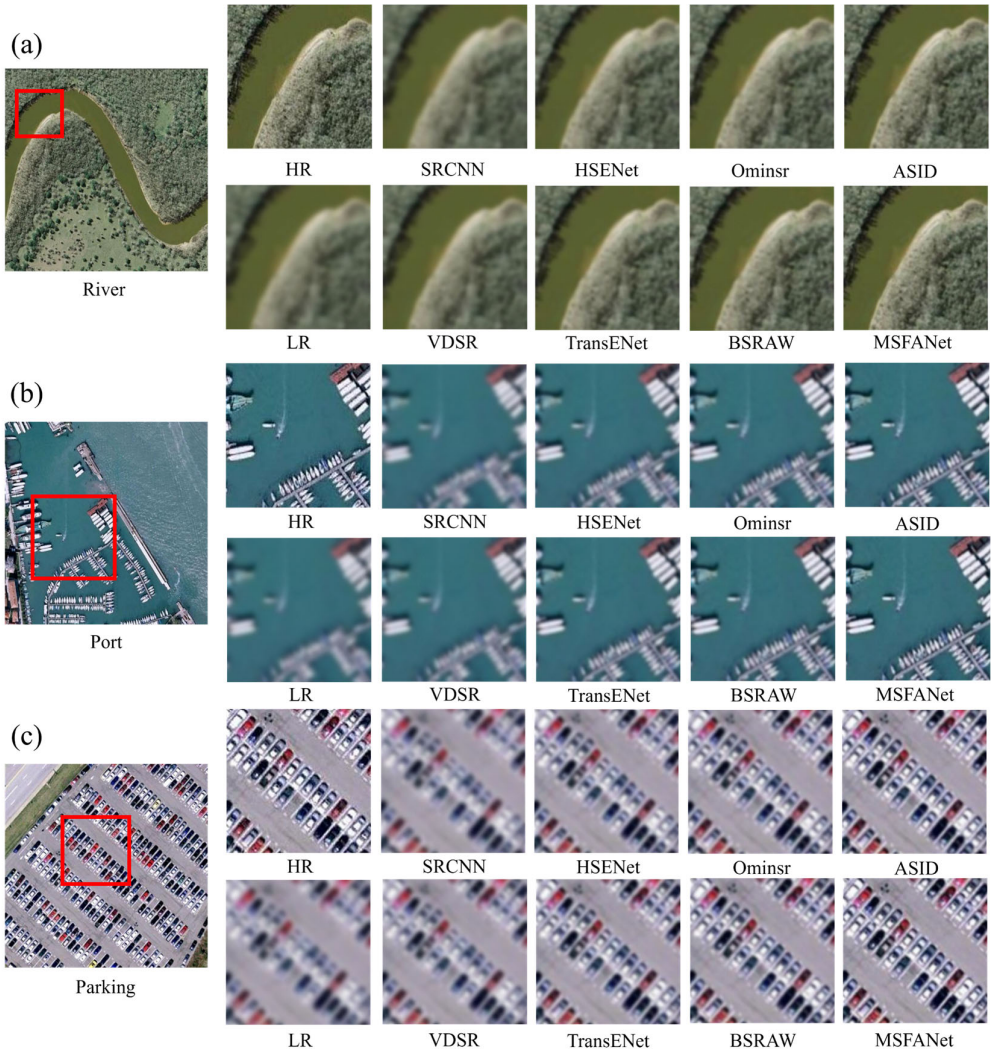

Figure 5. Comparative visualization on the RSSCN7 dataset. (**a**) River, (**b**) Port, (**c**) Parking lot. The red boxes highlight the focus regions for illustration.

In this study, MSFANet is developed based on the Swin Transformer architecture. To determine the optimal window size and attention head count for super-resolution tasks in remote sensing imagery, we systematically evaluated the impact of varying window sizes (2×2, 4×4, 8×8, and 16×16) and attention head counts (2 and 4) on model performance. MSFANet served as the baseline model, and experiments were conducted across two benchmark remote sensing datasets: AID and WHU-RS19. Multiple comparative experiments were performed, with PSNR and SSIM as the primary evaluation metrics (Table 5). The results demonstrate that both window size and attention head count significantly influence model performance, with the optimal configuration achieved at a window size of 8×8 and 4 attention heads. Additionally, the analysis revealed that increasing the number of attention heads generally improves PSNR and SSIM values but incurs higher memory consumption. Based on these findings, all subsequent experiments in this study adopted a window size of 8×8 and 4 attention heads for MSFANet to balance performance and computational efficiency.

Table 4. Ablation study of the proposed model at 4×factor. The best results are highlighted in bold.

