remote sensing

MDPI

Article

An Efficient Hybrid CNN-Transformer Approach for Remote Sensing Super-Resolution

Wenjian Zhang [1,2,3,†], Zheng Tan [1,2,3,†], Qunbo Lv [1,2,3], Jiaao Li [1,2,3], Baoyu Zhu [1,2,3] and Yangyang Liu [1,2,3,*]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, No. 9 Dengzhuang South Road, Haidian District, Beijing 100094, China; zhangwenjian21@mails.ucas.ac.cn (W.Z.); tanzheng@aircas.ac.cn (Z.T.); lvqunbo@aoe.ac.cn (Q.L.); lijiaao21@mails.ucas.ac.cn (J.L.)
[2] School of Optoelectronics, University of Chinese Academy of Sciences, No. 19(A) Yuquan Road, Shijingshan District, Beijing 100049, China
[3] Department of Key Laboratory of Computational Optical Imagine Technology, Chinese Academy of Sciences, No. 9 Dengzhuang South Road, Haidian District, Beijing 100094, China
* Correspondence: liuyy@aircas.ac.cn
† These authors contributed equally to this work.

Abstract: Transformer models have great potential in the field of remote sensing super-resolution (SR) due to their excellent self-attention mechanisms. However, transformer models are prone to overfitting because of their large number of parameters, especially with the typically small remote sensing datasets. Additionally, the reliance of transformer-based SR models on convolution-based upsampling often leads to mismatched semantic information. To tackle these challenges, we propose an efficient super-resolution hybrid network (EHNet) based on the encoder composed of our designed lightweight convolution module and the decoder composed of an improved swin transformer. The encoder, featuring our novel Lightweight Feature Extraction Block (LFEB), employs a more efficient convolution method than depthwise separable convolution based on depthwise convolution. Our LFEB also integrates a Cross Stage Partial structure for enhanced feature extraction. In terms of the decoder, based on the swin transformer, we innovatively propose a sequence-based upsample block (SUB) for the first time, which directly uses the sequence of tokens in the transformer to focus on semantic information through the MLP layer, which enhances the feature expression ability of the model and improves the reconstruction accuracy. Experiments show that EHNet's PSNR on UCMerced and AID datasets obtains a SOTA performance of 28.02 and 29.44, respectively, and is also visually better than other existing methods. Its 2.64 M parameters effectively balance model efficiency and computational demands.

Keywords: remote sensing image super-resolution; convolution neural network; Swin Transformer; efficient hybrid network; sequence-based upsample

Citation: Zhang, W.; Tan, Z.; Lv, Q.; Li, J.; Zhu, B.; Liu, Y. An Efficient Hybrid CNN-Transformer Approach for Remote Sensing Super-Resolution. *Remote Sens.* **2024**, *16*, 880. https://doi.org/10.3390/rs16050880

Academic Editors: Jia Wan, Zhitong Xiong and Jiangbin Zheng

Received: 9 January 2024
Revised: 23 February 2024
Accepted: 28 February 2024
Published: 1 March 2024

1. Introduction

The technique of Single Image Super-Resolution (SISR) employs software algorithms to compensate for lost details in a low-resolution (LR) image, restoring it to a high-resolution (HR) counterpart. This technology has seen extensive application across various fields, notably in video surveillance [1], medical diagnosis [2], and remote sensing [3,4]. In remote sensing, high spatial resolution images are very important in many scenarios, such as target detection [5], change detection [6], and object tracking [7].

Image sensors are the main limiting factor in the spatial resolution of remotely sensed images, and increasing the pixel density of sensors will significantly increase the cost of hardware. The remote sensing image super-resolution (RSISR) reconstruction technique is a method that can obtain high-resolution remote sensing images more efficiently than upgrading imaging equipment to improve image spatial resolution.

Image super-resolution reconstruction is an ill-posed problem, resulting in a scenario where one low-resolution input image can correspond to multiple high-resolution solutions. To overcome this issue, image prior information is typically used to constrain the solution space for HR reconstruction. Deep learning methods, applied to SR tasks in recent years, have shown capabilities to reconstruct images with clearer textures and edges compared to earlier learning-based SR approaches like those based on sparse coding [8] and local linear regression [9]. Super-Resolution Convolutional Neural Network (SRCNN) [10], the first CNN-based image super-resolution method, initiated this trend by learning an end-to-end nonlinear mapping from LR to HR images through a three-layer convolutional network. Since then, numerous CNN-based SR methods have been proposed, emphasizing residual blocks [11,12], dense connections [13,14], and recursive structures [15,16].

Because the receptive field of the convolution kernel is limited, the convolutional neural network can only perceive local information of the image. Researchers used pooling methods to expand the receptive field by building deeper models. However, the process of reducing the resolution of feature maps will lose some information. By fusing feature map information of different resolutions, the receptive field can be expanded while avoiding information loss during pooling. UNet [17], a classic Convolution Neural Network(CNN) architecture typically employed in image segmentation tasks, uses an encoder to extract features and downsample to lower-resolution feature maps, followed by a decoder that incrementally upsamples and merges these features through skip connections. Inspired by the success of the UNet structure in image segmentation, researchers have proposed various UNet variants, such as UNet++ [18] and Attention U-Net [19]. RUNet [20] is the first model to adapt the UNet architecture for image super-resolution tasks.

In recent years, due to the success of transformers [21] in the field of natural language processing, transformers have also attracted great attention in computer vision. The Multi-Head Self-Attention (MSA) mechanism, capable of establishing long-range dependencies and adaptively weighting different positions in a sequence, has proven particularly adept at processing image details and grasping global semantics. Vision Transformer (ViT) [22] is the first pure transformer structure for image recognition, achieving comparable performance to other convolution-based state-of-the-art (SOTA) methods. Following the introduction of ViT, many vision tasks have started to incorporate transformer-based models, including object detection [23] and image segmentation [24,25]. Image Pretrained Transformer (IPT) [26] is the first model to apply transformers to low-level tasks such as image super-resolution and denoising.

For pixel-level visual tasks like image restoration and segmentation, the computation cost of ViT-based models increases significantly with the resolution of the input image. Additionally, ViT usually requires a fixed sequence length, while in practical vision tasks, image sizes are variable. As transformer models discard the inductive biases of CNN, a large amount of training data is typically required to achieve good accuracy with ViT. The swin transformer [27] is a model based on window attention and shifting windows that substantially overcomes these drawbacks. SwinIR [28] is a SOTA method for SISR tasks based on the swin transformer. It outperforms previous models based on both pure convolutional structures and ViT-based architectures in public datasets like DIV2K and DF2K.

Existing algorithms usually choose to increase the number of network layers in order to more adequately extract features from low-resolution images and improve the quality of reconstructed images. However, too many layers may bring other negative effects on the performance of the network, for example, gradient vanishing, network degradation, and overfitting. In recent years, there have been some lightweight super-resolution models such as feature enhancement networks (FeNet) [29] and Omnisr [30], all of which have less computational resource consumption but poor reconstruction quality. Larger network models generate higher-quality super-resolution images but consume a large number of resources. Achieving a balance between reconstruction quality and model complexity is an important goal in image super-resolution research.

This paper presents an Efficient Super-Resolution Hybrid Network (EHNet) based on a UNet-like architecture that adeptly fuses CNN and Swin Transformer. It also introduces a

novel sequence-to-sequence upsample method that focuses more on semantic information, diverging from the previous convolution-based method. The patch merge module in the original Swin Transformer is not used in SwinIR to downsample the feature maps to get feature maps with different resolutions to extract features at different scales, so the feature extraction ability of SwinIR is weakened compared to the original Swin Transformer. UNet's inherent encoder-decoder structure enables it to have better feature extraction capabilities. We use the encoder part to first downsample to extract features and then use the decoder part to upsample to recover the detailed information. We design a convolution-based Lightweight Feature Extraction Block (LFEB) as the fundamental module in the encoder, which gradually downsamples to extract semantic features. Convolutional structures, being more cost-effective than self-attention mechanisms, are better suited for extracting image features. To further reduce computation costs, we employ depthwise convolutions. For the decoder, we utilize Swin Transformer as the backbone network because it can establish long-range dependencies through self-attention, enhancing the restoration of image details. Its window attention mechanism also significantly reduces the model's computational cost. On the other hand, almost all super-resolution tasks utilize convolution-based upsampling methods, like the widely-used sub-pixel convolution method [31]. However, data in the transformer flow in the form of a sequence of tokens. Our experiments demonstrate that employing convolution-based upsampling methods between two transformer layers may inadvertently introduce extraneous semantic information unrelated to the target. This can potentially reduce the model's accuracy. Thus, we propose a new upsampling module tailored for the attention mechanism of sequential data, the sequence-based upsample block (SUB).

