remote sensing

MDPI

Article

Satellite Image Super-Resolution via Multi-Scale Residual Deep Neural Network

Tao Lu [1], Jiaming Wang [1], Yanduo Zhang [1], Zhongyuan Wang [2] and Junjun Jiang [3,*]

[1] Hubei Key Laboratory of Intelligent Robot, School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China
[2] School of Computer Science, Wuhan University, Wuhan 430072, China
[3] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
* Correspondence: jiangjunjun@hit.edu.cn

Received: 13 May 2019; Accepted: 2 July 2019; Published: 4 July 2019

Abstract: Recently, the application of satellite remote sensing images is becoming increasingly popular, but the observed images from satellite sensors are frequently in low-resolution (LR). Thus, they cannot fully meet the requirements of object identification and analysis. To utilize the multi-scale characteristics of objects fully in remote sensing images, this paper presents a multi-scale residual neural network (MRNN). MRNN adopts the multi-scale nature of satellite images to reconstruct high-frequency information accurately for super-resolution (SR) satellite imagery. Different sizes of patches from LR satellite images are initially extracted to fit different scale of objects. Large-, middle-, and small-scale deep residual neural networks are designed to simulate differently sized receptive fields for acquiring relative global, contextual, and local information for prior representation. Then, a fusion network is used to refine different scales of information. MRNN fuses the complementary high-frequency information from differently scaled networks to reconstruct the desired high-resolution satellite object image, which is in line with human visual experience ("look in multi-scale to see better"). Experimental results on the SpaceNet satellite image and NWPU-RESISC45 databases show that the proposed approach outperformed several state-of-the-art SR algorithms in terms of objective and subjective image qualities.

Keywords: satellite imagery; super-resolution; residual network; multi-scale image; convolutional neural network

1. Introduction

Remote sensing satellites, which observe objects on the ground from outer space, are widely used in various real applications, such as environmental monitoring, resource exploration, disaster warning, and military applications. The observed images from satellites generally have low-resolution (LR) due to the limitations of spaceborne imaging equipment (Charge-coupled Device (CCD) sensors) and communication bandwidth. In addition, satellite images are affected by atmospheric turbulence, transmission noise, motion blur, and undersampling optical sensors. The quality and resolution of images from remote sensing satellites cannot meet the requirements of real satellite image analysis. Super-resolution (SR) technology can overcome hardware limitations and improve the spatial resolution of images through software manner. The first SR algorithm [1] was developed to improve the resolution of Landsat remote sensing images by fusing multi-frame complementary information. In the last decades, SR has been successfully applied to enhance the resolution and quality of remote sensing satellite images. A well known example is SPOT-5, which reaches 2.5 m resolution through the SR of two 5 m images that are sampled from shifting a double CCD array by

subpixel sampling interval [2,3]. Traditional SR image generation methods usually require multiple spatial/spectral/temporal low-resolution images of the same scene [4,5].

Existing image SR algorithms are divided into two categories, namely reconstruction- and learning-based algorithms [6]. Reconstruction-based algorithms fuse subpixel LR multi-frame information and reconstruct their latent high-resolution (HR) images. Previous satellite SR methods utilize reconstruction-based methods in solving the inverse problem of the degradation process. Reconstruction-based methods model the degradation process of imaging with mathematical formulas by using degradation factors, such as downsampling, optical blur, atmospheric disturbance, registration error, geometric deformation, and motion compensation [7–9]. Although reconstruction-based methods are simple and intuitive and can be flexibly combined with prior constraints, they rely on accurate subpixel precision estimation.

Inspired by the immense success of machine learning in object recognition and other tasks, learning-based SR methods have been highly valued and have become the mainstream direction of research. They aim to learn a mapping function between LR and HR image/patches through the prior information provided by a training dataset. Learning-based SR algorithms can obtain better subjective and objective reconstruction performance than reconstruction-based methods because external training databases provide considerable a priori information. In terms of the usage of prior training samples, learning-based SR algorithms can be divided into three categories, namely regression-, representation-, and deep-learning-based algorithms. Some representative regression-based [10–12] and representation-based SR algorithms [13–15] yield decent subjective and objective performance. These methods are efficient with flexible framework for using regularization terms.

Deep-learning-based approaches provide an end-to-end solution for learning complex mapping functions and are rapidly and successfully applied on SR tasks. A complex nonlinear mapping relationship between LR and HR patches is learned through convolutional neural networks (CNNs) [16,17], considering their excellent learning capability. Shi et al. [18] constructed a subpixel CNN, which provides a novel manner of directly and efficiently learning the mapping function from LR to HR images, which is further efficient. Kim et al. [19] stated that the construction of a deep network can effectively alleviate training difficulty with deep residual learning. Lai et al. [20] built a pyramid network for fusing multi-scale residuals in the feature domain. A generative adversarial network (GAN) [21,22], which comprises generator and discriminative networks, was used to generate fake details for simulating a good visual output. For satellite images, Luo et al. [23] replaced zero-padding with self-similarity to avoid the addition of unusable information and achieved good results. Wang et al. [24] proposed a multi-memory CNN for video SR to retain inter-frame temporal correlations.

The above-mentioned SR approaches mainly focus on the general nature images. As for satellite image, the object scale in the image is relatively different due to wide-range imaging, and it has important roles in vision tasks, such as segmentation, feature extraction, and object tracking. Some deep-learning algorithms designed for general images cannot efficiently handle satellite images because they do not specially consider the multi-scale nature of satellite images. Moreover, adequate high-frequency information, such as edges and textures, are crucial for satellite image detection [25] and object recognition [26–30]. The use of a single structure network in predicting and reconstructing objects without considering their different scales results in poor reconstruction performance. One practical solution is to explore the multi-scale information into deep neural networks. Zhang et al. [31] used multi-scale spatial structural self-similarity to learn multi-scale dictionaries. Fu et al. [32] utilized the multi-scale regions of an image to train a recurrent attention network for fine-grained recognition. Liu et al. [33] used multi-scale and multi-level network in a holistic manner to obtain hierarchical edge information. Similar to the inception network [34], Du et al. [35] fused different scale features from three varying filters.

The aforementioned CNN-based SR models build fine networks and have advanced the state-of-the-art performance on learning significant local detail information. The approach in [36]

points out that too small receptive field resulting in the lack of enough global information to yield good visual results. To obtain fine local detail information, they often use small image patches for training (e.g., 33 × 33 for SRCNN [17], 41 × 41 for VDSR [19], 32 × 32 for LapSRN [20], and 24 × 24 for SRResnet/SRGAN [21]). A small receptive field only considers a limited range of information during SR tasks. This model lacks the capability of obtaining global and contextual information for SR. On the contrary, Zeiler et al. [37] visualized convolutional network to indicate that different network layers have varying roles in representing the features that simulate the ventral pathway to enhance their performance [38–42]. They indicated that hierarchical features of different scales effectively improve the capability of acquiring global information.

Inspired by the observation of "look closer to see better" [32], we propose a flexible and versatile multi-scale residual deep neural network for satellite image SR, named MRNN, for the hierarchical reconstruction of satellite imagery with HR detail information. In this network, multi-scale receptive fields are similar to the observation from different distances by human eyes. We extract three scales of the image at large- (large-kernel-size network, for global information), middle- (middle-kernel-size network, for contextual information) and small-scale (small-kernel-size network, for fine local information) features to represent the multi-scale information of images. In comparison with traditional neural networks, MRNN fuses the residual information rather than intermediate features. Thus, the fusion network fuses all scales of residual information to improve the high-frequency details.

The contributions of this study are highlighted as follows: (i) The use of MRNN is proposed for satellite image SR. The proposed network contains three parts, namely multi-scale feature extraction; parallel small-, middle-, and large-scale; and residual fusion networks. The proposed multi-scale neural network leverages SR performance on the basis of "look in multi-scale to see better". (ii) The proposed residual enhancement and fusion networks effectively enhance the high-frequency information of satellite images in SR tasks. The fusion network refines fine edge/detail textures, thereby improving the details of the satellite image.

The remainder of this paper is organized as follows. In Section 2, we describe the framework of the proposed method. In Section 3, comparison is presented among the proposed method and some representative SR methods. The discussion and conclusion of this study are given in Sections 4 and 5, respectively.

