*remote sensing*

MDPI

*Article*

# Remote Sensing Image Super-Resolution via Residual-Dense Hybrid Attention Network

**Bo Yu [1], Bin Lei [2], Jiayi Guo [3,4,5], Jiande Sun [1], Shengtao Li [1,*] and Guangshuai Xie [2]**

[1] School of Information Science and Engineering, Shandong Normal University, Jinan 250399, China
[2] Shandong Institutes of Industrial Technology, Jinan 250102, China
[3] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[4] Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Beijing 100190, China
[5] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China
* Correspondence: saintaolee@sdnu.edu.cn

**Abstract:** Nowadays, remote sensing datasets with long temporal coverage generally have a limited spatial resolution, most of the existing research uses the single image super-resolution (SISR) method to reconstruct high-resolution (HR) images. However, due to the lack of information in low-resolution (LR) images and the ill-posed nature of SISR, it is difficult to reconstruct the fine texture of HR images under large-scale magnification factors (e.g., four times). To address this problem, we propose a new reference-based super-resolution method called a Residual-Dense Hybrid Attention Network (R-DHAN), which uses the rich texture information in the reference image to make up for the deficiency of the original LR image. The proposed SR model employs Super-Resolution by Neural Texture Transfer (SRNTT) as a backbone. Based on this structure, we propose a dense hybrid attention block (DHAB) as a building block of R-DHAN. The DHAB fuses the input and its internal features of current block. While making full use of the feature information, it uses the interdependence between different channels and different spatial dimensions to model and obtains a strong representation ability. In addition, a hybrid channel-spatial attention mechanism is introduced to focus on important and useful regions to better reconstruct the final image. Experiments show that compared with SRNTT and some classical SR techniques, the proposed R-DHAN method performs well in quantitative evaluation and visual quality.

**Keywords:** super-resolution; remote sensing; attention mechanism; dense connection mechanism

check for updates

**Citation:** Yu, B.; Lei, B.; Guo, J.; Sun, J.; Li, S.; Xie, G. Remote Sensing Image Super-Resolution via Residual-Dense Hybrid Attention Network. *Remote Sens.* **2022**, *14*, 5780. https://doi.org/10.3390/rs14225780

Academic Editor: Benoit Vozel

Received: 31 July 2022
Accepted: 10 November 2022
Published: 16 November 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

With the vigorous development of remote sensing technology, high-resolution (HR) remote sensing images play an important role in many fields, such as object detection [1,2], urban planning [3], semantic labeling [4] and object detection [5]. However, most accessible public remote sensing datasets cannot maintain long-term coverage and high spatial resolution at the same time. For example, the earliest remote sensing data of Sentinel-2 can be traced back to only seven years ago. For remote sensing datasets with a time coverage of more than 20 years, it is usually impossible to maintain a high spatial resolution. To avoid the huge cost of directly improving satellite imaging equipment, image super-resolution (SR) technology is proposed to improve the quality of low-resolution (LR) images. The SR methods based on the interpolation method [6,7] proposed earlier have poor reconstruction effects. In recent years, people have focused on the field of deep learning [8]. However, the single image super-resolution (SISR) method based on a convolutional neural network (CNN) [9–12] cannot accurately reconstruct the HR image texture that has been excessively damaged due to degradation and its final reconstruction effect is often fuzzy. Although

the SR method based on GAN [13] considers human subjective visual perception and effectively alleviates the appealing problem, the resulting artifacts have also become another thorny problem.

Considering that the details of LR image loss can be compensated by the rich information in its similar HR reference (Ref) image, the reference-based super-resolution (RefSR) method [14,15] came into being. Not only does it effectively avoid the ill-posed problem caused by the SISR method, but also the reconstructed texture is more realistic with the help of the rich detailed information of the Ref image. Image alignment and patch matching are two mainstream ideas of recent RefSR methods. In remote sensing SR tasks, HR-Ref images and LR images can be easily located at the same geographical location through longitude and latitude matching. Therefore, it can ensure that the image contents of Ref and LR images have a certain similarity, which further explains the adaptability of the RefSR method in the field of remote sensing. However, due to different shooting viewpoints and geographical coordinate deviation, the alignment degree between the Ref image and the LR image is still not ideal. Therefore, we choose the RefSR method based on patch matching. Because SRNTT [14] looks for the most similar LR-Ref patch pair in the global scope, it can deal with the dependence of long distance and ensure the robustness of the model in the case of serious dislocation between Ref and LR images.

Although the above-mentioned methods have good performance, their results can be further improved. Different from natural images, the spatial information of remote sensing images is very large and complex. Therefore, for most SR methods of remote sensing images, improving the representation ability of the network means that a higher level of abstraction and better data representation can be obtained. This is very important for the final reconstruction effect of the LR remote sensing image. To improve the performance of the model, previous methods usually redesign the model structure, such as deepening the depth [16], expanding the network width [17] and increasing the cardinality [18], while we achieve the goal through lightweight mechanisms (such as attention mechanism and dense connection mechanism) that do not need too much network engineering. Therefore, we propose a residual dense hybrid attention network (R-DHAN), which integrates feature information from different levels, reduces the role of unimportant channels and spatial regions and improves the effective utilization of features. The major contributions are as follows:

(1) We propose an end-to-end SR algorithm for remote sensing satellite images, called residual dense hybrid attention network (R-DHAN), which is superior to most classical algorithms in quantitative and qualitative evaluation.

(2) A spatial attention module (SA) and a channel attention module (CA) are added to the network. This helps the network have a more flexible discriminative ability for different local regions and re-examine the importance of different channels. It contributes to reconstructing the final image.

(3) Based on some lightweight mechanisms, we propose a new residual block named DHAB, which mainly includes the local feature fusion (LFF) module and convolution block attention module (CBAM). LFF module makes full use of the current intra-block features and the original input features, while CBAM uses the interdependence between different channels and spatial dimensions to re-weight the features with different degrees of importance. Both of them improve the characterization ability of the network.

In the rest of this article, we briefly review the relevant work in Section 2. The details of our proposed method are introduced in Section 3. The experimental setup and final results are provided in Section 4 and our work is summarized in Section 5.

## 2. Related Work

### *2.1. SISR*

In recent years, SISR algorithms based on deep learning has gradually become the mainstream. Compared with traditional SR methods, they have made full progress in improving performance. Dong et al. proposed SRCNN [19] which firstly adopted deep

learning in SISR by using a three-layer CNN to represent the mapping function, while SRCNN had outperformed traditional-based methods, Bicubic LR images made the network operate in a high-dimensional space and largely increased the computational cost. To alleviate the problem, a fast super-resolution convolutional neural network (FSRCNN) was proposed [20]. Then, Kim et al. proposed VDSR [9] and DRCN [21] successively, which provided a more easily achieved goal by learning residual mapping instead of directly generating HR images. The application of the residual blocks (RBs) and dense blocks (DBs) has raised the performance of SISR to a new level. Among them, the application of ResNet architecture in SISR evolved into SRResNet [22]. On this basis, EDSR [10] removed unnecessary batch normalization (BN) modules and added more convolution layers to the network. This makes the performance of EDSR better without taking up more computing resources.

However, the above method ignores human visual perception. GAN is an effective way to solve this problem. SRGAN [22] introduced GAN into the SR field for the first time and added perceptual loss [23] based on the common loss function. These operations greatly improved the visual aesthetics of images. On this basis, ESRGAN, proposed by Wang et al. [24], further improved the network structure, adversarial loss and perceptual loss and obtained a more realistic reconstruction effect. Compared with ordinary L1 or L2 loss functions, when the monitoring method is replaced by GAN, the model is often more visually sensitive, but the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [25] will be reduced accordingly.

### *2.2. RefSR*

Unlike SISR, the RefSR method requires additional input of images with similar content to LR images to assist the SR reconstruction process, which is called Ref images. They may come from images taken from different viewpoints, images of different frames in the same video, network search, etc.

