remote sensing

MDPI

Article

Hybrid-Scale Hierarchical Transformer for Remote Sensing Image Super-Resolution

Jianrun Shang [1], Mingliang Gao [1], Qilei Li [2], Jinfeng Pan [1,*], Guofeng Zou [1] and Gwanggil Jeon [1,3]

[1] School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China; 21404020515@stumail.sdut.edu.cn (J.S.); mlgao@sdut.edu.cn (M.G.); gfzou@sdut.edu.cn (G.Z.); gjeon@inu.ac.kr (G.J.)
[2] School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK; q.li@qmul.ac.uk
[3] Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, Republic of Korea
* Correspondence: pjfbysj@163.com

Abstract: Super-resolution (SR) technology plays a crucial role in improving the spatial resolution of remote sensing images so as to overcome the physical limitations of spaceborne imaging systems. Although deep convolutional neural networks have achieved promising results, most of them overlook the advantage of self-similarity information across different scales and high-dimensional features after the upsampling layers. To address the problem, we propose a hybrid-scale hierarchical transformer network (HSTNet) to achieve faithful remote sensing image SR. Specifically, we propose a hybrid-scale feature exploitation module to leverage the internal recursive information in single and cross scales within the images. To fully leverage the high-dimensional features and enhance discrimination, we designed a cross-scale enhancement transformer to capture long-range dependencies and efficiently calculate the relevance between high-dimension and low-dimension features. The proposed HSTNet achieves the best result in PSNR and SSIM with the UCMecred dataset and AID dataset. Comparative experiments demonstrate the effectiveness of the proposed methods and prove that the HSTNet outperforms the state-of-the-art competitors both in quantitative and qualitative evaluations.

Keywords: super-resolution; remote sensing image; convolutional neural network; transformer; self-similarity

Citation: Shang, J.; Gao, M.; Li, Q.; Pan, J.; Zou, G.; Jeon, G. Hybrid-Scale Hierarchical Transformer for Remote Sensing Image Super-Resolution. *Remote Sens.* **2023**, *15*, 3442. https://doi.org/10.3390/rs15133442

Academic Editors: Prashant K. Srivastava and Salah Bourennane

Received: 19 April 2023
Revised: 21 June 2023
Accepted: 30 June 2023
Published: 7 July 2023

1. Introduction

With the rapid progress of satellite platforms and optical remote sensing technology, remote sensing images (RSIs) have been broadly deployed in civilian and military fields, e.g., disaster prevention, meteorological forecast, military mapping, and missile warning [1,2]. However, due to hardware limitations and environmental restrictions [3,4], RSIs often suffer from low-resolution (LR) and contain some intrinsic noise. Upgrading physical imaging equipment to improve resolution is often plagued by high costs and long development cycles. Therefore, it is of utmost urgency to explore the remote sensing image super-resolution (RSISR).

Single-image super-resolution (SR) is a highly ill-posed visual problem which aims to reconstruct high-resolution (HR) images from corresponding degraded LR images. To this end, many representative algorithms have been proposed, which can be roughly divided into three categories, i.e., interpolation-based methods [5,6], reconstruction-based methods [7,8], and learning-based methods [9,10]. The interpolation-based methods generally utilize different interpolation operations, including bilinear interpolation, bicubic interpolation, and nearest interpolation, to estimate unknown pixel value [11]. These methods are relatively straightforward in practice, while the reconstructed images lack

essential details. In contrast, reconstruction-based methods improve image quality by incorporating prior information of the image as constraints into the HR image. These methods can restore high-frequency details with the help of prior knowledge, while they require substantial computational costs, making it difficult for them to be readily applied to RSIs [12]. Learning-based approaches try to produce HR images by learning the mapping relationship established between external LR–HR image training pairs. Compared with the aforementioned two lines of methods, learning-based methods achieve better performance and become the mainstream in this domain due to the powerful feature representation ability provided by convolutional neural networks (CNNs) [13]. However, learning-based methods generally adopt the post-upsampling framework [14], which solely exploits low-dimensional features while ignoring the discriminative high-dimensional feature information after the upsampling process.

In addition to utilizing nonlinear mapping between LR–HR image training pairs, the self-similarity of the image is also employed to improve the performance of SR algorithms. Self-similarity refers to the property of similar patches appear repeatedly in a single image and is broadly adopted in image denoising [15,16], deblurring [17], and SR [18–20]. Self-similarities are also an intrinsic property in RSIs, i.e., internal recursive information. Figure 1 illustrates the self-similarities in RSIs. One can see that the down-scaled image is on the left, and the original one is on the right. Similar highway patches with green box labels appear repeatedly in the same scale image, while the roof of factories with red box labels appear repeatedly across different scales, and these patches with similar edges and textures contain abundant internal recursive information. Previously, Pan et al. [21] employed dictionary learning to capture structural self-similarity features as additional information to improve the performance of the model. However, the sparse representation of SR has a limited ability to leverage the internal recursive information within the entire remote sensing image.

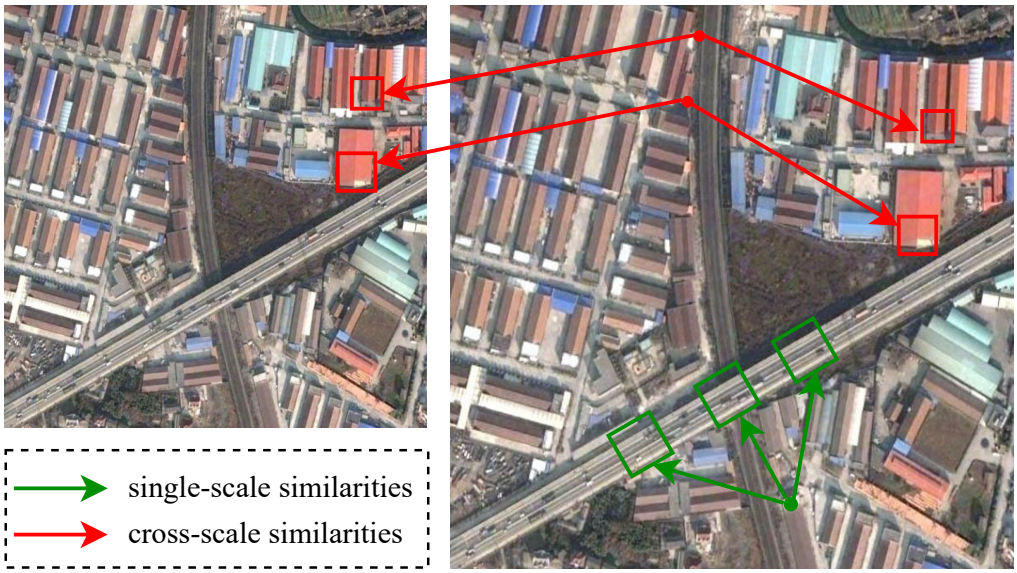


Figure 1. Illustration of self-similarities in RSIs with single-scale (green box) and cross-scale (red box).

In this paper, we propose a Hybrid-Scale Hierarchical Transformer Network (HSTNet) for RSISR. The HSTNet can enhance the representation of the high-dimensional features after upsampling layers and fully utilize the self-similarity information in RSIs. Specifically, we propose a hybrid-scale feature exploitation (HSFE) module to leverage the internal similar information both in single and cross scales within the images. The HSFE module contains two branches, i.e., a single-scale branch and a cross-scale branch. The former is employed to capture the recurrence within the same scale image, and the latter is utilized to learn the feature correlation across different scales. Moreover, we designed a

cross-scale enhancement transformer (CSET) module to capture long-range dependencies and efficiently model the relevance between high-dimension and low-dimension features. In the CSET module, the encoders are used to encode low-dimension features from the HSFE module, and the decoder is used to utilize to fuse the multiple hierarchies high-/low-dimensional features so as to enhance the representation ability of high-dimensional features. To sum up, the main contributions of this work are as follows:

1. We propose an HSFE module with two branches to leverage the internal recursive information from both single and cross scales within the images for enriching the feature representations for RSISR.
2. We designed a CSET module to capture long-range dependencies and efficiently calculate the relevance between high-dimension and low-dimension features. It helps the network reconstruct SR images with rich edges and contours.
3. Jointly incorporating the HSFE and CSET modules, we formed the HSTNet for RSISR. Extensive experiments on two challenging remote sensing datasets verify the superiority of the proposed model.

2. Related Literature

2.1. CNN-Based SR Models

Dong et al. [22] pioneered the adoption of an SR convolutional neural network (SRCNN) that utilizes three convolution layers to establish the nonlinear mapping relationship between LR–HR image training pairs. On the basis of the residual network introduced by He et al. [23], Kim et al. [24] designed a very deep SR convolutional neural network (VDSR) where residual learning is employed to accelerate model training and improve reconstruction quality. Lim et al. [25] built the enhanced deep super-resolution model to simplify the network and improve the computational efficiency via optimizing the initial residual block. Zhang et al. [26] designed a deep residual dense network in which the residual network with dense skip connections is used to transfer intermediate features. Benefiting from the channel attention (CA) module, Zhang et al. [27] presented a deep residual channel attention network to enhance the high-frequency channel feature representation. Dai et al. [28] designed a second-order CA mechanism to guide the model to improve the ability of discriminative learning ability and exploit more conducive features. Li et al. [29] proposed an image super-resolution feedback network (SRFBN) in which a feedback mechanism is adopted to transfer high-level feature information. The SRFBN could leverage high-level features to polish up the representation of low-level features.

Because of the impact of spatial resolution on the final performance of many RSI tasks, including instance segmentation, object detection, and scene classification, RSISR also raises significant research interest. Lei et al. [30] proposed a local–global combined network (LGCNet) which can enhance the multilevel representations, including local detail features and global information. Haut et al. [31] produced a deep compendium model (DCM), which leverages skip connection and residual unit to exploit more informative features. To fuse different hierarchical contextual features efficiently, Wang et al. [32] designed a contextual transformation network (CTNet) based on a contextual transformation layer and contextual feature aggregation module. Ni et al. [33] designed a hierarchical feature aggregation and self-learning network in which both self-learning and feedback mechanisms are employed to improve the quality of reconstruction images. Wang et al. [34] produced a multiscale fast Fourier transform (FFT)-based attention network (MSFFTAN), which employs a multi-input U-shape structure as the backbone for accurate RSISR. Liang et al. [35] presented a multiscale hybrid attention graph convolution neural network for RSISR in which a hybrid attention mechanism was adopted to obtain more abundant critical high-frequency information. Wang et al. [36] proposed a multiscale enhancement network which utilizes multiscale features of RSIs to recover more high-frequency details. However, the CNN-based methods above generally employ the post-upsampling framework that directly recovers HR images after the upsampling layer, ignoring the discriminative high-dimensional feature information after the upsampling process [14].