Scenarios	Baseline+	Parameter (M)	Memory (MB)	RSSCN7		AID		WHU-RS19	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
L0	**——**	1.37	5.26	24.84	0.5738	27.64	0.7308	25.95	0.6494
L1	**FRA**	1.36	5.23	24.83	0.5761	27.74	0.7348	25.91	0.6483
L2	**LSO**	0.64	3.6	24.86	0.5761	27.71	0.7324	25.99	0.6515
L3	**RFN**	8.93	12.10	25.07	0.5887	27.63	0.7318	26.28	0.6683
L4	**FRA+LSO**	0.94	3.64	24.85	0.5761	27.69	0.7322	25.98	0.6512
L5	**FRA+RFN**	8.92	12.05	25.03	0.5859	27.64	0.7317	26.21	0.6642
L6	**LSO+RFN**	10.61	14.22	25.31	0.6017	28.82	0.7766	26.69	0.6858
L7	**MSFANet**	**11.98**	**16.3**	**25.48**	**0.6093**	**29.1**	**0.7848**	**26.75**	**0.6873**

Table 5. Performance of MSFANet with different window sizes and attention heads on the AID dataset (×4 super-resolution task). The best results are highlighted in bold.

Model	Window Size	Attention Heads	AID		WHU-RS19	
			PSNR	SSIM	PSNR	SSIM
MSFANet	**2 × 2**	**2**	28.87	0.7753	26.58	0.6821
		4	28.95	0.7786	26.62	0.6835
	4 × 4	**2**	29.03	0.7822	26.65	0.6848
		4	29.05	0.7832	26.70	0.6861
	8 × 8	**2**	29.02	0.7805	26.68	0.6853
		4	**29.10**	**0.7848**	**26.75**	**0.6873**
	16 × 16	**2**	29.08	0.7791	26.66	0.6842
		4	29.07	0.7818	26.69	0.6859

3.6. Hardware Resource Consumption Analysis

In addition to evaluating reconstruction quality, we conducted comparisons of hardware resource consumption across models (Figure 6). Compared to SOTA models, MSFANet achieves superior performance with lower memory usage (Figure 6a). Specifically, it reduces memory consumption by 56.6% versus ASID and 29.75% versus BSRAW, while simultaneously improving PSNR by 0.25 dB over BSRAW and 0.18 dB over ASID. When compared to lightweight models (e.g., SRCNN, VDSR, DCM), MSFANet maintains a slight memory overhead but delivers significant PSNR improvements (>0.75 dB). As evidenced by Figure 6b,c, MSFANet also demonstrates advantages in both parameter efficiency and computational speed, requiring fewer computational resources while achieving better super-resolution outcomes. These results collectively validate the model's exceptional efficiency-performance balance.

Unlike ASID's "Attention Distillation," which solely optimizes computational efficiency, MSFANet forms a closed-loop process of "selection-optimization-fusion" through FRA-enhanced multi-scale feature selection, LSO-reduced local computational costs, and RFN-mitigated feature decay. This achieves dual improvements in accuracy and efficiency. MSFANet achieves dual improvements in accuracy and efficiency by forming a closed-loop process of 'selection-optimization-fusion'. This is accomplished through FRA to enhance multi-scale feature selection, LSO to reduce local computational costs, and RFN to mitigate feature decay.

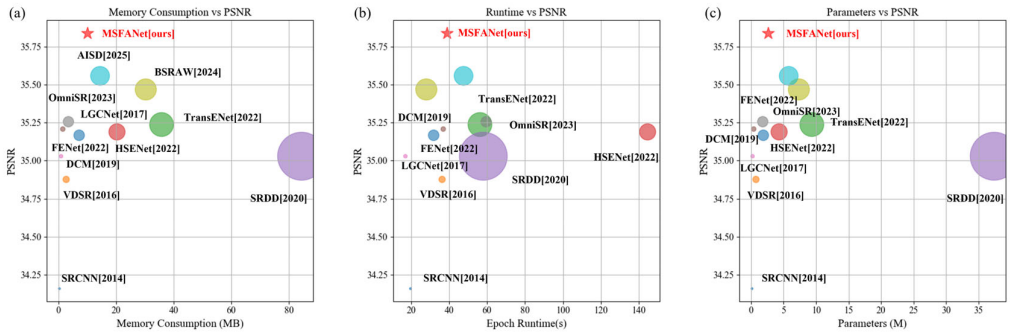

Figure 6. Computational resource analysis of models on AID dataset at 2× upscaling. (**a**) Memory consumption versus PSNR, (**b**) runtime versus PSNR, and (**c**) parameter count versus PSNR. Here, colorful circles denote the SOTA models, while the red star represents the proposed MSFANet model. Larger markers indicating higher consumption and smaller ones reflecting greater efficiency. Conventional CNN architectures (SRCNN, VDSR) are rendered in light blue, while Transformer-based approaches (TransENet, ASID) appear in medium blue.

