scientific reports

Check for updates

OPEN

Remote sensing image Super-resolution reconstruction by fusing multi-scale receptive fields and hybrid transformer

Denghui Liu, Lin Zhong✉, Haiyang Wu, Songyang Li & Yida Li

To enhance high-frequency perceptual information and texture details in remote sensing images and address the challenges of super-resolution reconstruction algorithms during training, particularly the issue of missing details, this paper proposes an improved remote sensing image super-resolution reconstruction model. The generator network of the model employs multi-scale convolutional kernels to extract image features and utilizes a multi-head self-attention mechanism to dynamically fuse these features, significantly improving the ability to capture both fine details and global information in remote sensing images. Additionally, the model introduces a multi-stage Hybrid Transformer structure, which processes features at different resolutions progressively, from low resolution to high resolution, substantially enhancing reconstruction quality and detail recovery. The discriminator combines multi-scale convolution, global Transformer, and hierarchical feature discriminators, providing a comprehensive and refined evaluation of image quality. Finally, the model incorporates a Charbonnier loss function and total variation (TV) loss function, which significantly improve training stability and accelerate convergence. Experimental results demonstrate that the proposed method, compared to the SRGAN algorithm, achieves average improvements of approximately 3.61 dB in Peak Signal-to-Noise Ratio (PSNR), 0.070 (8.2%) in Structural Similarity Index (SSIM), and 0.030 (3.1%) in Feature Similarity Index (FSIM) across multiple datasets, showing significant performance gains.

Keywords Remote sensing image, Image Super-resolution, GAN, Attention mechanism, Hybrid transformer, Multi-scale feature extraction, Comprehensive Discriminator

Remote sensing images contain rich details and perceptual information, which effectively support scene understanding and environmental analysis. In the field of remote sensing, image super-resolution (SR) reconstruction technology is particularly important and is widely applied in tasks such as environmental monitoring, object detection, and scene classification. In recent years, with the rapid development of blind super-resolution techniques, traditional bilinear downsampling models have gradually been replaced by more complex deep learning degradation models. Super-resolution networks based on Generative Adversarial Networks (GAN)[1]have shown excellent performance in image restoration. Since the introduction of the first attention-based super-resolution model, RCAN[2], in 2018, the application of attention mechanisms in super-resolution reconstruction has garnered widespread attention. However, due to the limitations of remote sensing image sensors and the effects of atmospheric disturbances, long-distance imaging, and spectral noise[3], traditional downsampling models struggle to accurately simulate the real degradation process, resulting in distortion in the reconstructed remote sensing images.

With the rapid advancement of attention mechanisms, their computational capabilities have significantly increased[4]. However, this also brings increased model complexity and computational burden, limiting the practicality of attention-based super-resolution models in real-world applications. Moreover, although GAN networks generate realistic images through learning strategies that make human perception more convincing, they still tend to produce noticeable artifacts when compensating for high-frequency details (such as image edges) and subtle feature representations[5]. For instance, in remote sensing images, detail edges may become blurred or unnaturally sharpened, making artifacts more easily noticeable to the human eye.

To address the challenges of insufficient detail extraction in low-resolution images, the ineffectiveness of degradation models, and the underperformance of reconstructed feature representations, it is necessary to

School of Electronics and Information, Xijing University, Xi'an 710123, China. ✉email: zhong_chen2@163.com

strike a balance between model complexity and real-world applicability[6]. Additionally, adopting more effective degradation models to better simulate the actual degradation process of remote sensing images is essential. In response to the challenges of high-frequency detail reconstruction in current remote sensing images, this paper proposes an improved remote sensing super-resolution (SR) reconstruction model with the following specific improvements:

1. **Multi-Scale Feature Extraction and Dynamic Feature Fusion**: Image features are extracted using multi-scale convolutional kernels (7×7, 5×5, 3×3) and dynamically fused using a multi-head self-attention mechanism. This enhances the ability to capture and represent both fine details and global information in remote sensing images, contributing to the generation of clearer and higher-quality high-resolution images.
2. **Multi-Stage Hybrid Transformer Structure**: After feature fusion, the image is processed through three custom Transformer modules. Each Transformer includes linear layers, positional encoding, and layer normalization, and uses self-attention mechanisms to process features at different resolutions. The first stage handles low-resolution features, the second stage refines them to medium resolution, and the third stage elevates them to high resolution, progressively improving image quality and detail.
3. **Comprehensive Discriminator for Multi-Dimensional Evaluation**: The comprehensive discriminator integrates multi-scale convolution, global Transformer, and hierarchical feature discriminators to provide a thorough evaluation of image quality. By combining the strengths of different discriminators, the model enhances the quality and realism of the generated images from multiple perspectives.

Related work

Degradation model

Current image super-resolution (SR) methods typically rely on traditional bicubic downsampling[7]and conventional degradation models[8,9], or their variants[10,11]. These models usually simulate the degradation process of images through blurring, downsampling, and noise addition. However, in real-world scenarios, noise may not only originate from the image itself but also be introduced by camera sensors or JPEG compression, exhibiting signal randomness and non-uniformity[12]. Even if the blurring part is accurately simulated, a mismatch between the noise and the actual image can significantly reduce the effectiveness of super-resolution reconstruction. Therefore, models like BSRGAN[13]and Real-ESRGAN[14], which are closer to the real degradation scenarios, are particularly important.

These advanced degradation models focus on three core factors: the blur kernel k[15], the downsampling kernel s, and the noise n. By randomly permuting the order of these factors (e.g., ksn, nks, snk) and combining different implementation methods (such as bicubic downsampling, nearest-neighbor, bilinear downsampling, etc.), a more complex degradation process in real scenes can be better simulated. Specifically, these factors are sequentially applied in a random order during the degradation process and then stacked once again in the final step. This also highlights the importance of JPEG noise in the degradation model, as it can further realistically reproduce the deterioration of image quality in the last step.

Natural image super-resolution

With the widespread application of deep learning in super-resolution (SR) image reconstruction tasks, performance has shown remarkable improvement, leading to the development of numerous SR reconstruction methods for natural images in recent years. Dong et al.[16]. proposed the Super-Resolution Convolutional Neural Network (SRCNN) model, which applies a three-layer convolutional neural network (CNN) to reconstruction tasks. This model uses CNNs to learn end-to-end feature mappings between low-resolution (LR) and high-resolution (HR) images, significantly reducing computational complexity compared to traditional methods. Subsequently, Dong et al.[17]. introduced the FSRCNN model based on SRCNN, increasing the network depth to effectively reconstruct more high-frequency details, though this also increased the difficulty of network training. Kim et al.[18]. proposed the Deeply-Recursive Convolutional Network for Super-Resolution (VDSR) model, which achieves multi-level feature cascading through the connection of multiple CNN layers, thereby enhancing the learning rate and accelerating convergence, effectively demonstrating the importance of network depth in SR reconstruction. Ledig et al.[19]. introduced the Super-Resolution Generative Adversarial Network (SRGAN) model, incorporating Generative Adversarial Networks (GAN)[20] into SR reconstruction tasks. In SRGAN, the generator and discriminator are trained collaboratively, making the network more attentive to the similarity in the feature space distribution and motivating the generation of natural images with high perceptual quality. Building on this foundation, Wang et al.[21]. proposed the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN), which optimizes the SRGAN framework by employing Residual Dense Blocks (RDB)[22] to reduce artifacts in the reconstructed results and sharpen edge textures.

Remote sensing image super-resolution

The structure of remote sensing images is more complex than that of natural images. Remote sensing images typically encompass a wide range of different scenes, such as buildings, farmlands, forests, and airports. A complete remote sensing image might consist of various scenes with significantly different textures and structural information, leading to inconsistent mapping relationships between high-resolution (HR) and low-resolution (LR) images across different scenes[23]. Additionally, the scale of objects in remote sensing images varies greatly. For example, objects like airplanes and vehicles may occupy only a few pixels in a remote sensing image, which is a stark contrast to natural images.