A kind of mainstream method in RefSR follows the idea of image alignment. These methods aim to improve the alignment of Ref and LR images as much as possible. Landmark [26] solves the problem of image alignment through global matching. Wang et al. [27] proposed a method of non-uniform distortion and repeatedly applied it before feature synthesis to optimize the Ref image. CrossNet [15] adopts the optical flow alignment method. When the Ref images and LR images are aligned at different scales, they are input into the decoder and spliced in the corresponding layer. However, these methods rely too much on the alignment quality of the image. When the alignment quality declines, the SR reconstruction effect of LR image is often not ideal, which does not fit well with the remote sensing task of images taken by different satellites at different viewpoints.

Another kind of mainstream method in RefSR follows the idea of patch matching. Boominathan et al. [28] first down-sampled Ref image and then matched its patch with the gradient feature of LR. Zheng et al. [29] applied the method of semantic matching, replacing simple gradient features with features for matching and then using the previous SISR method for feature synthesis. SRNTT [14] first extracts the Ref and LR features, which is completed by the pretraining network VGG. Then these features are divided into small patches and the most similar texture features are exchanged by calculating the similarity score between Ref patches and all LR patches. This ensures robustness when the Ref and LR images are significantly misaligned.

### *2.3. Methods of Improving Model Performance*

#### 2.3.1. Attention Mechanisms

Attention mechanisms first appeared in the field of machine translation. Later, people found that CNN has different degrees of importance in different spatial dimensions and channel dimensions. Using an attention mechanism, we can break the previous phenomenon of treating all dimensions equally by re-empowering different dimensions, so that neural networks can filter out irrelevant information.

Attention mechanisms include four main types: (1) channel attention mechanism [30,31], (2) spatial attention mechanism [32], (3) temporal attention mechanism [33], (4) hybrid attention mechanism [34,35]. These attention mechanisms can provide different weighted features according to different directions.

The core idea of the CA mechanism is to use the interdependence between channels to model and then the characteristics of different channels will be self-adaptive recalibrated. A typical network of CA mechanisms is Squeeze-and-Excitation Network (SE-Net) [36]. Because of its plug-and-play characteristics, its application in the field of SISR is also very common. For example, RCAN [37] and MSAN [38] applied the CA mechanism, which effectively improved the final reconstruction effect of the network, but they did not consider the role of the SA mechanism. CBAM [39] considers the role of CA and SA at the same time. It connects the two modules in series so that the network can give consideration to the influence of channel and space and retain the most useful "location" while paying attention to the more relevant "content".

#### 2.3.2. Dense Connection Mechanism

ResNet introduces a shortcut connection for the first time, which makes the structure available to the model deeper and reduces the difficulty of training. Densenet [40] introduces a dense connection mechanism, connecting each layer with each subsequent layer, rather than simply connecting a layer. Such an operation has a significant effect on slowing the disappearance of gradients and enhancing feature propagation. SRDenseNet [41] is the application of the dense connection mechanism in the SR field. Through dense connection, the features of the current layer are propagated to all subsequent layers, making full use of low-level features and high-level features, so that the reconstruction ability of the model has been greatly improved. In addition, RDN [42] proposed a new residual dense block (RDB), which combines the dense connection mechanism and the shortcut connection mechanism and can help the model extract local dense features.

## 3. Method

Given the good performance of SRNTT when LR image and Ref image are misaligned to a certain extent, SRNTT is used as the backbone structure in this method. However, we substantially redesigned the texture transfer structure in two aspects. Firstly, a hybrid channel-spatial attention mechanism (Figure 1) is added to the original network. This will be discussed in Section 3.1. Secondly, we replace the original RB with our proposed DHAB (Figure 2) to further improve the network performance. This will be discussed in Section 3.2.

As shown in Figure 1a, we retain the feature swapping part of SRNTT. First, we apply bicubic up-sampling on $I^{LR}$ to obtain the enlarged image $I^{LR\uparrow}$, which has the same size as $I^{HR}$. In order to obtain the Ref image with the same frequency band as $I^{LR\uparrow}$, we apply bicubic down-sampling and up-sampling on $I^{Ref}$ with the same scale and get $I^{Ref\downarrow\uparrow}$ with a blur degree similar to $I^{LR\uparrow}$. As for $I^{LR\uparrow}$ and $I^{Ref\downarrow\uparrow}$ patches, as shown in Figure 1b, we use the inner product in the neural feature space $\phi(I)$ to measure the similarity between neural features.

$$S_{i,j} = \left\langle P_i\left(\phi\left(I^{LR\uparrow}\right)\right), \frac{P_j\left(\phi\left(I^{Ref\downarrow\uparrow}\right)\right)}{\left\| P_j\left(\phi\left(I^{Ref\downarrow\uparrow}\right)\right)\right\|} \right\rangle, \tag{1}$$

where $P_i(\cdot)$ denotes sampling the $i$-th patch from the neural feature map and $S_{i,j}$ is the similarity between the $i$-th $I^{LR\uparrow}$ patch and the $j$-th $I^{Ref\downarrow\uparrow}$ patch. The similarity computation can be efficiently implemented as a set of convolution operations over all $I^{LR\uparrow}$ patches with each kernel corresponding to a $I^{Ref\downarrow\uparrow}$ patch. Where the position of the $I^{Ref\downarrow\uparrow}$ patch with the highest similarity score corresponding to each $I^{LR\uparrow}$ patch is denoted as $\mathrm{P}_{\max(x,y)}$. Each patch in $M$ centered at $(x, y)$ is defined as

$$P_{\omega(x,y)}(M) = \mathrm{P}_{\max(x,y)}\left(\phi\left(I^{Ref}\right)\right), \tag{2}$$

where $\omega(\cdot,\cdot)$ maps patch center to patch index.Here we replace the $I^{Ref\downarrow\uparrow}$ patch with a $I^{Ref}$ patch at the same position to preserve the reference information of the original HR. All the reference patches together constitute the exchange feature map $M$ at this scale.

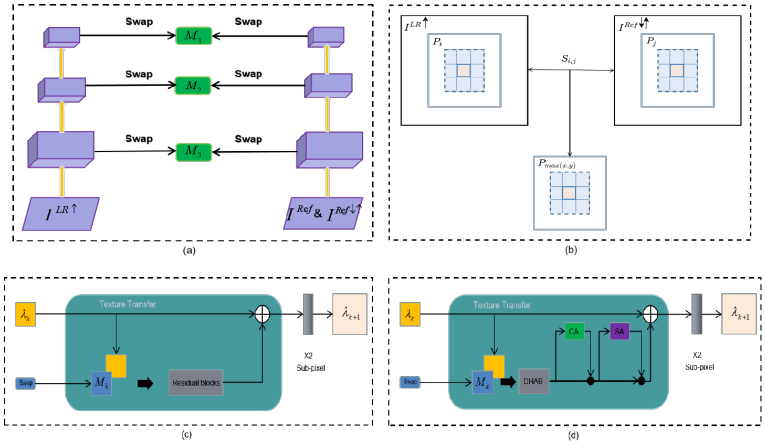


**Figure 1.** The network structure comparison of Super-Resolution by Neural Texture Transfer (SRNTT) and our Residual-Dense Hybrid Attention Network (R-DHAN). (**a**,**b**) The feature swapping network of SRNTT. (**c**) The texture transfer network of SRNTT. (**d**) The texture transfer network of our R-DHAN.

