*remote sensing*

MDPI

*Article*

# Deep Residual Squeeze and Excitation Network for Remote Sensing Image Super-Resolution

**Jun Gu [1,2,3], Xian Sun [1,2], Yue Zhang [1,2], Kun Fu [1,2,3] and Lei Wang [1,2,*]**

1 Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China
2 Key Laboratory of Network Information System Technology (NIST), Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China
3 School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
* Correspondence: wanglei@mail.ie.ac.cn

Received: 30 June 2019; Accepted: 1 August 2019; Published: 3 August 2019


**Abstract:** Recently, deep convolutional neural networks (DCNN) have obtained promising results in single image super-resolution (SISR) of remote sensing images. Due to the high complexity of remote sensing image distribution, most of the existing methods are not good enough for remote sensing image super-resolution. Enhancing the representation ability of the network is one of the critical factors to improve remote sensing image super-resolution performance. To address this problem, we propose a new SISR algorithm called a Deep Residual Squeeze and Excitation Network (DRSEN). Specifically, we propose a residual squeeze and excitation block (RSEB) as a building block in DRSEN. The RSEB fuses the input and its internal features of current block, and models the interdependencies and relationships between channels to enhance the representation power. At the same time, we improve the up-sampling module and the global residual pathway in the network to reduce the parameters of the network. Experiments on two public remote sensing datasets (UC Merced and NWPU-RESISC45) show that our DRSEN achieves better accuracy and visual improvements against most state-of-the-art methods. The DRSEN is beneficial for the progress in the remote sensing images super-resolution field.

**Keywords:** remote sensing; single image super-resolution; convolutional neural network

## 1. Introduction

High-resolution (HR) images with rich detailed textures and critical information play an essential part in later remote sensing image analysis, such as target detection, object recognition, land cover classification, etc. However, due to hardware limitation and large detection distance, the spatial resolution of these satellite imageries in ordinary civilian applications is often low-resolution (LR). Instead of enhancing the physical imaging technology, many researchers aim to reconstruct a visually pleasing HR remote sensing image from existing LR observed images, which is called single image super-resolution (SISR) [1].

In the past few years, a series of SR techniques based on the sparsity prior of image statistics have been proposed to recover HR remote sensing images. A dictionary of image edges and contours were utilized by Yang et al. [2] and Dong et al. [3]. The compressive sensing and structural self-similarity of the remote sensing images were used for the super-resolution task by Pan et al. [4]. The sparse properties in the spectral and spatial domain were explored by Li et al. [5] to recover the resolution of hyperspectral images. However, these methods all utilize low-level features of the remote sensing images.

With the immense popularity of deep learning, convolutional neural network (CNN) stands out as a powerful image super-resolution basic method. These deep learning methods learn high-level

feature representation automatically from data to provide significantly improved resolution restoration performance. Among them, Dong et al. firstly introduce a three-layer CNN into image SR named SRCNN [6], and achieved considerable improvement. Kim et al. increased the network depth in VDSR [7] and DRCN [8], achieving notable improvements over SRCNN. Tai et al. [9] later introduced recursive blocks in DRRN for deeper networks. These methods would have to first interpolate the original LR images to the desired size and then apply them into the neural network. This pre-processing method increases computation greatly and inevitably loses some details of the original LR inputs.

To deal with the pre-processing problem above, FSRCNN [10] extracts the features from the original LR images and then introduces a transposed convolution layer to up-scale the spatial resolution at the tail of the network. An efficient sub-pixel convolution layer was proposed in ESPCN [11] to up-scale the final LR feature maps into the HR output. Then, this sub-pixel convolutional layer became the main choice for deep architecture. SRResNet [12] took advantage of residual learning to construct a deeper network and achieved better performance. By removing unnecessary modules in conventional residual networks, Lim et al. [13] proposed EDSR and MDSR by removing the batch normolization layer in SRResNet, which achieves significant improvement.

In order to make the image more visual pleasing, many generative adversarial networks (GAN) [14] based models were proposed for single image super resolution. Leding et al. first introduce the GAN framework and the perceptual loss [15] into SRGAN [12], which achieves more visually pleasing images. Compared with the L1 or L2 loss function supervision method, GAN can produce visually sensible samples, but the accuracy of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [16] evaluation decreases.

With regard to the super-resolution of remote sensing images, Lei et al. proposed an algorithm named local-global combined networks (LGCNet) [17] to learn multilevel representations of remote sensing images. Haut et al. [18] learn the distribution of the image based on GAN and proposed an unsupervised SISR method. Xu et al. [19] propose a novel method named deep memory connected network (DMCN) to ease time consumption of reconstructing the remote sensing images resolution.

Recently, many super resolution methods tend to grow the network depth for better performance. As the depth grows, the features in the deep neural network would be hierarchical with different receptive fields. In addition, objects in remote sensing images have different scales due to factors such as the angle of view and the scale of the zoom. Therefore, hierarchical features from the network will provide more clues for image resolution reconstruction tasks. However, most methods based on the neural networks neglect to use hierarchical features and blindly increase network depth for super resolution. The highly complex spatial distribution of remote sensing images indicates that higher level abstraction and better data representation are essential for applications. Meanwhile, the ground objects of remote sensing images usually share a wider range of their scales, saying that the object itself and its surrounding environment are mutually coupling in the joint distribution of their image patterns [17], which is highly different from those of natural images. Therefore, for the super-resolution task of remote sensing images, a more powerful representational ability is of crucial importance for achieving better performance.

Aiming to promote remote sensing image super-resolution tasks, we design a novel network structure named deep residual squeeze and excitation network (DRSEN) with the inspiration of some newly emerging concepts in deep learning. Based on the new concepts, we have designed an efficient network structure and achieved satisfactory results. The whole network can be divided into the identity branch and the residual branch, which can be regarded as a variant of residual learning based on global thought. The identity branch directly takes an LR RGB image as input and outputs the HR counterpart. Unlike the global residual path of the linear stack of several convolutional layers in EDSR, our identity branch uses a single convolution layer as the feature extraction part and then uses the ESPCN [11] structure to output the image. Such identity branch reduces network parameters by absorbing some redundant convolution layers and guarantees the same accuracy. In particular, for the residual branch, the residual squeeze and excitation block (RSEB) is proposed as the basic

building block for our network. With RSEB, a high performance network can be built in a simple but rather effective manner. The RSEB can be regarded as a variant building block of ResNet [20]. RSEB consists of three parts: the feature extraction part, the local feature fusion (LFF) part and the squeeze and excitation (SE) part. The feature extraction module consists of two convolutional layers. Concatenating the states of preceding RSEB and some preceding layers within the current RSEB, LFF extracts local features inside each block and aggregates them simultaneously. We exploit the local feature fusion module to fuse different layer features and explore the most appropriate fusion method. The LFF part can make use of the features and adaptively preserve the information to improve the network representation ability. Squeeze and excitation module is first proposed in SENet [21] to improve network presentation capabilities and has also been rapidly applied in the field of image super-resolution. The authors in [22–25] have verified that the attention mechanism contributes to image super-resolution tasks. The output of one RSEB has direct access to the next RSEB. Furthermore, in contrast to EDSR, we remove redundant convolutional layers after ESPCN to reduce network parameters and not affect final result performance. In summary, the main contributions of this work are as follows:

1. We propose a deep residual squeeze and excitation network (DRSEN) for remote sensing satellite image SR reconstruction. Our DRSEN is in a convenient and effective end-to-end training manner and obtains a better accuracy and visual super-resolution performance.
2. We propose a modified residual block named RSEB, which contains local feature fusion (LFF) module and squeeze and excitation (SE) module based on some common concepts. The LFF module fuses the different level features in the current block and the SE module adaptively rescales features by considering interdependencies among feature channels. Both the LFF module and the SE module improve the representation ability of the network.
3. We propose an identity branch to replace the global residual path way and a more simplified upsampling module. These strategies can reduce network parameters and calculations without compromising the accuracy of the results.