2.2. Transformer-Based SR Models

Due to the strong long-range dependence learning ability of transformers, transformer-based image SR methods have been studied recently by many scientific researchers. Yang et al. [37] produced a texture transformer network for image super-resolution, in which a learnable texture extractor is utilized to exploit and transmit the relevant textures to LR images. Liang et al. [38] proposed SwinIR by transferring the ability of the Swin Transformer, which could achieve competitive performance on three representative tasks, namely image denoising, JPEG compression artifact reduction, and image SR. Fang et al. [39] designed a lightweight hybrid network of a CNN and transformer that can extract beneficial features for image SR with the help of local and non-local priors. Lu et al. [40] presented a hybrid model with a CNN backbone and transformer backbone, namely the efficient super-resolution transformer, which achieved impressive results with low computational cost. Yoo et al. [41] introduced an enriched CNN–transformer feature aggregation network in which the CNN branch and transformer branch can mutually enhance each representation during the feature extraction process. Due to the limited ability of multi-head self-attention to extract cross-scale information, cross-token attention is adopted in the transformer branch to utilize information from tokens of different scales.

Recently, transformers have also found their way into the domain of RSISR. Lei et al. [14] proposed a transformer-based enhancement network (TransENet) to capture features from different stages and adopted a multistage-enhanced structure that can integrate features from different dimensions. Ye et al. [42] proposed a transformer-based super-resolution method for RSIs, and they employed self-attention to establish dependencies relationships within local and global features. Tu et al. [43] presented a GAN that draws on the strengths of the CNN and Swin Transformer, termed the SWCGAN. The SWCGAN fully considers the characteristics of large size, a large amount of information, and a strong relevance between pixels required for RSISR. He et al. [44] designed a dense spectral transformer to extract the long-range dependence for spectral super-resolution. Although the transformer can improve the long-range dependence learning ability of the model, these methods do not leverage the self-similarity within the entire remote sensing image [45].

3. Methodology

3.1. Overall Framework

The framework of the proposed HSTNet is shown in Figure 2. It is built by the combination of three kinds of fundamental modules, i.e., a low-dimension feature extraction (LFE) module, a cross-scale enhancement transformer (CSET) module, and an upsample module. Specifically, the LFE module is utilized to extract high-frequency features across different scales, and the CSET module is employed to capture long-range dependency to enhance the final feature representation. The upsample module is adopted to transform the feature representation from a low-dimensional space to a high-dimensional space.

Given an LR image I_{LR}, a convolutional layer with a 3×3 kernel is utilized to extract the initial feature F_0. The process of shallow feature extraction is formulated as

$$F_0 = f_{\mathrm{sf}}(I_{\mathrm{LR}}), \tag{1}$$

where $f_{\mathrm{sf}}(\cdot)$ represents the operation of the convolutional operation and F_0 is the shallow feature.

As shown in Figure 3, the LFE module consists of five basic extraction (BE) modules, and each BE module contains two 3×3 convolution layers and one hybrid-scale feature exploitation (HSFE) module. As the core component of the BE module, the HSFE module is proposed to model image self-similarity. The whole low-dimensional feature extraction process is formulated as

$$F_{\mathrm{LFE}}^{i} = f_{\mathrm{lfe}}^{i}\left(F_{\mathrm{LFE}}^{i-1}\right) = f_{\mathrm{lfe}}^{i}\left(f_{\mathrm{lfe}}^{i-1}\left(\cdots f_{\mathrm{lfe}}^{1}(F_0)\cdots\right)\right), \quad i = 1, 2, 3, \tag{2}$$

where $f^i_{\text{lfe}}(\cdot)$ and F^i_{LFE} represent the operation of ith LFE module and its output. After the three cascaded LFE modules, a subpixel layer [46] is adopted to transform low-dimensional features into high-dimensional features, which is formulated as

$$F_{\text{up}} = \text{Subpixel}\left(F^3_{\text{LFE}}\right), \tag{3}$$

where F_{up} represents the high-dimension feature and $\text{Subpixel}(\cdot)$ denotes the function of the subpixel layer. The low-dimension features F^1_{LFE}, F^2_{LFE}, and F^3_{LFE} and the high-dimension feature F_{up} are fed into three cascaded CSET modules for feature hierarchical enhancement. To reduce the redundancy of the enhanced features, a 1×1 convolution layer is employed to reduce the feature dimension. The complete process including the enhancement and dimension reduction is formulated as

$$F^i_{CSET} = \begin{cases} f^i_{\text{cset}}\left(F^i_{\text{LFE}}, F^{i+1}_{\text{CSET}}\right), & i = 1, 2, \\ f^i_{\text{cset}}\left(F^i_{\text{LFE}}, F_{\text{up}}\right), & i = 3, \end{cases} \tag{4}$$

where $f^i_{\text{cset}}(\cdot)$ and F^i_{CSET} represent the operation of ith CSET module and its output, respectively. Finally, one convolution layer is employed to obtain SR image I_{SR} from the enhanced features. A conventional L_1 loss function was employed to train the proposed HSTNet model. Given a training set $\left\{I^i_{\text{LR}}, I^i_{\text{HR}}\right\}_{i=1}^N$, the loss function is formulated as:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left\|F_{\text{HSTNet}}\left(I^i_{\text{LR}}\right) - I^i_{\text{HR}}\right\|_1, \tag{5}$$

where F_{HSTNet} denotes the proposed model parameterized by θ and N represents the number of training LR–HR pairs.

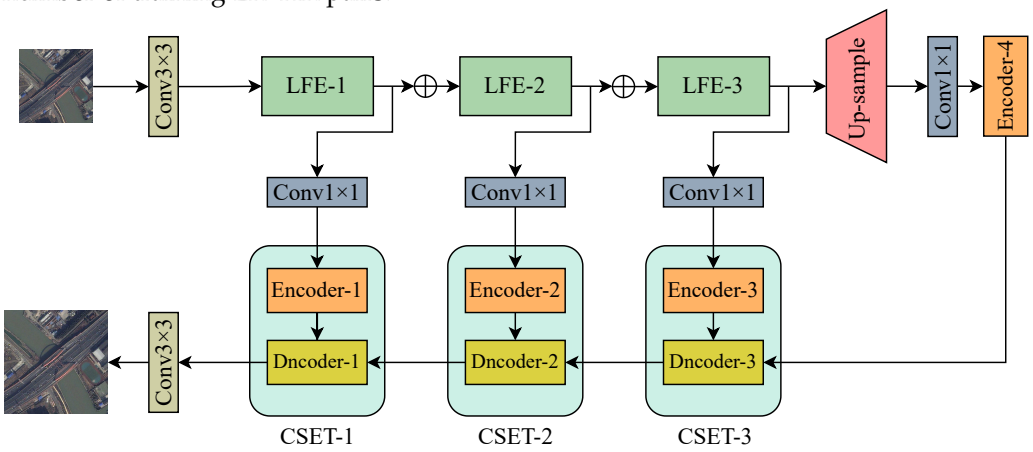


Figure 2. Architecture of the proposed HSTNet for remote sensing image SR.

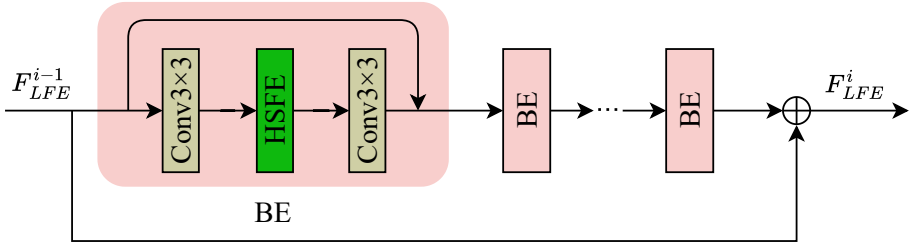


Figure 3. Architecture of the LFE module.

3.2. Hybrid-Scale Feature Exploitation Module

To explore the internal recursive information in single-scale and cross-scale, we propose an HSFE module. Figure 4 exhibits the architecture of the HSFE module, which

consists of a single-scale branch and a cross-scale branch. The single-scale branch aims to capture similar features within the same scale, and a non-local (NL) block [47] was utilized to calculate the relevance of these features. The cross-scale branch was applied to capture recursive features of the same image at different scales, and an adjusted non-local (ANL) block [45] was utilized to calculate the relevance of features between two different scales.