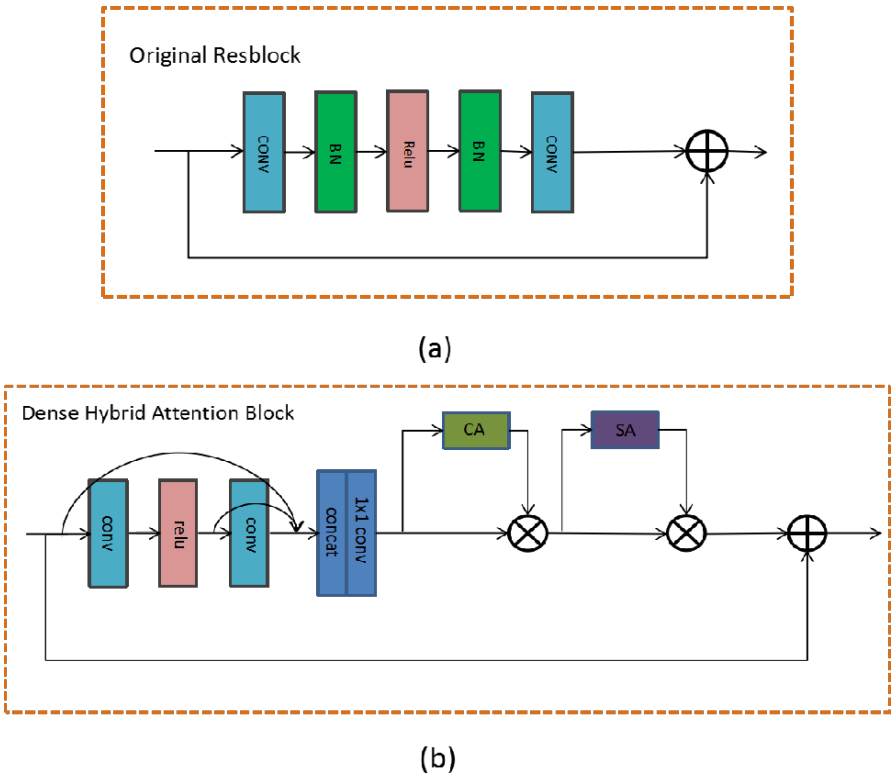


**Figure 2.** The comparison of the original residual block (RB) and our dense hybrid attention block (DHAB). (**a**) RB structure of SRNTT; (**b**) our DHAB architecture.

The texture transfer network of SRNTT is shown in Figure 1c. The base generative network takes the RBs as the main body and uses skip connections [16,43]. The network output $\lambda_k$ of layer $k$ can be expressed:

$$\lambda_{k+1} = [R(\lambda_k \| M_k) + \lambda_k] \uparrow_{2\times}, \tag{3}$$

where $\lambda_k$ denotes $I^{LR}$, $M_k$ denotes $I^{Ref}$ and $R(\bullet)$ denotes the RBs. The channel connection symbol is represented by $\|$ and the upscaling sub-pixel convolution [44] with $2\times$ scale is represented by $\uparrow_{2\times}$. The final reconstruction result SR image is expressed as:

$$I^{SR} = [R(\lambda_k \| M_k) + \lambda_k], \tag{4}$$

Our texture transfer network is shown in Figure 1d. Firstly, RBs in SRNTT are replaced with DHAB. More details about DHAB are given in Section 3.2. In addition, after $\lambda_k$ and $M_k$ are extracted by DHAB, the weighted feature map is generated by adding the hybrid channel-spatial attention mechanism and it is merged with the target content by using skip connection.

### *3.1. Channel-Spatial Attention Mechanism*

Bottleneck Attention Modules (BAMs) and CBAMs are two representative examples of channel-spatial attention mechanisms. Although they both involve the SA module and CA module, they are different in the arrangement and combination of these two modules. BAM keeps the two modules in a parallel structure, while CBAM keeps the two modules in a series structure. Relevant ablation experiments show that connecting CA and SA in sequence can bring optimal performance enhancement. Therefore, here we use the idea of CBAM for reference. After DHAB extracts relevant features, we add channel attention and spatial attention in sequence. The relevant features extracted by DHAB are affected by the original features $\lambda_k$ to focus on more important and useful areas and content.

In the CBAM module, the feature output $f_1$ after CA module is expressed as:

$$f_1 = Ca(f) \otimes f, \tag{5}$$

The feature output $f_2$ after the SA module is expressed as:

$$f_2 = Sa(f_1) \otimes f_1, \tag{6}$$

where $f \in \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times\mathrm{c}}$ denotes the input feature maps of the CBAM. $Ca \in \mathbb{R}^{1\times1\times\mathrm{c}}$ denotes the CA module. $Sa \in \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times1}$ denotes the SA module. Moreover, The height, width and channels of the feature map are represented by $h$, $w$ and $c$, respectively. $\otimes$ denotes multiplication element-wise.

$$Ca(f) = \sigma\Big(MLP\Big(F^c_{\mathrm{arg}}(f)\Big) + MLP(F^c_{max}(f))\Big) \tag{7}$$

$$= \sigma\Big(\Big(W_2\Big(W_1\Big(F^c_{\mathrm{avg}}(f)\Big)\Big) + (W_2(W_1(F^c_{max}(f))))\Big)\Big), \tag{8}$$

where $F^c_{\mathrm{avg}}$ denotes the operation of average-pooling in the CA module. $F^c_{\mathrm{max}}$ denotes the operation of max-pooling in the CA module. $MLP$ denotes a multilayer perception network. The weights of $MLP$ are denoted by $W_1 \in \mathbb{R}^{\mathrm{c}\backslash\mathrm{t}\times\mathrm{c}}$ and $W_2 \in \mathbb{R}^{\mathrm{c}\times\mathrm{c}\backslash\mathrm{t}}$, where $t$ denotes the scale of the number of channels, $\sigma$ denotes the sigmoid activation function.

$$\mathrm{Sa}(f_1) = \sigma\Big(F^{7\times7}\Big(CAT\Big(F^s_{\mathrm{mean}}\Big(F^{'}\Big), F^s_{\mathrm{max}}(f_1)\Big)\Big)\Big), \tag{9}$$

where $F^s_{\mathrm{mean}}$ denotes the operation of obtaining the mean value of the feature maps. $F^s_{\mathrm{max}}$ denotes the operation of obtaining the maximum value of the feature maps. Their output results are two SA maps. $CAT$ denotes the connection operation of SA maps. $F^{7\times7}$ means convolution with a filter of size $7 \times 7$. $\sigma$ denotes the sigmoid activation function.

### *3.2. Dense Hybrid Attention Block*

The improvements made by our DHAB compared with the original RB are shown in Figure 2, mainly reflected in LFF and CBAM.

Since the existence of the BN layer will not have a substantial impact on the super-resolution task, we deleted the BN layer in the original network to lighten the network

and further release the memory space of GPU. In addition, extracting and aggregating features can maximize the use of these features by the network, which can enhance the characterization ability of the network and further improve the final SR reconstruction effect. Thus, we have added the LFF module to DHAB. The LFF module can adaptively fuse some layers in the current DHAB with the preceding DHAB. However, because the features of the $j - l_{th}$ DHAB are directly introduced into the $j_{th}$ DHAB, resulting in too many features, we introduce a $1 \times 1$ convolution layer that performs the dimensionality reduction operation to ensure the constant output dimension. The above operations are expressed as:

$$f_{j,LLF} = S^{j}_{LLF}\left(\left[f_{j-1}; f_{j,\sigma}; f_{j,conv2}\right]\right), \tag{10}$$

where $S^{j}_{LFF}$ denotes the function of $1 \times 1$ convolution layer in the $j_{th}$ DHAB. $f_{j-1}$ denotes the output of the $j_{th}$ DHAB. $f_{j,\sigma}$ denotes the feature maps produced by the activation function. $f_{j,conv2}$ denotes the feature maps generated by the second convolution layer in the $j_{th}$ DHAB. The symbol $[\bullet]$ denotes the concatenation of the feature maps. After the LFF module, we further apply CBAM to distinguish the importance of different contents and regions.

## 4. Experiments

### *4.1. Experiment Settings*

In this section, the datasets used in this paper and the specific experimental details will be introduced.