The remainder of this paper is organized as follows. Section 2 introduces the details of our proposed method. Section 3 verifies the effectiveness of DRSEN by performing comparisons with the state-of-the-art image super-resolution methods. In Section 4, we discuss the issues of our network according to the experimental results. Section 5 concludes the discussions of the study.

## 2. Proposed Method

In the following, we will demonstrate the architecture of the proposed DRSEN, including the interior structure and mathematical expressions. Then, the local feature fusion module and the squeeze and excitation module within the RSEB will be illustrated in detail. Finally, the implementation details of our network will be introduced.

### *2.1. Network Architecture*

As shown in Figure 1, our DRSEN consists of a residual branch and an identity branch. The residual branch consists of three parts: shallow feature extraction module, residual squeeze and excitation blocks (RSEBs) and up-sampling module. Let's denote $I_{LR}$ and $I_{SR}$ as the input and the output of DRSEN. As the same with EDSR, we use only one convolutional layer to extract the shallow feature $F_0$ from the LR input:

$$F_0 = H_{SF}(I_{LR}), \tag{1}$$

where $H_{SF}(\cdot)$ denotes convolutional operation of the shallow feature extraction layer. $F_0$ is then used for deep feature extraction with our RSEBs. Supposing there are $N$ residual blocks, the output $F_i$ of the $i$th ($1 \leq i \leq N$) RSEB can be obtained by:

$$F_i = H_{RSEB,i}(F_{i-1}), \tag{2}$$

where $H_{RSEB,i}$ denotes the operations of the $i$th RSEB. $H_{RSEB,i}$ can be a composite function of operations. More details about RSEB will be given in Section 2.2.

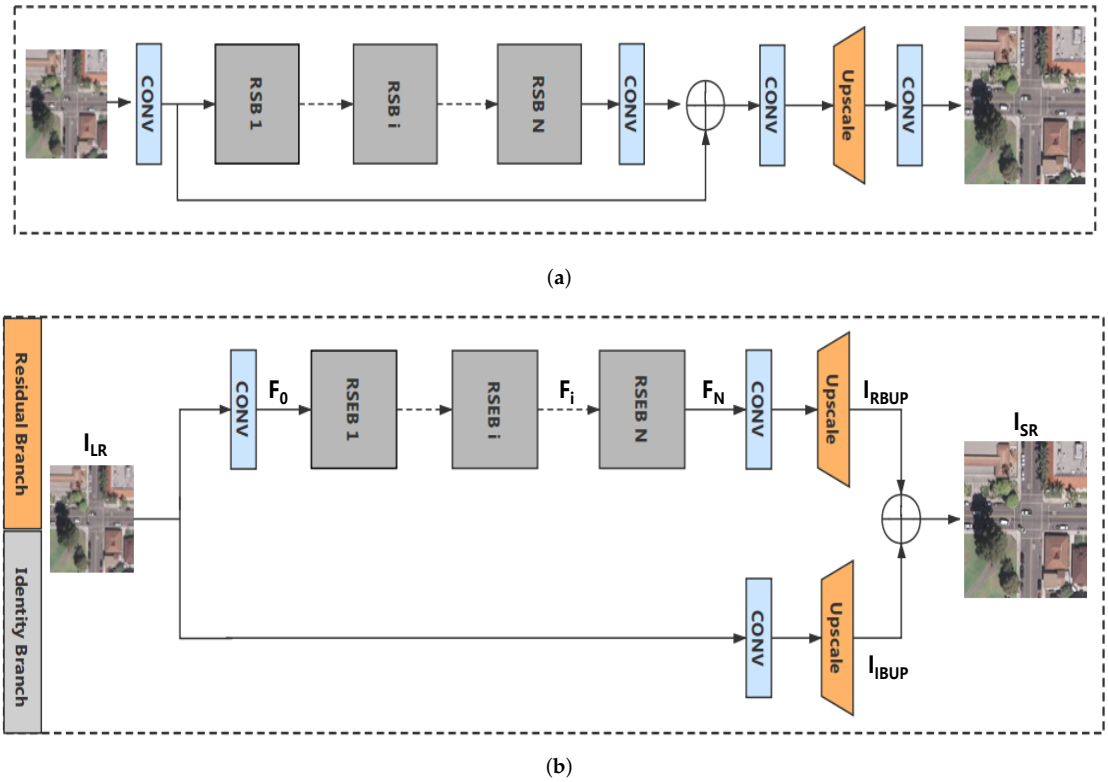


**Figure 1.** The network structure comparison of the EDSR and our DRSEN. (**a**) network architecture of EDSR; (**b**) network architecture of our deep residual attention network (DRSEN).

After extracting the features in the LR space, we use a convolutional layer to adjust the depth of the feature maps to $3 \times S^2$. Then, we up-scale these features via an up-sampling module. Inspired by EDSR, we also utilize the ESPCN as our up-scale part. In contrast to EDSR, there are no extra convolutional layers inserted after up-sampling in our network to improve the speed and reduce the parameters. The output of the residual branch can be formulated as:

$$I_{RBUP} = H_{UP}(H_A(F_N)), \tag{3}$$

where $H_{UP}(\cdot)$ and $I_{RBUP}$ denote the ESPCN and up-scaled images of the residual branch. $H_A$ and $F_N$ denote the convolutional layer before the up-scale module and the output of the last RSEB, respectively.

For the identity branch, we use a single convolutional layer and ESPCN module to directly output the corresponding high-resolution image, which reduces the convolution operations after the up-sampling layer in the EDSR. According to our experiment result, the removal of these convolutional layers has essentially no effect on the performance of the network while improving the speed. The process can be formulated as:

$$I_{IBUP} = H_{UP}(H_{identity}(I_{LR})), \tag{4}$$

where $H_{identity}$ and $I_{IBUP}$ denote the single convolutional layer and up-scaled image of the identity branch.