The principal contributions of this paper are summarized as follows:

1. We propose the Efficient Super-Resolution Hybrid Network (EHNet), a lightweight RSISR network that efficiently fuses CNN and Swin Transformer within a UNet-like structure. This hybrid model is capable of utilizing both the inductive bias of convolution and the long-range modeling capability. On the other hand, the multi-scale capability of UNet and skip connection can reconstruct images with richer details;
2. We design a lightweight and efficient convolutional block as the fundamental unit for image feature extraction. The dual-branch design of CSP enables the integration of features from different stages, aiding the model in understanding and utilizing these varied stage features. In addition, we found that SELayer can also realize channel feature combinations with much less computational cost than pointwise convolution;
3. In the decoder, we innovatively propose an upsampling method SUB based on a sequence of tokens. Compared with convolution-based upsampling methods, our SUB is more suitable for transformer-based models and can improve image detail recovery capabilities by focusing on semantic information.

2. Related Works

We divide the existing SR methods into two categories: natural images and remote sensing images, according to application scenarios. Nature images contain objects, scenes, and people from everyday life. These images often have more detail and are more accessible. Remote sensing images typically come from satellites or aircraft to obtain information about the Earth's surface. These images can contain features such as topography, land use, vegetation cover, etc. Additionally, remote sensing images are often difficult to obtain and have a small amount of data. In recent years, there have been many advanced SR models applied to natural images, and most of the SR models of remote sensing images are improved from advanced models of natural images. Table 1 lists the current SOTA methods for some of the SR tasks.

Table 1. Some SISR SOTA methods in recent years. The application scenarios of these methods, the number of parameters, and a brief description of the method are listed in the table.

Method	Application Scenarios	Params	Description
SRCNN [10]	Natural Image	69 K	The first SISR method using deep learning
VDSR [11]	Natural Image	671 K	A 20 layers model with residual learning
RCAN [32]	Natural Image	15.2 M	A 200 layers model with channel attention
IPT [26]	Natural Image	115.5 M	An SISR method using standard transformer
SwinIR [28]	Natural Image	3.87 M	An SISR method using Swin Transformer except for patch merge
HAT [33]	Natural Image	5.29 M	A SISR method activating more pixels based on SwinIR
LGCNet [34]	Remote sensing	193 K	The first RSISR method combining local and global features
DCM [35]	Remote sensing	1.84 M	An RSISR model with a network-in-network structure
CTNet [36]	Remote sensing	349 K	An RSISR method using lightweight convolution
HSENet [37]	Remote sensing	5.29 M	A hybrid-scale self-similarity exploitation network
TransENet [38]	Remote sensing	37.3 M	A transformer-based enhancement network

2.1. SISR Methods of Natural Images

SRCNN [10] was the pioneering method utilizing deep learning to establish a nonlinear mapping between LR and HR images, achieving state-of-the-art performance on some public datasets with only three convolutional layers. Later, many scholars proposed deeper CNN models to obtain better performance. Very Deep Super-Resolution (VDSR) [11] expanded the network depth to 20 layers through residual learning, achieving better results. Fast Super-Resolution Convolutional Neural Networks (FSRCNN) [39] gave up the idea of interpolating the image to the target image size in advance but greatly reduced the parameters and calculation amount by adding a deconvolution layer at the end of the network. Efficient Sub-pixel Convolutional Neural Network (ESPCN) [31] proposed an efficient sub-pixel convolution module to achieve upsampling. Sub-pixel convolution is used by a large number of super-resolution models due to its excellent performance. Residual Channel Attention Network (RCAN) [32] considers the relationship between each channel and constructs a deep network with up to 200 residual blocks. The excellent performance of RCAN, a model based on channel attention, has led more researchers to start focusing on the attention mechanism. Second-order Attention Network (SAN) [40] proposed a second-order attention mechanism, which establishes feature relationships by calculating second-order feature statistics so that the model has better feature representation capabilities. Holistic Attention Network (HAN) [41] not only utilizes channel attention and spatial attention to learn the channel and spatial interdependence of features of each layer but also introduces a layer of attention to explore the correlation between layers.

IPT [26] is an image restoration model based on standard transformer, but the excellent performance of IPT requires a large amount of data (1.1 M images) for training and a complex model (115.5 M parameters). SwinIR [28] proposed an image super-resolution method based on Swin Transformer [27], which is mainly composed of W-MSA and SW-MSA. Unlike the VIT-style model, window attention greatly reduces the computation and parameters of the model because the calculation of attention only needs to be performed within each window. Nevertheless, such window-limited attention computations impinge upon the transformer's intrinsic capability for modeling long-range dependencies. However, the sliding window attention mechanism adeptly compensates for this shortcoming, endowing swin with the advantages of both CNN and Transformer. Hybrid Network of CNN and Transformer (HNCT) [42] proposed a lightweight image super-resolution model that mixes CNN and Transformer. HNCT considers both local and non-local priors and extracts deep features that are beneficial to super-resolution reconstruction, while maintaining the model is lightweight enough. A Hybrid Attention Transformer (HAT) [33] adds channel attention based on SwinIR, which makes up for the shortcomings of insufficient utilization of information between Transformer channels. In addition, HAT also introduces an overlapping cross-attention module to better aggregate cross-window information.

2.2. SISR Methods of Remote-Sensing Images

Spatial resolution is a crucial metric for assessing the performance of remote sensing satellites. Remote sensing images with higher spatial resolution are capable of containing more target information and enhancing the accuracy of subsequent tasks like classification, segmentation, or detection. Merely interpolating images can only increase the resolution without adding additional effective information. Recently, learning-based image super-resolution methods have become mainstream for enhancing the resolution of remote-sensing images. Inspired by natural image SR networks, Lei et al. [34] first proposed an SR network that combines local and global features using deep learning, termed LGCNet. Haut et al. [35] introduced the Deep Compendium Model (DCM), which integrates residual blocks, skip connections, and a network-in-network structure. Pan et al. [43] presented the Residual Dense Backprojection Network (RDBPN) to address higher super-resolution magnifications, using a residual backprojection block structure for utilizing residual learning both globally and locally. Dong et al. [44] proposed a Second-order Multi-scale network (SMSR), which captures multi-scale information by reusing features learned at varying depths. Zhang et al. [45] extracted features of different scales using convolutions with varying kernel sizes and channel attention modules. Huan et al. [46] developed a new Pyramid-style Multi-Scale Residual Network (PMSRN) by merging hierarchical features to construct a Multi-Scale Dilated Residual Block (MSDRB). Leveraging the self-similarity of remote sensing images, Lei et al. [37] devised a Hybrid-scale Self-similarity Exploitation Network (HSENet), utilizing a Single-scale Self-similarity Exploitation Module (SSEM) to learn feature correlations at the same scale and also designed a Cross-scale Connection Structure (CCS) for capturing recurrences at different scales.

Lei et al. [38] proposed a Transformer-based Enhancement Network (TransENet), where the transformer is employed to extract features at different stages, and the multi-stage design allows for the fusion of high-dimensional and low-dimensional features. Tu et al. [47] combined the Swin Transformer with generative adversarial networks (GANs) to propose SWCGAN, where the generator is composed of both convolution and swin and the discriminator consists solely of the Swin Transformer. Shang et al. [48] designed a hybrid-scale hierarchical transformer network (HSTNet) to acquire long-range dependencies and effectively compute the correlations between high-dimensional and low-dimensional features. Wang et al. [36] created a lightweight convolution called the contextual transformation layer (CTL) to replace 3×3 convolutions, which can efficiently extract rich contextual features. Zhang et al. [29] proposed a FeNet that strikes a balance between performance and model parameters, where the lightweight lattice block (LLB) acts as a nonlinear extraction module to improve expressive ability.

3. Methodology

In this section, we first introduce the overall architecture of EHNet. Then, we introduce our proposed lightweight feature extraction module (LFEB) and a new sequence-based upsample block (SUB) in detail.

3.1. Network Architecture

Figure 1 displays the overall architecture of our EHNet, which designs an advanced encoder-decoder pattern based on the UNet structure. The encoder part uses efficient convolutional layers designed by us to capture the low-level features and spatial context information of the image, while the decoder part uses swin transformer to reconstruct image details. Additionally, following the Swin Transformer, there is a specialized upsampling module designed for the sequence of tokens. This module can more richly express the characteristics of the sequence of tokens, as it operates directly at the sequence level, avoiding the potential information compression and loss caused by convolutional layers. Moreover, it can perform SR reconstruction of images based on semantic information during upsampling. To compensate for the possible spatial information loss when reshaping feature maps into sequences, we have incorporated skip connections between the encoder and decoder. This network architecture design of EHNet not only facilitates the effective

integration of local details with global information but also enhances the performance of the model in performing image super-resolution reconstruction by utilizing the focused semantic information. This leads to significant improvements in image clarity and richness, making our model particularly suitable for application scenarios requiring high-quality image reconstruction.