In the field of remote sensing, super-resolution (SR) is a severely ill-posed problem, and image quality is influenced by numerous factors such as atmospheric disturbance, ultra-long-range imaging, and spectral noise. To address these challenges, researchers have proposed various innovative models. Zhang et al.[24]. introduced

the Multi-Scale Attention Network (MSAN) model, incorporating scale attention networks to enhance scene adaptability and improve detail reconstruction in diverse remote sensing scenarios. Dong et al.[25]. proposed the Dense Sampling Super-Resolution (DSSR) model, utilizing a dense sampling mechanism to upsample multiple low-dimensional features, allowing the network to integrate multiple prior features during reconstruction. Subsequently, Dong et al.[26]. introduced the Second-Order Multi-Scale Super-Resolution (SMSR) model, leveraging a two-stage learning process to aggregate global and local large-scale and small-scale feature information, thereby strengthening the capability of multi-scale feature extraction.

For reconstruction methods based on Generative Adversarial Networks (GANs), Jiang et al.[27]. proposed the Edge-Enhanced GAN (EEGAN) model, which employs a super-dense sub-network and an edge enhancement network to improve SRGAN, making it suitable for remote sensing reconstruction tasks and enhancing the edge reconstruction capability of remote sensing images. Lei et al.[28]. proposed the Coupled Discriminator GAN (CDGAN) model, which adopts a coupled discriminator network structure to enhance the local details of the reconstructed images. Lin et al.[29]. introduced channel attention to achieve high-frequency focus on local contours and designed an edge loss to constrain the training process, ensuring that the edge details of the generated images remain more complete. Sui et al.[30]. integrated an additional Noise Discriminator (ND), employing an adversarial learning strategy in data distribution learning to enhance the diversity of generated data and the detailed texture prediction of the diffusion model. These methods address some critical issues in remote sensing image super-resolution reconstruction to varying degrees, thereby improving the reconstruction quality of remote sensing images in complex scenes.

Methods

To address the challenges in reconstructing high-frequency details in remote sensing images, this paper proposes a GAN-based improved remote sensing super-resolution (SR) reconstruction model. The model mainly consists of a generator and a discriminator. The generator network produces SR images at a specified upscaling factor, which are then inputted into the discriminator network along with the HR images. The discriminator assesses the authenticity of the images, determining whether the input is real or generated.

Generator model building

The generator network is composed of three main modules: a multi-scale feature extraction module, a dynamic feature fusion module, and a multi-stage Hybrid Transformer, as shown in Fig. 1. These modules work in synergy, enabling the generator to progressively extract, fuse, and refine features from the low-resolution (LR) image, ultimately generating a high-quality high-resolution (HR) image.

1. **Multi-Scale Feature Extraction Module**: Similar to traditional multi-scale feature extraction methods, this model employs convolutional kernels of different scales (7×7, 5×5, 3×3) to extract multi-level features from the low-resolution image. This approach effectively captures various details and semantic features within the image, enhancing the network's perceptual ability. The input consists of images of size $x \in R^{N*3*H*W}$, where N represents the batch size, 3 is the number of input channels (RGB channels), and H and W denote the height and width of the image, respectively. In this module, the input image is processed through multiple convolutional kernels to extract multi-scale features. These features are then concatenated along the channel dimension, followed by a 1×1 convolution for channel fusion, resulting in high-dimensional multi-scale features $f_{ms} = Conv_{1*1} \in R^{N*C*H*W}$. The concatenated features merge local and global information across the channel dimension, thereby providing a richer input for subsequent processing.
2. **Dynamic Feature Fusion Module**: This module primarily utilizes a multi-head self-attention mechanism[31] to dynamically fuse multi-scale features. First, the input feature f_{ms} is reshaped into a sequential form for use in the subsequent self-attention mechanism. The multi-head self-attention mechanism reweights and fuses the features. For each position i, its output is computed as a weighted sum of queries Q, keys K, and values V. The calculation formula is as follows:

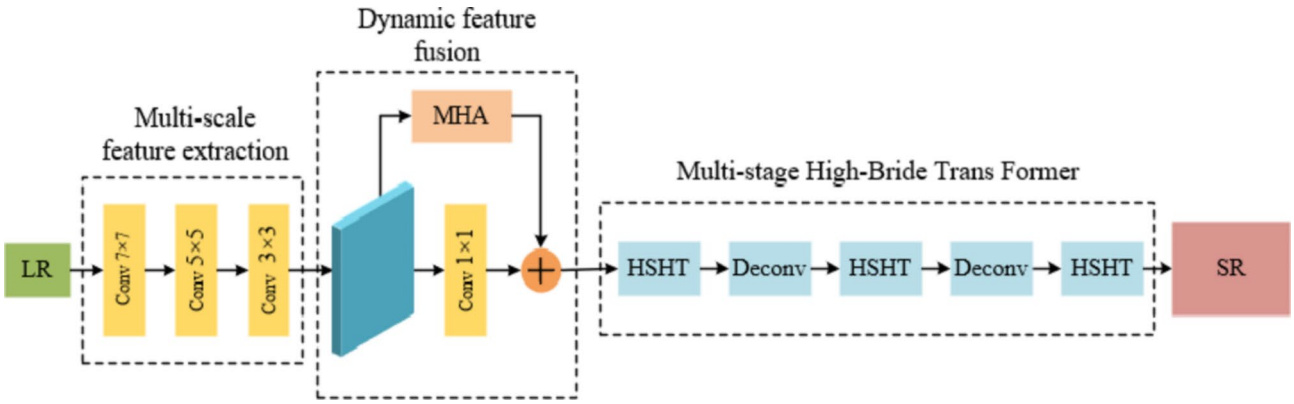


Fig. 1. Structure of generating network.

$$Q, K, V = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

In the formula, $\mathrm{Q}, \mathrm{K}, \mathrm{V}$ represents queries, keys, and values, while the matrix $\mathrm{d_k}$ is the scaling factor. The reweighted features are further fused through a 1×1 convolution to generate a new feature map:

$$f_{fused} = Conv_{1*1}(f_{ms} + Attention\_out) \tag{2}$$

The feature fusion module combines the local feature capturing ability of convolution operations with the global dependency modeling capability of the self-attention mechanism. By dynamically fusing features, it enhances the flexibility of feature representation and adaptively selects the most beneficial features for the final image generation.

3.**Multi-Stage Hybrid Transformer**: After feature fusion, $\mathrm{f_{fused}}$ is processed through a multi-stage Hybrid Transformer structure. This structure is divided into three stages, each consisting of a custom Transformer module. Each Transformer module includes a linear layer (Linear) and a standard Transformer architecture. The linear layer converts the channel dimension of the input features from C to the embedding dimension of the Transformer model. The input sequence is then processed through the Transformer modules, each of which includes multi-head self-attention and a feed-forward neural network. The Transformer in the first stage processes low-resolution features, maintaining the shape; the second stage processes medium-resolution features, refining the features; and the third stage processes high-resolution features, providing rich feature representations for subsequent image generation. Through progressive processing across multiple stages, features are progressively refined and enhanced to adapt to different resolution levels. The specific results of the HSHT module are shown in Fig. 2.

After feature fusion, the model enters the multi-stage Hybrid Transformer structure, which consists of three custom Transformer modules. Each Transformer module includes a linear layer (Linear) for feature mapping, a position encoding layer (PE) to enhance positional information of the sequence, and a layer normalization layer (LN) to stabilize the training process. Each stage of the Transformer module uses self-attention mechanisms and contextual modeling to optimize feature representation.

Specifically, the first stage processes low-resolution features by mapping $\mathrm{f_{fused}}$ through the linear layer and position encoding, followed by layer normalization before inputting into the Transformer module, with the output restored to the original image size. The second stage upsamples the image by a factor of two using a deconvolution layer, with the reshaped features fed into the second Transformer module to further enhance image details. The third stage again upsamples the image by a factor of two using a deconvolution layer, and the reshaped features are input into the third Transformer module for final high-resolution processing. This multi-stage processing effectively enhances the image resolution and quality.

After multi-stage feature processing, the generator concatenates these features along the channel dimension and finally uses a 1×1 convolution to fuse the channels, generating the final RGB image.