The outputs of the residual branch and the identity branch are combined finally via an element-wise summation to estimate the HR image, which can be formulated as:

$$I_{SR} = I_{RBUP} + I_{IBUP} = H_{DRSEN}(I_{LR}), \tag{5}$$

where $H_{DRSEN}$ denotes the function of our DRSEN.

Then, the proposed DRSEN is optimized with $L_1$ loss function, which has been demonstrated to be powerful for SR [26]. Given a training set $\{I^i_{LR}, I^i_{HR}\}^n_{i=1}$, which contains $n$ LR inputs and their HR counterparts. The goal of training DRSEN is to minimize the $L_1$ loss function

$$L(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\left\|H_{DRSEN}(I^i_{LR}) - I^i_{HR}\right\|_1 \tag{6}$$

where $\Theta = \{W_i, b_i\}$ denotes the parameter set of our proposed network. The loss is averaged over the training set. More of the training details would be shown in Section 3.1.3.

### 2.2. Residual Squeeze and Excitation Block

Our residual squeeze and excitation block (RSEB) is built upon the local feature fusion (LFF) and the squeeze and excitation (SE) module. The details of our proposed RSEB are illustrated in Figure 2.

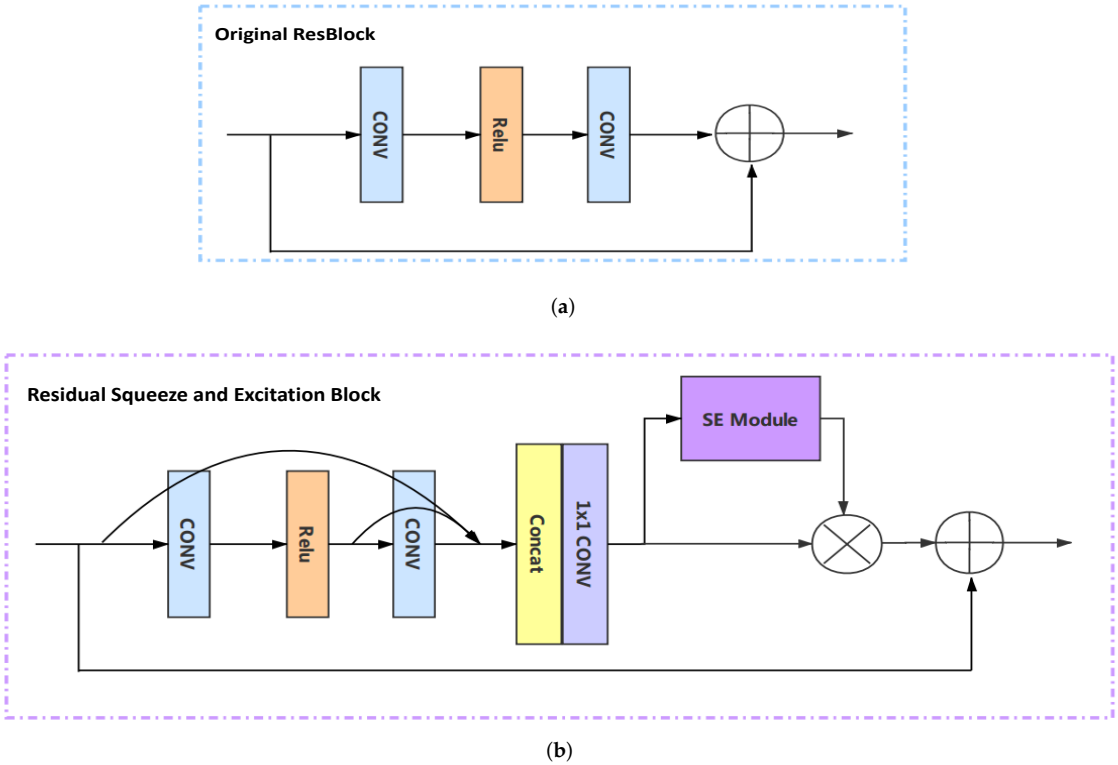


**Figure 2.** The comparison of the original Residual Block and our RSEB. (**a**) residual block structure of EDSR; (**b**) residual squeeze and excitation block (RSEB) architecture.

#### 2.2.1. Local Feature Fusion Module

The network representation power is significant for image SR. Extracting and aggregating the features among the whole network can make full use of the features and the network representation power will be enhanced. Thus, we propose a local feature fusion (LFF) module in the RSEB. The states from the preceding RSEB and some layers within current RSEB are adaptively fused by the LFF module. It is indispensable to reduce the feature number as the features of the $i$th RSEB are directly introduced

into the $(i-1)$th RSEB in series. We introduce a $1 \times 1$ convolutional layer to adaptively control the output information. This operation can be formulated as:

$$F_{i,LLF} = H^i_{LFF}([F_{i-1}; F_{i,\sigma}; F_{i,conv2}]), \tag{7}$$

where $H^i_{LFF}$ denotes the function of the $1 \times 1$ convolutional layer in the $i$th RSEB. $F_{i-1}$ is the output of the $i$th RSEB. $F_{i,\sigma}$ and $F_{i,conv2}$ refer to the feature maps produced by the activation funcation and the second convolutional layers in the $i$th RSEB. The symbol $[\cdot]$ denotes the concatenation of the feature maps.

#### 2.2.2. Sequeeze and Excitation Module

Squeeze and excitation module enhances the network representation ability by exploiting channel dependencies. The details of the SE module are illustrated in Figure 3. Through this module, important features are emphasized among the channels while suppressing useless features. The squeeze function in RSEB is shown as below:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W}, u_c(i,j) \tag{8}$$

where $z_c$ is the $c$th element of the squeezed channels and $F_{sq}$ denotes the squeeze function. $u_c$ is the $c$th channel of the input. $H$ and $W$ denote the height and width of the input.

Then, an excitation function follows the squeeze operation, which aims to fully capture the channel-wise dependencies. The excitation can be formulated as follows:

$$s = F_{ex}(z, W) = \sigma(W_u \delta(W_d z)), \tag{9}$$

where $F_{ex}$ denotes the excitation function and $z$ is the input squeezed signal from the previous layer. $\delta$ refers to the ReLU activation function, $W_d \in \mathbb{R}^{C \times \frac{C}{r}}$ is the channel downscaling with $1 \times 1$ kernel size and the dimensionality reduction ratio $r$. $W_u \in \mathbb{R}^{\frac{C}{r} \times C}$ up-scale the channel with ratio $r$ after being activated by ReLU. The final output of the block $x_c$ is rescaled with the channel statistics $s$.