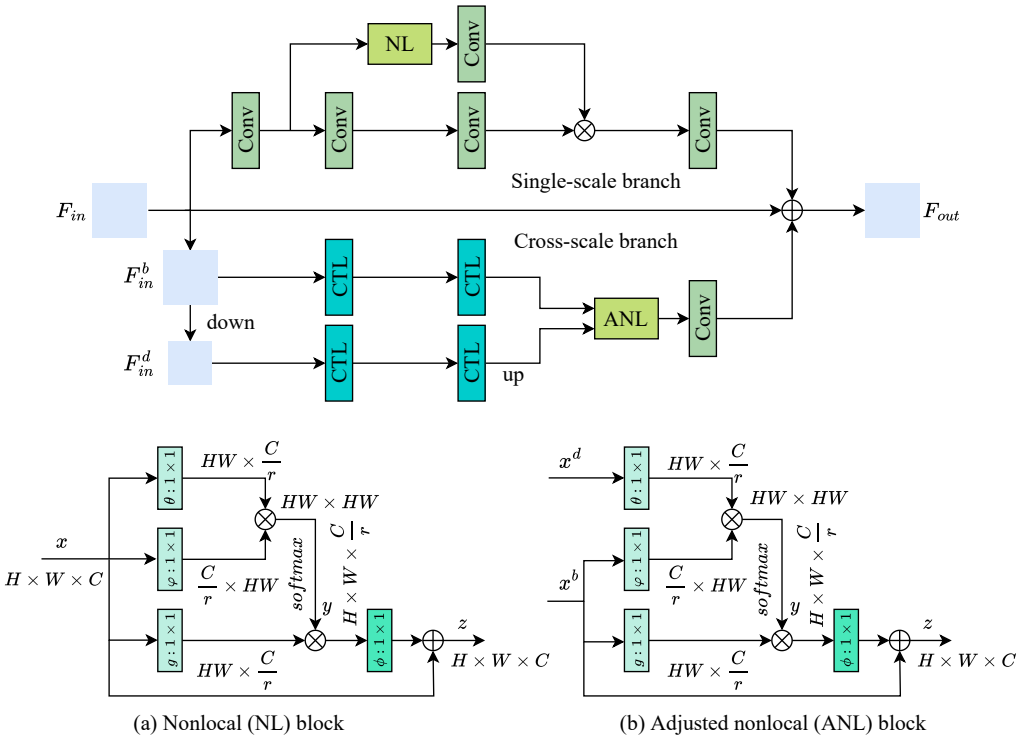


Figure 4. Architecture of the proposed HSFE module.

Single-scale branch: As depicted in Figure 4, we built the single-scale branch to extract single-scale features. Specifically, in the single-scale branch, several convolutional layers are applied to capture recursive features, and an NL block is employed to guide the network to concentrate on informative areas. As shown in Figure 4a, an embedding function is utilized to mine the similarity information as

$$f(x_i, x_j) = e^{\left(\theta^T(x_i)\varphi(x_j)\right)} = e^{\left((W_\theta x_i)^T(W_\varphi x_j)\right)}, \tag{6}$$

where i is the index of the output position, j is the index that enumerates all positions, and x denotes the input of the NL block. W_θ and W_φ are the embeddings weight matrix. The non-local function is symbolized as

$$y_i = \left(\sum_{\forall j} f(x_i, x_j) g(x_j)\right) \Big/ \sum_{\forall j} f(x_i, x_j). \tag{7}$$

The relevance between x_i and all x_j can be calculated by pairwise function $f(\cdot)$. The feature representation of x_j can be obtained by the function $g(\cdot)$. Eventually, the output of the NL block is obtained by

$$z_i = W_\phi y_i + x_i, \tag{8}$$

where W_ϕ is a weight matrix.

The convolution layer following the NL block transforms the input into an attention diagram, which is then normalized with a sigmoid function. In addition, the main branch

output features are multiplied by the attention diagram, where the activation values for each space and channel location are rescaled.

Cross-scale branch: As depicted in Figure 4, the cross-scale branch is utilized to perform cross-scale feature representation. Specifically, the input of the HSFE module is considered the basic scale feature, which is symbolized as F_{in}^{b}. To exploit the internal recursive information at different scales, the downsampled scale feature F_{in}^{d} is formulated as

$$F_{in}^{d} = f_{\text{down}}^{s}\left(F_{in}^{b}\right), \tag{9}$$

where $f_{\text{down}}^{s}(\cdot)$ denotes the operation of downsampling with scale factor s.

Two contextual transformation layers (CTLs) [48] are employed to extract feature with two different scales F_{in}^{b} and F_{in}^{d}. To align the spatial dimension of the features in different scales, the downsampled feature is firstly upsampled with the scale factor of s. x^{b} and x^{b} represent the output of the basic scale and the downsampled scale through the two CTLs, and this process is formulated as

$$\begin{aligned} x^{b} &= f_{ctl}\left(F_{in}^{b}\right) \\ x^{d} &= f_{up}^{s}\left(f_{ctl}\left(F_{in}^{b}\right)\right), \end{aligned} \tag{10}$$

where $f_{ctl}(\cdot)$ denotes the operation of two CTLs and $f_{up}^{s}(\cdot)$ represents the operation of upsample with scale factor s.

Similar to the single-scale branch, an ANL block [45] was introduced to exploit the feature correlation between two different scales RSIs. As shown in Figure 4b, the ANL block is improved compared to the NL block, and they have different inputs. Thus, z_i in Equation (8) for ANL block can be rewritten as

$$f\left(x_i^{d}, x_j^{b}\right) = e^{\left(\theta^{T}\left(x_i^{d}\right)\varphi\left(x_j^{b}\right)\right)} = e^{\left(\left(W_\theta x_i^{d}\right)^{T}\left(W_\varphi x_j^{b}\right)\right)}, \tag{11}$$

$$y_i = \left(\sum_{\forall j} f\left(x_i^{d}, x_j^{b}\right) g\left(x_j^{b}\right)\right) \Big/ \sum_{\forall j} f\left(x_i^{b}, x_j^{d}\right) \tag{12}$$

$$z_i = W_\phi y_i + x_i. \tag{13}$$

In the cross-scale branch, we employ the ANL block to fuse multiple scale features, therefore fully utilizing the self-similarity information. The HSFE module can be formulated as

$$F_{out} = f_{sin}(F_{in}) + f_{cro}(F_{in}) + F_{in}, \tag{14}$$

where F_{in} is the input of the HSFE module and F_{out} is the output of the HSFE module. $f_{sin}(\cdot)$ and $f_{cro}(\cdot)$ are the operation of the single-scale branch and cross-scale branch, respectively.

3.3. Cross-Scale Enhancement Transformer Module

The cross-scale enhancement transformer module is designed to learn the dependency relationship across long distances between high-dimension and low-dimension features and enhance the final feature representation. The architecture of the CSET module is shown in Figure 5a. Specifically, we introduced the cross-scale token attention (CSTA) module [41] to exploit the internal recursive information within an input image across different scales. Moreover, we use three CSET modules to utilize different hierarchies of feature information. Figure 5a illustrates in detail the procedure of feature enhancement using CSET-3 module as an example.

Transformer encoder: The encoders are used to encode different hierarchies of features from LFE modules. As shown in Figure 5a, the encoder is mainly composed of a multi-headed self-attention (MHSA) block and a feed-forward network (FFN) block, which

is similar to the original design in [49]. The FFN block contains two multilayer perceptron (MLP) layers with an expansion ratio r and a GELU activation function [50] in the middle. Moreover, we adopted layer normalization (LN) before the MHSA block and FFN block, and employed a local residual structure to avoid the gradient vanishing or explosion during gradient backpropagation. The entire process of the encoder can be formulated as

$$\begin{aligned} {F_{EN}^{i}}' &= f_{mhsa}\left(f_{ln}\left(F_{LFE}^{i}\right)\right) + F_{LFE}^{i} \\ F_{EN}^{i} &= f_{ffn}\left(f_{ln}\left({F_{EN}^{i}}'\right)\right) + {F_{EN}^{i}}', \end{aligned} \tag{15}$$

where $f_{mhsa}(\cdot)$, $f_{ln}(\cdot)$, and $f_{ffn}(\cdot)$ denote the function of the MHSA block, layer normalization, and FFN block, respectively. ${F_{EN}^{i}}'$ is the intermediate output of the encoder. F_{LFE}^{i} and F_{EN}^{i} are the input and output of the encoder in the ith CSET module.

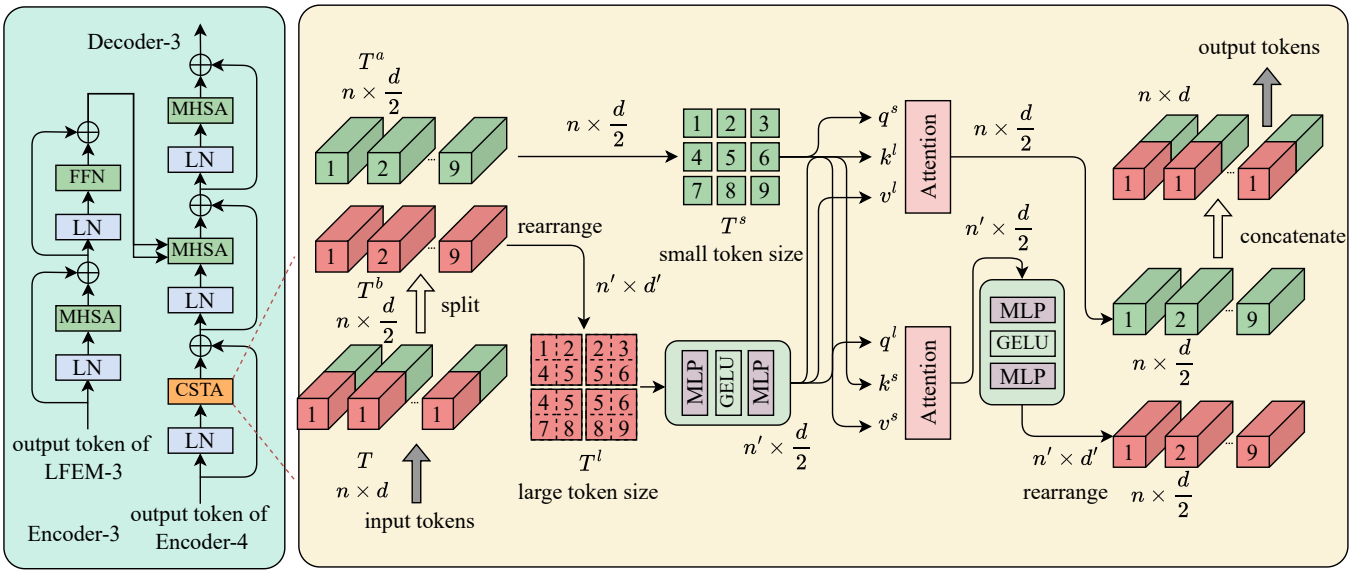
Transformer decoder: The decoders are utilized to fuse high-/low-dimensional features from multiple hierarchies to enhance the representation ability of high-dimensional features. As shown in Figure 5a, the decoder contains two MHSA blocks and a CSTA block [41]. With the CSTA block, the decoder can exploit the recursive information within an input image across different scales. The operation of the decoder can be formulated as

$$\begin{aligned} {F_{DE}^{i}}'' &= f_{csta}\left(f_{ln}\left(F_{up}\right)\right) + F_{up} \\ {F_{DE}^{i}}' &= f_{mhsa}\left(f_{ln}\left({F_{DE}^{i}}''\right), {F_{EN}^{i}}'\right) + {F_{DE}^{i}}'' \\ F_{CSET}^{i} &= f_{mhsa}\left(f_{ln}\left({F_{DE}^{i}}'\right)\right) + {F_{DE}^{i}}' \end{aligned} \tag{16}$$

where $f_{csta}(\cdot)$ denotes the process of the CSTA block and F_{up} is the output of Encoder-4. Each CSET module has two inputs, and the composition of the inputs is determined by the location of the CSET module. ${F_{DE}^{i}}'$ and ${F_{DE}^{i}}''$ represent the intermediate outputs of the decoder. F_{CSET}^{i} represents the output of ith CSET module.