2. Satellite Imagery SR Based on Multi-Scale Residual Neural Network

We use image saliency to show the difference among various image sizes and emphasize the role of multi-scale images. Image saliency [43] is an important visual feature in an image and emphasizes the importance degree of a region for human eye perception. The brightness of a saliency map represents the importance of object parts. The saliency map S is formulated as:

$$S(x,y) = \left\| \boldsymbol{I}_{\mu} - \boldsymbol{I}_{\omega hc}(x,y) \right\|_2^2, \quad (1)$$

where $\boldsymbol{I}_{\mu}$ is the mean image feature vector, and $\boldsymbol{I}_{\omega hc}(x,y)$ is the corresponding image pixel vector value at position (x,y) in the Gaussian blurred version (using a 5 × 5 separable binomial kernel) of the original image.

Figure 1 displays three sizes of image patches, namely large- (91 × 91), middle- (61 × 61), and small-sized (41 × 41) image patches. In the 91 × 91 image patch, the saliency map focuses on the global information in the image, such as the outline of a building. For the 61 × 61 image patch, which contains further contextual information, the saliency map focuses on building parts and street lines. For the 41 × 41 image patch, which has a small receptive field, only local information is observed, and global information is neglected. Here, long-distance observation experience can be reviewed; global configuration information, such as position and outward appearance, can be observed when we are far from the observed objects, and no detailed information is included. For additional details, we focus on local information, such as the decoration and color of a building, as we approach. This observation

is a good illustration of the role of multi-scale information in visual observation. Therefore, image reconstruction on only single-scale image patches cannot simultaneously and effectively recover the global and local information of the object.

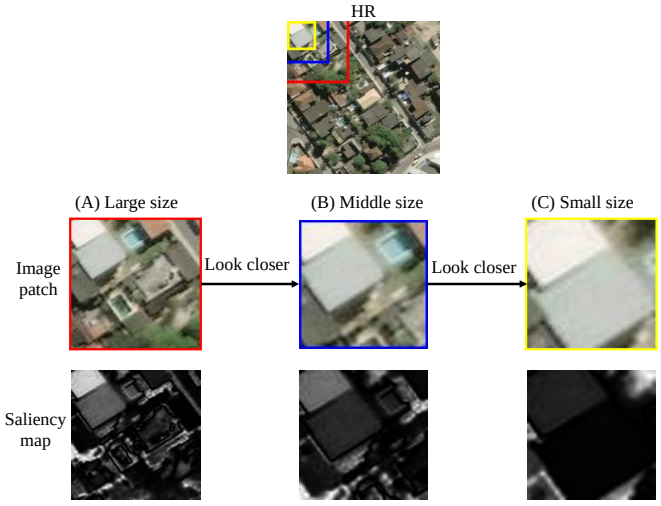

Figure 1. Saliency maps of multi-size image patches: (**A**) large-scale focuses on global configuration, such as edge position; (**B**) middle-scale focuses on subject parts as contextual information, such as building parts; and (**C**) small-scale focuses on detailed edges and textures. Saliency maps reveal the role of differently scaled image patches in SR reconstruction. There are different feature representation manners in different size image patch. This experiment is in line with human visual experience ("look in multi-scale to see better"). Large-, middle-, and small-scale networks are used to simulate different size receptive fields for acquiring relative global, contextual, and local information for prior representation.

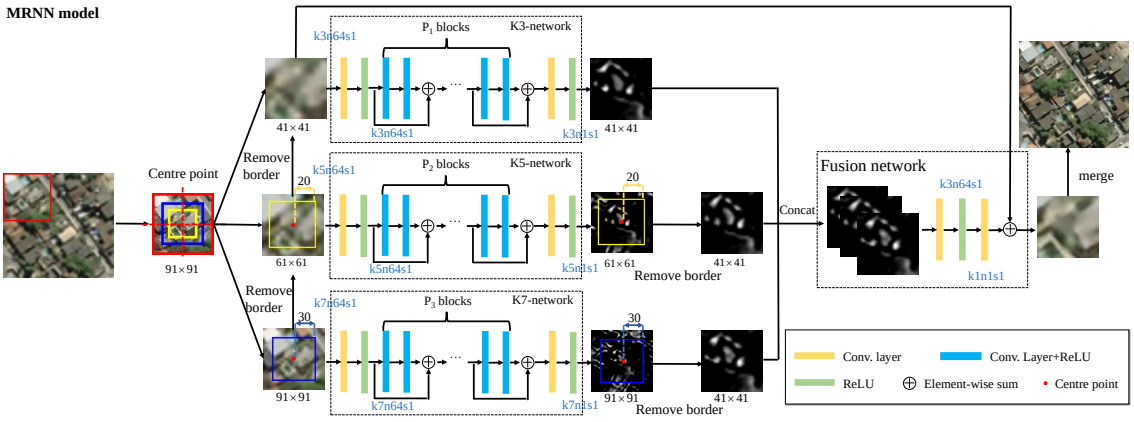

Figure 2. Network architecture of MRNN. The network includes three SR subnetworks and a residual fusion network (k is the convolution kernel size; n denotes the number of convolution kernels; and s indicates the stride size). The residual block is configured to two convolutional layers with multi-scale kernels followed by ReLU, and a skip connection. $p = \lfloor (D-2)/2 \rfloor$, where $\lfloor \rfloor$ is the floor function. We add a convolution layer + ReLU behind the residual structure when D is an odd number. The merge means converting image patches into an image.

We propose a novel multi-scale residual network, whose structure is shown in Figure 2. We establish three adaptive networks with different scale features to predict their high-frequency residual information in different scales for satellite images. Thus, we use residual images with varying scales to merge their high-frequency by utilizing a residual fusion network. As the pixel value in the residual image is small, we use the ImageEnhance module of Python Imaging Library https://github.com/

python-pillow/Pillow to conduct enhanced contrast processing of images. The enhanced image *blend_img* is given by

$$blend\_img = img1 \times (1 - \lambda) + img2 \times \lambda, \tag{2}$$

where $img1$ is the original image, and the enhancement factor $\lambda = 10$ represents the weight of the image blend. $img2$ is a generated image, whose pixel value is 0.5 plus the average value of $img1$. The greater is the λ, the greater is the contrast of the image.

For a pair of training datasets $\{\bar{\boldsymbol{X}}_i, \boldsymbol{Y}_i\}_{i=1}^{M}$, where LR image $\bar{\boldsymbol{X}}_i \in \Re^{h \times w}$ and HR image $\boldsymbol{Y}_i \in \Re^{ht \times wt}$, t is the amplification factor, i denotes the sample index, and M refers to the number of training samples. The LR image $\bar{\boldsymbol{X}}_i \in \Re^{h \times w}$ is interpolated to the HR image size with bicubic kernel as $\boldsymbol{X}_i \in \Re^{ht \times wt}$, and the tensor version of the training dataset is rewritten as $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{M}$. The superscript represents the type of network, and the subscript indicates the number of layers. Superscripts K3, K5, K7, C, and F represent the K3-network, K5-network, K7-network, Concat operation, and residual fusion network, respectively. The sampling of patches with different sizes results in various numbers of patches in each scale. However, all training sample sets share the same training set $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{M}$. The number of image patches is calculated as follows:

$$N_{S_D} = \lfloor ht/S_D \rfloor * \lfloor wt/S_D \rfloor * M, \tag{3}$$

where $\lfloor \rfloor$ is the floor function and S_D indicates the size of receptive field of the D-layer network. Image patches 41×41 and 61×61 are acquired on the basis of the center point of the 91×91 image patch (for additional details, see Figure 2). LR and HR image patch pairs with different scales are defined as $\{\boldsymbol{x}_j^{K3}, \boldsymbol{y}_j^{K3}\}_{j=1}^{N_{41}}$, $\{\boldsymbol{x}_j^{K5}, \boldsymbol{y}_j^{K5}\}_{j=1}^{N_{61}}$, and $\{\boldsymbol{x}_j^{K7}, \boldsymbol{y}_j^{K7}\}_{j=1}^{N_{91}}$, which have patch sizes of 41×41, 61×61 and 91×91 pixels, respectively. j is the index of the image patches, and N_{41}, N_{61}, and N_{91} denote the numbers of patches. Considering residual fusion, we use the patch center point to anchor three different size patches; thus, $N_{41} = N_{61} = N_{91}$.