4. Discussion

4.1. Advances in Other Remote Sensing Applications

While the current study focuses on remote sensing super-resolution improvement, recent advances in other domains are also driving the field toward greater intelligence and precision, thereby facilitating the in-depth exploration and application of geospatial knowledge.

In the domain of remote sensing image classification, cross-modal cross-attention Transformers, built upon the Transformer architecture, have significantly enhanced the performance of land cover classification and target recognition tasks by effectively integrating multimodal remote sensing data. Representative approaches such as the MFT [64] model propose an mCrossPA mechanism, which treats image patches from different modalities as queries and keys/values for cross-attention computation, enabling efficient multimodal fusion. The MCAITN [65] model introduces a multi-feature cross-attention induction mechanism, combining it with a Transformer backbone to fuse multi-scale features of hyperspectral images, demonstrating excellent performance in tasks like vegetation type identification and soil classification. Furthermore, models like CCFormer [66] and CM-Net [67] have validated the superiority of this mechanism on benchmark datasets such as Houston and Trento.

In the field of anomaly detection, memory-augmented autoencoders with adaptive reconstruction enhance the reconstruction of normal samples while suppressing the reconstruction of anomalous regions under unsupervised conditions [68]. This makes them suitable for tasks such as land cover anomaly identification and disaster monitoring in hyperspectral remote sensing images. The MAAE [69] model proposes superpixel-guided adaptive weight computation, sample affiliation mining, and entropy-based sparse addressing mechanisms, significantly improving anomaly detection capabilities in complex backgrounds. Meanwhile, the MAENet [70] model enhances the generalization and robustness of the model in noisy data, making it applicable to scenarios such as mineral exploration and disaster assessment.

For change detection tasks, dual-domain attention models effectively capture fine-grained change information in bi-temporal remote sensing images, such as building expansion and land cover type changes, by integrating spatial and frequency domain features. Representative methods include D^2Former [71], which employs a hybrid CNN-Transformer architecture with a U-Net-style encoder-decoder to extract multi-scale features; DDCD-Net [72], which uses a weight-shared Swin Transformer encoder for semantic feature extrac-

tion; and models like FTransDF-Net [73] and DML-UNet [74], which further demonstrate diverse implementations and application potential of this mechanism in change detection.

4.2. Potential Applications and Challenges

MSFANet's capability to reconstruct high-quality, high-resolution images from low-resolution inputs through algorithmic processing endows it with broad and profound application potential across numerous fields. It represents not merely a simple enhancement in image quality, but a key to unlocking more detailed and accurate geospatial information, thereby driving paradigm shifts in scientific research, commercial applications, and public services.

High-resolution foundational geospatial products form the cornerstone of national economic and social development. MSFANet can effectively compensate for the resolution deficiencies in historical archives or partially updated contemporary data. For large-scale mapping projects, acquiring sub-meter imagery across entire regions is both costly and time-consuming. By enhancing widely available medium-resolution imagery (such as Landsat and Sentinel series), MSFANet can generate base maps with enriched details, accelerating the production and updating of global-scale 1:5000 or larger-scale maps—particularly beneficial for remote areas lacking current surveying and mapping data.

In Earth observation applications, details often directly determine decision-making accuracy [1,2,5,6]. MSFANet enables the identification of small agricultural plots, irrigation channels, and even early signs of crop stress (such as initial pest and disease patches) that are indistinguishable in medium-resolution imagery. This supports more scientific decisions in variable-rate fertilization and precision irrigation, simultaneously increasing yields while reducing agrochemical use, thereby promoting sustainable agricultural development. The technology also aids in precise tree species distribution mapping, monitoring small-scale illegal deforestation, and assessing forest health conditions (such as canopy loss). In biodiversity conservation, it can even assist in identifying habitat ranges and behavioral patterns of large wildlife. Furthermore, it enables more accurate delineation of water body boundaries, monitoring subtle water volume changes in reservoirs and lakes, and identifying detailed distribution of algal blooms caused by eutrophication, providing early warnings for water resource management and pollution control [8–12].

During disasters and crises, high-resolution imagery translates into invaluable decision-making time and more effective response measures [1–6]. Following major natural disasters like floods, earthquakes, or wildfires, weather conditions or smoke often compromise imaging quality, resulting in suboptimal resolution. MSFANet can rapidly reconstruct clearer visualizations, helping rescue forces accurately locate damaged buildings, road disruptions, and temporary settlements of affected populations, significantly enhancing rescue efficiency. In national and public security domains, the super-resolution technology can enhance low-quality imagery from surveillance areas, assisting in identifying key targets, vehicle models, or vessel characteristics. In urban management, it also helps identify governance "blind spots" such as illegal constructions and unauthorized waste disposal sites.