CSTA block: The CSTA block [41] is introduced to utilize the recurrent patch information of different scales in the input image. The feature information flow of the CSTA module is illustrated in Figure 5b. Specifically, the input token embeddings $T \in \mathbb{R}^{n \times d}$ of the CSTA block are split into $T^a \in \mathbb{R}^{n \times \frac{d}{2}}$ and $T^b \in \mathbb{R}^{n \times \frac{d}{2}}$ along the channel axis. Then, $T^s \in \mathbb{R}^{n \times \frac{d}{2}}$ including n tokens from $T^a$and $T^l \in \mathbb{R}^{n' \times d'}$ including n' tokens by rearranging T^b are generated. The number of tokens in T^l can be set to $n' = \left[\frac{h-t'}{s'} + 1\right] \times \left[\frac{w-t'}{s'} + 1\right]$, where t' and s' represent the stride and token size. To improve efficiency, T^s is replaced by T^a, and T^l is tokenized with a larger token size and overlapping. Numerous large-size tokens can be obtained by overlapping, enabling the transformer to actively learn patch recurrence across scales.

To effectively exploit self-similarity across different scales, we computed cross-scale attention scores between tokens in both T^s and T^l. Specifically, the queries $q^s \in \mathbb{R}^{n \times \frac{d}{2}}$, keys $k^s \in \mathbb{R}^{n \times \frac{d}{2}}$, and values $v^s \in \mathbb{R}^{n \times \frac{d}{2}}$ were generated from T^s. Similarly, the queries $q^l \in \mathbb{R}^{n' \times \frac{d}{2}}$, keys $k^l \in \mathbb{R}^{n' \times \frac{d}{2}}$, and values $v^l \in \mathbb{R}^{n' \times \frac{d}{2}}$ were generated from T^l. The reorganized triples $\left(q^s, k^l, v^l\right)$ and $\left(q^l, k^s, v^s\right)$ were obtained by swapping their key–value pairs to each other. Then, the attention operation was executed using the reorganized triples. It should be noted that the projection of attention operations reduces the last dimension of queries, keys, and values in T^l from d' to $d/2$. Subsequently, we re-projected the attention results of T^l to the dimension of $n' \times d'$ then transformed to the dimension of $n \times \frac{d}{2}$. Finally, we concatenated the attention results to obtain the output of the CSTA block.



(a) Transformer module (b) Cross-scale token attention (CSTA) block

Figure 5. Architecture of the CSET module.

4. Experiments

4.1. Experimental Dataset and Settings

We evaluate the proposed method on two widely adopted benchmarks [30,31,51], namely the UCMecred dataset [52] and AID dataset [53], to demonstrate the effectiveness of the proposed HSTNet.

UCMerced dataset: This dataset consists of 2100 images belonging to 21 categories of varied remote sensing image scenes. All images exhibit a pixel size of 256 × 256 and a spatial resolution of 0.3 m/pixel. The dataset is divided equally into two distinct sets, one comprising 1050 images for training and the other for testing.

AID dataset: This dataset encompasses 10,000 remote sensing images, spanning 30 unique categories. In contrast to the UCMerced dataset, all images in this dataset have a pixel size of 600 × 600 and spatial resolution of 0.5 m/pixel. A selection of 8000 images from this dataset was randomly chosen for the purpose of training, while the remaining 2000 images were used for testing. In addition, a validation set consisting of five arbitrary images from each category was established.

To verify the generalization of the proposed method, we further adapted the trained model to the real-world images of Gaofen-1 and Gaofen-2 satellites. We downsampled HR images through bicubic operations to obtain LR images. Two mainstream metrics, namely peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM), were calculated on the Y channel of the YCbCr space for objective evaluation. They are formulated as

$$PSNR(I_{SR}, I_{HR}) = 10 \cdot \log 10 \times \left(\frac{L^2}{\frac{1}{N}\sum_{i=1}^{N}(I_{SR}(i) - I_{HR}(i))^2} \right), \tag{17}$$

where L represents the maximum pixel, and N denotes the number of all pixels in I_{SR} and I_{HR}.

$$SSIM(x, y) = \frac{2u_x u_y + k_1}{u_x^2 + u_y^2 + k_1} \cdot \frac{\sigma_{xy} + k_2}{\sigma_x^2 + \sigma_y^2 + k_2}, \tag{18}$$

where x and y represent two images. σ_{xy} symbolizes the covariance between x and y. u and σ^2 represent the average value and variance. k_1 and k_2 denote constant relaxation terms. Multi-adds and model parameters were utilized to evaluate the computational

complexity [32,54]. In addition, the natural image quality evaluator (NIQE) was adopted to validate the reconstruction of real-world images from Gaofen-1 and Gaofen-2 satellites [55].

4.2. Implementation Details

We conducted experiments on remote sensing image data with scale factors of $\times 2$, $\times 3$, and $\times 4$. During training, we randomly cropped 48×48 patches from LR images and extracted ground-truth references from corresponding HR images. We also employed horizontal flipping and random rotation (90°, 180° and 270°) to augment training samples. Table 1 presents the comprehensive hyperparameter setting of the cross-scale enhancement transformer (CSET) module.

Table 1. Parameter setting of the CSET module in the HSTNet.

	Heads	Head Dim	Hidden Size D	MLP Dim	Layers
Transformer Encoder	6	32	512	512	8
Transformer Decoder	6	32	512	512	1

We adopted the Adam optimizer [56] to train the HSTNet with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$. The initial learning rate was set to 10^{-4}, and the batch size was 16. The proposed model was trained for 800 epochs, and the learning rate decreased by half after 400 epochs. Both the training and testing stages were performed using the PyTorch framework, utilizing CUDAToolkit 11.4, cuDNN 8.2.2, Python 3.7, and two NVIDIA 3090 Ti GPUs.

4.3. Comparison with Other Methods

To verify the effectiveness of the proposed HSTNet, we conducted comparative experiments with some state-of-the-art (SOTA) competitors, namely SC [12], SRCNN [22], FSRCNN [57], VDSR [24], LGCNet [30], DCM [31], CTNet [48], ESRT [40], ACT [41], and TransENet [14]. Among these methods, SC [12], SRCNN [22], FSRCNN [57], VDSR [24], ESRT [40], and ACT [41] are the methods proposed for natural image SR. LGCNet [30], DCM [31], CTNet [48], and TransENet [14] are designed for RSISR. The experimental results for the UCMerced dataset and AID dataset with the scale factors of $\times 2$, $\times 3$ and $\times 4$ are reported in Table 2.

4.3.1. Quantitative Evaluation

Evaluation with UCMerced dataset: Table 2 shows that the proposed HSTNet achieves first place among competitors for the UCMerced dataset for all scale factors. Specifically, the HSTNet improves the PSNR comparatively by 0.71 dB, 0.54 dB, and 0.60 dB for scale factor $\times 2$ for LGCNet [30], DCM [31] and CTNet [48], respectively. The average PSNR values of the proposed HSTNet over the second-best TransENet that employs a transformer module are 0.16 dB, 0.15 dB and 0.12 dB when the scale factors are $\times 2$, $\times 3$ and $\times 4$, respectively. Additionally, the HSTNet outperforms LGCNet [30], DCM [31], and CTNet [48] in terms of SSIM by 0.0183, 0.0027, and 0.0102 for scale factor $\times 3$. Compared to ACT [41], which also uses a transformer structure, the average PSNR obtained by the proposed method increased by 0.31 dB, 0.27 dB, and 0.35 dB at scale factors of $\times 2$, $\times 3$ and $\times 4$, respectively. Moreover, Table 3 lists the mean PSNR of different methods on all 21 classes (All these 21 classes of UCMerced dataset: 1—Agricultural, 2—Airplane, 3—Baseballdiamond, 4—Beach, 5—Buildings, 6—Chaparral, 7—Denseresidential, 8—Forest, 9—Freeway, 10—Golfcourse, 11—Harbor, 12—Intersection, 13—Mediumresidential, 14—Mobilehomepark, 15—Overpass, 16—Parkinglot, 17—River, 18—Runway, 19—Sparseresidential, 20—Storagetanks, and 21—Tenniscourt) of the UCMerced dataset when the scale factor is $\times 3$. One can see that the proposed HSTNet performs best in 14 scene classes, ranks second in 5 scene classes, and third in 2 scene classes. The DCM [31] obtains the best PSNR in the other seven categories. It is worth mentioning that the HSTNet shows more effective performance in some scenes comprising prominent contours and rich edges, such as "Baseballdiamond",

"Buildings", and "Overpass". Overall, the mean PSNR in all 21 class scenes of the proposed HSTNet is 0.55 dB higher than DCM [31].

Table 2. Comparative results for the UCMerced dataset and AID dataset. The best and the second-best results are marked in red and blue, respectively.