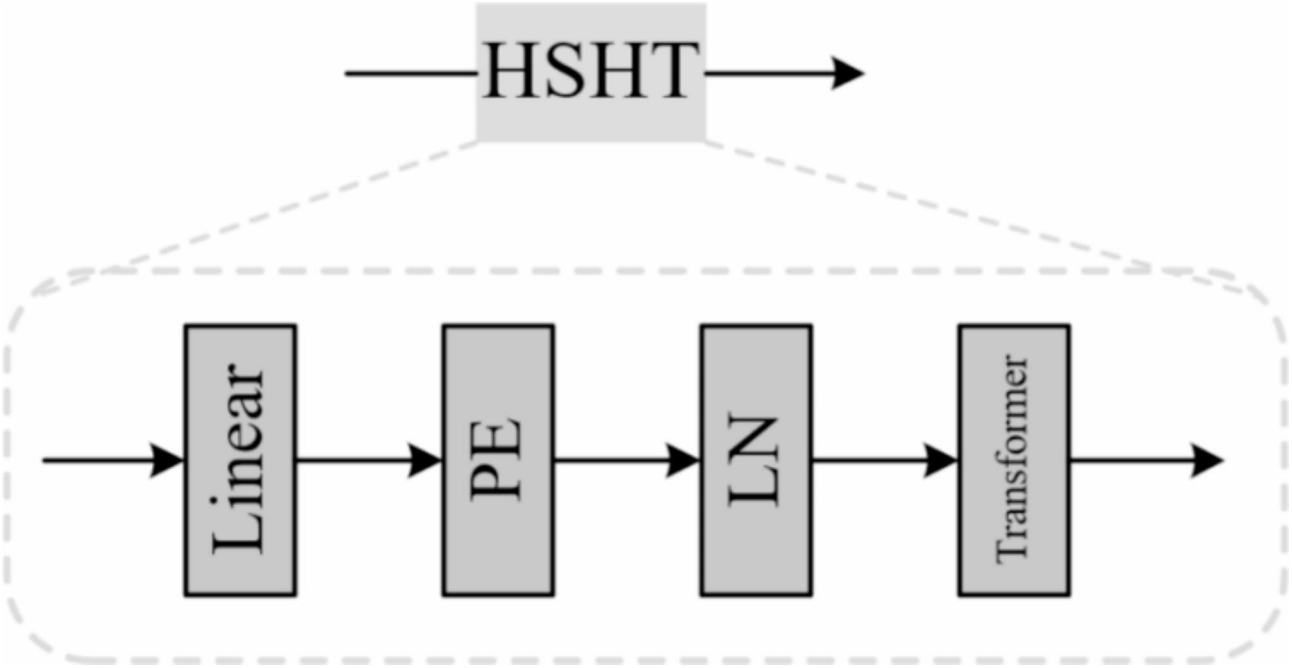


Fig. 2. HSHT module structure.

Discriminator model building

After generating the SR image, it is evaluated along with the HR image using the comprehensive discriminator module, as shown in Fig. 3. This module comprises three independent discriminators, each responsible for handling different feature scales and patterns, with the aim of providing a comprehensive assessment of the quality of the generated image.

1. Multi-Scale Convolution Discriminator: This discriminator begins by processing the input image through three progressively larger convolutional layers to extract features at different scales. Initially, the image is downsampled from 256×256 to 128×128, followed by 64×64 and 32×32. Each convolutional layer incrementally increases the number of feature channels and applies the ReLU activation function to introduce non-linearity. Subsequently, a series of residual convolutional blocks further refine the features at the 32×32 scale. Finally, the processed feature maps are flattened and passed through a fully connected layer to output the discrimination score. This module captures local details and overall structure by processing features at various scales, thus providing a comprehensive assessment of the generated image's quality.
2. Global Transformer Discriminator: This discriminator employs a pre-trained Swin Transformer to extract global contextual features. The Swin Transformer model, based on the Swin Transformer architecture, captures the global structural information of the image through deep learning. The input image is first processed by the Swin Transformer for feature extraction, and the extracted features are then classified through a fully connected layer. Given its ability to handle both global and detailed features, this module primarily assesses the overall quality and structural consistency of the generated image.
3. Hierarchical Feature Discriminator: This module first applies a 3×3 convolutional kernel to increase the number of channels in the input image to 256, using the ReLU activation function to extract features. After convolution, the feature maps are flattened and passed to a fully connected layer to generate the discrimination score. This module is designed to capture detailed features of the image and assess the quality of the details and texture through flattening and classification operations.

Through the collaborative work of these discriminator modules, the generator is able to progressively enhance the quality of the generated images, ensuring that the final output possesses higher resolution and greater realism.

Loss function

High-resolution (HR) remote sensing images contain rich high-frequency information, perceptual details, and environmental content. To achieve more detailed reconstruction of the image, this experiment incorporates content loss[32], pixel loss[33], adversarial loss[34], and total variation (TV) loss[35] to jointly constrain the generator, enhancing the overall robustness of the model.

Content Loss. In the generative network, adversarial learning helps maintain the realism of generated images but often results in significant artifacts and uncertain details. To ensure that the generated super-resolved (SR) image better aligns with the content distribution of the real high-resolution (HR) image, improve reconstruction

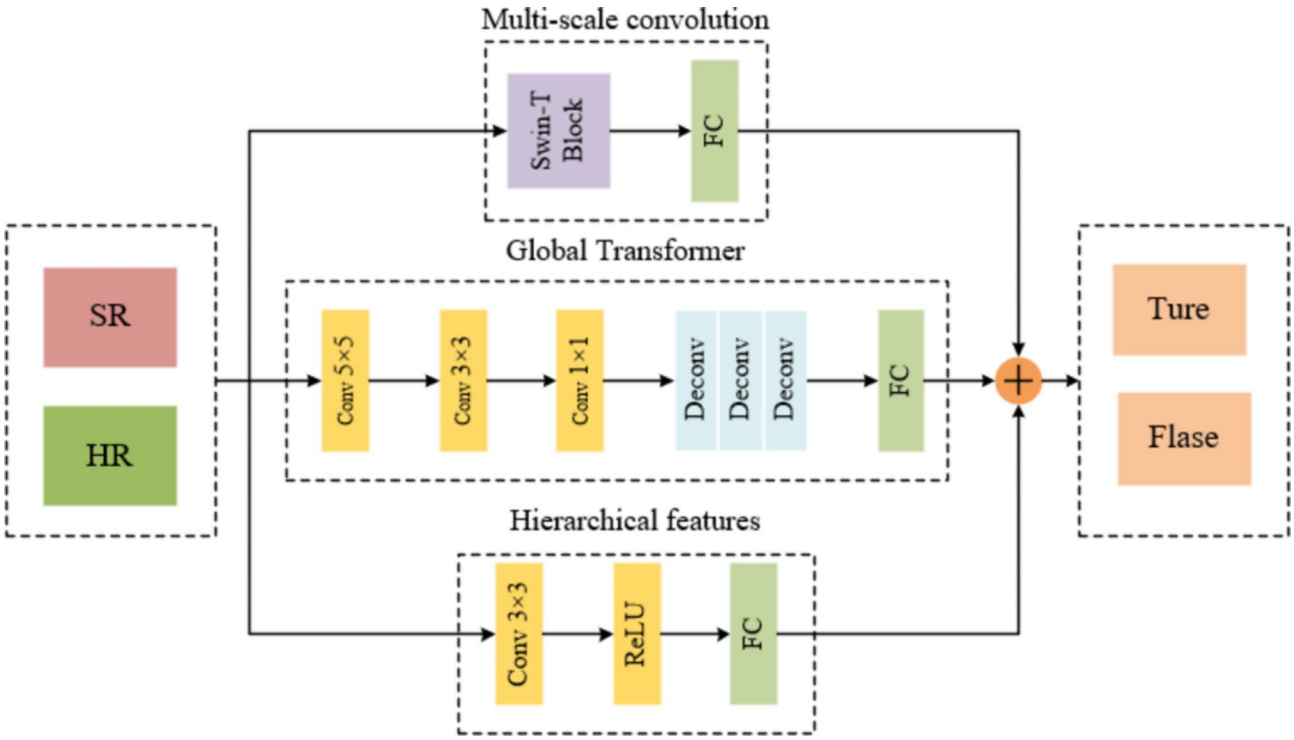


Fig. 3. Structure of Discriminator network.

quality, and limit excessive high-frequency content, this experiment further incorporates the Charbonnier loss function to enhance the consistency between the SR image and the HR image. The expression for content loss is:

$$L_{cont} = E_{I_{SR}} [\rho (I_{HR} - I_{SR})] \tag{3}$$

$$\rho (x) = (x^2 + \epsilon^2)^{1/2} \tag{4}$$

In the expression, ρ (*) denotes the Charbonnier loss function, with ϵ set to $10-3$.