#### 4.1.1. Datasets

Since there is no publicly available sensing dataset for super-resolution reference images, we choose to build our training set and test set using publicly available ArcGIS online maps and Google Earth images with high resolution. Among them, HR images are from ArcGIS online map images of Qingdao, China, and Jinan, China, taken in 2017, with a resolution of 0.8 m. The Ref image is from 2019 Google Earth images with a resolution of 0.5 m. Image acquisition and matching are based on longitude and latitude matching. We perform bicubic downsampling on HR images with a scale factor of 4 to obtain LR images. The size of HR and Ref images is $160 \times 160$ pixels, correspondingly, the size of the LR image is $40 \times 40$ pixels. Finally, we obtained 9160 pairs of samples, each pair of samples includes LR images, Ref images and HR ground real images. Figure 3 shows some examples in the dataset.

In order to further prove the effectiveness of the proposed model, we will compare it with the classical algorithm on the open benchmark natural dataset named CUFED5. It is composed of 11,871 training pairs. Each pair contains an original HR image and a corresponding reference image at $160 \times 160$ resolution. Figure 4 shows some examples in CUFED5.

In addition, we tested our method on the real remote sensing images of the GF-2 satellite, which has a spatial resolution of 2 m. The corresponding Ref images are still collected from Google Earth.

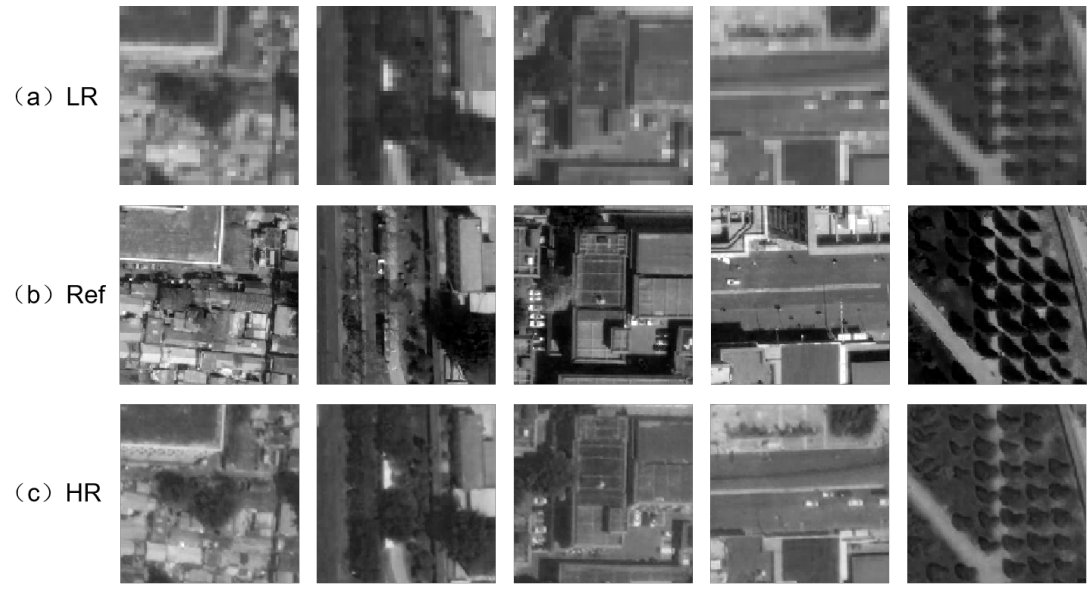


**Figure 3.** Examples from our private training dataset.


**Figure 4.** Examples from the CUFED5 dataset.

#### 4.1.2. Evaluation Details

During the training of R-DHAN, we set the batch size to 16. The Adam method [45] with $\beta_1 = 0.9$ and $\beta_1 = 0.99$ is used as the optimizer. We set the number of training iterations of the network to 20 and the initial learning rate to $5 \times 10^{-4}$, reduced by half every 10 epochs. The above parameter settings are close to SRNTT to ensure fairness. In addition, our experiments are conducted under the Pytorch framework and two NVIDIA GTX 1080Ti GPUs are used for model training. Like most SR methods, we use PSNR and SSIM as evaluation indicators. The higher the score, the better the performance.

### *4.2. Comparisons with the Other Methods*

In order to verify the effectiveness of our method, we compared our method with several other methods. The comparison methods include four SISR methods (namely SRCNN [19], FSRCNN [20], VDSR [9], EDSR [10]) and one RefSR method, namely SRNTT [14]. We use our private data set and public data set CUFED5 to train and test all models under the same conditions.

Table 1 shows the mean PSNR and SSIM values of the reconstructed HR images of our private test set on the ×2 and ×4 enlargement; Table 2 shows the mean PSNR and SSIM values of the reconstructed HR images of the CUFED5 test set on the ×2 and ×4 enlargement.

**Table 1.** Our private dataset ×2 and ×4 test results. Best results are in bold.

| | **SRCNN** | **FSRCNN** | **VDSR** | **EDSR** | **SRNTT** | **Ours** |
|---|---|---|---|---|---|---|
| | **PSNR/SSIM** | **PSNR/SSIM** | **PSNR/SSIM** | **PSNR/SSIM** | **PSNR/SSIM** | **PSNR/SSIM** |
| ×2 | 27.23/0.808 | 27.95/0.836 | 28.17/0.844 | 28.65/0.852 | 29.10/0.865 | **29.25/0.869** |
| ×4 | 26.69/0.793 | 27.08/0.805 | 27.33/0.814 | 27.77/0.825 | 28.01/0.836 | **28.12/0.839** |

**Table 2.** CUFED5 dataset ×2 and ×4 test results. Best results are in bold.

| | **SRCNN** | **FSRCNN** | **VDSR** | **EDSR** | **SRNTT** | **Ours** |
|---|---|---|---|---|---|---|
| | **PSNR/SSIM** | **PSNR/SSIM** | **PSNR/SSIM** | **PSNR/SSIM** | **PSNR/SSIM** | **PSNR/SSIM** |
| ×2 | 26.45/0.790 | 26.93/0.798 | 27.12/0.806 | 27.40/0.814 | 27.66/0.817 | **27.83/0.822** |
| ×4 | 25.33/0.745 | 25.55/0.756 | 25.72/0.765 | 25.93/0.778 | 26.24/0.784 | **26.37/0.788** |

As can be seen from the above table, our method achieved the best performance on both datasets, which proves the effectiveness of our method.

Figures 5 and 6 show the SR results with scaling factor ×2 on our private test set; Figures 7 and 8 show the SR results with scaling factor ×4 on our private test set. Through observation, it can be concluded that: (1) the reconstructed images obtained by the bicubic method are very fuzzy and a large amount of detailed information is lost; (2) although SRCNN, FSRCNN, VDSR and EDSR have made great improvements in content details, they cannot obtain deeper information from LR images, resulting in blurred image contours; (3) SRNTT can restore better texture details, but the final reconstruction result is still not ideal due to the low utilization of features between channels; (4) compared with other methods, our method can show better edge details.

Figures 9 and 10 show the SR results with scaling factor ×2 on CUEFD5 test set; Figures 11 and 12 show the SR results with scaling factor ×2 on CUEFD5 test set. Areas with obvious differences are locally amplified. As can be seen in Figure 9, R-DHAN better restores the detailed texture of the character's teeth and obtains significantly clearer edges; in Figure 10, the lines on the door frame are sharper and clearer. Similarly, in Figure 11, the reconstruction effect of the human nose and mouth is significantly better than other methods; In Figure 12, the URL watermark below the image is also restored to the greatest extent in our method. The appeal results show that our method is superior to other methods in visual quality.