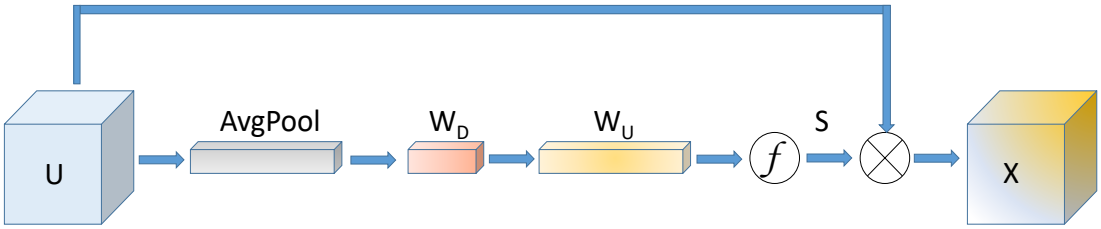


**Figure 3.** Squeeze and excitation module architecture.

### *2.3. Implementation Details*

Now, we introduce the implementation details of our proposed network DRSEN below. We build a residual branch by stacking 20 of the RSEBs. For residual branch, the size of the convolutional layer is $3 \times 3$ except for the SE module and the LFF module, whose kernel size is $1 \times 1$. The zero-padding strategy is used to keep the size fixed. To save the memory and decrease the computation, the shallow feature extraction of our residual branch has 32 filters. Within the RSEB, the number of channels of the first convolution layer is increased to 64 and then decreased to 32 with the second convolution layer. Later, the features are concatenated at the channel dimension and followed by a $1 \times 1$ convolution layer. In the SE module, one convolutional layer is utilized to reduce the number of output channels to 8 and then increase the number of channels to 64. We use the ESPCN to up-scale the coarse resolution

features to fine ones. For identity branch, we use a $3 \times 3$ convolution layer with the dilation of 2 to increase the number of channels and then up-scale the resolution with ESPCN.

## 3. Experiments and Results

### *3.1. Settings*

In this section, we first introduce the two datasets, the degradation model and the details of our training set. Then, we perform some ablation experiments to verify the effectiveness of the local feature fusion module and SE module. Finally, quantitative and visual results of our method and the state-of-the-art methods are shown.

#### 3.1.1. Datasets and Evaluation Metrics

We choose two datasets with different spatial resolutions to verify the robustness of our proposed method. There are some of the training images as shown in Figure 4.

**(1) UC MERCED [27]:** The UC Merced land-use dataset composed of 2100 land-use scene images measuring $256 \times 256$ pixels with high spatial resolution (0.3 m/pixel) in the RGB color space. This dataset is usually adopted for super-resolution task of remote sensing images. We randomly select 1700 images of the dataset for training and the other 400 samples as the testing set.

**(2) NWPU-RESISC45 [28]:** This dataset is a public benchmark created by Northwestern Polytechnical University (NWPU), which contains images with spatial resolutions varying from 30 m to 0.2 m per pixel. This dataset contains 45 scenes with a total number of 31,500 images, 700 per class. The size of each image is $256 \times 256$ pixels. We randomly select 4500 images for training and 180 images for testing.

We use the peak signal-to-noise ratio (PSNR) [dB] and structural similarity index measure (SSIM) as criteria to evaluate the performance of our proposed model.

**Figure 4.** Examples of images in two datasets. The first line is the UC Merced dataset, and the second line is the NWPU-RESISC45 dataset.

#### 3.1.2. Degradation Model

In order to demonstrate the effectiveness of our proposed DRSEN, we use bicubic down-sampling by adopting the Matlab function imresize with the option bicubic (denote as BI for short) to simulate LR images. We use this BI degradation model to simulate LR images with scaling factor $\times 2$, $\times 3$, and $\times 4$.

#### 3.1.3. Training Settings

Following settings of EDSR, in each training batch, we randomly extract LR RGB patches with the size of $48 \times 48$ as inputs. We randomly augment the patches by flipping horizontally and randomly rotating by $90^\circ, 180^\circ$ and $270^\circ$. The batch-size is set to 16. Our models are implemented with the Pytorch [29] framework and optimized with Adam [30] by setting $\beta 1 = 0.9$, $\beta 2 = 0.999$ and $\epsilon = 10^{-8}$. We set the initial learning rate to $1 \times 10^{-4}$ and halve it decreased every 100 epochs. All convolutional

filters are initialized by the method of He et al.'s initialization [31]. Training a DRSEN roughly takes 12 h with a NVIDIA Tesla P100 GPU which is manufactured through United States NVIDIA for 200 epochs.

### *3.2. Residual Squeeze and Excitation Block Analysis*

The residual squeeze and excitation block is the most critical property in our proposed network. To demonstrate the effect of each component in the block and verify the usefulness of this block, we carry out ablation experiments of super resolution tasks on the UC Merced dataset. In order to avoid the randomness of network training and the influence of noise, the network is trained 10 times under the same parameters and hardware conditions. Our experimental results are the mean of multiple training results. Table 1 shows the ablation investigation on the effects of local feature fusion (LFF) and SE module. For fairness, the number of blocks and features are the same. The experiment results measured by the mean PSNR(dB) of the testing dataset. We can conclude from the results that the model with both the LFF and SE module achieves the best performance.

**Table 1.** Comparative experiments of our model on UC Merced for $\times 2$ SR. Removing each component will degrade the final performance.

| LFF Module | SE Module | PSNR(dB) |
|---|---|---|
| √ | √ | 34.792 |
| √ | × | 34.690 |
| × | √ | 34.687 |
| × | × | 34.451 |

**Local Feature Fusion Module.** The LFF module adaptively aggregates the features to improve the representation capability of our deep network. To demonstrate the effect of this module, we remove the concatenation operator and the followed convolutional layers. The second and the last rows of Table 1 demonstrate that this LFF component can improve the performance of the baseline about 0.239 dB. This is mainly because LFF contributes to the power of the network representation ability.

As shown in Figure 5, we have done some research on three different structures of the local feature fusion module. Structure (a) combines only the input and output features of our novel blocks. Structure (b) combines the input features and each convolutional layer output features in the RSEB. Structure (c) is slightly different from structure (b), using the feature activated by the activation function to replace the output of the first convolutional layer in structure (b). The experimental result is the mean of multiple training results.

The experimental results are shown in Table 2. By comparing the results of structure (a) and the other two structures, we can find that the full utilization of the features in the block is more conducive to the improvement of the performance. Comparing the results of structure (b) and structure (c), we can conclude that the features activated by the activation function can obtain higher accuracy.