Method	Scale	UCMerced Dataset		AID Dataset	
		PSNR	SSIM	PSNR	SSIM
Bicubic	×2	30.76	0.8789	32.39	0.8906
SC [12]	×2	32.77	0.9166	32.77	0.9166
SRCNN [22]	×2	32.84	0.9152	34.49	0.9286
FSRCNN [57]	×2	33.18	0.9196	34.11	0.9228
VDSR [24]	×2	33.47	0.9234	35.05	0.9346
LGCNet [30]	×2	33.48	0.9235	34.80	0.9320
DCM [31]	×2	33.65	0.9274	35.21	0.9366
CTNet [48]	×2	33.59	0.9255	35.13	0.9354
ESRT [40]	×2	33.70	0.9270	35.15	0.9358
ACT [41]	×2	33.88	0.9283	35.17	0.9362
TransENet [14]	×2	34.03	0.9301	35.28	0.9374
Ours	×2	34.19	0.9338	35.35	0.9387
Bicubic	×3	27.46	0.7631	29.08	0.7863
SC [12]	×3	28.26	0.7971	28.26	0.7671
SRCNN [22]	×3	28.66	0.8038	30.55	0.8372
FSRCNN [57]	×3	29.09	0.8167	30.30	0.8302
VDSR [24]	×3	29.34	0.8263	31.15	0.8522
LGCNet [30]	×3	29.28	0.8238	30.73	0.8417
DCM [31]	×3	29.52	0.8394	31.31	0.8561
CTNet [48]	×3	29.44	0.8319	31.16	0.8527
ESRT [40]	×3	29.52	0.8318	31.34	0.8562
ACT [41]	×3	29.80	0.8395	31.39	0.8579
TransENet [14]	×3	29.92	0.8408	31.45	0.8595
Ours	×3	30.07	0.8421	31.61	0.8613
Bicubic	×4	25.65	0.6725	27.30	0.7036
SC [12]	×4	26.51	0.7152	26.51	0.7152
SRCNN [22]	×4	26.78	0.7219	28.40	0.7561
FSRCNN [57]	×4	26.93	0.7267	28.03	0.7387
VDSR [24]	×4	27.11	0.7360	28.99	0.7753
LGCNet [30]	×4	27.02	0.7333	28.61	0.7626
DCM [31]	×4	27.22	0.7528	29.17	0.7824
CTNet [48]	×4	27.41	0.7512	29.00	0.7768
ESRT [40]	×4	27.41	0.7485	29.18	0.7831
ACT [41]	×4	27.54	0.7531	29.19	0.7836
TransENet [14]	×4	27.77	0.7630	29.38	0.7909
Ours	×4	27.89	0.7694	29.57	0.7983

Evaluation with AID dataset: Table 2 reports the averaged evaluation results of the proposed method in comparison to other methods for AID datasets for scale factors of ×2, ×3, and ×4. One can see that the proposed HSTNet outperforms SRCNN [22], FSRCNN [57], and VDSR [24] by 1.17 dB, 1.54 dB, and 0.58 dB for scale factors ×4 in terms of PSNR values. It proves that the HSTNet ranks first with PSNR scores that are higher than LGCNet [30] by 0.55 dB, 0.88 dB, and 0.96 dB for scale factors ×2, ×3, and ×4, respectively. Compared to ESRT [40], which adopts a transformer structure, the average PSNR obtained by the proposed method increased by 0.20 dB, 0.27 dB, and 0.39 dB at scale factors of ×2, ×3, and ×4, respectively. Compared to the second-best method, TransENet [14], the HSTNet achieves a performance improvement of 0.16 dB and 0.0013 in PSNR and SSIM

scores, respectively, for scale factor ×3. In contrast to the UCMerced dataset, the AID dataset comprises 30 categories of scenes and a significantly larger number of images. Table 4 reports a detailed performance comparison of different methods for scale factor ×4 on all 30 scene classes (All these 30 classes of AID dataset: 1—Airport, 2—Bareland, 3—Baseballdiamond, 4—Beach, 5—Bridge, 6—Center, 7—Church, 8—Commercial, 9—Denseresidential, 10—Desert, 11—Farmland, 12—Forest, 13—Industrial, 14—Meadow, 15—Mediumresidential, 16—Mountain, 17—Park, 18—Parking, 19—Playground, 20—Pond, 21—Port, 22—Railwaystation, 23—Resort, 24—River, 25—School, 26—Sparseresidential, 27—Square, 28—Stadium, 29—Storagetanks, 30—Viaduct) of the AID dataset. It can be seen that the proposed HSTNet outperforms the other methods in 28 scene classes, while TransENet [14] obtains the best PSNR scores in the remaining 2 categories. Although the HSTNet ranks second in those two scene classes, its PSNR values are very close to the TransENet [14]. Notably, the HSTNet has an overall average PSNR that is 0.48 dB higher than TransENet [14].

4.3.2. Qualitative Evaluation

To further verify the advantages of the proposed method, the subjective results of SR images reconstructed by the aforementioned methods are shown in Figures 6 and 7. Figure 6 shows the reconstruction results of the above methods for the UCMerced dataset by taking "airplane" and "runway" scenes as examples. Figure 7 shows the visual results of the "stadium" and "medium-residential" scenes in the AID dataset. In general, the SR results reconstructed by the proposed method possess sharper edges and clearer contours compared with other methods, which verifies the effectiveness of the HSTNet.

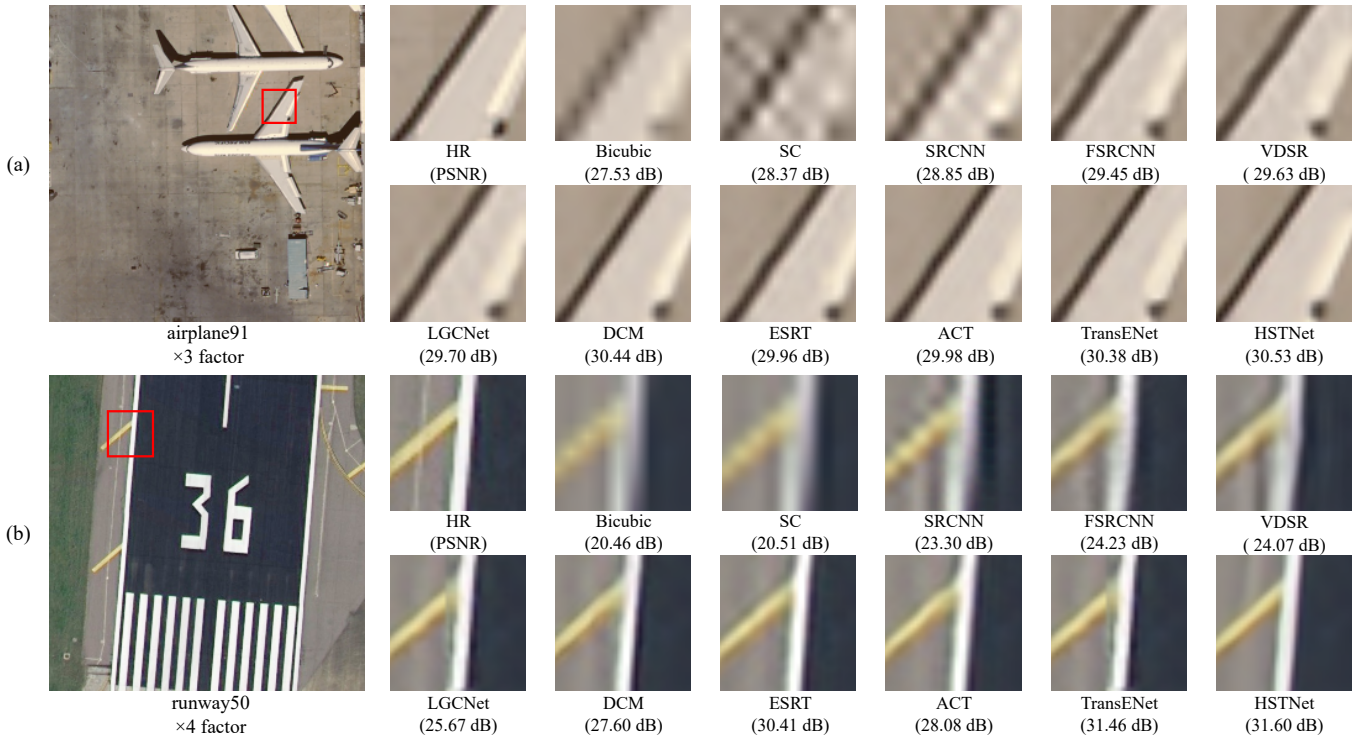


Figure 6. Subjective results for UCMerced dataset: (**a**) "Airplane91" scene with ×3 factor. (**b**) "Runway50" scene with ×4 factor.

Table 3. Average PSNR of per-category for UCMerced dataset with the scale factor of ×3. The best and the second-best results are marked in red and blue, respectively.

Class	Bicubic	SC [12]	SRCNN [22]	FSRCNN [57]	LGCNet [30]	DCM [31]	CTNet [48]	ESRT [40]	ACT [41]	TransENet [14]	Ours
1	26.86	27.23	27.47	27.61	27.66	29.06	28.53	28.13	27.86	28.02	27.93
2	26.71	27.67	28.24	28.98	29.12	30.77	29.22	29.45	29.78	29.94	29.98
3	33.33	34.06	34.33	34.64	34.72	33.76	34.81	34.88	35.05	35.04	35.13
4	36.14	36.87	37.00	37.21	37.37	36.38	37.38	37.45	37.55	37.53	37.76
5	25.09	26.11	26.84	27.50	27.8 1	28.51	27.99	28.18	28.66	28.81	29.12
6	25.21	25.82	26.11	26.21	26.39	26.81	26.40	26.43	26.62	26.69	26.78
7	25.76	26.75	27.41	28.02.	28.25	28.79	28.42	28.53	28.97	29.11	29.27
8	27.53	28.09	28.24.	28.35	28.44	28.16	28.48	28.47	28.56	28.59	28.65
9	27.36	28.28	28.69	29.27	29.52	30.45	29.60	29.87	30.25	30.38	30.65
10	35.21	35.92	36.15	36.43	36.51	34.43	36.46	36.54	36.63	36.68	36.69
11	21.25	22.11	22.82	23.29	23.63	26.55	23.83	23.87	24.42	24.72	24.91
12	26.48	27.20	27.67	28.06	28.29	29.28	28.38	28.53	28.85	29.03	29.32
13	25.68	26.54	27.06	27.58	27.76	27.21	27.87	27.93	28.30	28.47	28.64
14	22.25	23.25	23.89	24.34	24.59	26.05	24.87	24.92	25.32	25.64	25.74
15	24.59	25.30	25.65	26.53	26.58	27.77	26.89	27.17	27.76	27.83	28.31
16	21.75	22.59	23.11	23.34	23.69	24.95	23.59	23.72	24.11	24.45	24.53
17	28.12	28.71	28.89	29.07	29.12	28.89	29.11	29.14	29.28	29.25	29.32
18	29.30	30.25	30.61	31.01	31.15	32.53	30.60	30.98	31.21	31.25	31.21
19	28.34	29.33	29.40	30.23	30.53	29.81	31.25	31.35	31.55	31.57	31.71
20	29.97	30.86	31.33	31.92	32.17	29.02	32.29	32.42	32.74	32.71	32.98
21	29.75	30.62	30.98	31.34	31.58	30.76	31.74	31.99	32.40	32.51	32.77
AVG	27.46	28.23	28.66	29.09	29.28	29.52	29.41	29.52	29.80	29.92	30.07

Table 4. Average PSNR of per-category for AID dataset with the scale factor of ×4. The best and the second-best results are marked in red and blue, respectively.