2.1. Multi-Scale SR

We use three different scales of networks to simulate SR with different depths. The network depths are D_{k3}, D_{k5}, and D_{k7}. Parameter D is fine tuned according to the method in Section 3. In the K3-network, the convolution filter is defined as $k = 3$. The residual map of the K3-network at the patch level is defined as follows:

$$f^{K3}(\boldsymbol{x}_j^{K3}) = \boldsymbol{W}_{20}^{K3} * H_{19}^{K3}(\boldsymbol{x}_j^{K3}) + b_{20}^{K3}, \tag{4}$$

where $f^{K3}(\boldsymbol{x}_j^{K3})$ is the predicted residual patch with size 41×41; $\boldsymbol{W}_{20}^{K3}$ indicates the weight matrix with size $64 \times 3 \times 3 \times 1$; $\boldsymbol{b}_{20}^{K3}$ denotes the bias with size 1×1; H_{19}^{K3} represents the generated feature maps of the 19th layers by an activation ReLU, which is composed of 64 feature maps; and j refers to the index of image patches.

For the K5-network, the size of its convolution kernel is 5×5 pixels. For K7-network, the filter kernel size is 7×7 pixels. We use the same method to calculate the size of the input image patch. Their residual maps are calculated as follows:

$$f^{K5}(\boldsymbol{x}_j^{K5}) = \boldsymbol{W}_{15}^{K5} * H_{14}^{K5}(\boldsymbol{x}_j^{K5}) + b_{15}^{K5}, \tag{5}$$

$$f^{K7}(\boldsymbol{x}_j^{K7}) = \boldsymbol{W}_{15}^{K7} * H_{14}^{K7}(\boldsymbol{x}_j^{K7}) + b_{15}^{K7}, \tag{6}$$

where $\boldsymbol{W}_{15}^{K5}$ has a size of $64 \times 5 \times 5 \times 1$; $\boldsymbol{W}_{15}^{K7}$ has a size of $64 \times 7 \times 7 \times 1$; the size of b_{15}^{K7} and b_{15}^{K5} is 1×1; and j denotes the index of the image patches. H_{14}^{K5} and H_{14}^{K7} represent the feature maps of the 14th layers by the K5- and K7-networks, respectively.

2.2. Residual Fusion Network

To realize the complementarity of different scales of information, the global information of an object is described by large-scale information, and, the closer you look, the better the hierarchical details become. We use a fusion network for multi-scale residual fusion.

$$f^C(\boldsymbol{x}) = Concat(f^{K3}(\boldsymbol{x}_j^{K3})_r, f^{K5}(\boldsymbol{x}_j^{K5})_r, f^{K7}(\boldsymbol{x}_j^{K7})_r) = [f^{K3}(\boldsymbol{x}_j^{K3})_r, f^{K5}(\boldsymbol{x}_j^{K5})_r, f^{K7}(\boldsymbol{x}_j^{K7})_r], \tag{7}$$

where $f^{K3}(\boldsymbol{x}_j^{K3})_r$, $f^{K5}(\boldsymbol{x}_j^{K5})_r$, and $f^{K7}(\boldsymbol{x}_j^{K7})_r$ are the residual maps with the removal of border from the outputs of the three differently scaled networks. $f^C(\boldsymbol{x})$ represents the combined three layers of residual maps. The Concat function cascades the multi-scale residual maps in the third dimension (connect three tensors). Regardless of the same input $\boldsymbol{x}$, the outputs of K3-, K5-, and K7-networks are different because they reconstruct their residual information through their own scales. To fuse different scales of residual information, we use a simple two-layer network to fuse three channel information. A 1×1 convolution kernel is a linear combination of each pixel on different channels. The 1×1 convolution kernel is used to fuse the residual feature maps. The cross-channel information interaction among different scales of information is consistent with the hierarchical visual cognition mechanism. We can obtain the final fusion residual as follows:

$$R^F(f^C(\boldsymbol{x})) = \boldsymbol{W}_2^F * H_1^F(f^C(\boldsymbol{x})) + b_2^F, \tag{8}$$

where $\boldsymbol{W}_2^F$ is the second layer weight matrix, b_2^F represents its bias, $f^C(\boldsymbol{x})$ denotes the input multi-scale residual maps, and $R^F(\boldsymbol{x})$ indicates the final fused output residual map. Thus, the final HR image $\hat{y}$ is as follows:

$$\hat{y} = R^F(f^C(\boldsymbol{x})) + \boldsymbol{x}. \tag{9}$$

2.3. Loss Function

We define the loss function with mean squared error (MSE) as the objective function. In MRNN, we formulate the overall loss function as follows:

$$\begin{aligned} \text{Loss} = \alpha \textstyle\sum_{j=1}^{N_{91}} \left\| \boldsymbol{y}_j^{K3} - \boldsymbol{x}_j^{K3} - f^{K3}(\boldsymbol{x}_j^{K3}) \right\|_2^2 + \beta \textstyle\sum_{j=1}^{N_{91}} \left\| \boldsymbol{y}_j^{K5} \right. \\ \left. - \boldsymbol{x}_j^{K5} - f^{K5}(\boldsymbol{x}_j^{K5}) \right\|_2^2 + \chi \textstyle\sum_{j=1}^{N_{91}} \left\| \boldsymbol{y}_j^{K7} - \boldsymbol{x}_j^{K7} - f^{K7}(\boldsymbol{x}_j^{K7}) \right\|_2^2 \\ + \delta \left\| \boldsymbol{y}_j^{K3} - \boldsymbol{x}_j^{K3} - R^F(f^C(\boldsymbol{x})) \right\|_2^2, \end{aligned} \tag{10}$$

where the first three terms are the losses of the multi-scale residual networks (K3-, K5-, and K7-networks). The last term represents the residual fusion loss. We simply set $\alpha = \beta = \chi = \delta = 1$. We use a two-step method to train the network. Initially, we parallel-train three SR networks with differently-scaled patches. Then, we determine the fusion loss for the second time on the basis of the contacted residual maps.

A gradient descent method is used to optimize the network parameters by back propagation. Convolution operations reduce the size of the feature map. We maintain many edge pixels by padding zero to infer the center pixel accurately and ensure that all feature maps have the same size to preserve the information on the edge of the image patch.

3. Experiments

3.1. Experimental Data

The learning-based super-resolution methods learn the missing high-frequency information of LR images from the prior information provided in the training data. Generally, the more training data there are, the better reconstruction effect can be obtained by SR methods. In addition, the performance of the SR reconstruction method is also related to the similarity of the test image to the training image. If the test image is close to the statistical characteristics of the training images, it is more likely to get a

good reconstruction result. At this point, there may be fewer training samples to get good results. On the contrary, when the statistical characteristics of the test image and the training image are greatly different, it is difficult to achieve a satisfactory result even using a large-scale training set. To verify the performance of MRNN, we conducted experiments on two satellite image datasets, namely, SpaceNet image and NWPU-RESISC45, to ensure that all algorithms used the same amount of training data. The SpaceNet satellite image dataset https://spacenetchallenge.github.io/AOI_Lists/AOI_1_Rio.html includes five areas in Rio de Janeiro, Paris, Las Vegas, Shanghai, and Khartoum, which are collected from DigitalGlobe's WorldView-2 satellite and published publicly at Amazon. The complete satellite image of Rio de Janeiro (the spatial resolution is 0.5 m) has the highest resolution image with 2.8 M $\times$ 2.6 M pixels, and is divided into 6540 non-overlapping HR image patches with 436 $\times$ 404 pixels, and the main contents of interest in the image are buildings and roads. In total, 2080 images of buildings were randomly selected from these image patches, of which 2000, 40, and 40 images were used as the training set, validation set, and test samples, respectively.

The NWPU-RESISC45 dataset http://pan.baidu.com/s/1mifR6tU [44] is a publicly available benchmark for remote sensing image scene classification (RESISC), created by Northwestern Polytechnical University (NWPU). This dataset covers 45 classes with 700 images in each class. We randomly selected 52 images from each class, of which 50 were used for training and the rest for testing. The HR image size is 256 $\times$ 256 pixels. The spatial resolution of NWPU-RESISC45 varies from approximately 30 m to 0.2 m [44]. Images in the NWPU-RESISC45 dataset, compared with the SpaceNet dataset, have complicated and erratic imaging conditions, including various weather, seasons, and lighting conditions. These factors pose a huge difficulty for SR methods.

Image degradation is a very complex process to be modeled by some filter and down-sampling operators. Here, we interpolated the HR image with bicubic kernel into its LR version with scaling factor t. In the current works (for example, all the comparison methods in our work [17,20,21,23,45]), the most commonly used image degradation is the bicubic downsampling. Since learning-based super-resolution algorithms learn the mapping relationship between low-resolution and high-resolution images, the bicubic degradation is the fairest approach for comparison. Complex imaging degradation model will be investigated in future research. In the testing process, the images did not need to be partitioned.