Perceptual Loss. In the SRGAN model, perceptual loss is computed using features extracted from the DenseNet network before activation layers, rather than after. As the network deepens, most of the feature information after activation gradually fades, so the pre-activation features retain more information. Perceptual loss encourages the network to recover more high-frequency details to achieve perceptual alignment. In this experiment, perceptual loss is defined using the Charbonnier loss between the DenseNet features of the SR image and the HR image, both before activation. Its expression is:

$$\mathrm{L_{percep}} = \mathrm{E_{I_{SR}}} \{\rho [\upsilon_{\mathrm{feat(n)}} (\mathrm{I_{SR}}) - \upsilon_{\mathrm{feat(n)}} (\mathrm{I_{HR}})]\} \tag{5}$$

In the equation, $\upsilon_{\mathrm{feat(n)}}$ (*) represents the feature information extracted from the DenseNet model before the activation layer. In this experiment, the features are chosen from just before the maximum pooling layer in the 4th layer and just after the convolutional layer in the 3rd layer of the DenseNet-201 model.

Adversarial Loss. The loss feedback from the discriminator network can optimize the generator network, encouraging it to produce more natural images while simultaneously improving the performance of the discriminator network. Therefore, the adversarial loss needs to consider both the generator and the discriminator networks. The expression for adversarial loss is:

$$\mathrm{L_{adv}} = -\mathrm{E_{I_{HR}}} \{\log[1 - \mathrm{D_{Ra}}(\mathrm{I_{HR}}, \mathrm{I_{SR}})]\} - \mathrm{E_{I_{SR}}} \{\log[\mathrm{D_{Ra}}(\mathrm{I_{SR}}, \mathrm{I_{HR}})]\} \tag{6}$$

The expression for the discriminator loss is:

$$\mathrm{L_{adv}} = -\mathrm{E_{I_{HR}}} \{\log[\mathrm{D_{Ra}}(\mathrm{I_{HR}}, \mathrm{I_{SR}})]\} - \mathrm{E_{I_{SR}}} \{\log[1 - \mathrm{D_{Ra}}(\mathrm{I_{SR}}, \mathrm{I_{HR}})]\} \tag{7}$$

Total Variation (TV) Loss. Remote sensing images are inevitably affected by noise during acquisition, and the reconstruction process can amplify this noise while also introducing new noise. Noise with false information tends to have higher total variation in the image. The total variation of noisy images is significantly higher than that of non-noisy images. By minimizing the total variation loss, it is possible to reduce noise in the image while preserving edges. Therefore, TV loss is introduced, and its expression is:

$$\mathrm{L_{TV}} = \sum_{\mathrm{i,j}} \| \mathrm{I^{SR}_{i+1,j}} - \mathrm{I^{SR}_{i,j}} \|_1 + \| \mathrm{I^{SR}_{i,j+1}} - \mathrm{I^{SR}_{i,j}} \|_1 \tag{8}$$

In the expression, $\mathrm{I^{SR}_{i,j}}$ represents the pixel value at point (i, j), while i and j denote the corresponding point coordinates in the horizontal and vertical directions of the SR image. The introduction of TV loss helps suppress the generation of artifacts in SR images and prevents overfitting of the model during training.

Integrating the aforementioned loss functions, the total expression for the generation loss is:

$$\mathrm{L_G} = \lambda \mathrm{L_{cont}} + \mathrm{L_{percep}} + \eta \mathrm{L_{adv}} + \gamma \mathrm{L_{TV}} \tag{9}$$

Experimental configuration

Dataset and parameter settings

To validate the applicability of the proposed method under varying spatial resolutions, different sensor acquisition conditions, and minor perturbations, five distinct remote sensing image datasets were selected: PatternNet[36], AID[37], WHU-RS19[38], NWPURESISC45[39], and UCMERCED[40]. All these datasets consist of multi-class RGB images, covering a range of land cover types and scenarios, aiming to thoroughly evaluate the super-resolution performance of the model across different resolutions and imaging conditions.

PatternNet: Created by Xidian University, this dataset serves as a benchmark for large-scale high-resolution remote sensing image classification and retrieval. The images are sourced from the National Agricultural Imagery Program (NAIP) with a resolution of 1 m per pixel and a size of 256×256 pixels. It includes 30,400 images across 38 land cover categories, such as buildings, airports, farmland, and forests.

AID: Developed by Wuhan University, this dataset is designed for remote sensing image classification tasks. It contains 10,000 high-resolution aerial images with resolutions ranging from 0.5 m to 8 m and image dimensions of 600×600 pixels. The dataset covers 30 categories of scenes, including airports, commercial areas, agricultural zones, and forests, reflecting a wide range of surface cover types.

WHU-RS19: Released by Wuhan University, this high-resolution remote sensing dataset focuses on remote sensing image classification tasks. It includes 950 images across 19 land cover types, with each type having 50 images, and each image is 600×600 pixels. The dataset provides diverse land cover types, suitable for assessing the model's generalization capability.

NWPURESISC45: Released by Northwestern Polytechnical University, this dataset supports research in remote sensing image classification. It comprises 31,500 images across 45 different scene categories, with 700 images per category and a size of 256×256 pixels. The dataset offers rich scene information for evaluating model performance in diverse scenarios.

UCMERCED: A classic high-resolution remote sensing image dataset released by the University of California, Merced, widely used in remote sensing image classification research. It contains 2,100 images across 21 categories, with each category having 100 images. The images have a resolution of 0.3 m and a size of 256 × 256 pixels. This dataset is commonly used as a benchmark for validating model accuracy and robustness.

In the experiments, the training dataset consists of 20,000 images randomly selected from these datasets. All images from AID and WHU-RS19 were included in the training set, while images from PatternNet and NWPURESISC45 were randomly sampled to ensure diversity and coverage in the training data. The 2,100 images from the UCMERCED dataset were used exclusively as high-resolution (HR) images for the test set, providing a comprehensive evaluation of the performance and generalization capability of the super-resolution (SR) models.

Before training, all high-resolution images were resized to a uniform dimension of 256 × 256 pixels. To enhance data diversity and model generalization, all training images underwent random cropping, rotation, and were processed to generate corresponding low-resolution (LR) images using Gaussian blur, bicubic downsampling, and the addition of Gaussian noise to the HR samples. This data preprocessing approach enables the model to better handle various levels of image blurring and noise disturbances.

The model is implemented using the PyTorch framework and trained with mixed precision on two NVIDIA A100 80GB GPUs. During training, the Adam optimizer is used with the following parameters: $\beta 1 = 0.9$ and $\beta 2 = 0.99$. The overall loss function weight λ is set to 10, η to $5 \times 10 - 3$, and γ to $10 - 6$. The total number of training iterations is 200,000, with a batch size of 16 and an initial learning rate of $10 - 6$. The learning rate is halved every 50,000 iterations.

Evaluation metrics

In this experiment, the evaluation metrics used are Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Feature Similarity Index (FSIM)[41]. The detailed explanation of these metrics is as follows:

Peak Signal-to-Noise Ratio (PSNR): A higher PSNR value indicates better quality of the reconstructed image. PSNR is commonly used to quantify the quality of image reconstruction and reflects the difference between the reconstructed image and the reference high-resolution image. The specific formula for PSNR is as follows:

$$PSNR = log_{10}\left(\frac{MAX^2}{MSE}\right) * 10 \tag{10}$$

In the formula, MAX represents the maximum possible pixel value in the image, and MSE denotes the Mean Squared Error between the reconstructed image and the reference image.

Structural Similarity (SSIM): The closer the SSIM value is to 1, the higher the similarity in structural information between the reconstructed super-resolution (SR) image and the reference high-resolution (HR) image. SSIM measures image similarity by comparing luminance, contrast, and structure, serving as a comprehensive quality assessment metric. The formula is as follows:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{11}$$

Where μ_x and μ_y are the mean values of images x and y, respectively; σ_x^2 and σ_y^2 are the variances of images x and y, respectively; σ_{xy} is the covariance between images x and y; and c_1 and c_2 are constants used for stability.