**Table 2.** Comparative experiments of the local feature fusion module with three different structures on UC Merced for $\times 2$ SR.

| Structure | (a) | (b) | (c) |
|---|---|---|---|
| PSNR(dB) | 34.529 | 34.647 | 34.683 |

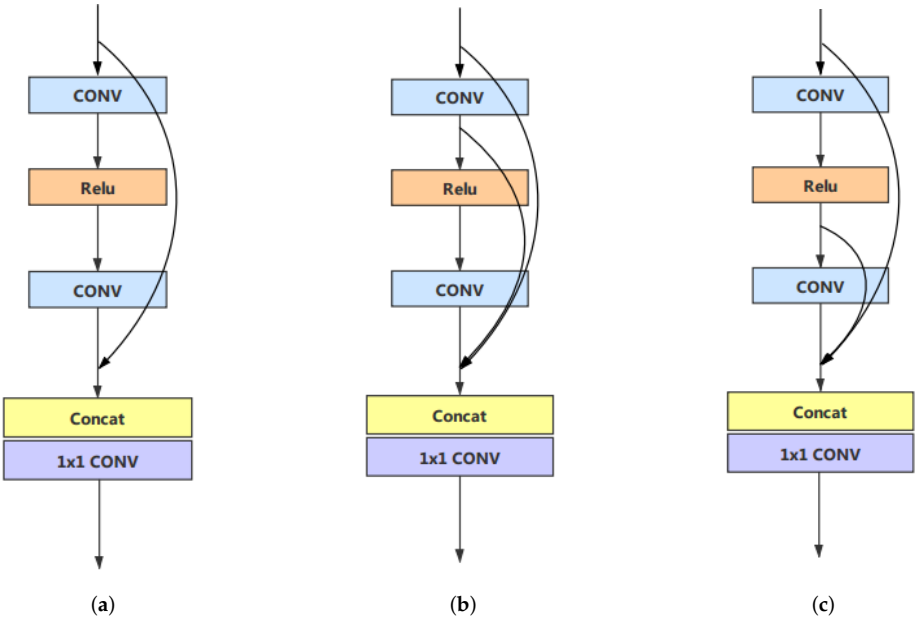


**Figure 5.** We researched about three different structures of the local feature fusion module. (**a**) has no connections in the block, (**b**) adopt a long-distance skip connection and (**c**) use a short path connection.

**Squeeze and excitation Module.** In order to evaluate the effect of the squeeze and excitation module, we run an ablation study on this component. Comparing the first and third columns of the results shown in Table 1, the SE module can improve the performance from 34.451 dB (last row) to 34.687 dB (3th row). These comparisons firmly demonstrate the effectiveness of the SE module and indicate that recalibrating the channel importance of features really enhance the performance.

### *3.3. Quantitative Results*

We evaluate the performance of our proposed remote sensing image SR network DRSEN on the two test datasets with three different up-sampling factors $\times 2$, $\times 3$ and $\times 4$. We compare our method with Bicubic [32] and five other state-of-the-art methods: SRCNN [6], FSRCNN [10], VDSR [7], LGCNet [17] and EDSR [13]. For a fair and convincing comparison, we retrain these methods under our experimental datasets. For EDSR, in order to fairly compare the performance of the network, we reduce the number of residual blocks to the same number as our DRSEN and set the convolution filters to 64. The parameter settings for the other methods are the same as in the paper. Table 3 presents the ultimate mean PSNR and SSIM over the test images in two datasets. For the rigor and credibility of the experiment, we independently train all these network 10 times under the same conditions. We use the average of multiple experimental results as the final result. Figure 6 shows the results of our multiple experiments with the proposed DRSEN. As illustrated in Table 3, our method achieves the best performance with the highest PSNR and SSIM.

For the UC Merced dataset, our proposed method outperforms EDSR with the average 0.260 dB increase on three scale factors in the terms of PSNR. Since the resolution of the original NWPU-RESISC45 dataset is slightly worse, the reconstruction performance is slightly worse than the UC Merced dataset. The PSNR of our method is 0.178 dB higher than EDSR. However, the amount of parameters in our network is much less than that of EDSR. The number of parameters for our DRSEN is 8.6 M, while the EDSR is 16.6 M. Figure 7 shows the number of network parameters for different CNN-based methods and the scaling factor $\times 2$ reconstruction quality on our dataset. In the case where the parameter amount is almost half of the EDSR, our performance indicators are even better.

**Table 3.** Evaluation of state-of-the-art SR methods on remote sensing datasets UC Merced and NWPU-RESISC45. Average PSNR(dB) and SSIM for scale ×2, ×3 and ×4. The bold numbers indicates the best performance. For the same scale, the upper row is the PSNR and the bottom row is the SSIM.

| Dataset | Scale | Bicubic | SRCNN | FSRCNN | VDSR | LGCNet | EDSR | Ours |
|---|---|---|---|---|---|---|---|---|
| UC Merced | ×2 | 30.781 | 32.803 | 33.241 | 33.498 | 33.503 | 34.572 | **34.792** |
| | | 0.8316 | 0.8750 | 0.8783 | 0.8943 | 0.8968 | 0.9439 | **0.9470** |
| | ×3 | 27.209 | 28.654 | 28.979 | 29.355 | 29.274 | 30.349 | **30.643** |
| | | 0.7502 | 0.8030 | 0.8124 | 0.8129 | 0.8112 | 0.8762 | **0.8821** |
| | ×4 | 25.732 | 26.793 | 27.013 | 27.141 | 27.158 | 27.887 | **28.147** |
| | | 0.6725 | 0.7226 | 0.7285 | 0.7358 | 0.7335 | 0.8055 | **0.8153** |
| NWPU-RESISC45 | ×2 | 30.746 | 31.755 | 32.281 | 32.757 | 32.576 | 34.184 | **34.400** |
| | | 0.8362 | 0.8749 | 0.8834 | 0.8951 | 0.8937 | 0.9361 | **0.9385** |
| | ×3 | 27.658 | 28.450 | 28.936 | 29.268 | 29.277 | 30.335 | **30.471** |
| | | 0.7592 | 0.7940 | 0.7998 | 0.8083 | 0.8081 | 0.8548 | **0.8637** |
| | ×4 | 26.378 | 27.358 | 27.513 | 27.839 | 27.742 | 28.364 | **28.543** |
| | | 0.7066 | 0.7466 | 0.7591 | 0.7625 | 0.7613 | 0.7832 | **0.7846** |

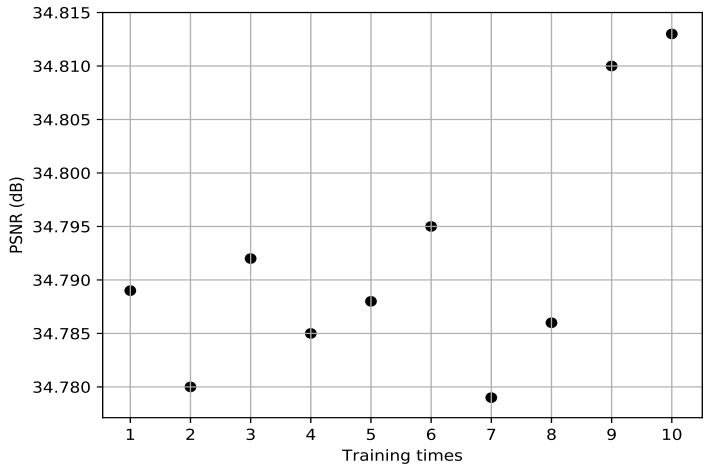


**Figure 6.** The index value of 10 training results. The results are evaluated with the UC Merced test dataset for ×2 SR.