Peak signal to noise ratio (PSNR) and structural similarity (SSIM) [46] (with default parameters) describe the similarity between the reconstructed and original images in terms of the image. Recent studies [47] have shown that feature similarity (FSIM) [47] and visual information fidelity (VIF) [48] are further consistent with the subjective results. Rectangular-normalized superpixel entropy index (RSEI) [49] (with default parameters) https://github.com/jiaming-wang/RSEI obtains further accurate image evaluation results by introducing the spatial structure of the image. Mutual information (MI) can express the dependence degree of the information between the images in terms of information. The higher is the MI score, the more substantial is the dependence and the higher is the similarity between images. The mutual information between patches $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ is defined as follows:

$$MI(\hat{\boldsymbol{y}};\boldsymbol{y}) = \sum_{q\in\boldsymbol{y}}\sum_{g\in\hat{\boldsymbol{y}}} P(q,g)\log\frac{P(q,g)}{P(q)P(g)}, \tag{11}$$

where q and g represent the gray-scale values, $P(g)$ denotes the ratio of the number of pixels of the gray value that is g to the increased image, and $P(q,g)$ is the joint distribution function of q and g.

We define the information gain between SR image $\hat{\boldsymbol{y}}$ and LR image $\boldsymbol{x}$ relative to HR image $\boldsymbol{y}$ as follows:

$$GMI(\hat{\boldsymbol{y}};\boldsymbol{y}) = \frac{MI(\hat{\boldsymbol{y}};\boldsymbol{y})}{MI(\boldsymbol{x};\boldsymbol{y})}. \tag{12}$$

All image quality assessment metrics only consider the Y component of the YCbCr color space.

3.2. Training Parameters

The proposed network is an end-to-end network, where each sub-network must train for 80 epochs as the pre-training network. The entire network is trained for 10 epochs.

Considering the deep network layer, the algorithm uses learning rate attenuation. We followed Kim et al. [19] for setting hyper-parameters: the learning rate was initialized to 0.1, the learning rate decreased by 1/10 every 20 epochs, and the network's momentum was 0.9. To avoid over-fitting, we used regularized ℓ_2-norm, and its weight decay was 0.0001. For the K3-residual learning network, we set the step size to 1 with a padding size of 1. For the K5-network, the step size was equal to 1 with a padding size of 2. For the K7-network, we set the step size to 1 and padding size to 3. We applied the MSRA method [50] to initialize the weights, that is, satisfying the Gaussian distribution whose mean value is 0, utilizing a variance of $\sqrt{\frac{2}{n}}$ (n is the batch size), and a constant to initialize the bias term with initial value 0. We initially converted the RGB image to the YCbCr color space and then reconstructed the Y channel. After the reconstruction, the Y channel image was restored to the RGB color space. We implemented the MRNN model using the Caffe library [51]. Training the MRNN roughly took 10 h with four 1080Ti GPUs.

3.3. Complementarity Analysis of Multi-Scale Residual

If there were less overlap between different scale residual information, it would mean that the complementarity of residual information between different scales is better [52]. Therefore, in this section, we show the distributions of residual information on different scales. We selected 15 representative LR images $\{\boldsymbol{x}_i\}(i = 1, ..., 15)$ and corresponding HR image $\{\boldsymbol{y}_i\}$ from SpaceNet image datasets with the same configuration of Section 3.6. The reconstruction residual maps of multi-scale networks are $\{f^j(\boldsymbol{x}_i)\}(j = K3, K5, K7)$ for a total of 45 residual images. The estimation residual error map was defined as $\boldsymbol{erm}_i^j = f^j(\boldsymbol{x}_i) - (\boldsymbol{y}_i - \boldsymbol{x}_i) \quad (i = 1, ..., 15 \quad and \quad j = K3, K5, K7)$, and we projected them into 3D and 2D residual feature spaces through principal component analysis (PCA), as shown in Figure 3. The distribution maps of 2D and 3D feature space show that multi-scale networks provide different estimation residual errors. This observation also proved that they are complementary. The overlap observed in Figure 3B covers a sufficiently large feature space, even if only three parallel networks are used. Therefore, additional parallel networks would only increase overlap.

The distribution maps in Figure 3 cannot clearly describe the complementary patterns of multi-scale residual. Therefore, we implemented the clustering of data by k-means and obtained their distribution of 2D feature space, as shown in Figure 4. We name the four patterns as "$s + m + l$", "$s + m/l$", "$m + s/l$", and "$l + s/m$". Pattern "$s + m + l$" represents the best case, that is, the high-frequency information of three scales is complementary between any two. The latter three patterns can be classified as: the information of two scales is considerably common, but a complementary relationship also exists, whereas the other scale complements them. This behavior effectively demonstrates the complementarity between multi-scale residuals.

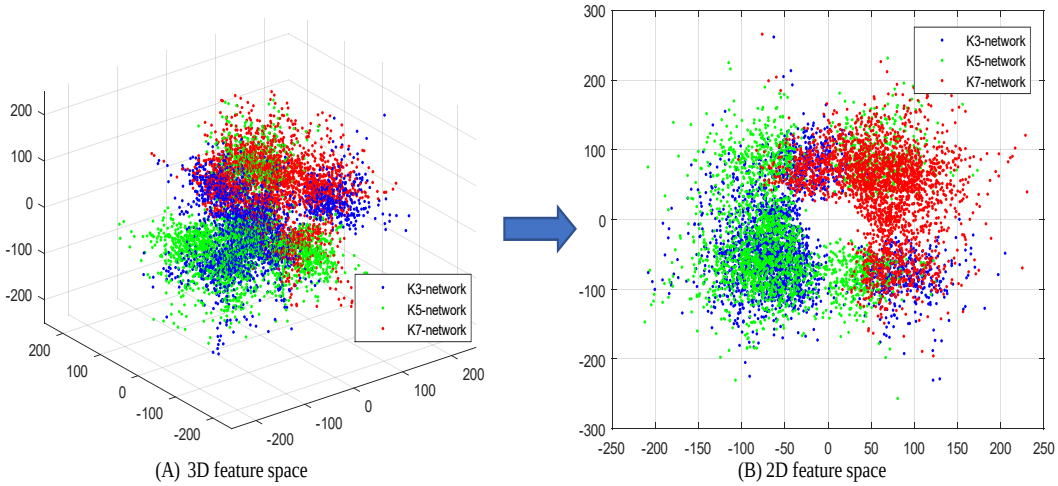

Figure 3. (**A**,**B**) are the visualizations of the estimation residual error map distributions in 3D and 2D feature spaces, respectively. Blue points represent the residual coming from the K3-network, while green and red points represent the K5- and K7- networks, respectively.

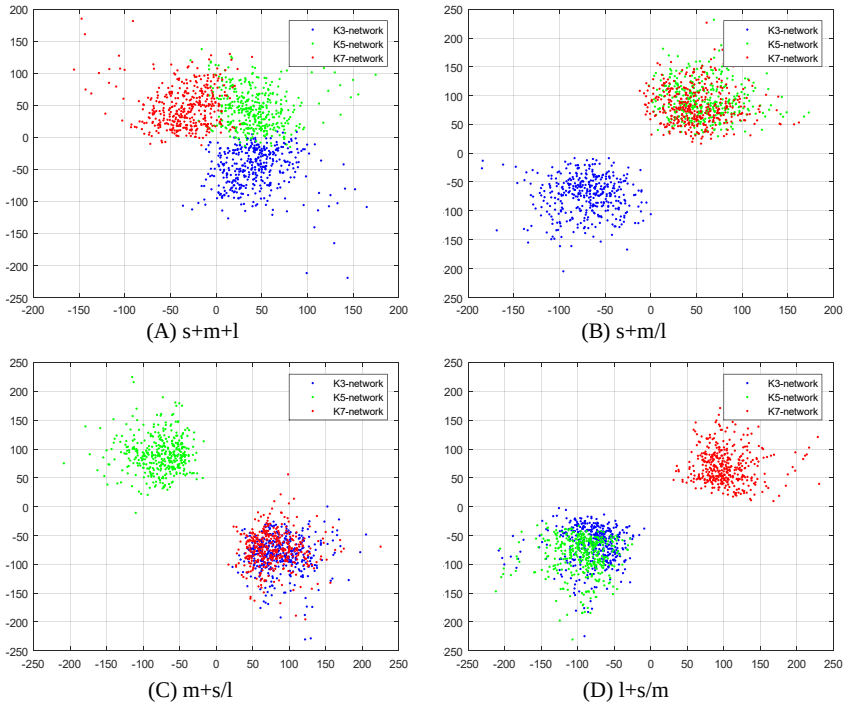

Figure 4. The complementary patterns of multi-scale residual map: (**A**) "$s+m+l$" indicates small- middle- and large-scale residuals are complementary each other; (**B**) "$s+m/l$" represents that small-scale residual information is complementary with both middle- and large- scale ones; (**C**) "$m+s/l$" means that middle-scale residual information is complementary with both small- and large- scale ones; and (**D**) "$l+s/m$" represents the pattern that large-scale residual information is complementary with both small- and middle- scale ones.