Class	Bicubic	SRCNN [22]	FSRCNN [57]	VDSR [24]	LGCNet [30]	DCM [31]	CTNet [48]	ESRT [40]	ACT [41]	TransENet [14]	Ours
1	27.03	28.17	27.70	28.82	28.39	28.99	28.80	28.98	29.01	29.23	29.29
2	34.88	35.63	35.73	35.98	35.78	36.17	36.12	36.15	36.15	36.20	36.45
3	29.06	30.51	29.89	31.18	30.75	31.36	31.15	31.35	31.37	31.59	31.69
4	31.07	31.92	31.79	32.29	32.08	32.45	32.40	32.47	32.45	32.55	32.61
5	28.98	30.41	29.83	31.19	30.67	31.39	31.17	31.42	31.42	31.63	31.75
6	25.26	26.59	25.96	27.48	26.92	27.72	27.48	27.73	27.75	28.03	28.23
7	22.15	23.41	22.74	24.12	23.68	24.29	24.10	24.29	24.32	24.51	24.56
8	25.83	27.05	26.65	27.62	27.24	27.78	27.63	27.78	27.79	27.97	28.06
9	23.05	24.13	23.69	24.70	24.33	24.87	24.70	24.88	24.89	25.13	25.32
10	38.49	38.84	38.84	39.13	39.06	39.27	39.25	39.25	39.24	39.31	39.45
11	32.30	33.48	32.95	34.20	33.77	34.42	34.25	34.41	34.43	34.58	34.59
12	27.39	28.15	28.19	28.36	28.20	28.47	28.47	28.53	28.47	28.56	28.76
13	24.75	26.00	25.49	26.72	26.24	26.92	26.71	26.93	26.94	27.21	27.19
14	32.06	32.57	32.50	32.77	32.65	32.88	32.84	32.89	32.87	32.94	33.26
15	26.09	27.37	26.84	28.06	27.63	28.25	28.06	28.25	28.25	28.45	28.54
16	28.04	28.90	28.70	29.11	28.97	29.18	29.15	29.20	29.18	29.28	29.42
17	26.23	27.25	26.98	27.69	27.37	27.82	27.69	27.84	27.84	28.01	28.34
18	22.33	24.01	23.47	25.21	24.40	25.74	25.27	25.80	25.75	26.40	26.38
19	27.27	28.72	28.09	29.62	29.04	29.92	29.66	29.96	29.96	30.30	30.52
20	28.94	29.85	29.50	30.26	30.00	30.39	30.25	30.39	30.38	30.53	30.79
21	24.69	25.82	25.40	26.43	26.02	26.62	26.41	26.62	26.61	26.91	27.18
22	26.31	27.55	27.12	28.19	27.76	28.38	28.19	28.40	28.40	28.61	28.76
23	25.98	27.12	26.77	27.71	27.32	27.88	27.72	27.90	27.89	28.08	28.22
24	29.61	30.48	30.22	30.82	30.60	30.91	30.83	30.92	30.92	31.00	31.27
25	24.91	26.13	25.66	26.78	26.34	26.94	26.75	26.96	26.99	27.22	27.43
26	25.41	26.16	25.88	26.46	26.27	26.53	26.46	26.55	26.54	26.63	26.87
27	26.75	28.13	27.62	28.91	28.39	29.13	28.94	29.17	29.15	29.39	29.72
28	24.81	26.10	25.50	26.88	26.37	27.10	26.86	27.14	27.10	27.41	27.68
29	24.18	25.27	24.73	25.86	25.48	26.00	25.82	26.01	26.02	26.20	26.43
30	25.86	27.03	26.54	27.74	27.26	27.93	27.67	27.92	27.95	28.21	28.48
AVG	27.3	28.4	28.03	28.99	28.61	29.17	29.03	29.18	29.19	29.38	29.57

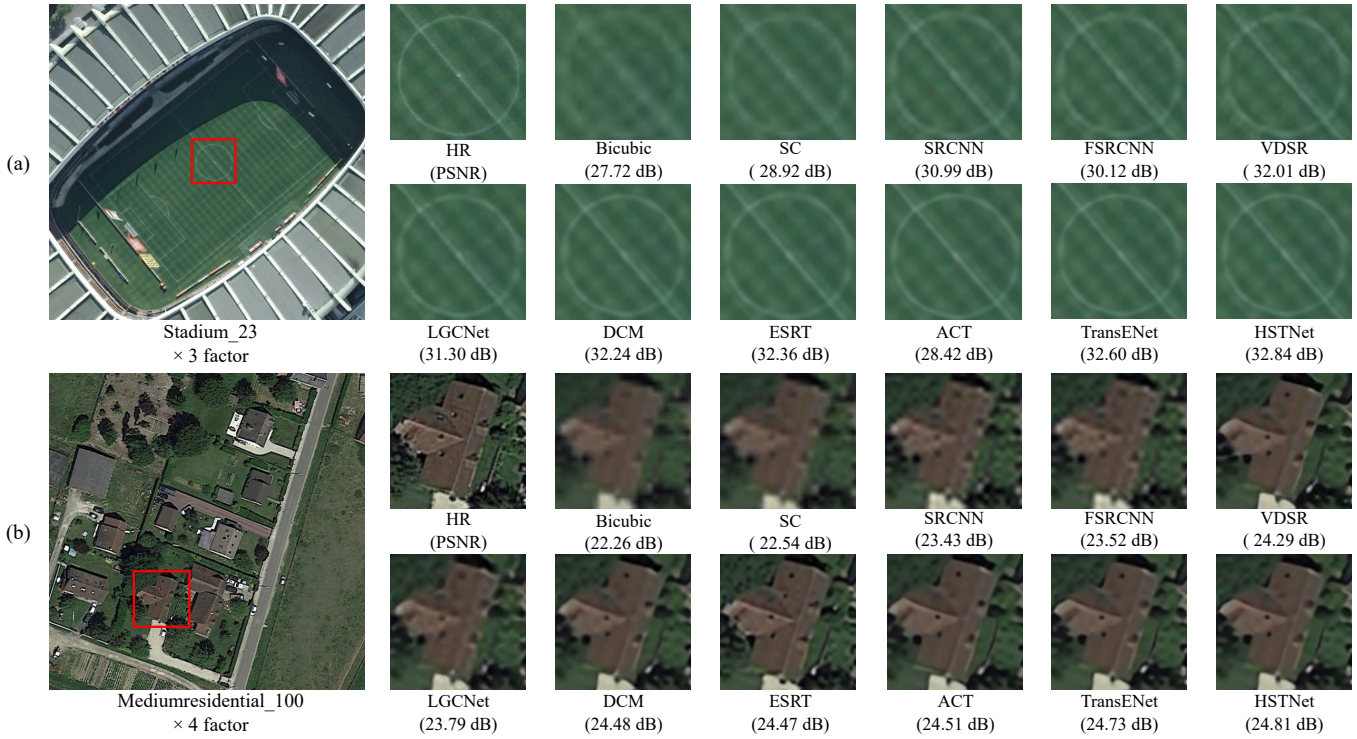


Figure 7. Subjective results for AID dataset: (**a**) "Stadium_23" scene with ×3 factor. (**b**) "Mediumresidential_100" scene with ×4 factor.

4.4. Results on Real Remote Sensing Data

Real images acquired by GaoFen-1 (GF-1) and GaoFen-2 (GF-2) satellites were employed to assess the robustness of the HSTNet. The spatial resolutions of GF-1 and GF-2 are 8 and 3.2 m/pixel, respectively. Three visible bands are selected from GF-1 and GF-2 satellite images to generate the LR inputs. The pre-trained DCM [31], ACT [41], and the proposed HSTNet models for the UCMerced dataset are utilized for SR image reconstruction. Figures 8 and 9 demonstrate the reconstruction results of the aforementioned methods on real data in some common scenes including river, factory, overpass, and paddy fields. One can see that the proposed HSTNet can obtain favorable results. Compared with DCM [31] and ACT [41], the reconstructed images of the proposed HSTNet achieved the lowest NIQE scores in all the four common scenes. Although the pixel size of these input images is different from the LR images in the training set, which are 600 × 600 and 256 × 256 for real-world images and training images, respectively, the HSTNet can still achieve good results in terms of visual perception qualities. It verifies the robustness of the proposed HSTNet.

4.5. Ablation Studies

Ablation studies with the scale factor of ×4 were conducted on the UCMerced dataset to demonstrate the effectiveness of the proposed fundamental modules in the HSTNet model.

4.5.1. Ablation Studies on the LFE Module

Number of LFE and HSFE modules: Table 5 presents a comparative analysis of varying quantities of LFE and HSFE modules. It indicates that when adopting two LFE and 2 HSFE modules, the model has the smallest number of parameters and computation, but the model has the lowest PSNR and SSIM values. The results indicate that the proposed HSTNet achieves the highest PSNR and SSIM when utilizing three LFE and five HSFE modules. When employing three LFE and eight HSFE modules, the model has the largest number of parameters and computation, and its performance is not optimal. Therefore,

considering the performance of the model and the computational complexity, we adopted three LFE and five HSFE modules in the proposed method. The results confirm the effectiveness of the LFE and HSFE modules in the proposed model, as well as the rationality of the number of LFE and HSFE modules.