In scientific research, by enhancing historical remote sensing data (such as Landsat imagery from decades ago), MSFANet enables the establishment of longer-term, more detailed records of glacier retreat, coastline changes, and polar ice melt, providing more accurate validation data for climate change models. It allows researchers to analyze urban morphology at finer scales—examining relationships between building density and energy consumption, correlations between neighborhood layouts and heat island effects, and impacts of green space distribution on resident health—thereby advancing theories in urban planning.

Although the technical aspects of the model have been thoroughly covered, MSFANet faces several challenges in practical deployment. The first challenge is computational resource constraints, particularly for deployment on edge devices. Although MSFANet reduces computational complexity through its hybrid attention mechanism, further optimization is still required when processing extremely large remote sensing images. Potential solutions include developing adaptive computation pathways that dynamically allocate computational resources based on image content complexity, or designing hierarchical super-resolution strategies that apply different reconstruction precision to regions of varying importance. The second challenge involves domain adaptation issues. In real-world application scenarios, distribution discrepancies between training and testing data can significantly impact model performance. This necessitates strong cross-domain generalization capabilities. Future work could explore strategies based on meta-learning or domain adaptation to enable MSFANet to quickly adapt to new remote sensing environments and data distributions. The third challenge lies in the super resolution model with other remote sensing processing tasks. Remote sensing analysis typically involves multi-task collaborative workflows, making the seamless integration of MSFANet into complete "super-resolution—classification—detection—change analysis" pipelines a critical issue to address. We plan to develop jointly optimized frameworks that enable end-to-end training of super-resolution with other tasks, thereby maximizing overall performance.

4.3. Future Works

Efficiency and lightweight design are pivotal for deploying models in practical scenarios, particularly on edge devices. Neural Architecture Search (NAS) can be employed to automatically discover optimal network architectures under specific computational constraints (e.g., latency and memory usage), replacing manual design to achieve an optimal balance between performance and efficiency. The implementation of "early-exit" mechanisms or conditional computation pathways enables the model to dynamically allocate computational resources based on the complexity of image regions (e.g., employing simplified processing for flat areas and more intensive computation for texture-rich regions). Furthermore, model pruning, quantization, and knowledge distillation methods tailored for super-resolution tasks should be explored. For instance, channel-level pruning strategies can eliminate redundant feature channels, while knowledge distillation allows a compact student model to approximate the performance of a larger teacher network.

Moving beyond conventional metrics such as PSNR and SSIM, future efforts should prioritize achieving visually realistic and physically plausible results. The integration of diffusion models can help capture the distribution of high-frequency details, generating richer and more authentic textures while mitigating issues like GAN mode collapse and training instability. This approach effectively alleviates the over-smoothing artifacts commonly associated with deterministic models. In the context of remote sensing imagery, it is crucial to preserve the spectral characteristics of ground objects during reconstruction. Future models should incorporate measures such as the Spectral Angle Mapper (SAM) as part of the loss function to ensure that reconstructed outputs are suitable for subsequent quantitative remote sensing analysis, rather than merely enhancing visual appearance. Additionally, leveraging large-scale pre-trained semantic segmentation networks (e.g., CLIP) to construct perceptual losses can ensure semantic consistency in reconstructed images and prevent structurally implausible object representations.

Real-world image degradation is far more complex than simple bicubic downsampling. Future research should shift its focus from non-blind super-resolution to blind super-resolution, developing models capable of handling unknown, complex, and potentially spatially varying degradation kernels, as well as real-world scenarios involving mixed

noise, blur, and compression artifacts. This necessitates the use of generative adversarial networks or physical models to simulate more realistic degradation processes for training. Reformulating the super-resolution problem as a maximum a posteriori (MAP) estimation task and incorporating plug-and-play frameworks that combine deep learning-based denoising priors with iterative optimization algorithms can further enhance model robustness and adaptability.