Feature Similarity (FSIM): A higher FSIM value, closer to 1, indicates greater similarity between the reconstructed SR image and the HR image in terms of feature information. FSIM primarily assesses image similarity through phase congruency and gradient magnitude, which helps capture finer details and texture information in the image more effectively. The specific formula is as follows:

$$FSIM(x, y) = \frac{\sum_i PC_i \bullet S_L(i) \bullet S_P(i)}{\sum_i PC_i} \tag{12}$$

In the formula, PC_i represents the phase congruency of the image pixels, which reflects the local structural information of the image; $S_L(i)$ indicates the luminance similarity of the image pixels; and $S_P(i)$ represents the gradient similarity of the image pixels.

Experimental results

Comparative experiments

The proposed algorithm was quantitatively compared with four super-resolution reconstruction methods: Bicubic, SRGAN[19], ESRGAN[21], and SRTransGAN[42] on the PatternNet, AID, WHU-RS19, NWPURESISC45, and UCMERCED datasets. Tables 1, 2 and 3 show the average PSNR, SSIM, and FSIM values for these five algorithms across different datasets.

The results indicate that deep learning-based reconstruction algorithms significantly outperform the traditional Bicubic algorithm. The proposed method exhibits superior performance compared to the other comparative algorithms in 2×, 3×, and 4× upscaling tasks. Specifically, for the 4× upscaling task across the three test sets, the proposed algorithm achieves a PSNR approximately 3.61 dB higher than SRGAN, an SSIM approximately 0.070 higher (about 8.2% improvement), and an FSIM approximately 0.030 higher (about 3.1% improvement). These results demonstrate that the proposed algorithm outperforms existing classical methods on multiple metrics, especially showing significant advantages in high-magnification reconstruction tasks.

Dataset	Scale	Bicubic	SRGAN	ESRGAN	SRTransGAN	Proposed algorithm
PatternNet	×2	29. 01	36. 58	36. 91	37. 76	37. 99
	×3	26. 03	31. 32	32. 55	33. 76	34. 10
	×4	24. 34	29. 74	30. 16	31. 23	31. 67
AID	×2	29. 43	37. 25	38. 37	37. 44	39. 55
	×3	26. 82	33. 73	35. 02	36. 33	36. 52
	×4	23. 34	31. 96	33. 18	34. 35	34. 73
WHU-RS19	×2	28. 59	36. 57	37. 15	38. 75	39. 06
	×3	24. 55	33. 81	35. 84	37. 80	38. 08
	×4	22. 96	31. 63	33. 94	35. 71	36. 08
NWPURESISC45	×2	28. 61	37. 04	38. 31	37. 81	38. 74
	×3	25. 73	32. 28	35. 52	35. 07	37. 15
	×4	23. 25	30. 98	32. 09	33. 73	34. 86
UCMERCED	×2	29. 13	36. 91	37. 14	37. 92	38. 69
	×3	24. 81	34. 36	34. 61	34. 94	37. 33
	×4	22. 18	32. 51	32. 55	32. 24	34. 94

Table 1. Average PSNR of different algorithms on PatternNet, AID, WHU-RS19,NWPURESISC45,UCMERCED.

Dataset	Scale	Bicubic	SRGAN	ESRGAN	SRTransGAN	Proposed algorithm
PatternNet	×2	0. 856	0.937	0. 960	0. 962	0. 972
	×3	0. 794	0.883	0. 906	0. 918	0. 935
	×4	0. 737	0.814	0. 848	0. 881	0. 897
AID	×2	0. 812	0.892	0. 902	0. 990	0. 998
	×3	0. 760	0.917	0. 957	0. 954	0. 967
	×4	0. 713	0.874	0. 908	0. 924	0. 936
WHU-RS19	×2	0. 800	0.920	0. 944	0. 934	0. 954
	×3	0. 742	0.861	0. 897	0. 896	0. 922
	×4	0. 689	0.898	0. 926	0. 964	0. 988
NWPURESISC45	×2	0 0.797	0.841	0. 916	0. 953	0 0.971
	×3	0 0.716	0.813	0. 899	0. 922	0 0.965
	×4	0 0.669	0.874	0. 946	0. 972	0 0.983
UCMERCED	×2	0 0.813	0.914	0. 944	0. 941	0 0.969
	×3	0 0.736	0.843	0. 883	0. 916	0 0.932
	×4	0 0.621	0.871	0. 923	0. 968	0 0.979

Table 2. Average SSIM of different algorithms on PatternNet, AID, WHU-RS19,NWPURESISC45,UCMERCED.

Ablation experiments

To comprehensively evaluate the contribution of each module in the generator and discriminator networks to the final image quality and network performance, we conducted systematic ablation experiments. The results are shown in Tables 4 and 5:

Generator network ablation experiments

Multi-Scale Feature Extraction Module: Removing the multi-scale convolution kernels and using only a single-scale (3×3) convolution kernel for feature extraction. The results indicate that the high-resolution images generated with this configuration lack detail and global structure, highlighting the importance of multi-scale feature extraction in capturing image details and enhancing the network's perceptual capability.

Dynamic Feature Fusion Module: Replacing the multi-head self-attention mechanism with simple feature concatenation. The results show a significant decline in the detail and quality of the generated images after removing dynamic feature fusion, underscoring the critical role of the multi-head self-attention mechanism in optimizing feature representation and dynamically adjusting feature weights.

Multi-Stage Hybrid Transformer: Gradually removing different stages of the Hybrid Transformer structure (retaining only the first stage Transformer, and retaining the first two stages Transformers). The experiments demonstrate that each stage is crucial for the gradual enhancement of image details and resolution. Removing any stage affects the quality of the final generated image, confirming the importance of each stage in the overall process.

Dataset	Scale	Bicubic	SRGAN	ESRGAN	SRTransGAN	Proposed algorithm
PatternNet	×2	0. 861	0.988	0. 994	0. 997	0. 997
	×3	0. 850	0.989	0. 990	0. 993	0. 996
	×4	0. 834	0.976	0. 981	0. 986	0. 993
AID	×2	0. 824	0.989	0. 904	0. 908	0. 986
	×3	0. 814	0.896	0. 900	0. 903	0. 903
	×4	0. 798	0.887	0. 892	0. 897	0. 899
WHU-RS19	×2	0. 832	0.907	0. 910	0. 914	0. 915
	×3	0. 822	0.902	0. 906	0. 909	0. 912
	×4	0. 806	0.894	0. 898	0. 903	0. 909
NWPURESISC45	×2	0.841	0.913	0.920	0.926	0.947
	×3	0.830	0.901	0.914	0.921	0.912
	×4	0.809	0.895	0.906	0.917	0.929
UCMERCED	×2	0.827	0.903	0.917	0.935	0.974
	×3	0.804	0.891	0.911	0.929	0.938
	×4	0.796	0.874	0.904	0.923	0.942

Table 3. Average FSIM of different algorithms on PatternNet, AID, WHU-RS19,NWPURESISC45,UCMERCED.

Experimental configuration	PSNR (dB)	SSIM	FSIM
Baseline Model (Full Model)	37.99	0.972	0.981
Removed the multi-scale feature extraction module	33.52	0.791	0.815
Removed the dynamic feature fusion module	34.83	0.835	0.874
Removal of multi-stage Highbride Transform	35.15	0.897	0.906

Table 4. Generator network ablation experiments.

Experimental configuration	Performance score
Baseline Model (Full Model)	0.998
Remove multi-scale convolutional discriminators	0.949
Removed the global Transformer discriminator	0.902
Remove hierarchical feature discriminators	0.971

Table 5. Discriminator network ablation experiment.

Discriminator network ablation experiments

Multi-Scale Convolution Discriminator: Removing the progressively increasing convolution layers and using only fixed-scale (3×3) convolution layers for processing the input images. The results indicate that this configuration lacks the ability to capture features at different scales, affecting the accurate assessment of the quality of generated images.

Global Transformer Discriminator: Replacing the pre-trained Swin Transformer with a traditional Convolutional Neural Network (CNN). Comparison results show that the Global Transformer Discriminator performs better than traditional CNNs in capturing global structure and consistency of images. Removing this module significantly decreases the overall quality and structural consistency evaluation of the generated images.