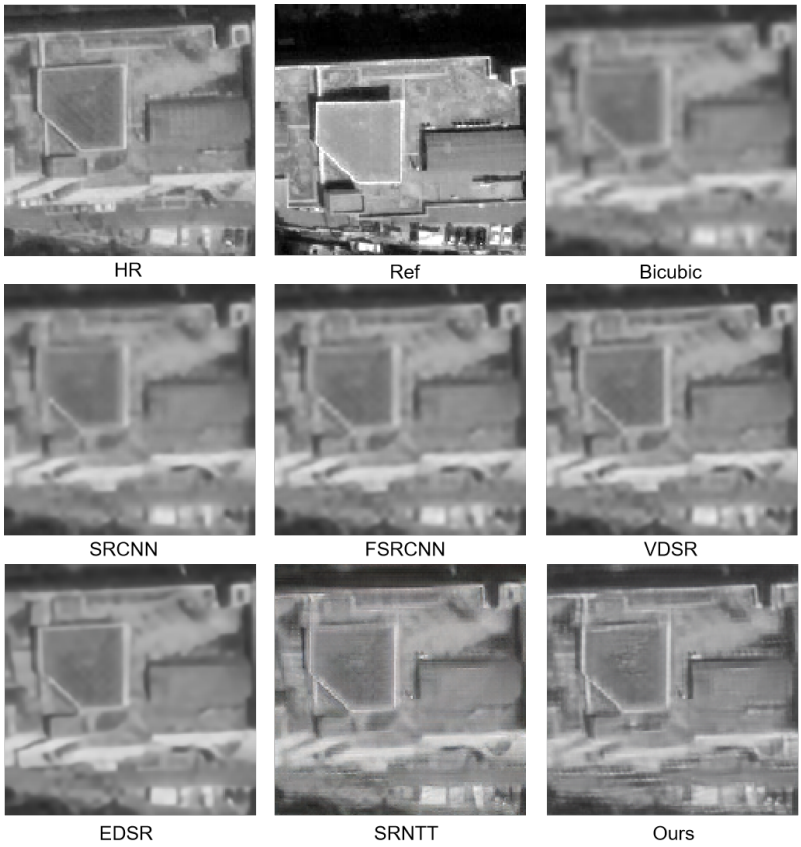


**Figure 5.** Visual comparison on example 1 of our private test set with a scaling factor of ×2.

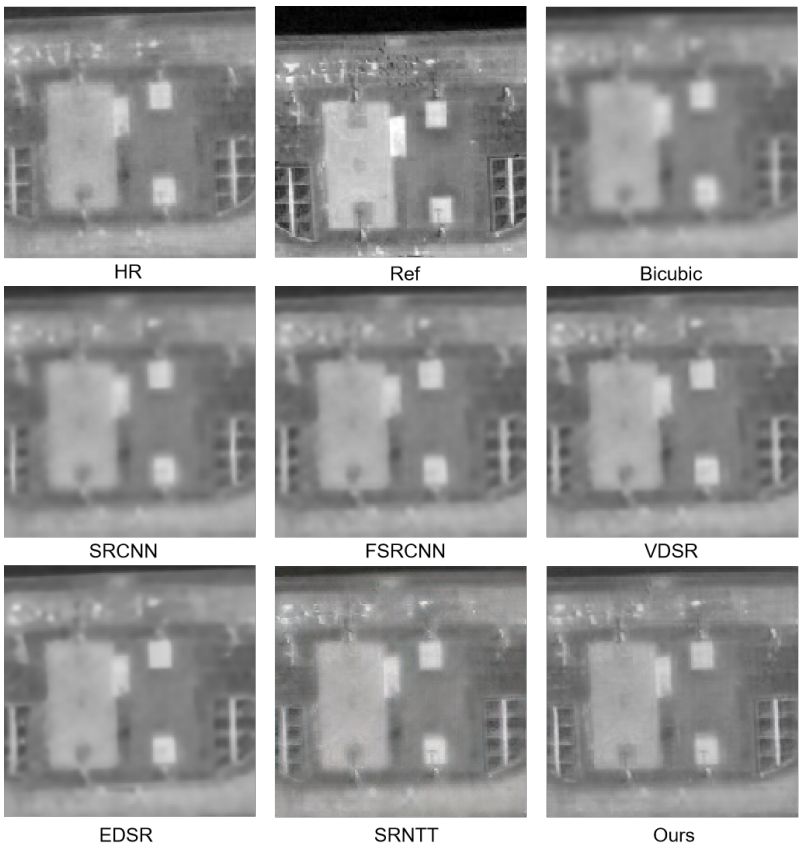


**Figure 6.** Visual comparison on example 2 of our private test set with a scaling factor of ×2.

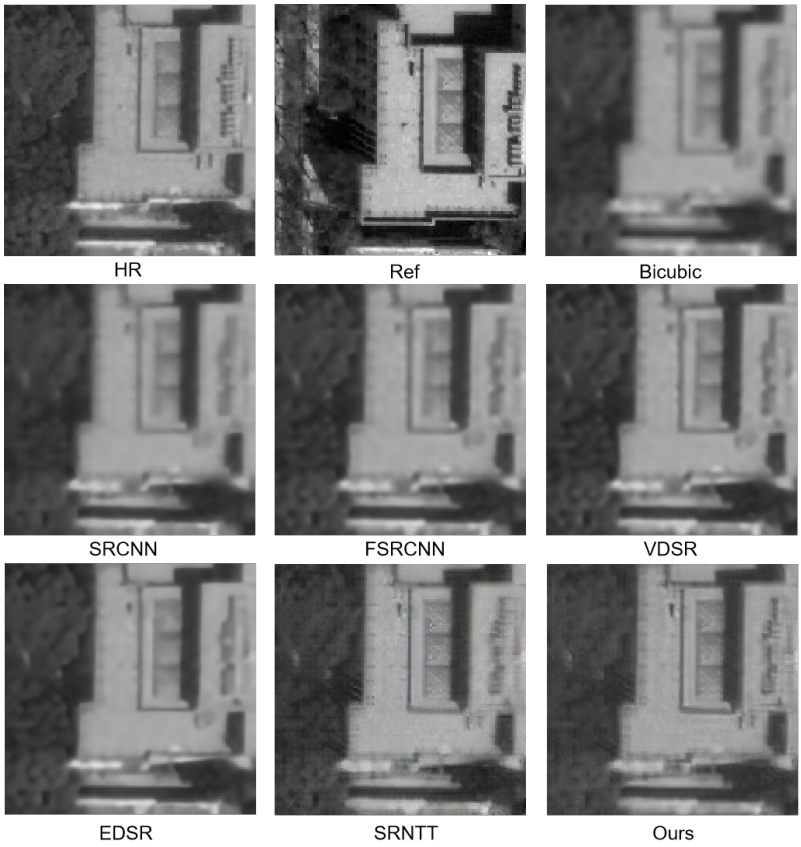


**Figure 7.** Visual comparison on example 1 of our private test set with a scaling factor of ×4.

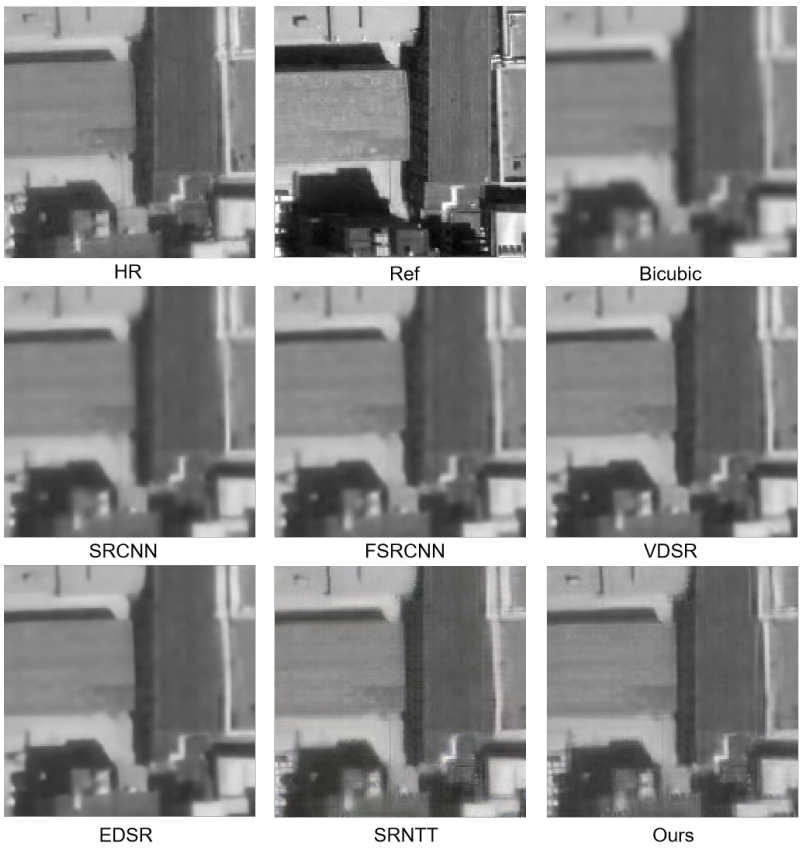


**Figure 8.** Visual comparison on example 2 of our private test set with a scaling factor of ×4.

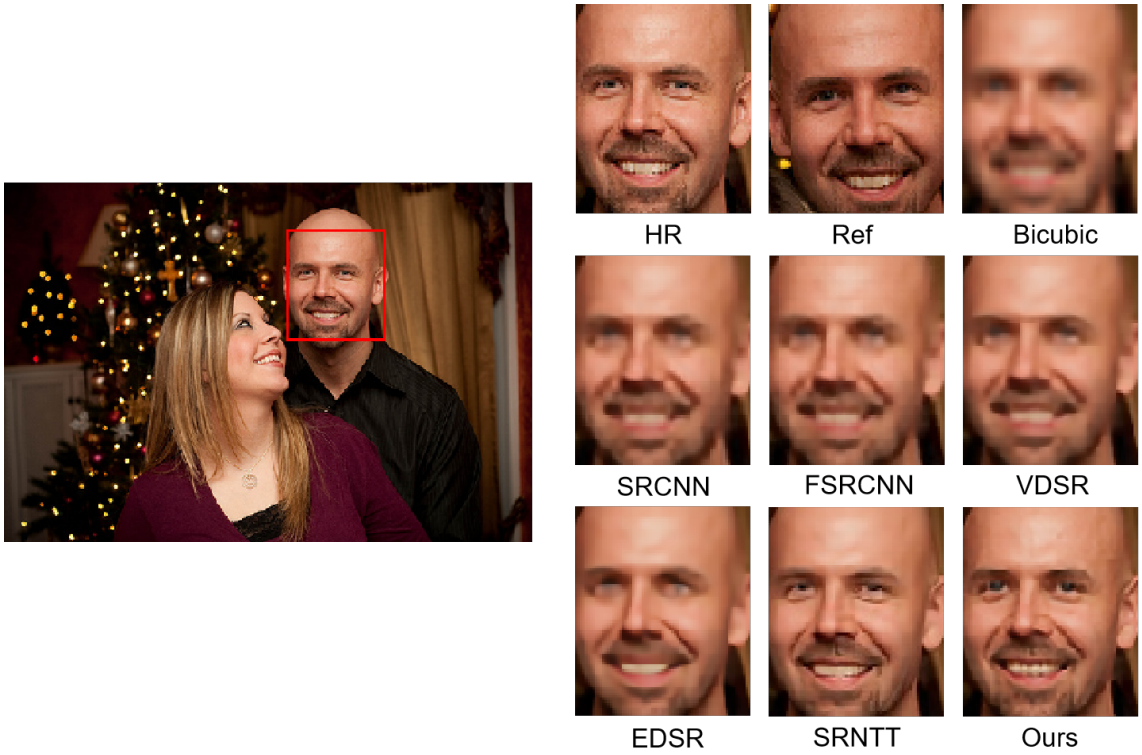


**Figure 9.** Visual comparison on example 1 of CUFED5 test set with a scaling factor of ×2.


**Figure 10.** Visual comparison on example 2 of CUFED5 test set with a scaling factor of ×2.


**Figure 11.** Visual comparison on example 1 of CUFED5 test set with a scaling factor of ×4.


**Figure 12.** Visual comparison on example 2 of CUFED5 test set with a scaling factor of ×4.

### *4.3. Ablation Studies*

In this section, based on the CUFED5 dataset, we conducted some ablation experiments to prove the effectiveness of the main components of our proposed method, including the DHAB module and the CBAM module. Table 3 shows the results of 2× and 4× SR. By observing that the DHAB module can improve the network performance by 0.14 and

0.11 dB on different scales, the CBAM can improve the network performance by 0.05 and 0.04 dB on different scales. It can be seen that the DHAB module contributes more to the improvement of the model than the CBAM and the best performance of the model can be obtained when the two modules are used at the same time.

**Table 3.** Ablation studies of different modules for different scales.

| Scale | DHAB | CBAM | PSNR/SSIM |
|---|---|---|---|
| ×2 | × | × | 27.66/0.817 |
| | × | √ | 27.71/0.818 |
| | √ | × | 27.80/0.820 |
| | √ | √ | 27.83/0.822 |
| Scale | DHAB | CBAM | PSNR/SSIM |
| ×4 | × | × | 26.24/0.784 |
| | × | √ | 26.28/0.784 |
| | √ | × | 26.35/0.786 |
| | √ | √ | 26.37/0.788 |

In order to further verify the effectiveness of the DHAB module, we replaced the DHAB module with the same number of RB, DB and RDB without using the CBAM. The performance comparison is shown in Table 4.

**Table 4.** Performance comparison of different residual blocks. Best results are in bold.

| | **With RB** | **With DB** | **With RDB** | **With DHAB** |
|---|---|---|---|---|
| | **PSNR/SSIM** | **PSNR/SSIM** | **PSNR/SSIM** | **PSNR/SSIM** |
| ×2 | 27.66/0.817 | 27.71/0.818 | 27.74/0.819 | **27.80/0.820** |