We performed quantitative validation as follows:

$$\begin{aligned} &C_{erm^j} = card(\left|\boldsymbol{erm}^j\right| > t), \\ &C_{overlap} = card(\left|\boldsymbol{erm}^{K3}\right| \,\&\, \left|\boldsymbol{erm}^{K5}\right| \,\&\, \left|\boldsymbol{erm}^{K7}\right| > t), \\ &where \quad j = K3, K5, K7, \end{aligned} \tag{13}$$

where $abs(.)$ represents the absolute value of the matrix in an element-wise manner. The function $card(.)$ can count the number of nonzero elements in a matrix. C_{erm^j} represents the number of elements whose values are greater than threshold t. $C_{overlap}$ denotes the number of above elements at the same locations in three error residual maps. We refer to Wang et al. [52] and set $t = 9$ to represent high-value components (high-frequency information signals). Figure 5 plots the bar. The blue bar represents the error only from the K3-network, and the green and red bars indicate the errors only from the K5- and K7-networks, respectively. $C_{overlap}$ is the purple bar. $\forall j \in \{\text{K3, k5, k7}\}$, $C_{overlap} < C_{erm^j}$, and networks of different scales play different roles in the proposed method.

Statistical data and qualitative assessments prove that high-frequency information learned by multi-scale networks is complementary. This case is the reason we fuse multi-scale residual maps for improving reconstruction performance.

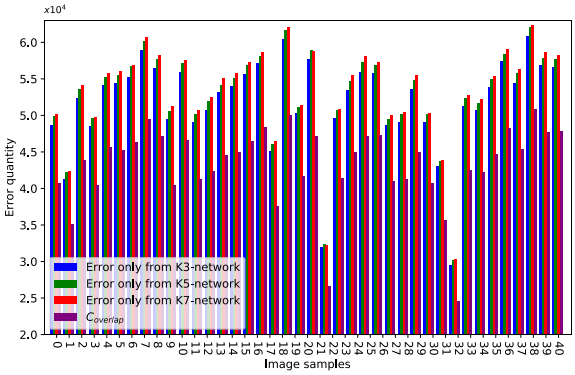

Figure 5. The quantities of estimation errors from multi-scale residual. The quantities of estimation errors from multi-scale more than overlap. Please zoom in to see the differences.

3.4. Performance and Model Trade-Offs

We configured the multi-scale residual network to different depths and compared their performance. We set D at 5, 10, 15, 20, and 25 to test the network performance. The input image patch size changed when the network depth changed. We used PSNR to measure the network performance, as shown in Figure 6. For the K3-network, the performance was optimal when D was 20. For K5- and K7- networks, the performance of networks was optimal when D was 15. The receptive field $S_D \times S_D$ of the D-layer network is defined as $S_D = (k-1) \times D + 1$, and k is the kernel size.

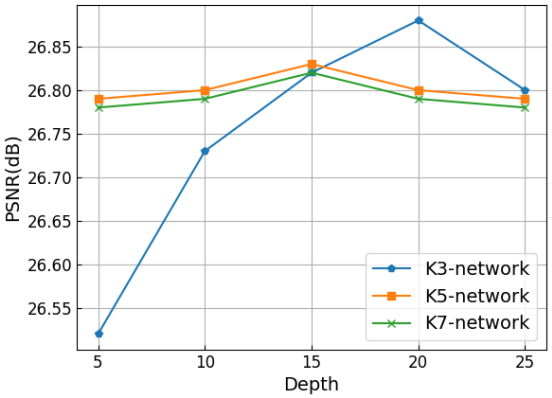

Figure 6. Network depth versus PSNR (dB) score.

3.5. Visualizing the Learned Filters and Feature Maps

The experiments presented in the previous section showed that three different depths of 3×3 networks can replace MRNN. The results prove that "deeper is not better" in certain low-level vision tasks. We would like K5- and K7-networks to learn contextual and global information to compensate for the lack of information in the K3-network. Therefore, we visualize the networks to consider the role of differently-scaled networks in this section.

In the recognition task, the features learned by the network exhibit hierarchical features. Deep features are more discriminative than shallow features, such as color and edge. Therefore, horizontal visualization is suitable for describing the recognition process from low to high level. The image restoration is different from the recognition task. To explore the role of differently scaled networks, we longitudinally visualize the MRNN, that is, the filters and feature maps of the penultimate layer of the differently scaled networks.

A large difference is observed in the complexity of patterns from the filters. Figure 7 represents the feature maps. The larger the filters are, the less local detail information is represented in the feature maps. The smaller are the filters, the more apparent is the detail information in the feature maps.

Overall, we observe that differently scaled networks have their own advantages on various scale objects. For example, a large-scale network performs efficiently on global configuration, a middle-scale network is good at contextual information, and a small-scale network performs well in local detail information. K3-, K5-, and K7-networks have different levels of functionality in the network. A single-scale network cannot simultaneously learn different scales of information. Thus, the multi-scale information should be fused to improve image reconstruction performance.

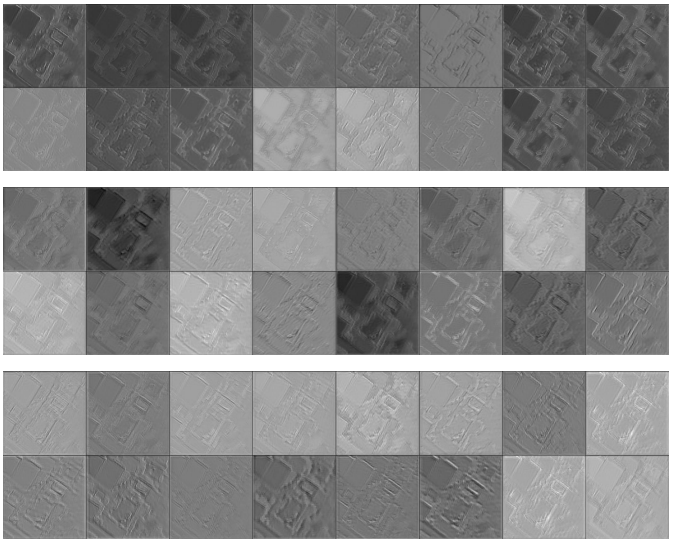

Figure 7. Visualization of the last but one layer feature maps with scale factor of 4. Feature maps from K3- (first two rows), K5- (third and fourth rows), and K7-networks (last two rows). Small-sized filters transport considerable local detailed information from feature maps. In addition, the first row has richer details than the second and the third rows, which can be seen as fine-grain network for SR. The third row has blurry edges and contains coarse-grain global information. The second row is the middle-grain network for contextual information.

3.6. Performance Comparison with State-Of-The-Art SR Algorithms

We conducted subjective qualitative and quantitative analyses on the reconstructed images by using PSNR, SSIM, FSIM, VIF, RSEI, and GMI. To verify the effectiveness of our algorithm, we compared MRNN with the following state-of-the-art SR algorithms:

- SelfExSR [45] is the best performing algorithm based on self-similarity based SR.
- SRCNN [17] is a classic deep-learning based approach, which first uses CNN for SR task.

- LapSRN [20] is the most famous multi-scale SR algorithm based on deep learning.
- VISR [23] is the best performing of satellite image SR algorithm via CNN.
- SRResnet [21] is an excellent depth network algorithm with high computing efficiency and high visual fidelity.

These algorithms were implemented using their public source codes and available parameters provided by the authors, and all images were down-sampled by using the same bicubic kernel of MATLAB. For a fair comparison, we trained all these algorithms with the same database configuration and evaluated the same satellite images with the proposed network.