Hierarchical Feature Discriminator: Using only the basic configuration of convolutional layers and fully connected layers, without hierarchical feature processing. The results show a significant reduction in the evaluation of image details and texture quality, validating the effectiveness of the hierarchical feature discriminator in capturing detailed features.

Quantitative analysis

For visual effect analysis, we selected six random images from different scenes in the 4× magnification results of the PatternNet, WHU-RS19, AID, and NWPURESISC45 datasets, and compared their local details through magnified views, as shown in Fig. 4. From these images, it is observed that the Bicubic algorithm produces relatively blurry results with a lack of detail information. In contrast, SRGAN and ESRGAN algorithms generate results with more detailed information; however, ESRGAN introduces artifacts and noise in some edge details. Although SRTransGAN generally provides superior results, it also exhibits artifacts and noise in certain details and performs poorly in edge sharpening.

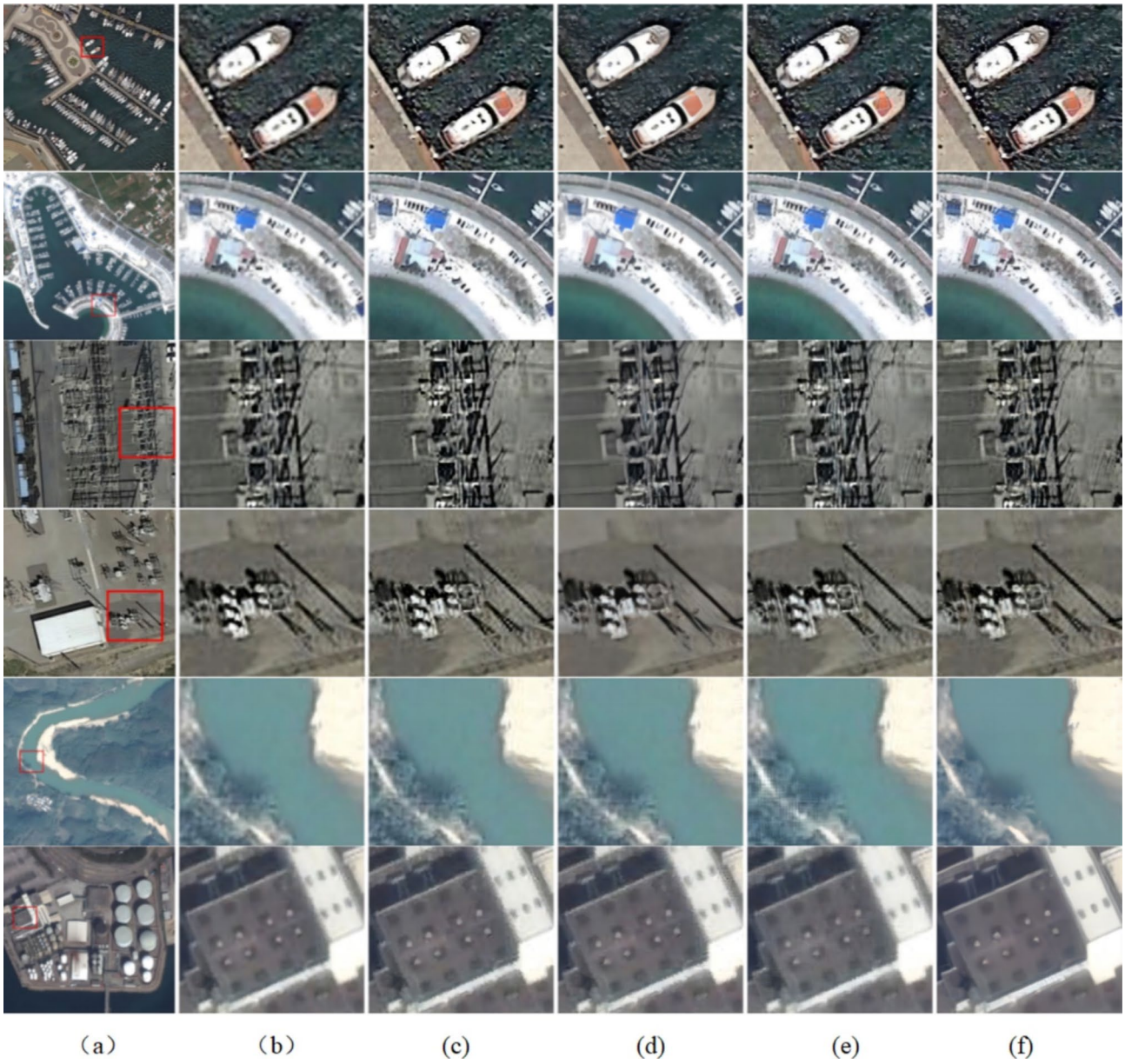


Fig. 4. 4x reconstruction results. (**a**) HR; (**b**) Bicubic; (**c**) SRGAN; (**d**) ESRGAN; (**e**) SRTransGAN; (**f**) Proposed algorithm.

The proposed algorithm demonstrates clearer detail and texture information in the reconstruction results, particularly excelling in edge sharpening. The magnified details in the images reveal that the reconstructed images from the proposed algorithm closely match the real high-resolution (HR) images in terms of color and brightness, with especially impressive handling of ship edges. Furthermore, the proposed algorithm produces significantly clearer building details and textures compared to the other algorithms.

Model efficiency analysis

The proposed algorithm was evaluated for 4× image magnification on the PatternNet, WHU-RS19, AID, and NWPURESISC45 datasets, and compared with four other algorithms in terms of runtime, as shown in Fig. 5. From Fig. 5, it is evident that the SRGAN algorithm, due to its network structure incorporating batch normalization (BN) layers, has the slowest reconstruction speed and requires the most time. The Bicubic algorithm, being an interpolation operation, has the shortest runtime. The proposed algorithm, due to the introduction of the multi-stage Hybrid Transformer, requires slightly more time compared to SRTransGAN.

However, as shown in Tables 1, 2 and 3, despite the slight increase in time for the proposed algorithm compared to SRTransGAN, it achieves higher values in PSNR, SSIM, and FSIM metrics. This indicates that the proposed algorithm outperforms other comparison algorithms in terms of image reconstruction quality.

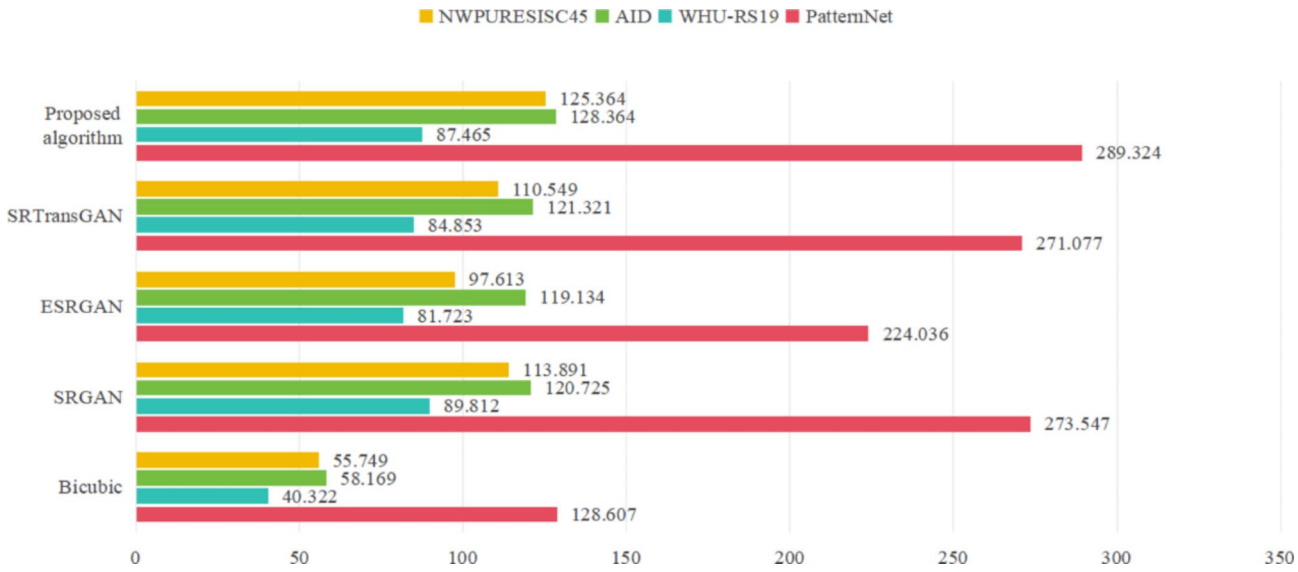


Fig. 5. Running time of different algorithms on PatternNet, WHU-RS19 ,AID and NWPURESISC45.