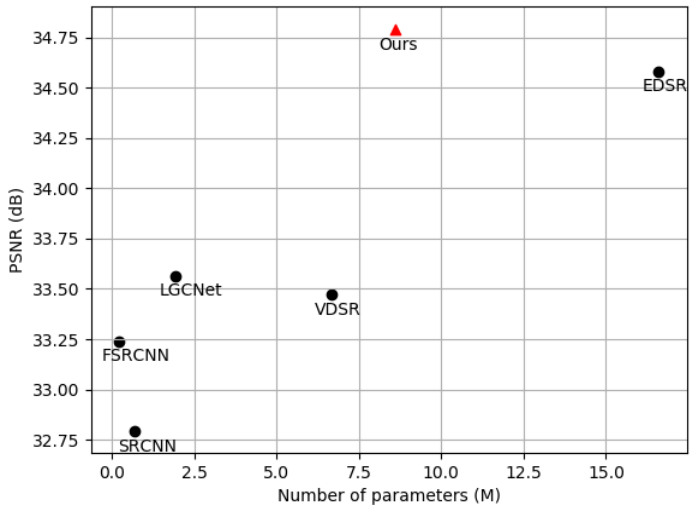


**Figure 7.** The number of network parameters versus performance. The results are evaluated with a UC Merced test dataset for ×2 SR. Our proposed models achieve better performance with relatively fewer parameters.

### *3.4. Visual Results*

In order to more fully demonstrate the effectiveness of our approach, we also show some of the visual comparisons on three scales $\times 2$, $\times 3$ and $\times 4$ in Figures 8–10. We observed that, on the different scale factors, our proposed DRSEN achieves better results, reduced sawtooth and ringing artifact and better reconstructed the structure of the objects in the picture.

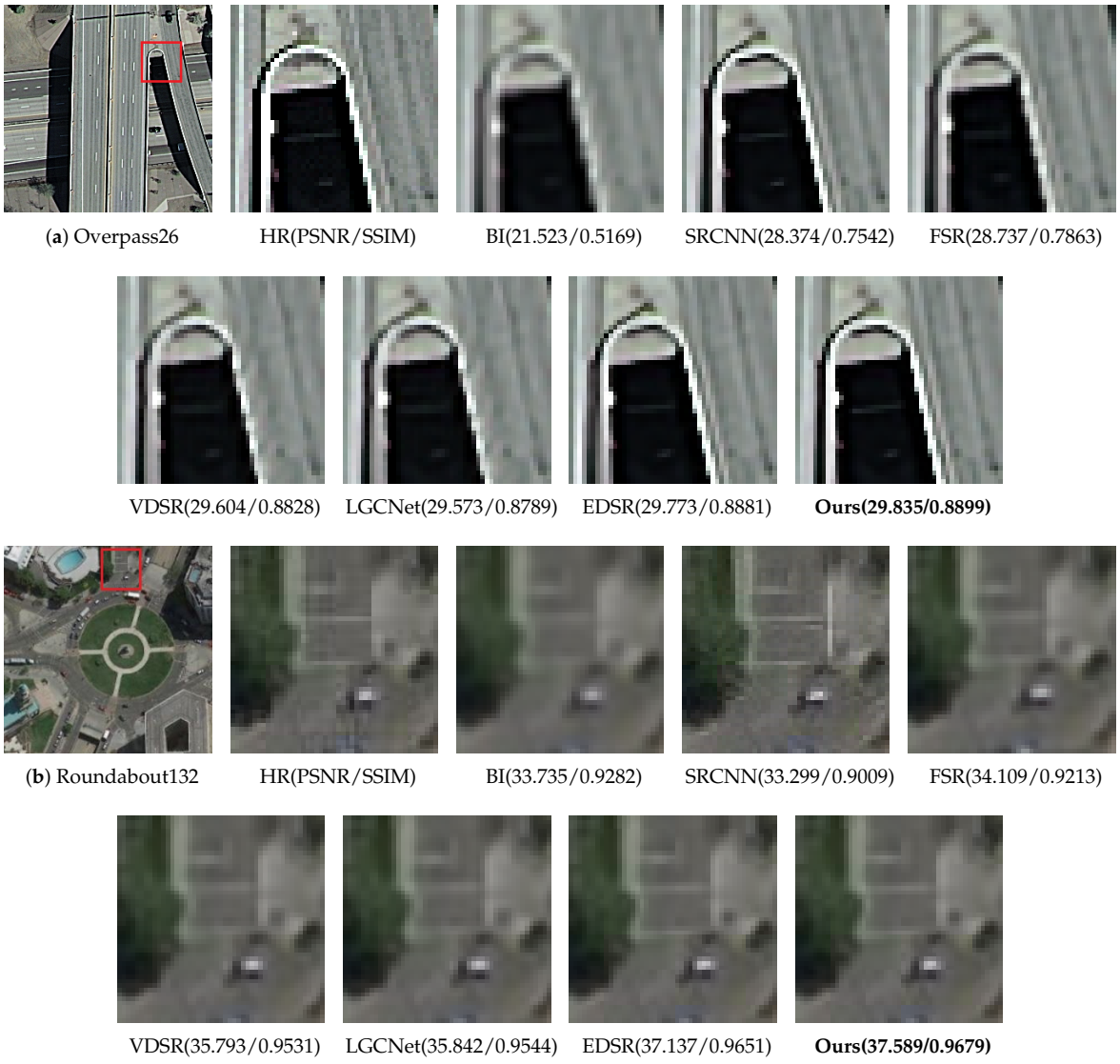


**Figure 8.** Super-resolution results of "overpass26" (UC Merced) and "roundabout132" (NWPU-RESISC45) with scale factor $\times 2$. The edges of the overpass and the lane line in our results are more clear. We refer to FSRCNN as FSR for short.