Figure 8 shows the PSNR, SSIM, FSIM, VIF, RSEI, and GMI of all 40 testing images. MRNN obtained improved reconstruction results. The corresponding significance levels were 100%, 100%, 97.5%, 100%, 100%, and 95%, respectively. The difference in score between MRNN and other methods was statistically significant. Tables 1 and 2 show a considerable quantitative advantage of the proposed method compared with cutting-edge deep learning based algorithms. This finding indicates that residual multi-scale networks are relatively effective in learning different scales of content and structure, and they restore image information effectiveness by using a deeper and flatter network than those used by competing algorithms.

For simple observation, we amplified the representative scale object in randomly selected reconstructed image for comparison. As shown in Figure 9, we selected a roof (small-scale object), building (middle-scale object), and street corner (large-scale object) to show the SR performance. For the examples shown in Figure 9, our method produced sharper edges and finer details than the other methods for all object scales. In addition, our method produced sharper edges and finer details than LapSRN for all object scales. This condition confirms that MRNN fuses multi-scale residual information to enhance visual performance. Figure 10 shows a further intuitive result that only our method can restore a clear outline.

Table 1. Average results of PSNR, SSIM, MI, and GMI on the SpaceNet dataset with scale factor of 4. **Bold** indicates the best performance.

Eval. Mat	Bicubic	SelfExSR	SRCNN	LapSRN	VISR	SRResnet	MRNN
PSNR	25.13	26.20	26.53	26.52	26.74	26.38	**27.02**
SSIM	0.7262	0.7613	0.7675	0.7672	0.7793	0.7738	**0.7894**
FSIM	0.9097	0.9448	0.9501	0.9503	0.9570	0.9546	**0.9575**
VIF	0.3272	0.3848	0.3917	0.3864	0.4021	0.3964	**0.4124**
RSEI	0.3590	0.3760	0.3782	0.3785	0.3828	0.3797	**0.3859**
MI	5.0582	5.1229	5.0980	5.1040	5.1114	5.0972	**5.1424**
GMI	1.0000	1.0128	1.0079	1.0091	1.0106	1.0078	**1.0167**

Table 2. Average results of PSNR, SSIM, MI, and GMI on the NWPU-RESISC45 dataset with scale factor of 4. **Bold** indicates the best performance.

Eval. Mat	Bicubic	SelfExSR	SRCNN	LapSRN	VISR	SRResnet	MRNN
PSNR	25.57	28.48	28.49	28.81	28.77	28.80	**28.93**
SSIM	0.6920	0.7403	0.7378	0.7578	0.7503	0.7564	**0.7580**
FSIM	0.7872	0.8384	0.8336	0.8400	0.8386	0.8419	**0.8461**
VIF	0.2968	0.3499	0.3464	0.3612	0.3661	0.3678	**0.3685**
RSEI	0.3572	0.3740	0.3760	0.3766	0.3777	0.3807	**0.3837**
MI	4.4972	4.5807	4.5507	4.5517	4.5478	4.5808	**4.6210**
GMI	1.0000	1.0191	1.0122	1.0125	1.0117	1.0191	**1.0283**

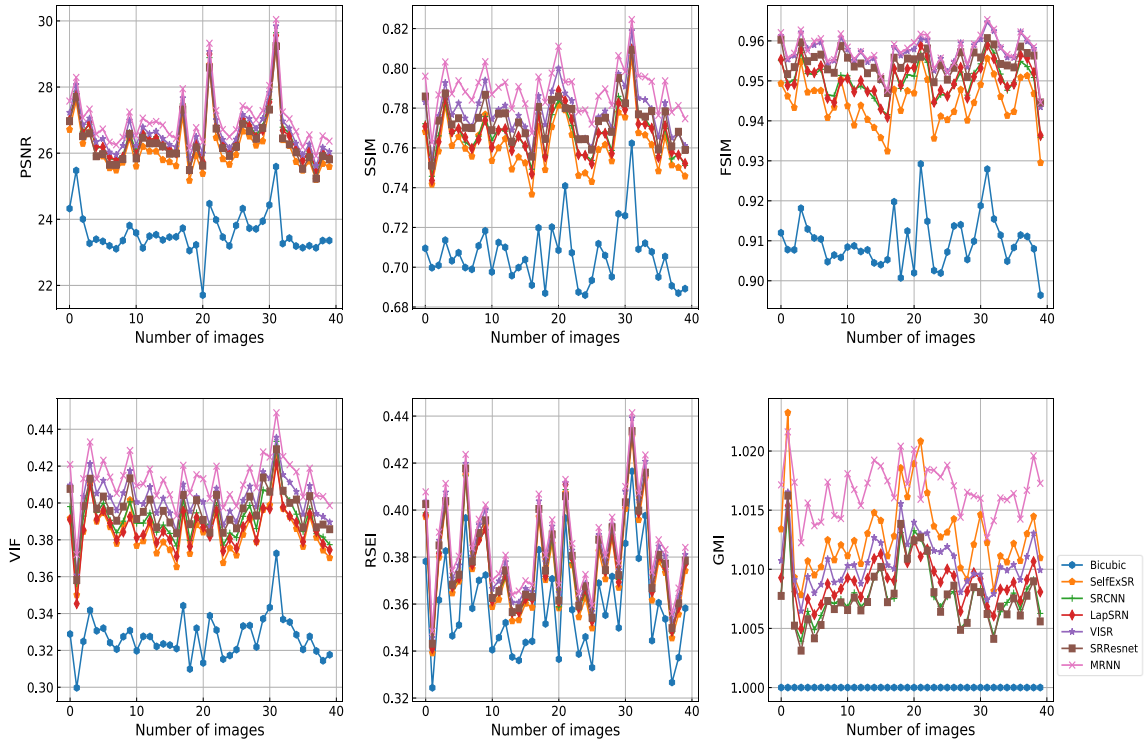

Figure 8. Objective results of SR algorithms over SpaceNet satellite images. X-axes represent the index of testing samples. Y-axes indicate evaluation index: PSNR, SSIM, FSIM, VIF, RSEI, and GMI.

Figure 9. Subjective performance of different SR algorithms over SpaceNet satellite images. We selected three objects with representative scales, i.e., roof (small-scale object), building (middle-scale object), and street corner (large-scale object), and MRNN recovered more texture information.

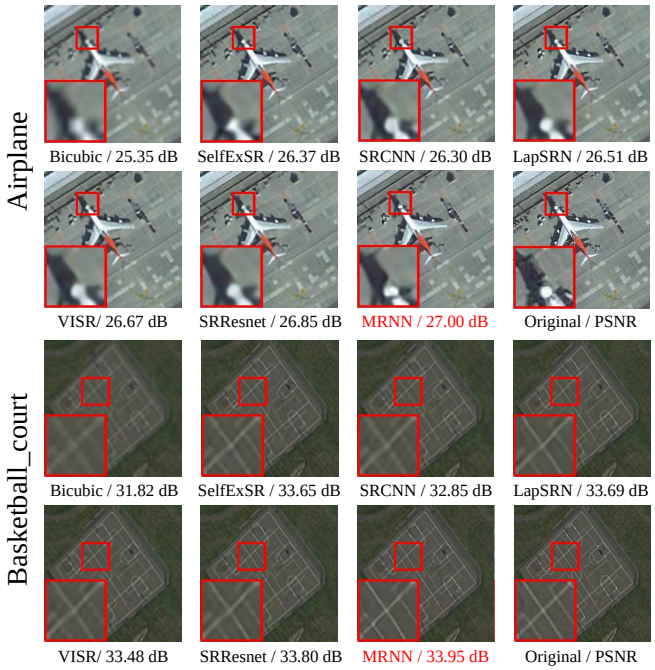

Figure 10. The images from NWPU-RESISC45 with scale factor 4×. Only MRNN successfully recovered the edge of the airplane's head. The contour in the image is sharp in the result of MRNN.

3.7. Time Complexity

Figure 11 shows the running time of all algorithms. The running time of the traditional algorithm is longer than that of deep learning algorithms and has no training phase. MRNN is a parallel network with three different scales and does not increase the time complexity of the network, especially when the network is complex. Although LapSRN has a better running time performance, its PSNR is lower than that of MRNN. Our method is slightly slower than VISR in terms of running time. However, MRNN has improved PSNR, SSIM, FSIM, VIF, RSEI, and GMI. We implemented all algorithms in the experiments under the same hardware configuration: Intel Core i7-6700 K CPU @4.00 GHz, NVIDIA GTX1080 8 GB RAM.