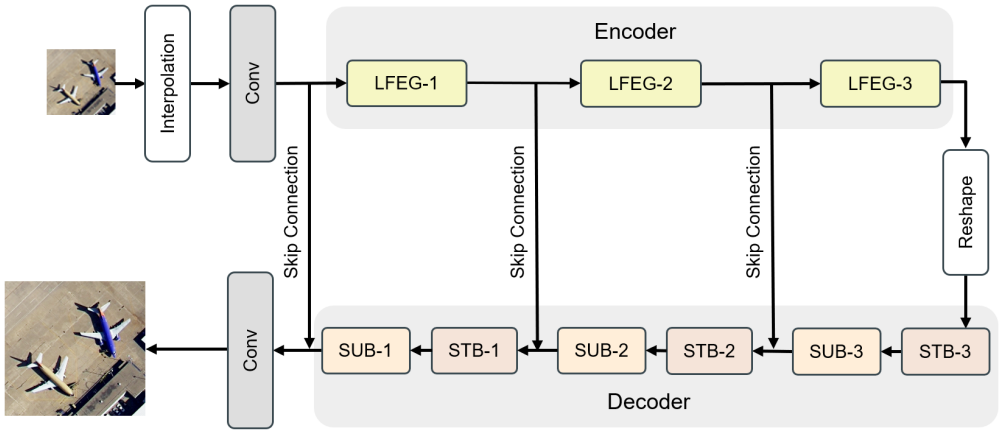


Figure 1. Architecture of the proposed EHNet.

Given an LR image, we first interpolate it to the target resolution and then use a 3×3 convolution to transform it into a feature map, thereby extracting the initial features f_0. This process can be expressed mathematically as follows:

$$\mathrm{f}_0 = \mathrm{Conv}(\mathrm{Interpolate}(\mathrm{I}_{\mathrm{LR}})) \tag{1}$$

where the Conv denotes a convolutional operation and f_0 represents the initial feature, which will be the input of the following feature extraction part.

We use three LFEGs to construct the encoder within the UNet structure. The primary function of these LFEGs is to extract low-level features at various scales from the image. Each LFEG is composed of multiple stacked LFEBs. The feature map is downsampled $\frac{1}{2}$ for each LFEG, so the resolution of the feature map after three LFEGs is $\frac{1}{8}$ of the HR. The output of the encoder part can be written as follows:

$$\mathrm{f}_\mathrm{n} = \mathrm{LFEG}_\mathrm{n}(\mathrm{f}_{\mathrm{n}-1}),\ \mathrm{n} = 1, 2, 3 \tag{2}$$

where $\mathrm{LFEG}_\mathrm{n}(\cdot)$ and f_n represent the operation of ith LFEG and its output.

After passing through the encoder composed of convolutional structures, we will use Swin Transformer Blocks (STB) and SUB to gradually upscale and restore image details.

STB is the basic module of Swin Transformer, which divides the image into a series of windows, and all the Attention is computed only within the window. This windowed attention mechanism greatly reduces the amount of computation. However, only calculating the attention within the windows weakens the long-term modeling ability of the transformer, so there is also a window sliding mechanism in the Swin Transformer to transfer the information between the windows.

In our EHNet, STB is mainly used to extract higher-dimensional semantic features for SUB, while our specially designed SUB uses these features to recover the image details and upsample the feature maps by a factor of 2. The output of each upsampling is concatenated with the corresponding output of the encoder part before being used as the input of the next layer. This feature fusion operation compensates for the loss of spatial information due to downsampling.

$$\mathrm{F}_\mathrm{n} = \begin{cases} \mathrm{SUB}_\mathrm{n}(\mathrm{STB}_\mathrm{n}(\mathrm{f}_\mathrm{n})), & \mathrm{n} = 3 \\ \mathrm{SUB}_\mathrm{n}(\mathrm{STB}_\mathrm{n}([\mathrm{F}_{\mathrm{n}+1}, \mathrm{f}_\mathrm{n}])), & \mathrm{n} = 2,\ 1 \end{cases} \tag{3}$$

where SUB_n and STB_n represent the operation of ith sequence-based upsample block and Swin Transformer block, and F_n represent the output after ith upsample.

Finally, after concatenating the output of the decoder, F_1, with f_0, and then passing it through another convolutional layer, we can obtain the final SR image.

3.2. Lightweight Feature Extraction Block (LFEB)

In this section, we design an efficient feature extraction module that can extract rich features for the decoder to use with low computation. The LFEB is the base unit of the encoder, and we stack multiple LFEBs and incorporate residual learning to form a residual-in-residual structure of the LFEG, which is capable of constructing deeper networks without gradient explosion. Each LFEG ends with a pooling layer to downsample the feature map. Finally, three LFEGs form the encoder part. The encoder of our EHNet is shown in Figure 2.

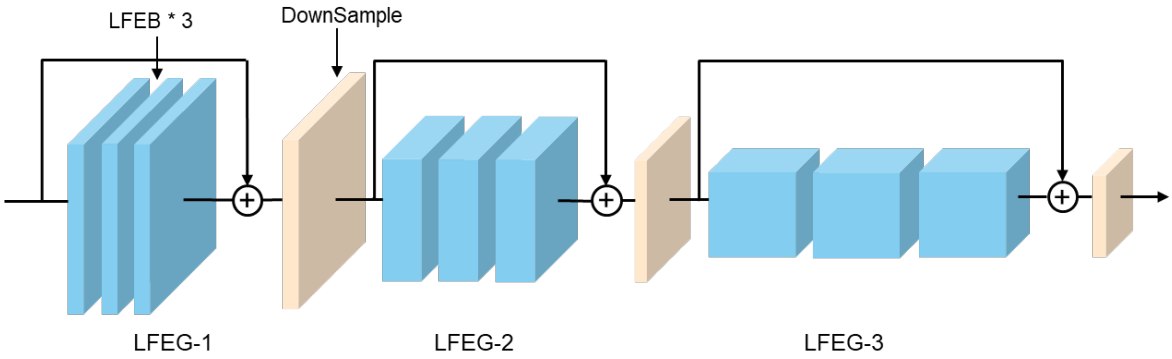


Figure 2. Architecture of encoder in EHNet.

LFEB's overall structural design concept is similar to the Residual Channel Attention Block (RCAB) [32]. RCAB mainly consists of standard convolution and Channel Attention (CA) in tandem with it. Our LFEB is mainly composed of CSP and lightweight convolution modules. The dual-branch design of CSP effectively integrates information from different stages with minimal computational cost. On the other hand, the lightweight convolution modules, consisting of depthwise convolution (dwconv) and Squeeze and Excitation layer [49] (SELayer), are able to extract features efficiently. The SELayer enables cross-channel feature fusion while reducing the computational cost caused by the pointwise convolution (pwconv) in separable convolutions. Whereas in our LFEB, we use depthwise convolutionin tandem with SELayer as the basic combination. In many lightweight convolutional designs, dwconv with pointwise convolution (pwconv) is a common combination, and pwconv is used to compensate for the lack of information fusion between channels in dwconv. However, in our experiments, it is demonstrated that this combination design is not necessarily helpful for super-resolution tasks, and SELayer can also take on the function of channel information fusion instead of pwconv, with lower computation effort.

SELayer adaptively recalibrates the feature responses between channels by explicitly modeling their interdependencies. Specifically, SELayer learns to automatically obtain the importance of each channel and then enhances useful features and suppresses features that are less useful according to this importance. The main operation of SELayer is to globally average pool the feature map to obtain $1 \times 1 \times C$ features (Squeeze) and then predict the importance of each channel through the fully connected layer, obtaining channel-level attention weights (Excitation), which are used to recalibrate the feature maps.

Because of the success of CSPDarknet in Yolov4 [50], we also add our own design of Cross Stage Partial (CSP) connection to extend the channel space in LEFB, and the addition of CSP hardly increases the computation and also improves the performance of the model to a certain extent. The structure of LFEB is shown in Figure 3.

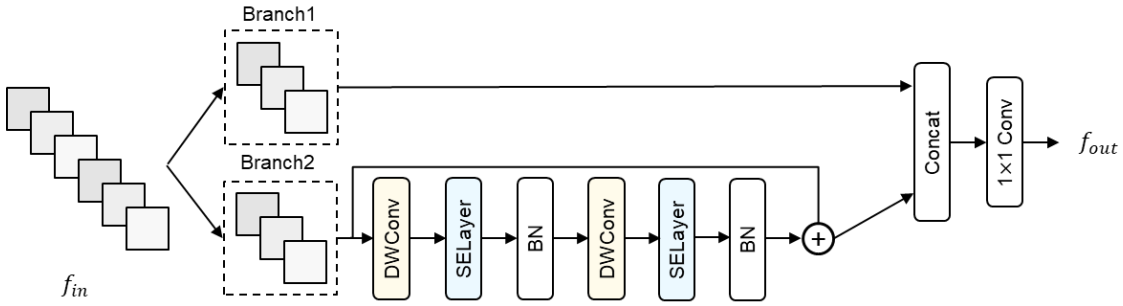


Figure 3. Lightweight Feature Extraction Block (LFEB).