| ×4 | 26.24/0.784 | 26.28/0.784 | 26.31/0.785 | **26.35/0.786** |

The results in Table 4 show that the DHAB achieves higher PSNR and SSIM, which proves the powerful characterization ability of the DHAB. Similarly, to further verify the effectiveness of the CBAM, we replace the DHAB with the original RB. In this case, we replace the CBAM with the CA block and the SA block, respectively. The performance comparison is shown in Table 5.

**Table 5.** Performance comparison of different attention modules. Best results are in bold.

| | **RB+CA** | **RB+SA** | **RB+CA+SA** |
|---|---|---|---|
| | **PSNR/SSIM** | **PSNR/SSIM** | **PSNR/SSIM** |
| ×2 | 27.68/0.818 | 27.67/0.817 | **27.71/0.818** |
| ×4 | 26.26/0.784 | 26.25/0.784 | **26.28/0.784** |

By observing the results in Table 5, it is easy to conclude that, in terms of model performance, the CBAM is stronger than the CA block and the SA block. There is no information regarding the training steps, for example, the amount of data required to train the network, convergence tests and complexity. In addition, in Table 6, we report the number of model parameters, the training time and the inference time of different SISR and RefSR methods. Compared with the results of SRNTT, we improved the network performance under the premise of better indicators, which proves the superiority of our method.

We report the number of model parameters, training time and reasoning time of different SISR and RefSR methods. Compared with the results of SRNTT, we improved the

network performance under the premise of better indicators, which proves the superiority of our method.

**Table 6.** Comparison of model parameters and inference runtime.

| Method | Param (M) | Training Time (h) | Inference Time (s) |
|---|---|---|---|
| SRCNN | 0.48 | 18.3 | 0.0035 |
| FSRCNN | 0.30 | 15.6 | 0.0027 |
| VDSR | 0.67 | 24.6 | 0.0046 |
| EDSR | 1.08 | 22.1 | 0.0133 |
| SRNTT | 4.20 | 28.0 | 6.8045 |
| Ours | 4.16 | 27.8 | 6.9232 |

### *4.4. Results on Real Remote Sensing Data*

In this section, we use remote sensing images from the real world to verify the robustness of our proposed method. The model was trained on our private training set and tested on the remote sensing images from GaoFen-2. Figures 13 and 14 show the results of ×2 and ×4 enlargements. Even if the spatial resolution of the input LR image is different from that of the LR image in the training data set, our method can still effectively improve the visual quality of remote sensing images.