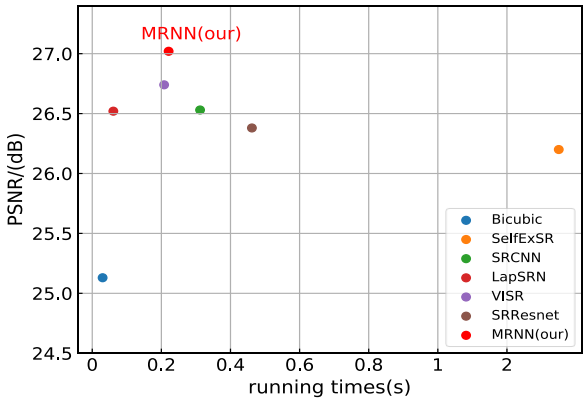

Figure 11. Mean running time (seconds) of all 40 testing samples for different SR algorithms.

4. Discussions

4.1. Multi-Scale Prior Information

Lai et al. [20] proposed a progressive SR method to super-resolve images gradually. A Laplacian pyramid is used in the generative network for SR. A residual recurrent network is adopted to predict the output information in each pyramid level. Here, LapSRN designs a multi-scale training strategy, which trains multi-scale combinations as $2\times$, $4\times$, and $8\times$ in one net. This process involves the addition of multi-scale training pairs to cover different scale samples. Many differences are observed between LapSRN and MRNN. First, LapSRN directly performs multi-scale information fusion in the feature domain, whereas MRNN constructs multi-scale parallel networks and performs multi-scale information fusion in the residual domain. The residual map is closely related to the high frequency information of the image, which is the purpose of SR. Second, LapSRN can perform SR tasks at different scales of factor one-shot, but it ignores the multi-scale information in the input image. MRNN completely investigates the multi-scale information of the input image in SR at fixed-scale factors. On the basis of the experimental results, the fusion of multi-scale residual information has a better performance than LapSRN at a scale factor of 4.

4.2. Residual Learning Versus Pixel Learning

VISR [23] uses a self-similar padding instead of zero padding to avoid the addition of unnecessary information. Therefore, VISR is performed in the pixel domain, such as SRCNN. The reconstruction in the pixel domain focuses on the low- and middle-frequency information in the image. However, SR infers the missing high-frequency information. Furthermore, the residual values of images are frequently small or zero, and the residual network has consistently less calculation burden than pixel learning. The recovery of high-frequency information on satellite images can improve recognition performance. At this point, residual learning is further suitable for satellite image SR scenarios. In addition, the experimental results confirm this inference in terms of subjective and objective image qualities.

4.3. Subpixel Network Versus Pixel Network

SRResnet [21] directly divides LR images into small image patches. Similar to LapSRN and ESPCNN [18], these networks directly use LR inputs (subpixels) to learn mapping functions. Subpixel networks simulate the degradation process and are more efficient than pixel-based networks. By contrast, pixel networks interpolate LR inputs into the same size of HR samples and use the residual information in networks. Residual recursive networks are assumed to overcome the vanishing problem for improving network performance. Thus, subpixel and pixel networks have their own advantages. On the basis of the experimental data in the SpaceNet database, the pixel network outperforms its subpixel competitors.

4.4. Applicability of the Proposed Method

We conducted experiments on the Jilin-1 satellite image to further illustrate the applicability of the proposed algorithm. The imaging environment and resolution of the test image are different from those of the training datasets (NWPU-RESISC45). The size of LR Jilin-1 satellite image is 408×204 pixels. Figure 12 shows the reconstruction results obtained from our proposed approaches and the comparison methods. Considering the absence of ground truth image, we introduce mean gradient (MG) to calculate the sharpness of the SR image. MG is defined as follows:

$$\mathrm{MG} = \left|grd_x(\boldsymbol{y})\right| + \left|grd_y(\boldsymbol{y})\right|, \tag{14}$$

where $grd_x(\boldsymbol{y})$ and $grd_y(\boldsymbol{y})$ are the gradients of image $\boldsymbol{y}$ on the x- and the y-axes, respectively. The proposed MRNN recovers sharp edges, and enjoys the first MG scores. The comparison results of real video satellite images show the applicability of the proposed method. Considering the image characteristics between satellites, we introduce GAN to learn the cross-domain degradation model for solving the real-world SR problems in the future.

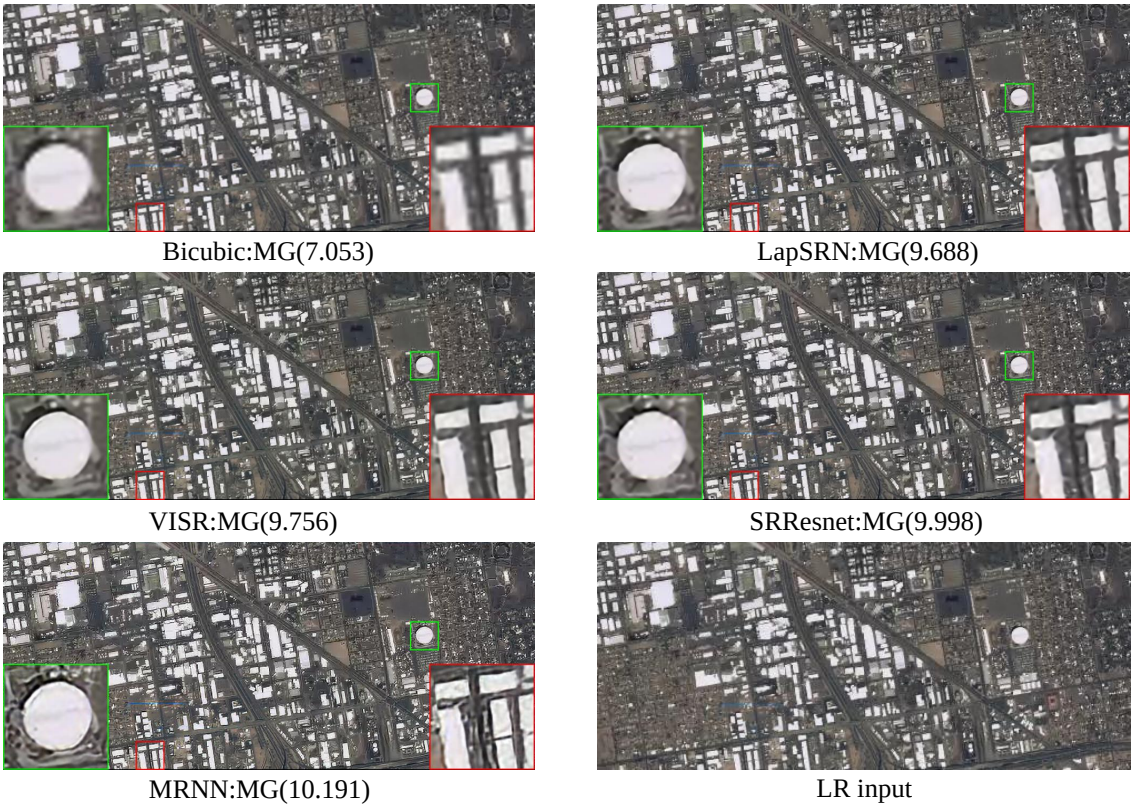

Figure 12. An example of the reconstruction results on the Jilin-1 imagery with a scale of 4. MRNN recovered sharp building edges.

5. Conclusions

This paper presents a multi-scale residual CNN, namely MRNN, based on the characteristics of satellite images, for enhancing SR performance. It first extracts different sizes of patches from LR satellite images. Then, multi-scale deep residual neural networks are applied to simulate differently sized receptive fields for acquiring different levels of information. Then, a fusion network is used to refine the multi-scale features. Based on the proposed novel network, reasonably accurate high-frequency information, such as edges and textures, can be obtained by complementing the residual information at different scales. The experimental results on the SpaceNet database show that the proposed MRNN effectively enhanced the high-frequency information in the reconstructed images. MRNN also exhibited better subjective and objective image qualities than several state-of-the-art deep-learning-based SR algorithms for satellite images. MRNN is mainly designed for true color satellite images SR. As is known, multi-spectral images have higher spectral resolution but lower spatial resolution. It would be very interesting to investigate the fusion of the multi-spectral images and the true color images in the proposed framework to improve the visualized quality of the multi-spectral images in the future.