CSP allows for the fusion of features at different network stages due to its dual-branch design. Doing so helps to integrate and propagate features from lower and higher levels more efficiently, improving the model's understanding and utilization of features from different levels. In super-resolution tasks, this fusion can help the network better understand image details and facilitate more accurate detail reconstruction. The CSP structure in our LFEB divides the input 2C feature maps f_{in} into two branches, each with a number of channels C. This process can be written in the following form using Equation (4):

$$f_{in} = [f_1, f_2] \tag{4}$$

where f_1 and f_2 denotes the feature map at the beginning of the two branches.

In branch1, features are extracted as usual through the subsequent two convolutional layers. In branch2, f_2 is directly concatenated with the features extracted in branch1. Finally, a 1×1 convolution is used for information fusion, producing the output feature f_{out}. This process can be mathematically described as follows:

$$f_{out} = Conv_{1\times1}([f_1, branch2(f_2)]) \tag{5}$$

where 'branch2' represents the convolutions, batch normalization (bn), SELayer, and all other operations within 'branch2'.

The branch2 of our LFEB consists mainly of a tandem stack of dwconv and SELayer, both of which have low computation cost, with a BN layer added to speed up convergence.

3.3. Sequence-Based Upsample Block

In super-resolution tasks, most of the models use convolution-based upsampling methods such as transposed convolution or sub-pixel convolution. The design inspiration for our SUB originally came from the patch expanding layer by Cao [51], which can achieve upsampling and feature dimension change without using convolution or interpolation. Compared with sub-pixel convolution and bilinear interpolation, this type of upsampling has achieved higher segmentation accuracy in segmentation tasks. Based on this sequence-based upsampling concept, we propose a new upsampling module SUB that is more suitable for super-resolution tasks. And our SUB can focus more on the semantic information of the image to obtain better reconstruction results, which is the first time that this sequence-based upsampling method is proposed for super-resolution tasks.

The structure of our SUB is shown in Figure 4. The input sequence of tokens is first dimensionally transformed through the MLP layer, where the MLP layer is able to introduce nonlinear transforms to enhance the model feature learning and expression capabilities, and also to double the channel dimension. The MLP is then followed by a layer of Swin Transformer to recover more details of the image. There are three layers of Swin Transformer in the decoder, each of which corresponds to three downsampling layers in the encoder part. After one layer of transformer, we rearrange the sequence of tokens into feature maps of $B \times 2C \times H \times W$ and then go through a Pixel Shuffle operation to change the resolution of the feature maps to $2\times$ of the input and the dimension of the channels to $\frac{1}{4}$ of the input. Finally, we change the sequence of tokens into the form of

feature maps mainly to facilitate the fusion with the features extracted from the convolution in the encoder.

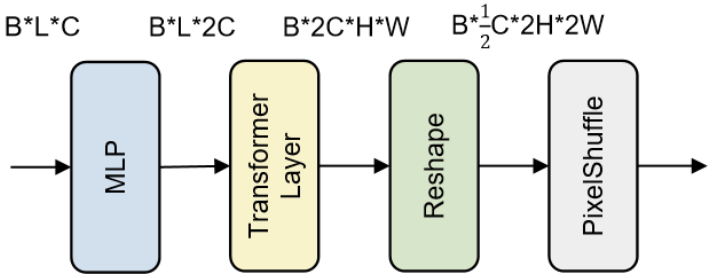
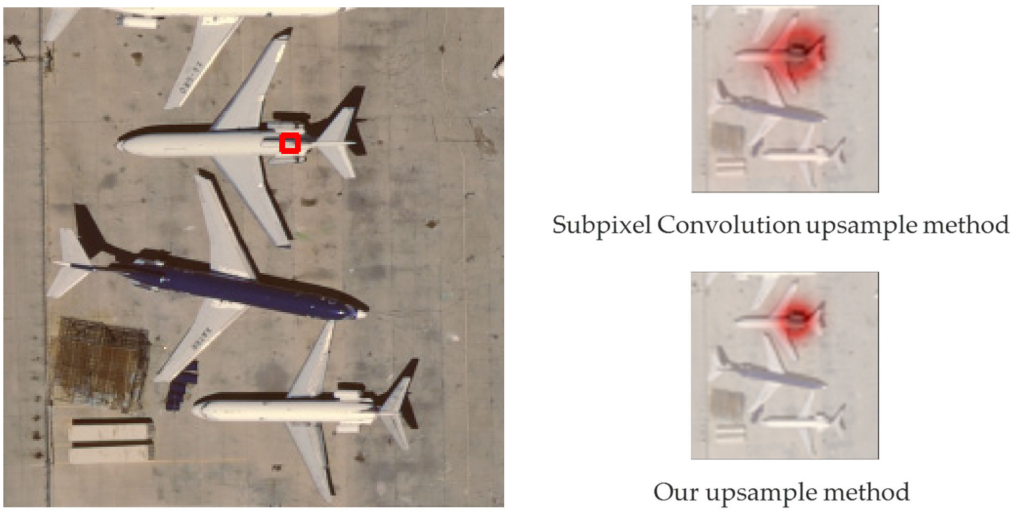


Figure 4. Sequence-based Upsample Block (SUB).

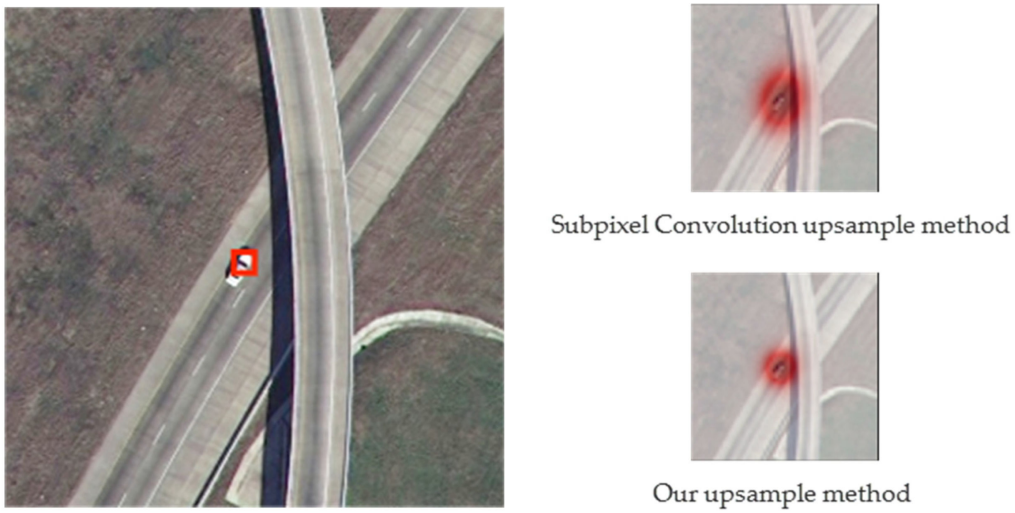
In summary, our SUB effectively upsamples the sequence of tokens in transformers and restores more precise and accurate details in super-resolution tasks. To demonstrate the effectiveness of our SUB module, we used Local Attribution Maps (LAM) [52] to analyze which pixels in the input LR contribute most to the SR (Super Resolution) reconstruction. LAM is a method for attribution analysis based on integrated gradients. By selecting a region of interest in the image, LAM can identify pixels that significantly contribute to the SR reconstruction of that area.

We applied LAM to analyze both the convolution upsampling method and our SUB, with results shown in Figure 5. In the airplane scene, we selected the engine part as the target region. It can be seen that there are many pixels in the LAM results sampled on the convolution that do not match the semantic information of the airplane, also have an impact on the SR results, and this additional introduction of extraneous pixel information degrades the quality of the SR reconstruction. While the LAM results of our method are more focused on the part that matches the target semantics, most of the pixels with large contributions are focused on the airplane engine part, and SR reconstruction based on the semantic information is an important reason why our EHNet can obtain higher performance. Similar results also appear in the overpass scene, we selected a car on the road as the target region, and our method also obtains results that are more focused on the car part, which leads to better reconstruction results.



(**a**) LAM results of airplane73

Figure 5. *Cont.*



(**b**) LAM results of overpass67

Figure 5. LAM results of two methods in two different scenes, where the red shaded area shows the degree of semantic focusing. The red box represents the target area that we have selected.

4. Experiments

4.1. Experiment Settings

To verify the effectiveness of our model, we trained on two widely used public remote sensing datasets, UCMerced [53] and AID [54], respectively.