Conclusion

This paper presents an improved remote sensing image super-resolution reconstruction model based on a multi-scale receptive field and Hybrid Transformer structure. The model significantly enhances remote sensing image super-resolution performance through innovative generator and discriminator designs, incorporating multi-scale feature extraction, self-attention mechanisms, Transformer modules, and a comprehensive discriminator.

The strategy of multi-scale feature extraction and dynamic fusion enables the model to better capture detailed information and global structures in remote sensing images. By introducing convolutional kernels of varying sizes (7×7, 5×5, 3×3) in the generator, the model simultaneously extracts features at different scales, thereby effectively enhancing its ability to perceive local details and global information. Coupled with the multi-head self-attention mechanism, the model further achieves dynamic feature fusion, which contributes to improved detail representation in the reconstructed images, particularly in preserving and enhancing high-frequency information.

The introduction of the multi-stage Hybrid Transformer structure greatly improves the quality of the generated images. This structure processes image features progressively through three custom Transformer modules, enhancing image quality from low resolution to high resolution. Through this step-by-step refinement, the model can more accurately capture feature representations at different resolutions and effectively model features using self-attention mechanisms, ultimately significantly enhancing image detail recovery and visual quality.

The design of the discriminator is also a notable innovation of the model. By integrating multi-scale convolution, global Transformer, and hierarchical feature discriminators, the comprehensive discriminator evaluates the quality of generated images from multiple dimensions. The multi-scale convolution discriminator excels in capturing local features, the Transformer discriminator focuses on global information modeling, and the hierarchical feature discriminator emphasizes different levels of feature representations. This multi-dimensional and multi-level evaluation approach allows the discriminator to more comprehensively and accurately assess image quality, providing valuable feedback to the generator and further improving the realism and detail representation of the generated images.

Additionally, this paper incorporates Charbonnier loss and Total Variation (TV) loss functions to improve model training stability and accelerate convergence. The use of these loss functions not only effectively mitigates instability during training but also significantly enhances the generator's ability to capture image details, ensuring superior visual and perceptual performance of the generated remote sensing images.

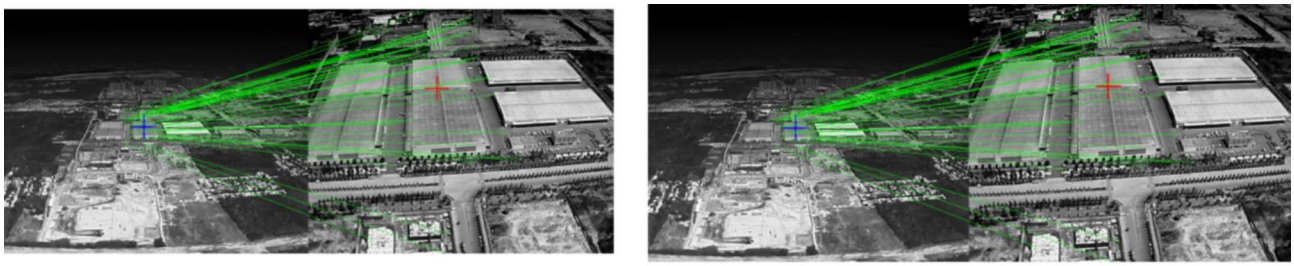
Experimental results demonstrate that the proposed model achieves significant improvements in PSNR, SSIM, and FSIM metrics, particularly in the restoration of high-frequency texture details and enhancement of fine features, showcasing its outstanding performance.

Future work

Image matching

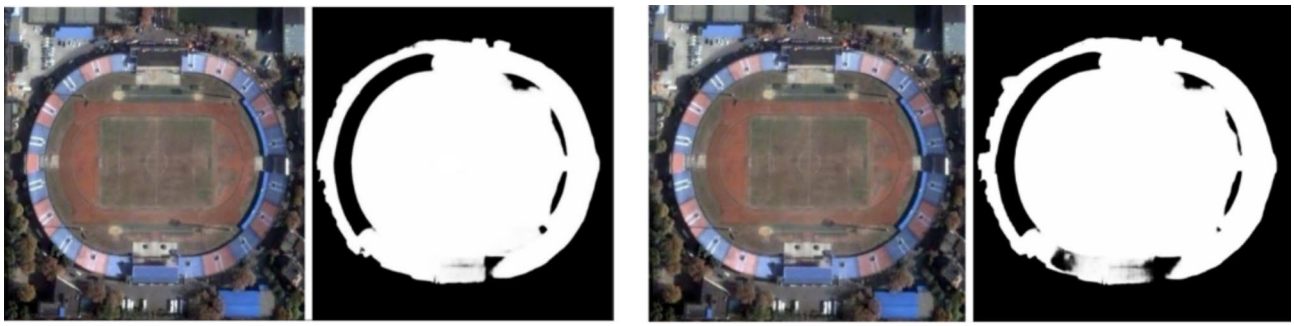
Applying the improved remote sensing image super-resolution reconstruction algorithm to images collected by drones has significantly enhanced the details and textures of the template images. This substantial improvement has notably increased the accuracy in subsequent image matching and target detection tasks.

In follow-up research, the Xfeat image matching method[43] was used to match feature points in images. Both the original template and the template reconstructed using the proposed algorithm were employed for matching the same target image. The results are shown in Fig. 6.



(a) (b)

Fig. 6. Comparison of image matching results.



(a) (b)

Fig. 7. Comparison of image segmentation results.

From Fig. 6(a), it can be observed that on the left is the template image, and on the right is the target image. Due to errors in detail extraction, there is a noticeable discrepancy in the matching results. However, in Fig. 6(b), after performing super-resolution processing on the template image using the proposed algorithm, the detail reconstruction is significantly improved. The detail extraction in the image matching process becomes more accurate, and the matching results are more precise. This demonstrates the feasibility of the proposed method and establishes a foundation for accuracy enhancement in future work.

Image segmentation

After applying the improved super-resolution reconstruction algorithm proposed in this paper to remote sensing images, the reconstructed images exhibit finer details and edges, with clearer gradient information between the foreground and background. In subsequent research, the BiRefNet al.gorithm[44] was used to separate the foreground and background of the remote sensing images. Both the original and reconstructed remote sensing images were segmented, and the segmentation results are shown in Fig. 7.

Figure 7(a) shows the original remote sensing image and its segmentation results. Due to the insufficient gradient information between the foreground and background, the details at the building edges are poor, leading to distortion at the edges after segmentation. In contrast, Fig. 7(b) uses the reconstructed remote sensing image, and the increased gradient differences between the foreground and background improve the segmentation results.

This research on image refinement lays a foundation for numerous future downstream applications. Future work will focus on high-resolution reconstruction of images in other domain scenarios to further enhance the generalization capability of the reconstruction algorithm.

Data availability

The data presented in this study are available at https://huggingface.co/datasets/blanchon/PatternNet.Code and models are publicly available at https://github.com/QJHyuntun/SR-net.git.

Received: 8 October 2024; Accepted: 10 January 2025
Published online: 16 January 2025

References

1. Gui, J., Sun, Z., Wen, Y., Tao, D. & Ye, J. A review on generative adversarial networks: algorithms, theory, and applications. *IEEE Trans. Knowl. Data Eng.* **35**, 3313–3332 (2020).