5. Conclusions

This study proposes MSFANet, a multi-scale feature fusion Transformer with hybrid attention, designed to address the insufficient spatial resolution of remote sensing images caused by sensor limitations, transmission constraints, and external interference. Extensive experiments on three public remote sensing datasets (RSSCN7, AID, and WHU-RS19) for $\times 2$, $\times 3$, and $\times 4$ super-resolution tasks demonstrate the superiority of MSFANet. It outperforms eight state-of-the-art models, including SRCNN, VDSR, ASID, and BSRAW, across five evaluation metrics. For example, on the RSSCN7 dataset at $\times 2$ scaling, MSFANet achieves a PSNR of 28.29 dB and an SSIM of 0.7683, surpassing ASID by 0.21% in both metrics. Ablation studies confirm that FRA, LSO, and RFN work synergistically to enhance performance, with the full MSFANet (L7) configuration yielding the best results. In terms of hardware efficiency, MSFANet reduces memory consumption by 56.6% compared to ASID and by 29.75% compared to BSRAW while maintaining higher reconstruction quality. Overall, MSFANet effectively balances efficiency and performance, offering a reliable solution for remote sensing image super-resolution and establishing a foundation for its application in domains such as urban planning and disaster assessment.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/s25216729/s1, Figure S1. Visual comparison of the proposed model and the SOTA methods on three test images from the AID dataset; Figure S2. Visual comparison of the proposed model and the SOTA methods on three test images from the WHU-RS19 dataset; Table S1. The performances on categories within the RSSCN7 dataset for the $\times 4$ upscaling; Table S2. The performances on categories within the AID dataset for the $\times 4$ upscaling; Table S3. The performances on categories within the WHU-RS19 dataset for the $\times 4$ upscaling.

Author Contributions: Conceptualization, Writing—original draft, J.Y.; Data curation, Formal analysis, C.L.; Methodology, Visualization, L.P.; Writing—review and editing, C.Z.; Project administration, Resources, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key R&D Program of Zhejiang (No. 2024C03236).

Data Availability Statement: The data and code are accessible at the following link: https://github.com/AlvinsaideYu/MSFANet (accessed on 19 September 2025).

Acknowledgments: We extend our sincere gratitude to the editor and the anonymous reviewers for their invaluable feedback and suggestions, which have significantly contributed to improving this paper. We also wish to thank our colleagues who participated in data processing and manuscript revision.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Rau, J.Y.; Jhan, J.P.; Hsu, Y.C. Analysis of Oblique Aerial Images for Land Cover and Point Cloud Classification in an Urban Environment. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1304–1319. [CrossRef]
2. Ghaffarian, S.; Kerle, N.; Filatova, T. Remote Sensing-Based Proxies for Urban Disaster Risk Management and Resilience: A Review. *Remote Sens.* **2018**, *10*, 1760. [CrossRef]
3. Shi, Z.; Zou, Z. Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image? *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3623–3634. [CrossRef]