UCMerced dataset: This dataset contains 21 types of remote sensing scenarios, including airports, highways, ports, etc. Each scene category has 100 images, each measuring 256 × 256 pixels, and the spatial resolution of these images is 0.3 m/pixel. This dataset is divided into two equal parts, one of which is used as a training set with a total of 1050 images, and the other part is used as a test set, with 20% of the training set being used as a validation set;

AID dataset: Compared with the UCMerced dataset, the AID dataset is a dataset with a larger number and size of images, containing 10,000 images and a total of 30 remote sensing scenes. The image size of the AID dataset is 600 × 600 pixels, and the spatial resolution of the image is 0.5 m/pixel. In this dataset, 8000 images were randomly selected as the training set images, and the remaining 2000 images were used as the test set images. In addition, we selected five images in each category for a total of 150 images as the validation set.

The images in both the UCMerced dataset and the AID dataset were used as HR images in the experiment, and their corresponding LR images were obtained by Bicubic interpolation. We trained and evaluated the model by constructing such paired HR-LR images.

We used peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) to evaluate the experimental results, and all evaluations of the super-resolution results were performed on the RBG channel. In general, SSIM is more reflective of image quality as perceived by the human eye but is computationally complex, whereas PSNR is computationally simple but does not necessarily fully reflect the human eye's perception of image quality. In our experiments we used a combination of these two metrics to more comprehensively assess image super-resolution quality. In our experiments, the original images in each dataset were treated as HR images, and the corresponding LR images were obtained by performing bicubic interpolation on the HR images. The PSNR and SSIM of a super-resolution image can be calculated by the following equation:

$$PSNR(I_{SR}, I_{HR}) = 10log_{10}\left(\frac{255}{MSE(I_{SR}, I_{HR})}\right) \quad (6)$$

$$MSE(I_{SR}, I_{HR}) = \frac{1}{N}\sum_{i=1}^{N}(I_{SR}(i) - I_{HR}(i))^2 \tag{7}$$

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)} \tag{8}$$

where $\mu_x(\mu_y)$ and $\sigma_x(\sigma_y)$ are the mean and variance, respectively. σ_{xy} is the covariance between x and y. c_1 and c_2 are constants. I_{SR} is the super-resolution image and I_{HR} is the high-resolution image.

Floating Point Operations (FLOPs) and model parameters are used to measure the computation cost of the model, where the input image size is 64×64 when calculating FLOPs.

Our loss function employs the L1 loss, which is most common in super-resolution tasks. Given a training set $\left\{I_{LR}^i, I_{HR}^i\right\}_{i=1}^N$, the loss function can be expressed as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|EHNet\left(I_{LR}^i\right) - I_{HR}^i\right\|_1 \tag{9}$$

We conducted experiments on remote sensing images with scale factors of $\times 2$ and $\times 4$. During training, we randomly cropped the image, and the size of the cropped image was 192×192. We also performed random flips and rotations on the training samples to increase sample diversity. We used the Adam optimizer, where $\beta_1 = 0.9$, $\beta_2 = 0.99$. We adopted the cosine annealing learning rate decay strategy with an initial learning rate of 5×10^{-5} and a minimum learning rate of 1×10^{-7}. During the training process, we used a batch size of 16 and trained 2000 epochs on the model. The entire training was performed on two NVIDIA 3080 Ti GPUs.

4.2. Ablation Studies

In this section, we performed a series of ablation experiments on the UCMerced dataset to explore the importance of each module in our model, where all models were trained on the same settings. For simplicity, all experiments had a super-resolution factor of 4.

4.2.1. Effects of LFEB

The LFEB is the most important component of the encoders, and we explored the effect of using this module with different settings. The number of LFEBs in each LFEG in our experiments is set to 9. Compared to RCAB, a benchmark module commonly used in super-resolution modeling, our LFEB is 0.11 dB higher in PSNR metrics. We compared the most commonly used combination of dwconv + pwconv with our dwconv + SELayer combination scheme and found that our approach has better performance. Also, the use of pwconv has a larger computation cost and memory usage, whereas SELayer is a lightweight feature calibration module using only fully connected layers. We also validated the effectiveness of the CSP dual-branch structure in LFEB and found that the PSNR improved by 0.06 after the introduction of the CSP; all results are shown in Table 2.

Table 2. PSNR and SSIM results with different components in LFEB. Bold data indicates that it is the best method.

	PSNR	**SSIM**
RCAB	27.90	0.7679
Dwconv + pwconv	27.93	0.7685
Dwconv + SELayer	27.96	0.7687
Dwconv + SELayer + CSP	**28.02**	**0.7711**

In recent years, there have also been some excellent Attention Modules that are often used in various super-resolution tasks, and we also compared SELayer with these methods. The Convolutional Block Attention Module (CBAM) [55] can perform Attention operations in both spatial and channel dimensions combining the Channel Attention Module And

by combining the channel attention module and spatial attention module together, the network can achieve better feature selection and reinforcement in both channel and spatial dimensions, improving the model's representation ability. Efficient Channel Attention (ECA) [56] proposes a local cross-channel interaction strategy without dimensionality reduction, which can be efficiently implemented by one-dimensional convolution with high efficiency. We tested several other popular convolutional attention methods with other parts of the LFEB fixed unchanged, and SELayer obtained the best performance in both PSNR and SSIM metrics. The experimental results are shown in Table 3.

Table 3. PSNR and SSIM results with different attention modules in LFEB.

	PSNR	SSIM
CBAM	27.98	0.7699
ECA	27.99	0.7697
SELayer	**28.02**	**0.7711**

4.2.2. Effects of SUB

SUB is a new sequence-based upsampling module we propose that can improve the detail restoration ability of the decoder composed of transformer by focusing on semantic information.

We explored the experimental performance of different components forming an SUB and found the most effective SUB settings. All experimental results are shown in Table 4. Judging from the experimental results, if we only use the MLP layer for dimension transformation, the effect is average, and after adding a layer of Swin Transformer, the PSNR increases by 0.1 dB. There are two ways to transform features from an extended channel dimension to a larger spatial resolution: one is to directly reshape the feature map to the target resolution, and the other is to reshape with the channel dimension unchanged and then use pixel shuffle to increase the spatial resolution. From the experimental results, the latter scheme can obtain higher reconstruction results.

Table 4. PSNR and SSIM results with different components in SUB.

	PSNR	SSIM
Mlp + pixel shuffle	27.92	0.7681
Mlp + transformer + reshape	27.99	0.7703
Mlp + transformer + pixel shuffle	**28.02**	**0.7711**

We also compared SUB with transposition convolution and subpixel convolution, which are commonly used as upsampling methods in other SOTA methods, and our SUB is higher than transposition convolution and subpixel convolution in PSNR by 0.23 dB and 0.12 dB, respectively, and SSIM is higher than them by 0.0051 and 0.0036, respectively. Our experimental results verified the validity of the SUB upsampling method. The experimental results are shown in Table 5.

Table 5. PSNR and SSIM results with different upsample methods.

	PSNR	SSIM
Transpose Convolution	27.79	0.7660
Subpixel Convolution	27.90	0.7675
SUB	**28.02**	**0.7711**

4.2.3. Ablation Study of Our EHNet

We performed ablation experiments on the whole EHNet, mainly including the number of layers of Swin Transformer and the number of layers of convolution, as well as the effect of feature dimensions on model accuracy and model complexity. We can see that when the LFEB, the number of swin layers, and the number of feature channels are set to

9, 2, and 96, respectively, the EHNet can obtain higher PSNR and SSIM and keep a low computational overhead. All the experimental results are shown in Table 6.

Table 6. PSNR and SSIM results with different settings in EHNet.

Number of LFEB in a LFEG	Number of Layers in Swin	Number of Features	PSNR	SSIM	Params
6	2	96	27.91	0.7686	2.54 M
12	2	96	27.99	0.7705	2.74 M
9	1	96	27.96	0.7698	1.93 M
9	3	96	27.97	0.7702	3.35 M
9	2	72	27.94	0.7691	2.47 M
9	2	96	**28.02**	**0.7711**	2.64 M
9	2	120	28.01	0.7706	2.84 M

4.3. Comparison with the State-of-the-Arts

To verify the effectiveness of the proposed EHNet, we conducted comparative experiments with some SOTA competitors, namely, SRCNN [10], VDSR [11], LGCNet [34], DCM [35], CTNet [36], HSENet [37], TransENet [38], SwinIR [28] and HAT [33]. Among these methods, SRCNN [10], VDSR [11], HAT [33], and SwinIR [28] are the methods proposed for natural image SR. LGCNet [34], DCM [35], HSENet [37], CTNet [36] and TransENet [38] are designed for RSISR. We retrained all of these methods based on open-source code and tested them under the same conditions.