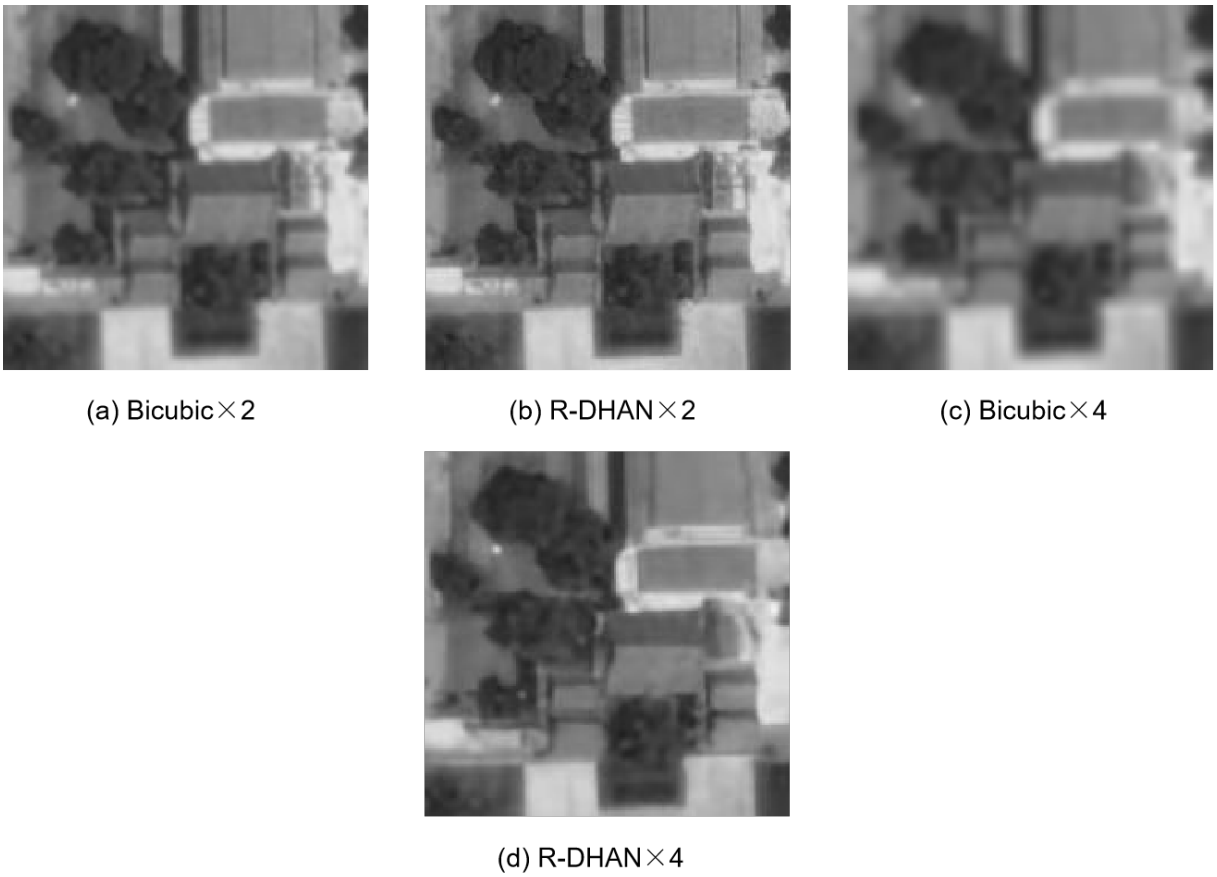


**Figure 13.** SR results of real of ×2 and ×4 scale factors for the real example 1 of the GaoFen-2 satellite. (**a**–**d**) The results of Bicubic×2, R-DHAN×2, Bicubic×4 and R-DHAN×4, respectively.

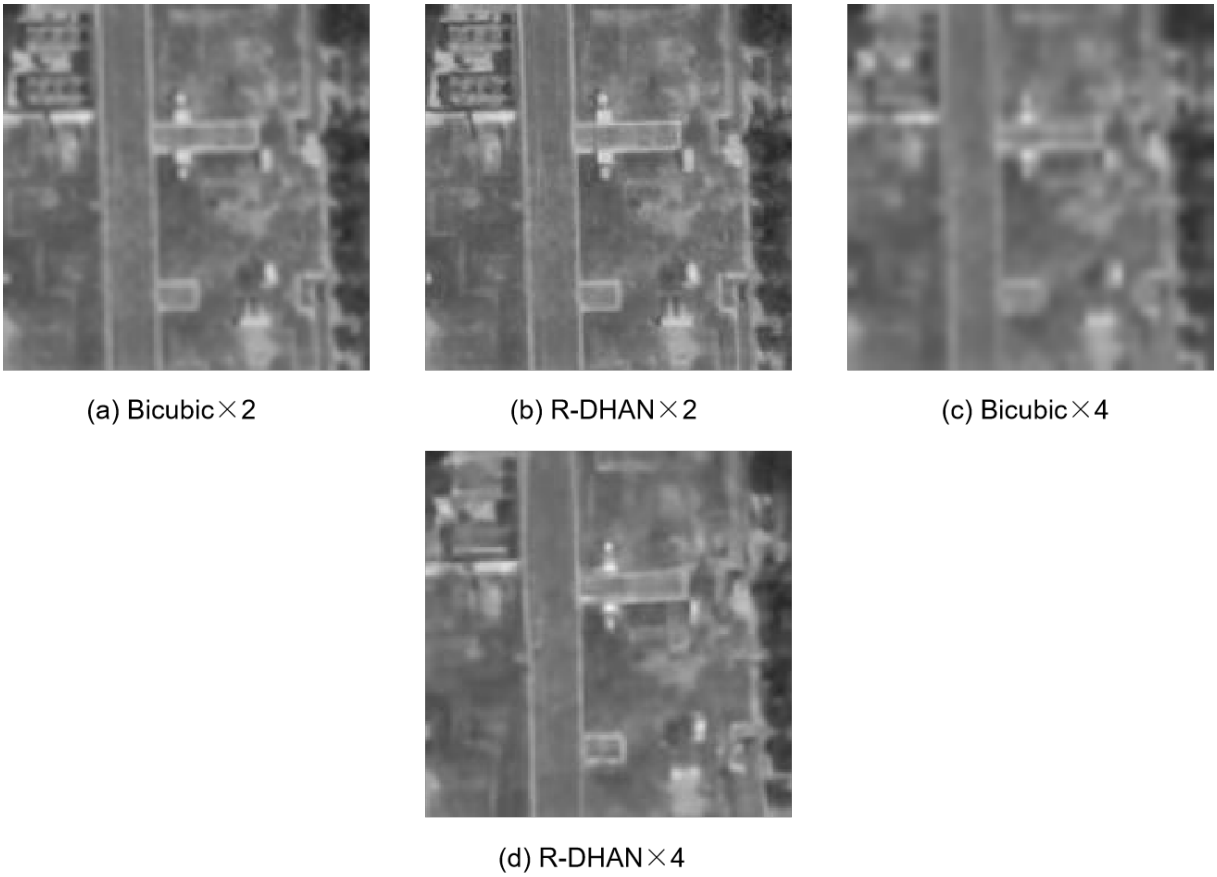


**Figure 14.** SR results of real of ×2 and ×4 scale factors for the real example 2 of the GaoFen-2 satellite. (**a**–**d**) The results of Bicubic×2, R-DHAN×2, Bicubic×4 and R-DHAN×4, respectively.