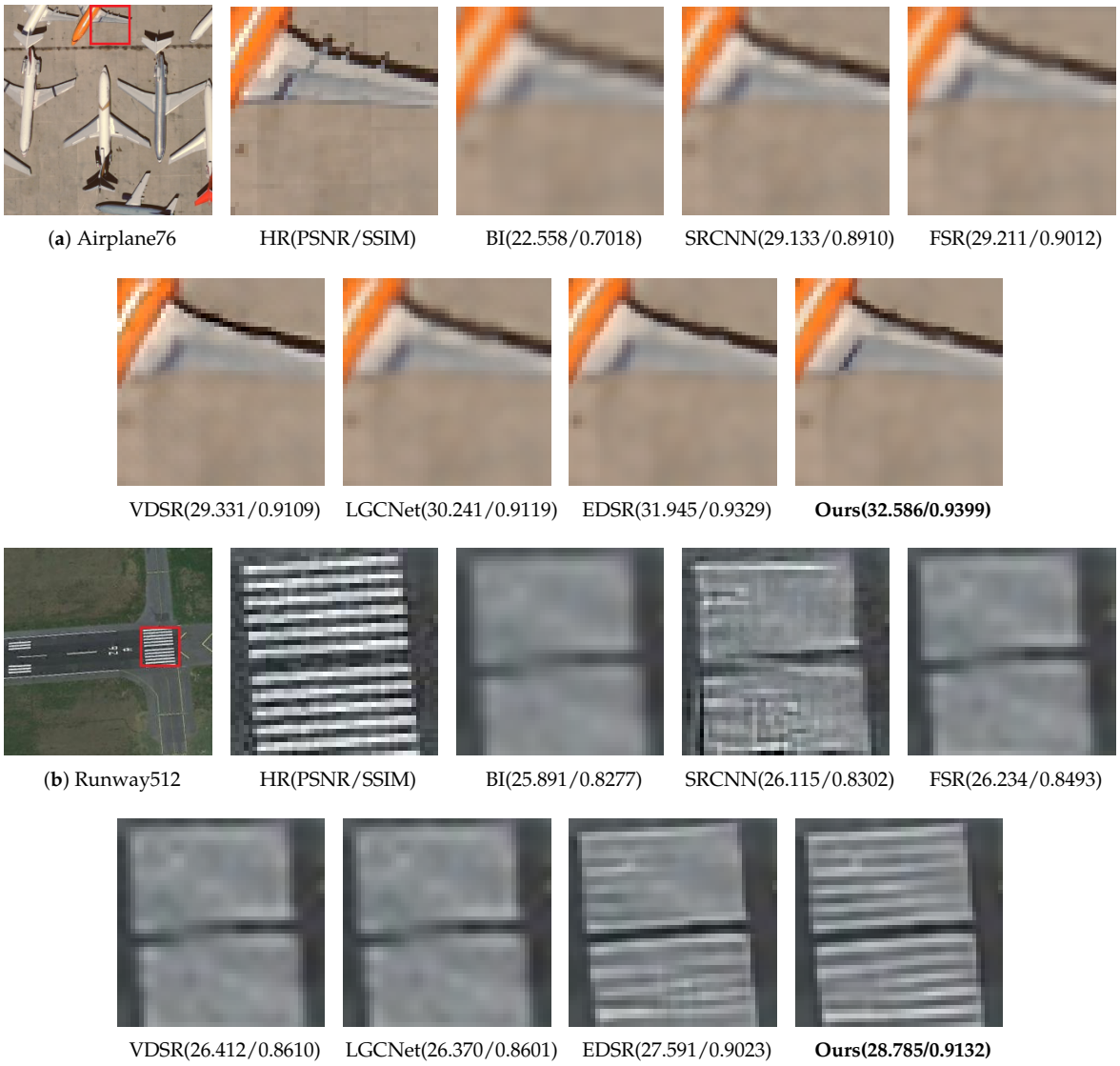


**Figure 9.** Super-resolution results of "airplane76" (UC Merced) (**a**) and "runway512" (NWPU-RESISC45) (**b**) with scale factor $\times 3$. The texture of the airplane and the lines in the sidewalk are observed in our methods, while others suffer from blurring artifacts. We refer to FSRCNN as FSR for short.

As shown in Figure 10, on the large scale factor, the bicubic up-sampling strategy results in loss of texture and structure and produces blurry SR results. Methods such as SRCNN, VDSR, and LGCNet that take such bicubic up-sampling results as network inputs will produce erroneous structural and texture information and fail to recover more details, ultimately resulting in poor SR image quality. Although EDSR uses the original LR image as the input to the network, it cannot restore the correct texture structure, while our DRSEN can recover more structural and texture information in the original corresponding HR image. This comparison apparently shows that our network has more powerful representation capabilities and can extract complex features from the LR space.

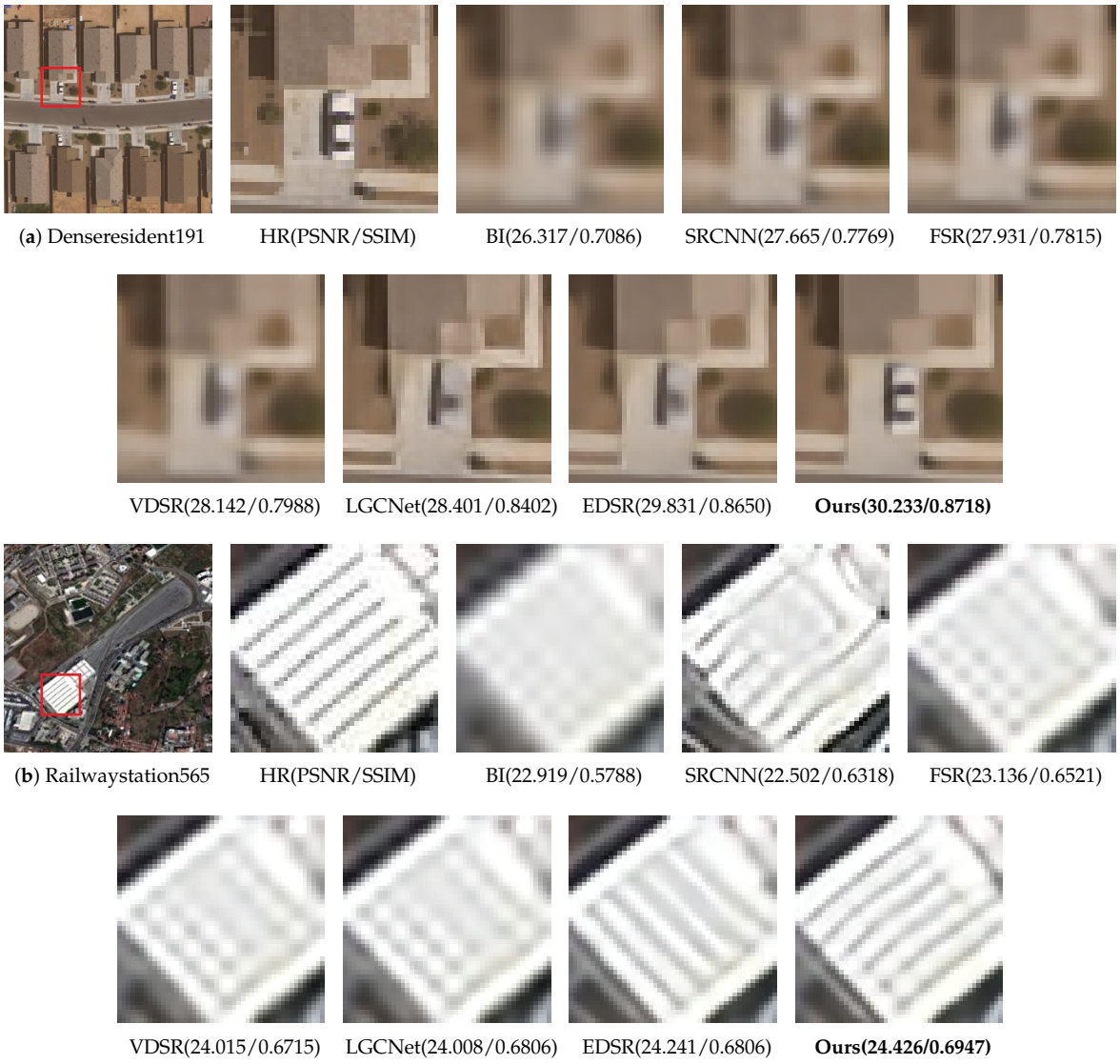


**Figure 10.** Super-resolution results of "denseresident191" (UC Merced) (**a**) and "railwaystation565" (NWPU-RESISC45) (**b**) with scale factor $\times 4$. The outline of the car is distinct and the lattices of the building roof are closer to the original image. We refer to FSRCNN as FSR for short.