4.3.1. Quantitative Evaluation

Quantitative Results on UCMerced Dataset: Table 7 presents a comparison of the latency and performance accuracy of various methods on the UCMerced dataset. The results indicate that our EHNet achieves a superior balance between the number of parameters and accuracy. In the case of $\times 2$ and $\times 4$ super-resolution factors, EHNet demonstrates the best performance in terms of PSNR. Compared to recent high-performing models such as SwinIR [28], TransENet [38], and HSENet [37], EHNet shows improvements in both parameter count and performance. Specifically, under the $\times 4$ super-resolution factor, EHNet's PSNR is higher than TransENet [38], SwinIR [28] and HAT [33] by 0.24 dB, 0.15 dB and 0.16 dB, respectively, while having only 7%, 58%, and 50% of their parameter sizes. In comparison with lightweight models like SRCNN [10], VDSR [11], and CTNet [36], our EHNet also maintains competitive performance in terms of model accuracy and efficiency.

Table 7. Comparative results for the UCMerced dataset.

Method	Scale	Params	FLOPs	PSNR	SSIM
Bicubic	$\times 2$	-	-	30.76	0.8789
SRCNN [10]	$\times 2$	69 K	0.028 G	32.84	0.9152
VDSR [11]	$\times 2$	671 K	0.275 G	33.47	0.9234
LGCNet [34]	$\times 2$	193 K	0.195 G	33.48	0.9235
DCM [35]	$\times 2$	1.84 M	1.299 G	33.65	0.9274
CTNet [36]	$\times 2$	349 K	0.105 G	33.59	0.9255
HSENet [37]	$\times 2$	5.29 M	3.886 G	34.22	**0.9327**
TransENet [38]	$\times 2$	37.3 M	1.012 G	34.03	0.9301
SwinIR [28]	$\times 2$	3.87 M	1.693 G	34.15	0.9307
HAT [33]	$\times 2$	5.12 M	2.247 G	34.17	0.9311
EHNet (ours)	$\times 2$	2.64 M	0.785 G	**34.29**	0.9320
Bicubic	$\times 3$	-	-	27.46	0.7631
SRCNN [10]	$\times 3$	69 K	0.028 G	28.97	0.8132
VDSR [11]	$\times 3$	671 K	0.275 G	29.75	0.8346
LGCNet [34]	$\times 3$	193 K	0.195 G	29.28	0.8238
DCM [35]	$\times 3$	1.84 M	1.299 G	29.86	0.8393
CTNet [36]	$\times 3$	349 K	0.105 G	29.44	0.8319
HSENet [37]	$\times 3$	5.29 M	3.886 G	30.04	0.8433
TransENet [38]	$\times 3$	37.3 M	1.012 G	29.90	0.8397

Table 7. *Cont.*

Method	Scale	Params	FLOPs	PSNR	SSIM
SwinIR [28]	×3	3.87 M	1.693 G	30.12	0.8487
HAT [33]	×3	5.12 M	2.247 G	**30.15**	**0.8489**
EHNet (ours)	×3	2.64 M	0.785 G	30.09	0.8465
Bicubic	×4	-	-	25.65	0.6725
SRCNN [10]	×4	69 K	0.028 G	26.78	0.7219
VDSR [11]	×4	671 K	0.255 G	27.54	0.7522
LGCNet [34]	×4	193 K	0.195 G	27.02	0.7333
DCM [35]	×4	2.16 M	0.754 G	27.22	0.7528
CTNet [36]	×4	360 K	0.180 G	27.41	0.7512
HSENet [37]	×4	5.43 M	4.136 G	27.73	0.7623
TransENet [38]	×4	37.46 M	2.599 G	27.78	0.7635
SwinIR [28]	×4	4.54 M	2.801 G	27.87	0.7659
HAT [33]	×4	5.29 M	2.497 G	27.86	0.7683
EHNet (ours)	×4	2.64 M	1.205 G	**28.02**	**0.7711**

Quantitative Results on AID Dataset: In Table 8, our proposed EHNet demonstrates exceptional performance across all metrics on the AID test dataset. However, due to its limited model capacity, the performance of our model deteriorates when trained on the larger AID training dataset. Despite this limitation, EHNet still achieves the best or second-best performance in terms of PSNR on the AID test dataset and obtains the optimal results in the SSIM metric, which is more aligned with human visual perception. Overall, the method we propose maintains competitive performance. To further analyze the reasons behind these phenomena, we conducted an in-depth discussion on the quantitative performance of different methods across various categories.

Table 8. Comparative results for the AID dataset.

Method	Scale	Params	FLOPs	PSNR	SSIM
Bicubic	×2	-	-	32.39	0.8906
SRCNN [10]	×2	69 K	0.028 G	34.49	0.9286
VDSR [11]	×2	671 K	0.275 G	35.11	0.9340
LGCNet [34]	×2	193 K	0.195 G	34.80	0.9320
DCM [35]	×2	1.84 M	1.299 G	35.21	0.9366
CTNet [36]	×2	349 K	0.105 G	35.13	0.9354
HSENet [37]	×2	5.29 M	3.886 G	**35.50**	0.9383
TransENet [38]	×2	37.3 M	1.012 G	35.40	0.9372
SwinIR [28]	×2	3.87 M	1.693 G	35.35	0.9370
HAT [33]	×2	5.12 M	2.247 G	35.49	0.9388
EHNet (ours)	×2	2.64 M	0.785 G	35.42	**0.9390**
Bicubic	×3	-	-	32.39	0.8906
SRCNN [10]	×3	69 K	0.028 G	30.55	0.8372
VDSR [11]	×3	671 K	0.275 G	31.17	0.8511
LGCNet [34]	×3	193 K	0.195 G	30.86	0.8498
DCM [35]	×3	1.84 M	1.299 G	31.31	0.8561
CTNet [36]	×3	349 K	0.105 G	31.16	0.8515
HSENet [37]	×3	5.29 M	3.886 G	31.49	0.8588
TransENet [38]	×3	37.3 M	1.012 G	31.50	0.8588
SwinIR [28]	×3	3.87 M	1.693 G	31.47	0.8600
HAT [33]	×3	5.12 M	2.247 G	**31.53**	**0.8612**
EHNet (ours)	×3	2.64 M	0.785 G	31.51	0.8609
Bicubic	×4	-	-	27.30	0.7036
SRCNN [10]	×4	69 K	0.028 G	28.40	0.7561
VDSR [11]	×4	671 K	0.255 G	28.99	0.7753

Table 8. *Cont.*

Method	Scale	Params	FLOPs	PSNR	SSIM
LGCNet [34]	×4	193 K	0.195 G	28.61	0.7626
DCM [35]	×4	2.16 M	0.754 G	29.17	0.7824
CTNet [36]	×4	360 K	0.180 G	29.00	0.7768
HSENet [37]	×4	5.43 M	4.136 G	29.32	0.7867
TransENet [38]	×4	37.46 M	2.599 G	29.44	0.7912
SwinIR [28]	×4	4.54 M	2.801 G	29.26	0.7863
HAT	×4	5.29 M	2.497 G	29.43	0.7921
EHNet (ours)	×4	2.64 M	1.205 G	**29.44**	**0.7922**

Table 9 lists the performance across the 30 categories in the AID dataset. The experiments demonstrate that our method performs well in scenes with rich textural details, such as airports, schools, parking lots, and sparse residential areas, achieving the best PSNR results in most cases. In contrast, the scenes where PSNR results are less satisfactory tend to be those with more uniform and less detailed environments, such as bare land, beaches, and deserts. These images lack sufficient feature information. Our method primarily relies on enhancing high-frequency details to improve image resolution, and in scenes with simple content, there may not be enough information for effective reconstruction. On the other hand, the PSNR evaluation metric may be more suited to assessing detail enhancement in richly textured scenes. In less textured environments, PSNR may not fully reflect the true improvement in image quality.

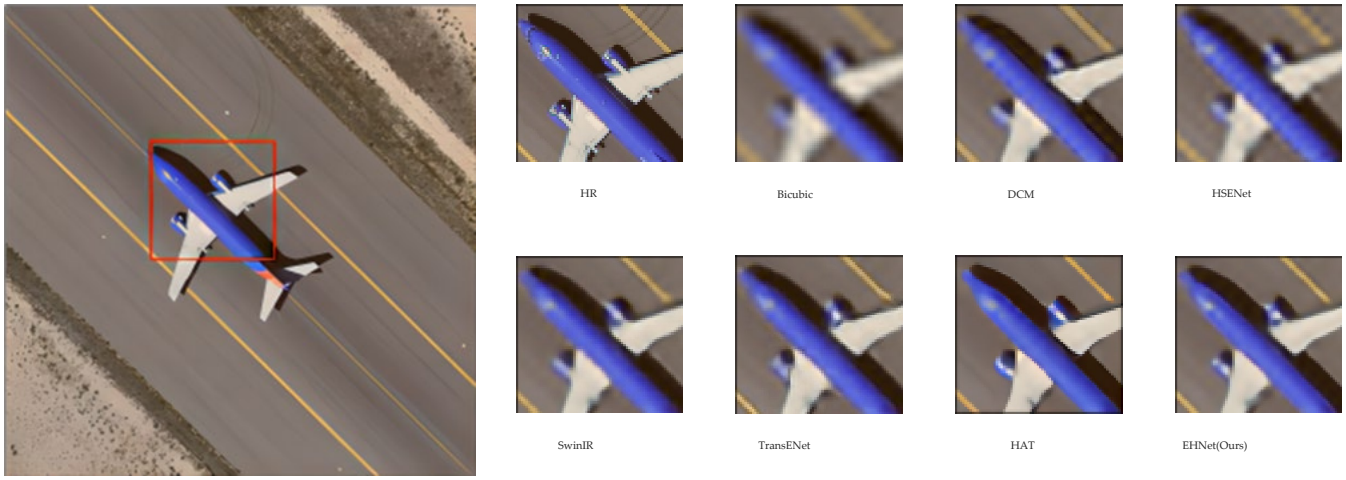
Table 9. Mean PSNR (dB) of each class for the scale factor of ×4 on the AID dataset.