2. Zhang, Y. et al. : Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision, pp. 286–301. Springer, Cham (2018).
3. Zhang, Z., Wang, J. & Su, Y. A survey on the optical remote sensing image super-resolution technology. *SSpacecr Recovery Remote Sens.* **41** (6), 21–33 (2020).
4. Ibrahim, M., Ramzy, R., Benavente, D. P. & Lumbreras, F. *SWViT-RRDB: Shifted Window Vision Transformer Integrating Residual in Residual Dense Block for Remote Sensing Super-Resolution* (VISAPP, 2024).
5. Yan, J., Su, X. H. Y., Zhang, Y., Shi, M. & Gao, Y. Camouflage target detection based on strong semantic information and feature fusion. *J. Electron. Imaging.* **32**, 063019–063019 (2023).
6. Jungil Kong, J. & Kim and Jaekyoung Bae. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 1428, 17022–17033. (2020).
7. Hirahara, D., Takaya, E., Kadowaki, M., Kobayashi, Y. & Ueda, T. Effect of the Pixel Interpolation Method for Downsampling Medical Images on Deep Learning Accuracy. Journal of Computer and Communications : n. pag. (2021).
8. Zhang, K., Zhou, X., Zhang, H. & Zuo, W. Revisiting Single Image Super-Resolution Under Internet Environment: Blur Kernels and Reconstruction Algorithms. Pacific Rim Conference on Multimedia (2015).
9. Liang, Jingyun, K., Zhang, S., Gu, L. V., Gool & Timofte, R. Flow-based Kernel Prior with Application to Blind Super-Resolution. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) : 10596–10605. (2021).
10. Cao, G., Tian, H., Yu, L., Huang, X. & Wang, Y. Accelerate Histogram-Based Contrast Enhancement by Selective Downsampling. ArXiv abs/1709.04583 : n. pag. (2017).
11. Ren, Y., Li, R. & Liu, Y. Super-resolution reconstruction of face images based on iterative upsampling and downsampling layers. International Symposium on Robotics, Artificial Intelligence, and Information Engineering (2022).
12. Plötz, T. & Roth, S. Benchmarking Denoising Algorithms with Real Photographs. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) : 2750–2759. (2017).
13. Zhang, K., Liang, J., Gool, L. V. & Radu Timofte. and. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) : 4771–4780. (2021).
14. Wang, X., Xie, L., Dong, C. & Shan, Y. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) : 1905–1914. (2021).
15. Yamawaki, K. & Han, X. Deep Image and Kernel Prior Learning for Blind Super-Resolution. Proceedings of the 4th ACM International Conference on Multimedia in Asia : n. pag. (2022).
16. Dong, C., Loy, C. C., He, K. & Tang, X. Image Super-resolution using deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**, 295–307 (2014).
17. Dong, C. & Tang, X. Chen Change Loy and Accelerating the Super-Resolution Convolutional Neural Network. European Conference on Computer Vision (2016).
18. Kim, J. & Lee, J. K. and Kyoung Mu Lee. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) : 1646–1654. (2015).
19. Ledig, C. et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) : 105–114. (2016).
20. Miyato, T., Kataoka, T., Koyama, M. & Yoshida, Y. Spectral normalization for generative adversarial networks. ArXiv abs/1802.05957 : n. pag. (2018).
21. Wang, X. et al. *Chen Change Loy, Yu Qiao and Xiaoou Tang* (Enhanced Super-Resolution Generative Adversarial Networks. ECCV Workshops, 2018).
22. Liu, S., Huang, D. & Wang, Y. Receptive Field Block Net for Accurate and Fast Object Detection. European Conference on Computer Vision (2017).
23. Zhang, J., Shao, M. Z., Yu, L. & Li, Y. Image super-resolution reconstruction based on sparse representation and deep learning. *Signal. Process. Image Commun.* **87**, 115925 (2020).
24. Zhang, S., Yuan, Q., Li, J., Sun, J. & Zhang, X. Scene-adaptive remote sensing image Super-resolution using a multiscale attention network. *IEEE Trans. Geosci. Remote Sens.* **58**, 4764–4779 (2020).
25. Dong, X., Jia, X. S. X., Xi, Z., Gao, L. & Zhang, B. Remote sensing image Super-resolution using Novel dense-sampling networks. *IEEE Trans. Geosci. Remote Sens.* **59**, 1618–1633 (2021).
26. Dong, X., Wang, L., Jia, X. S. X., Gao, L. & Zhang, B. Remote sensing image Super-resolution using second-order Multi-scale Networks. *IEEE Trans. Geosci. Remote Sens.* **59**, 3473–3485 (2021).
27. Jiang, K. et al. Edge-enhanced GAN for remote sensing image Superresolution. *IEEE Trans. Geosci. Remote Sens.* **57**, 5799–5812 (2019).
28. Lei, S., Shi, Z. & Zou, Z. Coupled adversarial training for remote sensing image Super-resolution. *IEEE Trans. Geosci. Remote Sens.* **58**, 3633–3643 (2020).
29. Lin, Z., Liu, Y., Ye, W., Lin, B. & Zhou, H. DAE2GAN: image super-resolution for remote sensing based on an improved edge-enhanced generative adversarial network with double-end attention mechanism. *J. Appl. Remote Sens.* **18**, 014521–014521 (2024).
30. Sui, J., Wu, Q. & Pun, M. O. Denoising Diffusion Probabilistic Model with Adversarial Learning for Remote sensing Super-resolution. *Remote Sens.* **16**, 1219 (2024).
31. Vaswani, A. et al. Lukasz Kaiser and Illia Polosukhin. Attention is all you need. *Neural Inform. Process. Syst.* (2017).
32. Shi, W., Yang, W. & Liao, Q. Boosting External-Reference Image Quality Assessment by Content-Constrain Loss and Attention-based Adaptive Feature Fusion. 2023 International Joint Conference on Neural Networks (IJCNN) : 1–8. (2023).
33. Guo, C., Chen, X. & Chen, Y. and Chuying Yu. Multi-Stage Attentive Network for Motion Deblurring via Binary Cross-Entropy Loss. Entropy 24 : n. pag. (2022).
34. Ouattara, T. & Abdoulaye Valère Carin Jofack Sokeng, Irié Casimir Zo-Bi, Koffi Fernand Kouamé, Clovis Grinand and Romuald Vaudry. Detection of Forest Tree Losses in Côte d'Ivoire Using Drone Aerial Images. Drones : n. pag. (2022).
35. Chen, M., Pu, Y. & Bai, Y. Low-dose CT image denoising using residual convolutional network with fractional TV loss. *Neurocomputing* **452**, 510–520 (2020).
36. de Souza, C. M., Bastos, D. S. & Leonardo, A. Souza Filho and Magali Rezende Gouvêa Meireles. A study of training approaches of a hybrid Summarisation Model Applied to Patent dataset. *J. Inf. Knowl. Manag.* **22**, 2350030:1–2350030 (2023).
37. Xia, G. S. et al. AID: a Benchmark Data Set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **55**, 3965–3981 (2016).
38. Xia, G. S., Yang, W., Delon, J. & Gousseau, Y. *Hong Sun and Henri Maître* (STRUCTURAL HIGH-RESOLUTION SATELLITE IMAGE INDEXING., 2010).
39. Cheng, G., Han, J. & Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. Proceedings of the IEEE 105 : 1865–1883. (2017).
40. Yang, Y. & Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. ACM SIGSPATIAL International Workshop on Advances in Geographic Information Systems (2010).
41. Zhang, L., Zhang, L., Mou, X. & Zhang, D. FSIM: a feature similarity index for Image Quality Assessment. *IEEE Trans. Image Process.* **20**, 2378–2386 (2011).
42. Baghel, N. & Shiv Ram Dubey and Satish Kumar Singh. SRTransGAN: Image Super-Resolution using Transformer based Generative Adversarial Network. ArXiv abs/2312.01999 : n. pag. (2023).

43. Potje, G. A., Cadar, F., Araujo, A., Martins, R. & Nascimento, E. R. XFeat: Accelerated Features for Lightweight Image Matching. ArXiv abs/2404.19174 : n. pag. (2024).
44. Zheng, P., Gao, D., Liu, D. P. F. L., Laaksonen, J. & Wanli, W. A. Ouyang and Niculae Sebe. Bilateral Reference for High-Resolution Dichotomous Image Segmentation. ArXiv abs/2401.03407 : n. pag. (2024).

Author contributions
Conceptualization, D.L and L.Z.; methodology, D.L.; software, D.L.; validation, D.L., S.L. and B.L.; formal analysis, H.W.; investigation, L.Z.; resources, D.L.; data curation, L.Z.; writing—original draft preparation, D.L; writing—review and editing, L.K.; visualization, D.L.; supervision, S.L.; project administration, Y.L.; All authors have read and agreed to the published version of the manuscript.

Funding
This research received no external funding.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-86446-5.

Correspondence and requests for materials should be addressed to L.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.