4. Zhong, C.; Chen, J.; Yi, B.; Li, H. Examining the Reliable Trend of Global Urban Land Use Efficiency from 1985 to 2020 Using Robust Indicators and Analysis Tools. *Habitat Int.* **2025**, *163*, 103477. [CrossRef]
5. Wang, B.; Zhao, Y.; Li, X. Multiple Instance Graph Learning for Weakly Supervised Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5613112. [CrossRef]
6. Wang, P.; Zhang, H.; Zhou, F.; Jiang, Z. Unsupervised Remote Sensing Image Super-Resolution Using Cycle CNN. In Proceedings of the IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 3117–3120. [CrossRef]
7. Chen, C.; Wang, Y.; Zhang, N.; Zhang, Y.; Zhao, Z. A Review of Hyperspectral Image Super-Resolution Based on Deep Learning. *Remote Sens.* **2023**, *15*, 2853. [CrossRef]
8. Wang, X.; Yi, J.; Guo, J.; Song, Y.; Lyu, J.; Xu, J.; Yan, W.; Zhao, J.; Cai, Q.; Min, H. A Review of Image Super-Resolution Approaches Based on Deep Learning and Applications in Remote Sensing. *Remote Sens.* **2022**, *14*, 5423. [CrossRef]
9. Geng, T.; Liu, X.Y.; Wang, X.; Sun, G. Deep Shearlet Residual Learning Network for Single Image Super-Resolution. *IEEE Trans. Image Process.* **2021**, *30*, 4129–4142. [CrossRef] [PubMed]
10. Woodcock, C.E.; Strahler, A.H. The Factor of Scale in Remote Sensing. *Remote Sens. Environ.* **1987**, *21*, 311–332. [CrossRef]
11. Jia, S.; Wang, Z.; Li, Q.; Jia, X.; Xu, M. Multiattention Generative Adversarial Network for Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5624715. [CrossRef]
12. Sdraka, M.; Papoutsis, I.; Psomas, B.; Vlachos, K.; Ioannidis, K.; Karantzalos, K.; Gialampoukidis, I.; Vrochidis, S. Deep Learning for Downscaling Remote Sensing Images: Fusion and Super-Resolution. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 202–255. [CrossRef]
13. Xiao, C.; Chen, Y.; Sun, C.; You, L.; Li, R. AM-ESRGAN: Super-Resolution Reconstruction of Ancient Murals Based on Attention Mechanism and Multi-Level Residual Network. *Electronics* **2024**, *13*, 3142. [CrossRef]
14. Lehmann, T.M.; Gonner, C.; Spitzer, K. Survey: Interpolation Methods in Medical Image Processing. *IEEE Trans. Med. Imaging* **2002**, *18*, 1049–1075. [CrossRef]
15. Patel, V.; Mistree, K. A Review on Different Image Interpolation Techniques for Image Enhancement. *Int. J. Emerg. Technol. Adv. Eng.* **2013**, *3*, 129–133.
16. Upchurch, P.; Gardner, J.; Pleiss, G.; Pless, R.; Snavely, N.; Bala, K.; Weinberger, K. Deep Feature Interpolation for Image Content Changes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7064–7073. [CrossRef]
17. Zhu, Y.; Dai, Y.; Han, K.; Wang, J.; Hu, J. An Efficient Bicubic Interpolation Implementation for Real-Time Image Processing Using Hybrid Computing. *J. Real-Time Image Process.* **2022**, *19*, 1211–1223. [CrossRef]
18. Fei, B.; Lyu, Z.; Pan, L.; Zhang, J.; Yang, W.; Luo, T.; Zhang, B.; Dai, B. Generative Diffusion Prior for Unified Image Restoration and Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2023; pp. 9935–9946.
19. Sayed, M.; Brostow, G. Improved Handling of Motion Blur in Online Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1706–1716. [CrossRef]
20. Ashar, A.A.K.; Abrar, A.; Liu, J. A Survey on Object Detection and Recognition for Blurred and Low-Quality Images: Handling, Deblurring, and Reconstruction. In Proceedings of the 2024 8th International Conference on Information System and Data Mining, Chongqing, China, 22–24 March 2024; pp. 95–100. [CrossRef]
21. Ben Yedder, H.; Cardoen, B.; Hamarneh, G. Deep Learning for Biomedical Image Reconstruction: A Survey. *Artif. Intell. Rev.* **2021**, *54*, 215–251. [CrossRef]
22. Gothwal, R.; Tiwari, S.; Shivani, S. Computational Medical Image Reconstruction Techniques: A Comprehensive Review. *Arch. Comput. Methods Eng.* **2022**, *29*, 5635–5662. [CrossRef]
23. Matos, F.; Bernardino, J.; Durães, J.; Cunha, J. A Survey on Sensor Failures in Autonomous Vehicles: Challenges and Solutions. *Sensors* **2024**, *24*, 5108. [CrossRef]
24. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a Deep Convolutional Network for Image Super-Resolution. In *Computer Vision–ECCV 2014, Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Proceedings, Part IV; Springer: Cham, Switzerland, 2014; pp. 184–199. [CrossRef]
25. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654. [CrossRef]
26. Ullah, S.; Song, S.H. SRResNet Performance Enhancement Using Patch Inputs and Partial Convolution-Based Padding. *Comput. Mater. Contin.* **2023**, *74*, 2999–3014. [CrossRef]
27. Tian, C.; Xu, Y.; Zuo, W.; Zhang, B.; Fei, L.; Lin, C.W. Coarse-to-Fine CNN for Image Super-Resolution. *IEEE Trans. Multimed.* **2020**, *23*, 1489–1502. [CrossRef]
28. Tian, C.; Yuan, Y.; Zhang, S.; Lin, C.W.; Zuo, W.; Zhang, D. Image Super-Resolution with an Enhanced Group Convolutional Neural Network. *Neural Netw.* **2022**, *153*, 373–385. [CrossRef]