Class	LGCNet [34]	DCM [35]	CTNet [36]	HSENet [37]	TransENet [38]	SwinIR [28]	HAT [33]	Ours
1	28.39	28.99	28.80	29.12	29.26	29.08	17.28	**29.27**
2	35.78	36.17	36.12	36.34	36.38	36.19	36.23	36.24
3	30.75	31.36	31.15	31.49	31.63	31.45	31.59	31.61
4	32.08	32.45	32.40	32.60	32.66	32.52	32.58	32.59
5	30.67	31.39	31.17	31.55	31.70	31.47	31.69	31.68
6	26.92	27.72	27.48	27.91	28.09	27.83	28.11	**28.11**
7	23.68	24.29	24.10	24.43	24.53	24.36	24.55	**24.57**
8	27.24	27.78	27.63	27.90	28.00	27.83	28.03	**28.03**
9	24.33	24.87	24.70	25.02	25.17	24.97	25.16	25.16
10	39.06	39.27	39.25	39.47	39.55	39.29	39.31	39.31
11	33.77	34.42	34.25	34.59	34.67	34.46	34.61	34.64
12	28.20	28.47	28.47	28.54	28.59	28.56	28.58	**28.58**
13	26.24	26.92	26.71	27.09	27.24	27.00	27.27	**27.27**
14	32.06	32.88	32.84	32.97	33.00	32.91	32.95	32.95
15	26.09	28.25	28.06	28.41	28.50	28.36	28.51	**28.53**
16	28.04	29.18	29.15	29.22	29.30	29.24	29.29	29.28
17	26.23	27.82	27.69	27.93	28.04	27.91	28.05	**28.04**
18	22.33	25.74	25.27	26.16	26.49	26.03	26.46	**26.52**
19	27.27	29.92	29.66	30.19	30.38	30.07	30.36	**30.38**
20	28.94	30.39	30.25	30.48	30.58	30.43	30.54	30.54
21	24.69	26.62	26.41	26.80	26.95	26.74	26.97	**26.98**
22	26.31	28.38	28.19	28.52	28.64	28.44	28.67	**28.66**
23	25.98	27.88	27.72	28.00	28.13	27.98	28.16	**28.16**
24	29.61	30.91	30.83	30.97	31.04	30.96	31.02	31.02
25	24.91	26.94	26.75	27.10	27.25	27.07	27.28	**27.28**
26	25.41	26.53	26.46	26.60	26.63	26.60	26.67	**26.68**
27	26.75	29.13	28.94	29.30	29.46	29.24	29.45	29.45
28	24.81	27.10	26.86	27.28	27.48	27.20	27.47	27.47
29	24.18	26.00	25.82	26.12	26.22	26.07	26.24	**26.26**
30	25.86	27.93	27.67	28.09	28.24	28.03	28.27	**28.26**
avg	28.61	29.17	29.03	29.32	29.44	29.26	29.43	**29.44**

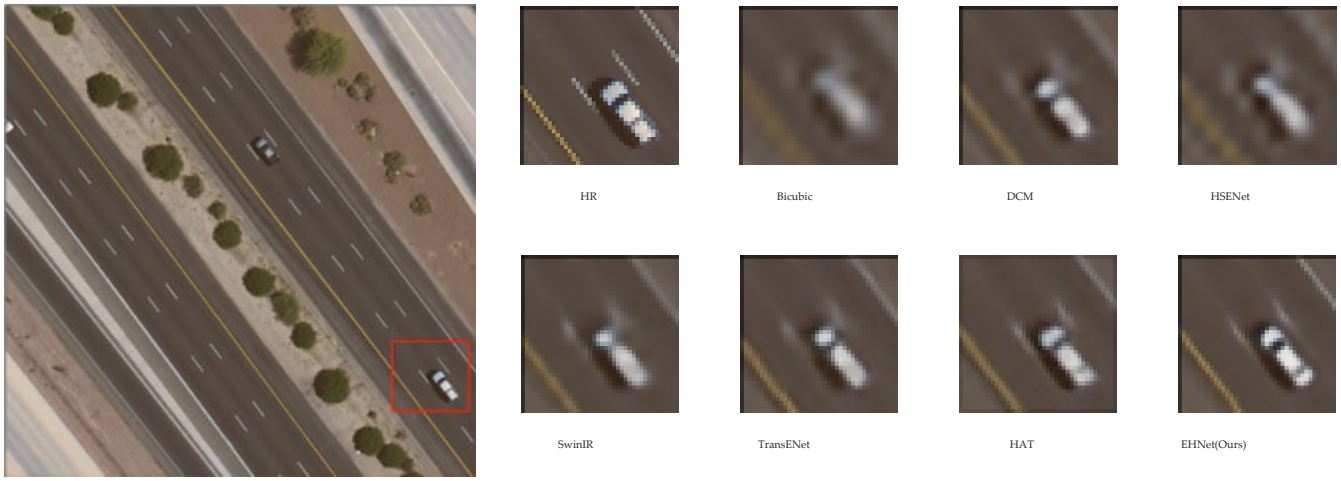
4.3.2. Quantitative Evaluation

In addition to the quantitative comparisons discussed above, we also conducted a qualitative analysis of super-resolved image quality. Figure 6 presents the visual results for two scenarios from the UCMerced dataset: airplane and freeway. In the case of 'airplane78',

our method successfully recovers the texture of the engine part while maintaining sharp edges. For 'freeway97', our EHNet uniquely restores the car windows, a detail not achieved by other methods. Moreover, the super-resolved image exhibits clearer lane lines, demonstrating EHNet's significant advantage in recovering image details.



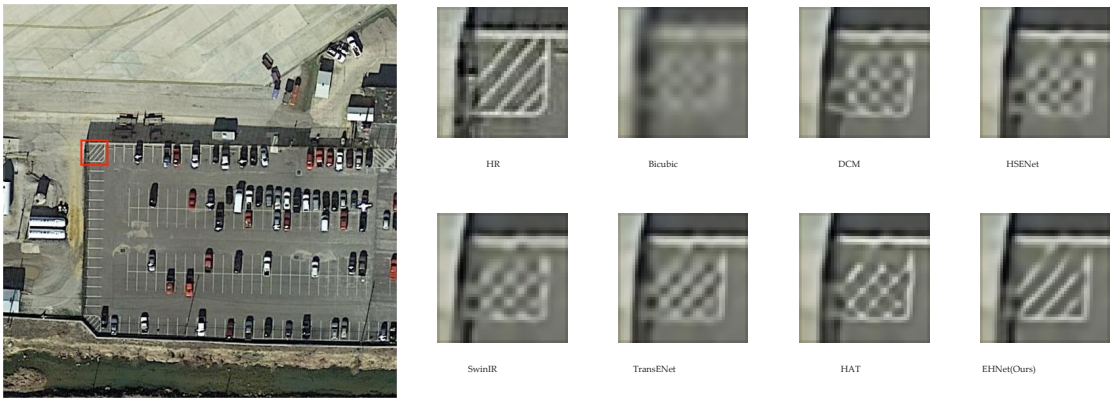
(**a**) SR results of airplane78



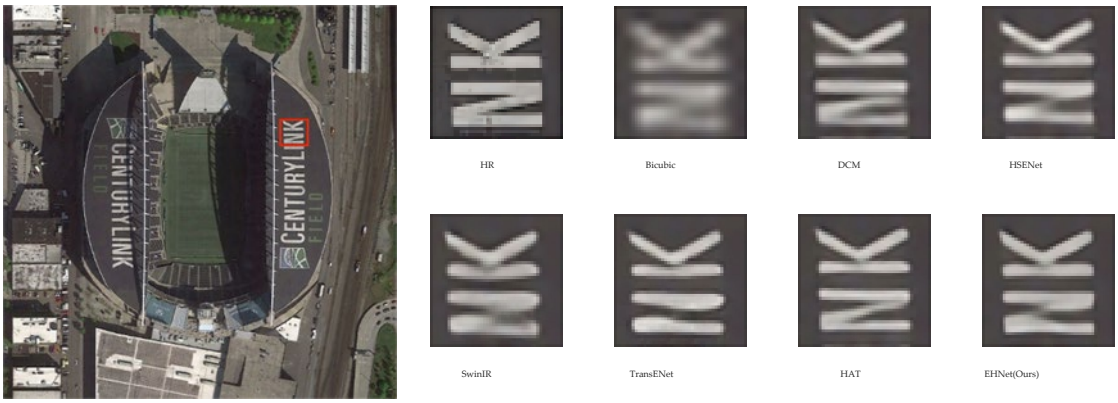
(**b**) SR results of freeway97

Figure 6. Visualization results of different RSISR methods on UCMerced dataset for $\times 4$ SR.