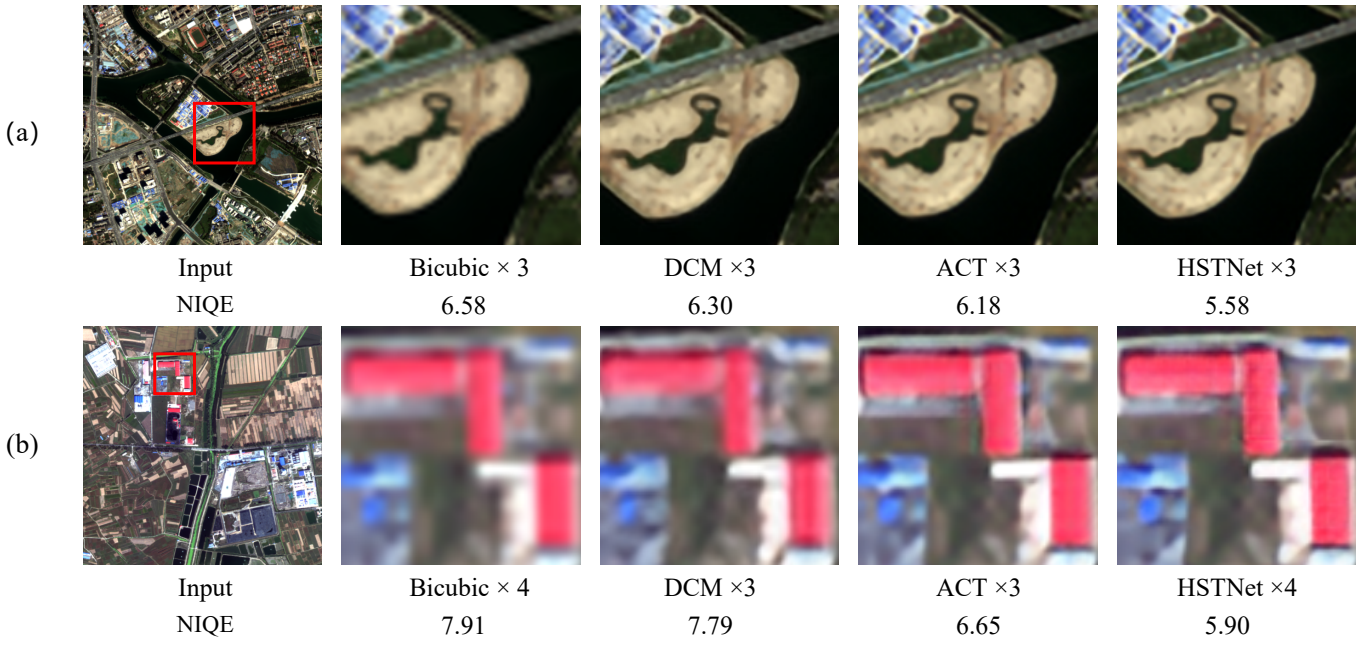


Figure 8. Subjective results on real GaoFen-1 satellite data: (**a**) "River" with ×3 factor. (**b**) "Factory" with ×4 factor.

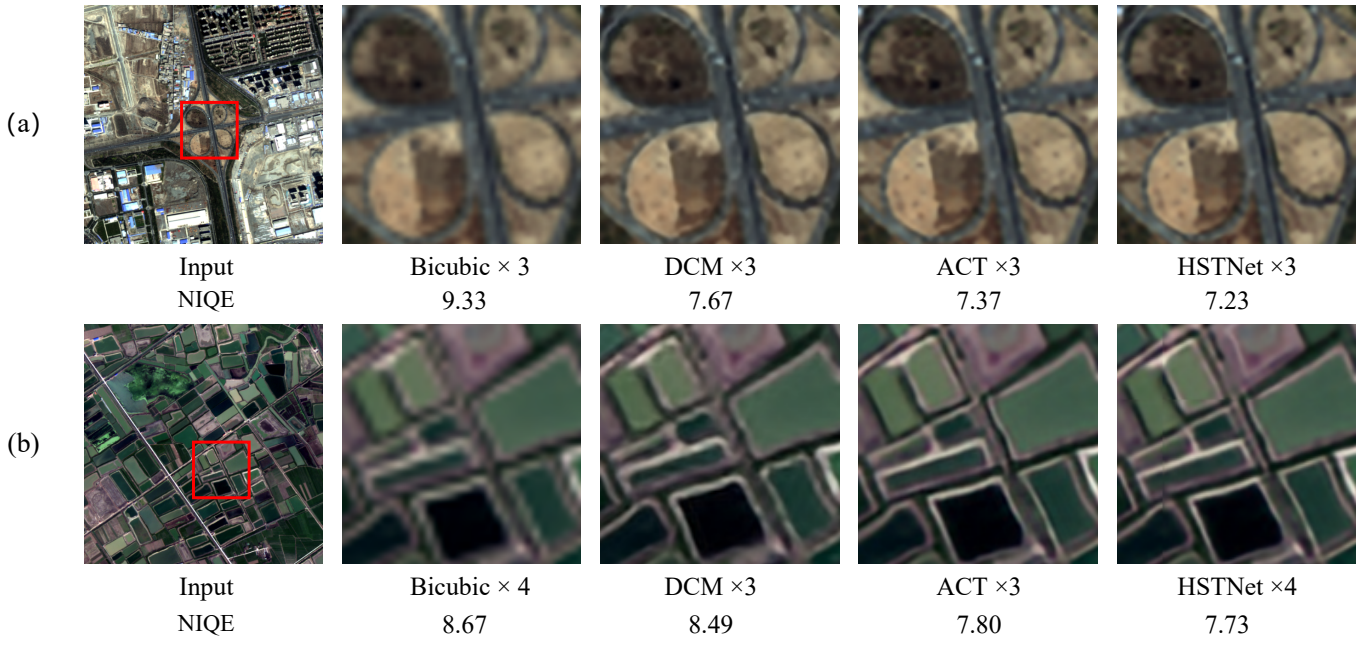


Figure 9. Subjective results on real GaoFen-2 satellite data: (**a**) "Overpass" with ×3 factor. (**b**) "Paddy fields" with ×4 factor.

Effects of the HSFE module: We devised the HSFE module in the proposed LFE module to exploit the recursive information inherent in the image. We conducted further ablation studies by substituting the HSFE module with widely used feature extraction modules in SR algorithms, namely RCAB [27], CTB [48], CB [58], and SSEM [45] to validate the effectiveness of the HSFE module. Among them, SSEM [45] is also used to mine

scale information. As presented in Table 6, the HSFE module outperforms the other feature extraction modules in terms of PSNR and SSIM, demonstrating its effectiveness in feature extraction. Meanwhile, it is also competitive in terms of parameter quantity and computational complexity.

Table 5. Ablation analysis of the number of LFE and HSFE modules (the best result is highlighted in bold).

Scale	Numbers of LFE	Numbers of HSFE	PSNR	SSIM	Params	Multi-Adds
×4	2	2	27.57	0.7546	30.2M	73.6G
×4	2	5	27.72	0.7603	31.9M	135.9G
×4	2	8	27.61	0.7566	33.6M	205.1G
×4	3	2	27.58	0.7542	40.8M	95.5G
×4	3	5	**27.89**	**0.7694**	43.4M	194.4G
×4	3	8	27.73	0.7608	46.0M	292.8G

Table 6. Ablation analysis of different feature extraction modules in LFE module (the best result is highlighted in bold).

Scale	RCAB	CTB	CB	SSEM	HSFE	PSNR	SSIM	Params	Multi-Adds
×4	✓	✗	✗	✗	✗	26.33	0.7010	41.2M	112.0G
×4	✗	✓	✗	✗	✗	27.36	0.7451	40.3M	75.1G
×4	✗	✗	✓	✗	✗	27.51	0.7510	45.7M	275.2G
×4	✗	✗	✗	✓	✗	27.61	0.7561	42.5M	160.0G
×4	✗	✗	✗	✗	✓	**27.89**	**0.7694**	43.4M	194.4G

4.5.2. Ablation Studies on the CSET Module

Number of CSET modules: The CSET module is designed to learn the dependency relationship across long distances between features of different dimensions. To validate the effectiveness of the proposed CSET modules, we conducted ablation experiments using varying numbers of CSET modules. Table 7 proves that the configuration of three CSET modules performs the best in terms of PSNR and SSIM. The features of low-dimension space are transmitted more to the high-dimension space, reducing the difficulty of optimization and facilitating the convergence of the deep model. The aforementioned results demonstrate the effectiveness of the CSET module in enhancing the representation of high-dimensional features.

Effects of the CSTA block: The CSTA [41] block is introduced to enable the CSET module to utilize the recurrent patch information of different scales in the input image. To verify the effectiveness of the CSTA module, we analyzed the composition of the transformer. Table 8 presents the comparative results of two different transformers. It proves that the CSTA block is beneficial to improve the performance of the HSTNet.

Table 7. Ablation analysis of different feature extraction modules in the LFE module (the best result is highlighted in bold).

Scale	Transformer-3	Transformer-2	Transformer-1	Transformer-0	PSNR	SSIM
×4	✗	✗	✗	✗	27.54	0.7522
×4	✓	✗	✗	✗	27.61	0.7562
×4	✓	✓	✗	✗	27.73	0.7618
×4	✓	✓	✓	✗	**27.89**	**0.7694**
×4	✓	✓	✓	✓	27.50	0.7509

Table 8. Ablation analysis of the CSTA block. The best performances are highlighted in **bold**.

Transformer	PSNR	SSIM
MHSA + FFN	27.77	0.7630
MHSA + FFN + CSTA	**27.89**	**0.7694**

5. Conclusions and Future Work

In this paper, we present a hybrid-scale hierarchical transformer network (HSTNet) for remote sensing image super-resolution (RSISR). The HSTNet contains two crucial components, i.e., a hybrid-scale feature exploitation (HSFE) module and a cross-scale enhancement transformer (CSET) module. Specifically, the HSFE module with two branches was built to leverage the internal recurrence of information both in single and cross scales within the images. Meanwhile, the CSET module was built to capture long-range dependencies and effectively mine the correlation between high-dimension and low-dimension features. Experimental results on two challenging remote sensing datasets verified the effectiveness and superiority of the proposed HSTNet. In the future, more efforts are expected to simplify the network architecture and design a more effective low-dimensional feature extraction branch to improve RSISR performance.