29. Tian, C.; Zhuge, R.; Wu, Z.; Xu, Y.; Zuo, W.; Chen, C.; Lin, C.W. Lightweight Image Super-Resolution with Enhanced CNN. *Knowl.-Based Syst.* **2020**, *205*, 106235. [CrossRef]
30. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690. [CrossRef]
31. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Loy, C.C. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018; pp. 63–79. [CrossRef]
32. You, C.; Li, G.; Zhang, Y.; Zhang, X.; Shan, H.; Li, M.; Ju, S.; Zhao, Z.; Zhang, Z.; Cong, W.; et al. CT Super-Resolution GAN Constrained by the Identical, Residual, and Cycle Learning Ensemble (GAN-CIRCLE). *IEEE Trans. Med. Imaging.* **2019**, *39*, 188–203. [CrossRef]
33. Zhang, K.; Hu, H.; Philbrick, K.; Conte, G.M.; Sobek, J.D.; Rouzrokh, P.; Erickson, B.J. SOUP-GAN: Super-Resolution MRI Using Generative Adversarial Networks. *Tomography* **2022**, *8*, 905–919. [CrossRef] [PubMed]
34. Park, J.K.; Son, S.; Lee, K.M. Content-Aware Local GAN for Photo-Realistic Super-Resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 10585–10594. [CrossRef]
35. Lei, S.; Shi, Z.; Mo, W. Transformer-Based Multistage Enhancement for Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5615611. [CrossRef]
36. Lu, Z.; Li, J.; Liu, H.; Huang, C.; Zhang, L.; Zeng, T. Transformer for Single Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 457–466. [CrossRef]
37. Zhang, M.; Zhang, C.; Zhang, Q.; Guo, J.; Gao, X.; Zhang, J. ESSAformer: Efficient Transformer for Hyperspectral Image Super-Resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 23073–23084. [CrossRef]
38. Lei, S.; Shi, Z. Hybrid-Scale Self-Similarity Exploitation for Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5401410. [CrossRef]
39. Wang, H.; Chen, X.; Ni, B.; Liu, Y.; Liu, J. Omni Aggregation Networks for Lightweight Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2023; pp. 22378–22387. [CrossRef]
40. Wang, J.; Wang, B.; Wang, X.; Zhao, Y.; Long, T. Hybrid Attention-Based U-Shaped Network for Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5612515. [CrossRef]
41. Conde, M.V.; Vasluianu, F.; Timofte, R. BSRAW: Improving Blind Raw Image Super-Resolution. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 8500–8510. [CrossRef]
42. Park, K.; Soh, J.W.; Cho, N.I. Efficient Attention-Sharing Information Distillation Transformer for Lightweight Single Image Super-Resolution. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 26 February–1 March 2025; Volume 39, pp. 6416–6424. [CrossRef]
43. Younesi, A.; Ansari, M.; Fazli, M.; Ejlali, A.; Shafique, M.; Henkel, J. A Comprehensive Survey of Convolutions in Deep Learning: Applications, Challenges, and Future Trends. *IEEE Access* **2024**, *12*, 41180–41218. [CrossRef]
44. Song, X.; Yan, L.; Liu, S.; Gao, T.; Han, L.; Jiang, X.; Jin, H.; Zhu, Y. Agricultural Image Processing: Challenges, Advances, and Future Trends. *Appl. Sci.* **2025**, *15*, 9206. [CrossRef]
45. Ferreira Rocha, P.R.; Fonseca Gonçalves, G.; dos Reis, G.; Guedes, R.M. Mechanisms of Component Degradation and Multi-Scale Strategies for Predicting Composite Durability: Present and Future Perspectives. Mechanisms of Component Degradation and Multi-Scale Strategies for Predicting Composite Durability: Present and Future Perspectives. *J. Compos. Sci.* **2024**, *8*, 204. [CrossRef]
46. Tamilarasan, S.; Wang, C.K.; Shih, Y.C.; Kuan, Y.D. Generative Adversarial Networks for Stack Voltage Degradation and RUL Estimation in PEMFCs under Static and Dynamic Loads. *Int. J. Hydrogen Energy* **2024**, *89*, 66–83. [CrossRef]
47. Tang, N.; Zhang, D.; Gao, J.; Qu, Y. FSRDiff: A Fast Diffusion-Based Super-Resolution Method Using GAN. *J. Vis. Commun. Image Represent.* **2024**, *101*, 104164. [CrossRef]
48. Wang, Y.; Yang, W.; Chen, X.; Wang, Y.; Guo, L.; Chau, L.P.; Liu, Z.; Qiao, Y.; Kot, A.C.; Wen, B. Sinsr: Diffusion-Based Image Super-Resolution in a Single Step. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 25796–25805. [CrossRef]
49. Moser, B.B.; Shanbhag, A.S.; Raue, F.; Frolov, S.; Palacio, S.; Dengel, A. Diffusion Models, Image Super-Resolution, and Everything: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 11793–11813. [CrossRef]
50. Wu, R.; Sun, L.; Ma, Z.; Zhang, L. One-Step Effective Diffusion Network for Real-World Image Super-Resolution. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 92529–92553. [CrossRef]