Our network is a fully convolutional network and the input image size can be arbitrary. In the network training process, the neural network learns the mapping relationship between the low-resolution image patch which is cropped from the original low-resolution input and the high-resolution image patch. In the testing phase, we only need to input the image to be processed into the network to get the corresponding super-resolution results. The upper limit of the input image size is only related to the memory of the device. If there is not enough memory, we can also divide the large image into several small images and finally stitch them together to get the final output.

## 4. Discussion

According to the experimental results and analysis in Sections 3.3 and 3.4, the proposed algorithm performs better than the other methods. However, there still remains some limitations. Our DRSEN is a PSNR-oriented method, which tends to output over-smoothed results without sufficient high-frequency details at large ratio scaling, as illustrated in Figure 10.

Although our DRSEN has acquired competitive PSNR and SSIM, there is still a certain distance between visual results and human visual perception. The possible reason is that the PSNR metric

fundamentally disagrees with the subjective evaluation of human observers. At the same time, when training the network, we need to use the paired low-resolution observations and high-resolution remote sensing images. However, in actual situations, it is often difficult to obtain pairs of high-resolution and degraded images. In the future, we will explore the perceptual loss to keep a balance between the value of PSNR and visual quality. Additionally, we will investigate the cycle-consistency concept of CycleGAN [33] to train the super-resolution network with the unpaired dataset. In this case, we can learn to convert the low-resolution remote sensing image into a high-resolution image by observing only the degraded image.

## 5. Conclusions

In this paper, we propose a novel network named DRSEN to improve the representation of deep networks and achieve better performance in remote sensing image super-resolution task. Specifically, the RSEB is proposed as the building module of our SR deep network. We use the local feature fusion module in the RSEB to make use of the features within the input and the block. Such module can improve the network representation capability and stabilize the training. Meanwhile, the squeeze and excitation module is used to adaptively recalibrate channel-wise feature responses by explicitly modelling interdependencies between channels to improve the ability of our network further. We modify the global residual path way and remove some redundant convolutional layers to decrease the parameters and computation. Our model is trained on two public benchmark remote sensing datasets with various spatial resolution. The experimental results demonstrate that our proposed method can obtain accurate results with fewer parameters and outperform most of the state-of-the-art methods regarding quality and accuracy. In the future, we will continue to focus on improving super resolution quality and addressing the problem of the paired dataset.

**Author Contributions:** Formal analysis, J.G.; Funding acquisition, Y.Z.; Investigation, J.G.; Methodology, J.G.; Supervision, X.S., Y.Z., K.F. and L.W.; Visualization, J.G.; Writing—original draft, J.G.; Writing—review and editing, J.G., X.S., Y.Z., K.F. and L.W.

**Funding:** This research was funded by the National Natural Science Foundation of China under Grant 41801349

**Acknowledgments:** The authors would like to thank all the colleagues in the lab, who generously provided their original images and helped to annotate the images. The authors are thankful for the anonymous reviewers for their helpful suggestions.

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

## References

1. Glasner, D.; Bagon, S.; Irani, M. Super-resolution from a single image. In *2009 IEEE 12th International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2009; pp. 349–356.
2. Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. Image super-resolution via sparse representation. *IEEE Trans. Image Process.* **2010**, *19*, 2861–2873. [CrossRef] [PubMed]
3. Dong, W.; Zhang, L.; Shi, G.; Wu, X. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Trans. Image Process.* **2011**, *20*, 1838–1857. [CrossRef] [PubMed]
4. Pan, Z.; Yu, J.; Huang, H.; Hu, S.; Zhang, A.; Ma, H.; Sun, W. Super-resolution based on compressive sensing and structural self-similarity for remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 4864–4876. [CrossRef]
5. Li, J.; Yuan, Q.; Shen, H.; Meng, X.; Zhang, L. Hyperspectral image super-resolution by spectral mixture analysis and spatial–spectral group sparsity. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1250–1254. [CrossRef]
6. Dong, C.; Chen, C.L.; He, K.; Tang, X. *Learning a Deep Convolutional Network for Image Super-Resolution*; Springer International Publishing: Cham, Switzerland, 2014; pp. 184–199.
7. 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.

8. Kim, J.; Kwon Lee, J.; Mu Lee, K. 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.
9. Tai, Y.; Yang, J.; Liu, X. Image super-resolution via deep recursive residual network. In Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3147–3155.
10. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2016; pp. 391–407.
11. 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.
12. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
13. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M.; Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced Deep Residual Networks for Single Image Super-Resolution. In Proceedings of the Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 1132–1140.
14. 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, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
15. Johnson, J.; Alahi, A.; Li, F.-F. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2016; pp. 694–711.
16. 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]
17. 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]
18. Haut, J.M.; Fernandez-Beltran, R.; Paoletti, M.E.; Plaza, J.; Plaza, A.; Pla, F. A new deep generative network for unsupervised remote sensing single-image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6792–6810. [CrossRef]
19. Xu, W.; Xu, G.; Wang, Y.; Sun, X.; Lin, D.; Wu, Y. Deep Memory Connected Neural Network for Optical Remote Sensing Image Restoration. *Remote Sens.* **2018**, *10*, 1893. [CrossRef]
20. 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.
21. 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, 18–22 June 2018; pp. 7132–7141.
22. 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.
23. Cheng, X.; Li, X.; Yang, J.; Tai, Y. SESR: Single image super resolution with recursive squeeze and excitation networks. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 147–152.
24. Hu, Y.; Li, J.; Huang, Y.; Gao, X. Channel-wise and Spatial Feature Modulation Network for Single Image Super-Resolution. *arXiv* **2018**, arXiv:1809.11130.
25. Kim, J.H.; Choi, J.H.; Cheon, M.; Lee, J.S. RAM: Residual Attention Module for Single Image Super-Resolution. *arXiv* **2018**, arXiv:1811.12043.
26. Zhao, H.; Gallo, O.; Frosio, I.; Kautz, J. Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imaging* **2017**, *3*, 47–57. [CrossRef]
27. 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, Seattle, WA, USA, 2–5 November 2010; pp. 270–279.

28. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
29. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in Pytorch. 2017. Available online: https://openreview.net/forum?id=BJJsrmfCZ (accessed on 1 August 2019).
30. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 1026–1034.
32. Keys, R. Cubic convolution interpolation for digital image processing. *IEEE Trans. Acoust. Speech Signal Process.* **1981**, *29*, 1153–1160. [CrossRef]
33. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.