Author Contributions: Conceptualization, J.S., M.G. and G.J.; methodology, J.S. and M.G.; software, J.S. J.P., and G.Z.; validation, J.S. Q.L. and M.G.; formal analysis, J.S. and M.G.; investigation, J.S. and Q.L.; resources, M.G. and J.S.; writing, J.S. and Q.L.; supervision, M.G. and G.J.; project administration, J.S., M.G. and G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the Natural Science Foundation of Shandong Province of China (ZR2022MF307) and the National Natural Science Foundation of China (Nos. 61601266 and 61801272).

Data Availability Statement: Not applicable.

Acknowledgments: This work is supported in part by the Natural Science Foundation of Shandong Province of China (ZR2022MF307) and the National Natural Science Foundation of China (Nos.61601266 and 61801272).

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

References

1. Harrie, L.; Oucheikh, R.; Nilsson, Å.; Oxenstierna, A.; Cederholm, P.; Wei, L.; Richter, K.F.; Olsson, P. Label Placement Challenges in City Wayfinding Map Production—Identification and Possible Solutions. *J. Geovisualization Spat. Anal.* **2022**, *6*, 16. [CrossRef]
2. Kokila, S.; Jayachandran, A. Hybrid Behrens-Fisher- and Gray Contrast–Based Feature Point Selection for Building Detection from Satellite Images. *J. Geovisualization Spat. Anal.* **2023**, *7*, 8. [CrossRef]
3. Shen, H.; Zhang, L.; Huang, B.; Li, P. A MAP Approach for Joint Motion Estimation, Segmentation, and Super Resolution. *IEEE Trans. Image Process.* **2007**, *16*, 479–490. [CrossRef] [PubMed]
4. Köhler, T.; Huang, X.; Schebesch, F.; Aichert, A.; Maier, A.K.; Hornegger, J. Robust Multiframe Super-Resolution Employing Iteratively Re-Weighted Minimization. *IEEE Trans. Comput. Imaging* **2016**, *2*, 42–58. [CrossRef]
5. Zhang, L.; Wu, X. An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE Trans. Image Process.* **2006**, *15*, 2226–2238. [CrossRef]
6. Hung, K.W.; Siu, W.C. Robust Soft-Decision Interpolation Using Weighted Least Squares. *IEEE Trans. Image Process.* **2012**, *21*, 1061–1069. [CrossRef]
7. Lu, X.; Yuan, H.; Yuan, Y.; Yan, P.; Li, L.; Li, X. Local learning-based image super-resolution. In Proceedings of the 2011 IEEE 13th International Workshop on Multimedia Signal Processing, Hangzhou, China, 17–19 October 2011; pp. 1–5.
8. Zhang, K.; Gao, X.; Tao, D.; Li, X. Single Image Super-Resolution With Non-Local Means and Steering Kernel Regression. *IEEE Trans. Image Process.* **2012**, *21*, 4544–4556. [CrossRef]
9. Schulter, S.; Leistner, C.; Bischof, H. Fast and accurate image upscaling with super-resolution forests. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3791–3799.
10. Wang, L.; Guo, Y.; Liu, L.; Lin, Z.; Deng, X.; An, W. Deep Video Super-Resolution Using HR Optical Flow Estimation. *IEEE Trans. Image Process.* **2020**, *29*, 4323–4336. [CrossRef]
11. Chang, K.; Ding, P.L.K.; Li, B. Single image super-resolution using collaborative representation and non-local self-similarity. *Signal Process.* **2018**, *149*, 49–61. [CrossRef]

12. Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. Image Super-Resolution Via Sparse Representation. *IEEE Trans. Image Process.* **2010**, *19*, 2861–2873. [CrossRef]
13. Li, Y.; Sixou, B.; Peyrin, F. A review of the deep learning methods for medical images super resolution problems. *Irbm* **2021**, *42*, 120–133. [CrossRef]
14. Lei, S.; Shi, Z.; Mo, W. Transformer-based Multi-Stage Enhancement for Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *60*, 1–11.
15. Buades, A.; Coll, B.; Morel, J.M. A non-local algorithm for image denoising. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 60–65.
16. Xu, J.; Zhang, L.; Zuo, W.; Zhang, D.; Feng, X. Patch Group Based Nonlocal Self-Similarity Prior Learning for Image Denoising. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 244–252.
17. Michaeli, T.; Irani, M. Blind Deblurring Using Internal Patch Recurrence. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
18. Freedman, G.; Fattal, R. Image and video upscaling from local self-examples. *ACM Trans. Graph.* **2011**, *30*, 12:1–12:11. [CrossRef]
19. Yang, J.; Lin, Z.L.; Cohen, S.D. Fast Image Super-Resolution Based on In-Place Example Regression. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 1059–1066.
20. Shocher, A.; Cohen, N.; Irani, M. "Zero-Shot" Super-Resolution Using Deep Internal Learning. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
21. Pan, Z.; Yu, J.; Huang, H.; Hu, S.; Zhang, A.; Ma, H.; Sun, W. Super-Resolution Based on Compressive Sensing and Structural Self-Similarity for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 4864–4876. [CrossRef]
22. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 184–199.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
24. 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.
25. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced Deep Residual Networks for Single Image Super-Resolution. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1132–1140.
26. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2472–2481.
27. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
28. Dai, T.; Cai, J.; Zhang, Y.; Xia, S.T.; Zhang, L. Second-order attention network for single image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11065–11074.
29. Li, Z.; Yang, J.; Liu, Z.; Yang, X.; Jeon, G.; Wu, W. Feedback network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3867–3876.
30. 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]
31. Haut, J.M.; Paoletti, M.E.; Fernández-Beltran, R.; Plaza, J.; Plaza, A.J.; Li, J. Remote Sensing Single-Image Superresolution Based on a Deep Compendium Model. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1432–1436. [CrossRef]
32. Wang, X.; Wang, Q.; Zhao, Y.; Yan, J.; Fan, L.; Chen, L. Lightweight Single-Image Super-Resolution Network with Attentive Auxiliary Feature Learning. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020.
33. Ni, N.; Wu, H.; Zhang, L. Hierarchical Feature Aggregation and Self-Learning Network for Remote Sensing Image Continuous-Scale Super-Resolution. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
34. Wang, Z.; Zhao, Y.; Chen, J. Multi-Scale Fast Fourier Transform Based Attention Network for Remote-Sensing Image Super-Resolution. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 2728–2740. [CrossRef]
35. Liang, G.M.; KinTak, U.; Yin, H.; Liu, J.; Luo, H. Multi-scale hybrid attention graph convolution neural network for remote sensing images super-resolution. *Signal Process.* **2023**, *207*, 108954. [CrossRef]
36. Wang, Y.; Shao, Z.; Lu, T.; Wu, C.; Wang, J. Remote Sensing Image Super-Resolution via Multiscale Enhancement Network. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
37. Yang, F.; Yang, H.; Fu, J.; Lu, H.; Guo, B. Learning Texture Transformer Network for Image Super-Resolution. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5790–5799.
38. 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.

39. Fang, J.; Lin, H.; Chen, X.; Zeng, K. A Hybrid Network of CNN and Transformer for Lightweight Image Super-Resolution. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–24 June 2022; pp. 1102–1111.
40. Lu, Z.; Li, J.; Liu, H.; Huang, C.; Zhang, L.; Zeng, T. Transformer for Single Image Super-Resolution. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–24 June 2022; pp. 456–465.
41. Yoo, J.; Kim, T.; Lee, S.; Kim, S.; Lee, H.S.; Kim, T.H. Enriched CNN-Transformer Feature Aggregation Networks for Super-Resolution. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023; pp. 4945–4954.
42. Ye, C.; Yan, L.; Zhang, Y.; Zhan, J.; Yang, J.; Wang, J. A Super-resolution Method of Remote Sensing Image Using Transformers. In Proceedings of the 2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Online, 22–25 September 2021; Volume 2, pp. 905–910.
43. Tu, J.; Mei, G.; Ma, Z.; Piccialli, F. SWCGAN: Generative Adversarial Network Combining Swin Transformer and CNN for Remote Sensing Image Super-Resolution. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 5662–5673. [CrossRef]
44. He, J.; Yuan, Q.; Li, J.; Xiao, Y.; Liu, X.; Zou, Y. DsTer: A dense spectral transformer for remote sensing spectral super-resolution. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *109*, 102773. [CrossRef]
45. Lei, S.; Shi, Z. Hybrid-Scale Self-Similarity Exploitation for Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–10. [CrossRef]
46. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
47. Wang, X.; Girshick, R.B.; Gupta, A.K.; He, K. Non-local Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
48. Wang, S.; Zhou, T.; Lu, Y.; Di, H. Contextual Transformation Network for Lightweight Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *60*, 1–13. [CrossRef]
49. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. *arXiv* **2017**, arXiv: abs/1706.03762.
50. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* **2016**, arXiv:1606.08415.
51. Qin, M.; Mavromatis, S.; Hu, L.; Zhang, F.; Liu, R.; Sequeira, J.; Du, Z. Remote Sensing Single-Image Resolution Improvement Using A Deep Gradient-Aware Network with Image-Specific Enhancement. *Remote. Sens.* **2020**, *12*, 758. [CrossRef]
52. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the ACM SIGSPATIAL International Workshop on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010.
53. 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.* **2016**, *55*, 3965–3981. [CrossRef]
54. Muqeet, A.; Hwang, J.; Yang, S.; Kang, J.H.; Kim, Y.; Bae, S.H. Multi-attention Based Ultra Lightweight Image Super-Resolution. In Proceedings of the ECCV Workshops, Glasgow, UK, 23–28 August 2020.
55. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Process. Lett.* **2013**, *20*, 209–212. [CrossRef]
56. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
57. Dong, C.; Loy, C.C.; Tang, X. Accelerating the Super-Resolution Convolutional Neural Network. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
58. Zhang, D.; Shao, J.; Li, X.; Shen, H.T. Remote Sensing Image Super-Resolution via Mixed High-Order Attention Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5183–5196. [CrossRef]