Figure 7 shows two examples of the AID dataset. For parking210, our proposed method successfully recovers clear marker lines, while the other methods are either very blurred or have checkerboard artifacts. Furthermore, in the super-resolution result of 'stadium262', our model achieves sharper edges around letters, further evidencing its superior performance in enhancing details.



(**a**) SR results of parking210



(**b**) SR results of stadium262

Figure 7. Visualization results of different RSISR methods on AID dataset for $\times 4$ SR.

5. Conclusions

In our work, we introduce a novel model named EHNet, an efficient single-frame SR model for remote sensing. EHNet ingeniously merges an encoder formed by LFEB with an improved Swin Transformer within a UNet architecture. The LFEB utilizes depthwise convolution to reduce computation cost, while the incorporation of SELayer enhances inter-channel information fusion, addressing the shortcomings of insufficient channel information integration in depthwise convolution. Additionally, we employ a CSP dual-branch structure to boost model performance without adding extra parameters. In the decoder part, we utilize Swin Transformer to restore image details and introduce a novel sequence-based upsampling method, SUB, to capture more accurate long-range semantic information. EHNet achieves state-of-the-art results on multiple metrics in the AID and UCMerced datasets and surpasses existing methods in visual quality. Its 2.64 M parameters effectively balance model efficiency and computation cost, highlighting its potential for broader application in SR tasks.

The results of the experiment show that our EHNet performs better on smaller datasets, but its performance is degraded for datasets such as AID, which has a larger image size and dataset size. We investigate the model's super-resolution reconstruction results for different scenes and find that our EHNet tends to underperform in those scenes with fewer details and smaller gradients. We speculate that the reason why the model does not perform well enough on large datasets may be that our model has a small number of parameters and cannot fully cope with all the scenes, especially those with smaller gradients. In addition, our model does not perform as well as the super-resolution factors of 2 and 4 on the super-resolution factor of 3, which may be due to the fact that our UNet architecture of EHNet adopts $2\times$ downsampling, so it does not work well enough for LR reconstruction with a super-resolution factor of 3.

In future research, we will focus on enhancing the model's performance in scenes with less texture, further improving its overall effectiveness.

Author Contributions: Conceptualization, W.Z., Y.L. and Z.T.; methodology, W.Z.; software, W.Z.; investigation, W.Z., J.L. and B.Z.; writing—original draft preparation, W.Z.; writing—review and editing, W.Z., Z.T. and Y.L.; project administration, Q.L.; funding acquisition, Z.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Program Project of Science and Technology Innovation of the Chinese Academy of Sciences (no. KGFZD-135-20-03-02), and this research was funded by the Innovation Foundation of Key Laboratory of Computational Optical Imaging Technology, CAS (no. CXJJ-23S016).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Aakerberg, A.; Nasrollahi, K.; Moeslund, T.B. Real-world super-resolution of face-images from surveillance cameras. *IET Image Process.* **2022**, *16*, 442–452. [CrossRef]
2. Ahmad, W.; Ali, H.; Shah, Z.; Azmat, S. A new generative adversarial network for medical images super resolution. *Sci. Rep.* **2022**, *12*, 9533–9552. [CrossRef] [PubMed]
3. Zhang, J.; Lei, J.; Xie, W.; Fang, Z.; Li, Y.; Du, Q. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [CrossRef]
4. Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Wang, Y.; Zhang, L. From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution. *Inf. Fusion* **2023**, *96*, 297–311. [CrossRef]
5. Wang, Y.; Bashir, S.M.A.; Khan, M.; Ullah, Q.; Wang, R.; Song, Y.; Guo, Z.; Niu, Y. Remote sensing image super-resolution and object detection: Benchmark and state of the art. *Expert Syst. Appl.* **2022**, *197*, 116793–116807. [CrossRef]
6. Wang, L.; Wang, L.; Wang, Q.; Bruzzone, L. RSCNet: A residual self-calibrated network for hyperspectral image change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5529917. [CrossRef]
7. Habibi, Y.; Sulistyaningrum, D.R.; Setiyono, B. A new algorithm for small object tracking based on super-resolution technique. *AIP Conf. Proc.* **2017**, *1867*, 256–270.
8. Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. Image super-resolution via sparse representation. *IEEE Trans. Image Process.* **2010**, *19*, 2861–2873. [CrossRef]
9. Timofte, R.; De Smet, V.; Van Gool, L. Anchored neighborhood regression for fast example-based super-resolution. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1920–1927.
10. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; pp. 184–199.
11. 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.
12. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
13. Tong, T.; Li, G.; Liu, X.; Gao, Q. Image super-resolution using dense skip connections. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4799–4807.
14. Li, J.; Du, S.; Wu, C.; Leng, Y.; Song, R.; Li, Y. Drcr net: Dense residual channel re-calibration network with non-local purification for spectral super resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1259–1268.
15. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1637–1645.
16. Lv, X.; Wang, C.; Fan, X.; Leng, Q.; Jiang, X. A novel image super-resolution algorithm based on multi-scale dense recursive fusion network. *Neurocomputing* **2022**, *489*, 98–111. [CrossRef]
17. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
18. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018; pp. 3–11.
19. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.

20. Hu, X.; Naiel, M.A.; Wong, A.; Lamm, M.; Fieguth, P. RUNet: A robust UNet architecture for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 15–20 June 2019; pp. 505–507.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
23. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
24. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
25. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
26. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12299–12310.
27. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
28. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1833–1844.
29. Wang, Z.; Li, L.; Xue, Y.; Jiang, C.; Wang, J.; Sun, K.; Ma, H. FeNet: Feature enhancement network for lightweight remote-sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
30. Wang, H.; Chen, X.; Ni, B.; Liu, Y.; Liu, J. Omni Aggregation Networks for Lightweight Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 22378–22387.
31. 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 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
32. 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.
33. Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; Dong, C. Activating more pixels in image super-resolution transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 22367–22377.
34. 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]
35. Haut, J.M.; Paoletti, M.E.; Fernández-Beltran, R.; Plaza, J.; Plaza, A.; Li, J. Remote sensing single-image superresolution based on a deep compendium model. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1432–1436. [CrossRef]
36. 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*, 5615313. [CrossRef]
37. Lei, S.; Shi, Z. Hybrid-scale self-similarity exploitation for remote sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5401410. [CrossRef]
38. Lei, S.; Shi, Z.; Mo, W. Transformer-based multistage enhancement for remote sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5615611. [CrossRef]
39. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 391–407.
40. 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.
41. Niu, B.; Wen, W.; Ren, W.; Zhang, X.; Yang, L.; Wang, S.; Zhang, K.; Cao, X.; Shen, H. Single image super-resolution via a holistic attention network. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 191–207.
42. Fang, J.; Lin, H.; Chen, X.; Zeng, K. A hybrid network of cnn and transformer for lightweight image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1103–1112.
43. Pan, Z.; Ma, W.; Guo, J.; Lei, B. Super-resolution of single remote sensing image based on residual dense backprojection networks. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7918–7933. [CrossRef]
44. Dong, X.; Wang, L.; Sun, X.; Jia, X.; Gao, L.; Zhang, B. Remote sensing image super-resolution using second-order multi-scale networks. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 3473–3485. [CrossRef]
45. 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.* **2020**, *58*, 4764–4779. [CrossRef]

46. Huan, H.; Li, P.; Zou, N.; Wang, C.; Xie, Y.; Xie, Y.; Xu, D. End-to-end super-resolution for remote-sensing images using an improved multi-scale residual network. *Remote Sens.* **2021**, *13*, 666–690. [CrossRef]
47. 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]
48. 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. [CrossRef]
49. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Munich, Germany, 8–14 September 2018; pp. 7132–7141.
50. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
51. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218.
52. Gu, J.; Dong, C. Interpreting super-resolution networks with local attribution maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9199–9208.
53. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
54. Xia, G.-S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
55. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
56. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.