51. Dutta, D.; Chetia, D.; Sonowal, N.; Kalita, S.K. State-of-the-Art Transformer Models for Image Super-Resolution: Techniques, Challenges, and Applications. *arXiv* **2025**, arXiv:2501.07855. [CrossRef]
52. Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Lin, C.W.; Zhang, L. TTST: A Top-k Token Selective Transformer for Remote Sensing Image Super-Resolution. *IEEE Trans. Image Process.* **2024**, *33*, 738–752. [CrossRef]
53. Kang, X.; Duan, P.; Li, J.; Li, S. Efficient Swin Transformer for Remote Sensing Image Super-Resolution. *IEEE Trans. Image Process.* **2024**, *33*, 6367–6379. [CrossRef]
54. Zu, B.; Cao, T.; Li, Y.; Li, J.; Ju, F.; Wang, H. SwinT-SRNet: Swin Transformer with Image Super-Resolution Reconstruction Network for Pollen Images Classification. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108041. [CrossRef]
55. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Gool, L.V.; Timofte, R. SwinIR: Image Restoration Using Swin Transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 1833–1844. [CrossRef]
56. Xu, P.; Xiang, Z.; Fu, J.; Pu, T.; Wang, K.; Ji, C.; Bai, T.; Liu, E. Self-Supervised Stereo Matching with Multi-Baseline Contrastive Learning. *arXiv* **2025**, arXiv:2508.10838. [CrossRef]
57. Zhang, J.; Tu, Y. SwinFR: Combining SwinIR and Fast Fourier for Super-Resolution Reconstruction of Remote Sensing Images. *Digit. Signal Process.* **2025**, *159*, 105026. [CrossRef]
58. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep Learning Based Feature Selection for Remote Sensing Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325. [CrossRef]
59. 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 Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
60. Dai, D.; Yang, W. Satellite Image Classification via Two-Layer Sparse Coding with Biased Image Representation. *IEEE Geosci. Remote Sens. Lett.* **2010**, *8*, 173–176. [CrossRef]
61. Maeda, S. Image Super-Resolution with Deep Dictionary. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 464–480. [CrossRef]
62. Haut, J.M.; Paoletti, M.E.; Fernandez-Beltran, R.; Plaza, J.; Plaza, A.; Li, J. Remote Sensing Single-Image Superresolution Based on a Deep Compendium Model. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1432–1436. [CrossRef]
63. Lei, S.; Shi, Z.; Zou, Z. Super-Resolution for Remote Sensing Images via Local–Global Combined Network. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1243–1247. [CrossRef]
64. Roy, S.K.; Deria, A.; Hong, D.; Rasti, B.; Plaza, A.; Chanussot, J. Multimodal Fusion Transformer for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5515620. [CrossRef]
65. Li, Z.; Liu, R.; Sun, L.; Zheng, Y. Multi-Feature Cross Attention-Induced Transformer Network for Hyperspectral and LiDAR Data Classification. *Remote Sens.* **2024**, *16*, 2775. [CrossRef]
66. Guo, H.; Tian, B.; Liu, W. CCFormer: Cross-Modal Cross-Attention Transformer for Classification of Hyperspectral and LiDAR Data. *Sensors* **2025**, *25*, 5698. [CrossRef]
67. Wang, H.; Li, Z.; Wu, L. Cross—Transformer Fusion Network for Multimodal Remote Sensing Image Classification. *Photogramm. Rec.* **2025**, *40*, e70014. [CrossRef]
68. Aguila, A.L.; Liu, P.; Puonti, O.; Iglesias, J.E. Conditional Diffusion Models for Guided Anomaly Detection in Brain Images Using Fluid-Driven Anomaly Randomization. *arXiv* **2025**, arXiv:2506.10233. [CrossRef]
69. Huo, Y.; Cheng, X.; Lin, S.; Zhang, M.; Wang, H. Memory-Augmented Autoencoder with Adaptive Reconstruction and Sample Attribution Mining for Hyperspectral Anomaly Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5931010. [CrossRef]
70. Zhao, Z.; Sun, B. Hyperspectral Anomaly Detection via Memory-Augmented Autoencoders. *CAAI Trans. Intell. Technol.* **2023**, *8*, 1274–1287. [CrossRef]
71. Zheng, H.; Liu, H.; Lu, L.; Li, S.; Lin, J. D2Former: Dual-Domain Transformer for Change Detection in VHR Remote Sensing Images. *Electronics* **2024**, *13*, 2204. [CrossRef]
72. Xu, Z.; Zhang, C.; Qi, J.; Li, X.; Yao, B.; Wang, L. A Dual-Difference Change Detection Network for Detecting Building Changes on High-Resolution Remote Sensing Images. *Geocarto Int.* **2024**, *39*, 2322080. [CrossRef]
73. Li, Z.; Zhang, Z.; Li, M.; Zhang, L.; Peng, X.; He, R.; Shi, L. Dual Fine-Grained Network with Frequency Transformer for Change Detection on Remote Sensing Images. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *136*, 104393. [CrossRef]
74. Zhan, M.J.; Xie, X.Y.; Wang, H.; Guo, Z.Y. Change Detection of Remote Sensing Building Images Based on Dual-Domain Strip Attention DML-UNet Network. *J. Supercomput.* **2025**, *81*, 854. [CrossRef]

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