## 5. Discussion

Compared with SRNTT, we obtain better performance without introducing more parameters. This paper proves that the mixed use of the dense connection mechanism and the attention mechanism is effective in terms of improving the SR performance of remote sensing images. However, due to the lack of datasets of multiband remote sensing images, we did not carry out relevant experiments on them. Therefore, the direction of our future work is to make datasets of multiband remote sensing images and further verify the robustness of our methods.

## 6. Conclusions

In this paper, we propose a new remote sensing image SR network named R-DHAN to solve the problems of insufficient resolution and blurred details of remote sensing images. Specifically, we design a dense hybrid attention block (DHAB), which makes good use of the rich high-frequency information in the multi-level feature map. In addition, we added a channel-spatial attention block to pay more attention to the more important and difficult reconstruction areas. A large number of experiments show that our method is superior to many classical methods in quality and accuracy. In addition, experiments on real satellite data (GF-2) verify the robustness of R-DHAN.

**Author Contributions:** Conceptualization, B.Y. and B.L.; methodology, B.Y. and J.G.; software, B.Y.; validation, B.Y. and J.S.; investigation, S.L. and J.G.; resources, J.S. and S.L.; data curation, J.G. and G.X.; writing—original draft preparation, B.Y.; writing—review and editing, J.G. and J.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by the Scientic Research Leader Studio of Jinan (No. 2021GXRC081), Joint Project for Smart Computing of Shandong Natural Science Foundation (No. ZR2020LZH015).

**Data Availability Statement:** Data sharing not applicable.

**Acknowledgments:** The author Bo Yu would like to thank all teachers who provided guidance and help for me to complete this project.

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

## Abbreviations

| | |
|---|---|
| BAM | Bottleneck Attention Module |
| BN | Batch normalization |
| CA | Channel attention |
| CBAM | Convolution block attention module |
| CNN | Convolutional Neural Network |
| DB | Dense block |
| DHAB | Dense hybrid attention block |
| DRCN | Deeply-Recursive Convolutional Network |
| EDSR | Enhanced Deep Super-Resolution Networks |
| ESRGAN | Enhanced Super-Resolution Generative Adversarial Networks |
| FSRCNN | Accelerating the Super-Resolution Convolutional Neural Network |
| GAN | Generative Adversarial Network |
| LFF | Local feature fusion |
| MSAN | Multiscale Attention Network |
| PSNR | Peak signal-to-noise ratio |
| RB | Residual block |
| RCAN | Residual Channel Attention Networks |
| RDB | Residual dense block |
| R-DHAN | Residual-Dense Hybrid Attention Network |
| RefSR | Reference based super-resolution |
| SA | Spatial attention |
| SE-Net | Squeeze-and-Excitation Network |
| SISR | Single image super-resolution |
| SRCNN | Image Super-Resolution Using Deep Convolutional Networks |
| SRGAN | Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network |
| SRNTT | Super-Resolution by Neural Texture Transfer |
| SSIM | Structural similarity ratio |
| VDSR | Accurate Image Super-Resolution Using Very Deep Convolutional Networks |

## References

1. Dong, Z.; Wang, M.; Wang, Y.; Zhu, Y.; Zhang, Z. Object detection in high resolution remote sensing imagery based on convolutional neural networks with suitable object scale features. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 2104–2114. [CrossRef]
2. Amit, S.N.K.B.; Shiraishi, S.; Inoshita, T.; Aoki, Y. Analysis of satellite images for disaster detection. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 5189–5192.
3. Mathieu, R.; Freeman, C.; Aryal, J. Mapping private gardens in urban areas using object-oriented techniques and very high-resolution satellite imagery. *Landsc. Urban Plan.* **2007**, *81*, 179–192. [CrossRef]
4. Pan, B.; Shi, Z.; Xu, X.; Shi, T.; Zhang, N.; Zhu, X. CoinNet: Copy initialization network for multispectral imagery semantic segmentation. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 816–820. [CrossRef]
5. Bai, Y.; Zhang, Y.; Ding, M.; Ghanem, B. Sod-mtgan: Small object detection via multi-task generative adversarial network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 206–221.
6. Thurnhofer, S.; Mitra, S.K. Edge-enhanced image zooming. *Opt. Eng.* **1996**, *35*, 1862–1870. [CrossRef]
7. Fekri, F.; Mersereau, R.M.; Schafer, R.W. A generalized interpolative VQ method for jointly optimal quantization and interpolation of images. In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), Seattle, WA, USA, 15 May 1998; IEEE: Piscataway, NJ, USA, 1998; Volume 5, pp. 2657–2660.
8. Yang, W.; Zhang, X.; Tian, Y.; Wang, W.; Xue, J.H.; Liao, Q. Deep learning for single image super-resolution: A brief review. *IEEE Trans. Multimed.* **2019**, *21*, 3106–3121. [CrossRef]
9. 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.

10. 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.
11. Liu, B.; Zhao, L.; Li, J.; Zhao, H.; Liu, W.; Li, Y.; Wang, Y.; Chen, H.; Cao, W. Saliency-Guided Remote Sensing Image Super-Resolution. *Remote Sens.* **2021**, *13*, 5144. [CrossRef]
12. Song, H.; Xu, W.; Liu, D.; Liu, B.; Liu, Q.; Metaxas, D.N. Multi-stage feature fusion network for video super-resolution. *IEEE Trans. Image Process.* **2021**, *30*, 2923–2934. [CrossRef] [PubMed]
13. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; Volume 27.
14. Zhang, Z.; Wang, Z.; Lin, Z.; Qi, H. Image Super-Resolution by Neural Texture Transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7982–7991.
15. Zheng, H.; Ji, M.; Wang, H.; Liu, Y.; Fang, L. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 88–104.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
17. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
18. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
19. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [CrossRef] [PubMed]
20. 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; pp. 391–407.
21. 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.
22. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
23. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 694–711.
24. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
25. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
26. Yue, H.; Sun, X.; Yang, J.; Wu, F. Landmark image super-resolution by retrieving web images. *IEEE Trans. Image Process.* **2013**, *22*, 4865–4878. [PubMed]
27. Wang, Y.; Liu, Y.; Heidrich, W.; Dai, Q. The light field attachment: Turning a dslr into a light field camera using a low budget camera ring. *IEEE Trans. Vis. Comput. Graph.* **2016**, *23*, 2357–2364. [CrossRef] [PubMed]
28. Boominathan, V.; Mitra, K.; Veeraraghavan, A. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In Proceedings of the 2014 IEEE International Conference on Computational Photography (ICCP), Evanston, IL, USA, 13–15 May 2016; pp. 1–10.
29. Zheng, H.; Ji, M.; Han, L.; Xu, Z.; Wang, H.; Liu, Y.; Fang, L. Learning Cross-scale Correspondence and Patch-based Synthesis for Reference-based Super-Resolution. In Proceedings of the BMVC, London, UK, 4–7 September 2017; p. 2.
30. Haut, J.M.; Fernandez-Beltran, R.; Paoletti, M.E.; Plaza, J.; Plaza, A. Remote sensing image superresolution using deep residual channel attention. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9277–9289. [CrossRef]
31. Tong, W.; Chen, W.; Han, W.; Li, X.; Wang, L. Channel-attention-based DenseNet network for remote sensing image scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4121–4132. [CrossRef]
32. Shi, W.; Du, H.; Mei, W.; Ma, Z. (SARN) spatial-wise attention residual network for image super-resolution. *Vis. Comput.* **2021**, *37*, 1569–1580. [CrossRef]
33. Tran, D.T.; Iosifidis, A.; Kanniainen, J.; Gabbouj, M. Temporal attention-augmented bilinear network for financial time-series data analysis. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *30*, 1407–1418. [CrossRef] [PubMed]
34. Chen, J.; Wan, L.; Zhu, J.; Xu, G.; Deng, M. Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 681–685. [CrossRef]
35. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.-S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.

36. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
37. 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.
38. 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]
39. 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.
40. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
41. 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.
42. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 2480–2495. [CrossRef] [PubMed]
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 630–645.
44. 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.
45. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.