Author Contributions: Conceptualization, T.L. and J.W.; Data curation, J.W.; Formal analysis, J.W. and J.J.; Funding acquisition, T.L.; Investigation, T.L.; Methodology, Z.W.; Project administration, T.L.; Resources, T.L.; Supervision, Y.Z.; Validation, T.L. and J.J.; Visualization, J.W.; Draft Preparation, J.W.; Writing, review and editing, T.L. and J.J.

Funding: This work is supported by the NSFC grants (61502354, 61671332, 41501505, 61771353), the Central Government Guides Local Science and Technology Development Projects (2018ZYYD059), the Natural Science Foundation of Hubei Province of China (2018CFA024, 2018ZYYD059, 2012FFA099, 2012FFA134, 2013CF125, 2014CFA130, 2015CFB451), Provincial teaching research projects in Hubei Universities (2017324), Scientific Research Foundation of Wuhan Institute of Technology (K201713), Wuhan Institute of Technology Key teaching and construction projects (Z2017009).

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

References

1. Tsai, R.Y.; Huang, T.S. Multiframe image restoration and registration. *Adv. Comput. Vis. Image Process.* **1984**, *1*, 317–339.
2. Lim, K.H.; Kwoh, L.K. Super-resolution for SPOT5—Beyond supermode. In Proceedings of the 30th Asian Conference on Remote Sensing, Beijing, China, 18–23 October 2009.
3. Nasrollahi, K.; Moeslund, T.B. Super-resolution: A comprehensive survey. *Mach. Vis. Appl.* **2014**, *25*, 1423–1468. [CrossRef]
4. Garzelli, A. A Review of Image Fusion Algorithms Based on the Super-Resolution Paradigm. *Remote Sens.* **2016**, *8*, 797. [CrossRef]
5. Shao, Z.; Cai, J. Remote Sensing Image Fusion with Deep Convolutional Neural Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 1656–1669. [CrossRef]
6. Yue, L.; Shen, H.; Li, J.; Yuan, Q.; Zhang, H.; Zhang, L. Image super-resolution: The techniques, applications, and future. *Signal Process.* **2016**, *128*, 389–408. [CrossRef]
7. Shen, H.; Zhang, L.; Huang, B.; Li, P. A MAP approach for joint motion estimation, segmentation, and super resolution. *IEEE Trans. Image Process.* **2007**, *16*, 479–490. [CrossRef]
8. Zhong, Y.; Zhang, L. Remote sensing image subpixel mapping based on adaptive differential evolution. *IEEE Trans. Syst. Man Cybern. Part B* **2012**, *42*, 1306–1329. [CrossRef] [PubMed]
9. Kohler, T.; Huang, X.; Schebesch, F.; Aichert, A.; Maier, A.; Hornegger, J. Robust Multiframe Super-Resolution Employing Iteratively Re-Weighted Minimization. *IEEE Trans. Comput. Imaging* **2016**, *2*, 42–58. [CrossRef]
10. Huang, H.; He, H. Super-resolution method for face recognition using nonlinear mappings on coherent features. *IEEE Trans. Neural Netw.* **2011**, *22*, 121–130. [CrossRef]
11. Romano, Y.; Isidoro, J.; Milanfar, P. RAISR: Rapid and accurate image super resolution. *IEEE Trans. Comput. Imaging* **2017**, *3*, 110–125. [CrossRef]
12. Zhang, H.; Huang, B. Support vector regression-based downscaling for intercalibration of multiresolution satellite images. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 1114–1123. [CrossRef]
13. Dang, C.; Radha, H. Fast Single-Image Super-Resolution Via Tangent Space Learning of High-Resolution-Patch Manifold. *IEEE Trans. Comput. Imaging* **2017**, *3*, 605–616. [CrossRef]
14. Elbakary, M.; Alam, M. Superresolution Construction of Multispectral Imagery Based on Local Enhancement. *IEEE Geosci. Remote Sens. Lett.* **2008**, *5*, 276–279. [CrossRef]
15. Zhang, K.; Gao, X.; Tao, D.; Li, X. Single image super-resolution with multiscale similarity learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2013**, *24*, 1648–1659. [CrossRef] [PubMed]
16. Kappeler, A.; Yoo, S.; Dai, Q.; Katsaggelos, A.K. Video Super-Resolution with Convolutional Neural Networks. *IEEE Trans. Comput. Imaging* **2016**, *2*, 109–122. [CrossRef]
17. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [CrossRef]
18. 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.
19. Kim, J.; Kwon Lee, J.; Mu Lee, K. 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.
20. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep laplacian pyramid networks for fast and accurate superresolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; Volume 2, p. 5.

21. 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.
22. Jiang, K.; Wang, Z.; Yi, P.; Wang, G.; Lu, T.; Jiang, J. Edge-Enhanced GAN for Remote Sensing Image Superresolution. *IEEE Geosci. Remote Sens.* **2019**, *19*, 1–14. [CrossRef]
23. Luo, Y.; Zhou, L.; Wang, S.; Wang, Z. Video Satellite Imagery Super Resolution via Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2398–2402. [CrossRef]
24. Wang, Z.; Yi, P.; Jiang, K.; Jiang, J.; Ma, J. Multi-Memory Convolutional Neural Network for Video Super-Resolution. *IEEE Trans. Image Process.* **2018**, *28*, 2530–2544. [CrossRef]
25. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
26. Lu, T.; Ming, D.; Lin, X.; Hong, Z.; Bai, X.; Fang, J. Detecting building edges from high spatial resolution remote sensing imagery using richer convolution features network. *Remote Sens.* **2018**, *10*, 1496. [CrossRef]
27. Zhao, P.; Liu, K.; Zou, H.; Zhen, X. Multi-stream convolutional neural network for SAR automatic target recognition. *Remote Sens.* **2018**, *10*, 1473. [CrossRef]
28. Zhang, W.; Witharana, C.; Liljedahl, A.; Kanevskiy, M. Deep convolutional neural networks for automated characterization of arctic ice-wedge polygons in very high spatial resolution aerial imagery. *Remote Sens.* **2018**, *10*, 1487. [CrossRef]
29. Xu, Y.; Zhu, M.; Li, S.; Feng, H.; Ma, S.; Che, J. End-to-end airport detection in remote sensing images combining cascade region proposal networks and multi-threshold detection networks. *Remote Sens.* **2018**, *10*, 1516. [CrossRef]
30. Ma, J.; Zhao, J.; Jiang, J.; Zhou, H.; Guo, X. Locality preserving matching. *Int. J. Comput. Vis.* **2019**, *127*, 512–531. [CrossRef]
31. Zhang, Y.; Liu, J.; Bai, W.; Guo, Z. Exploiting multi-scale spatial structures for sparsity based single image super-resolution. In Proceedings of the 2014 IEEE International Conference on Image Processing, Paris, France, 27–30 October 2014; pp. 3877–3881.
32. Fu, J.; Zheng, H.; Mei, T. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4476–4484.
33. Liu, Y.; Cheng, M.; Hu, X.; Wang, K.; Bai, X. Richer Convolutional Features for Edge Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5872–5881.
34. 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.
35. Du, X.; Qu, X.; He, Y.; Guo, D. Single Image Super-Resolution Based on Multi-Scale Competitive Convolutional Neural Network. *Sensors* **2018**, *18*, 789. [CrossRef]
36. Zhang, X.; Yang, W.; Hu, Y.; Liu, J. DMCNN: Dual-Domain Multi-Scale Convolutional Neural Network for Compression Artifacts Removal. In Proceedings of the 2018 25th IEEE International Conference on Image Processing, Athens, Greece, 7–10 October 2018.
37. Zjournaleiler, M.D.; Fergus, R. *Visualizing and Understanding Convolutional Networks*; Springer: Berlin, Germany, 2014; pp. 818–833.
38. Andrews, T.J.; Watson, D.M.; Rice, G.E.; Hartley, T. Low-level properties of natural images predict topographic patterns of neural response in the ventral visual pathway. *J. Vis.* **2015**, *15*, 3. [CrossRef]
39. Zeng, K.; Lu, T.; Liang, X.; Liu, K.; Chen, H.; Zhang, Y. Face super-resolution via bilayer contextual representation. *Front. Signal Process. Image Commun.* **2019**, *75*, 147–157. [CrossRef]
40. Tschechne, S.; Neumann, H. Hierarchical representation of shapes in visual cortex from localized features to figural shape segregation. *Front. Comput. Neurosci.* **2014**, *8*, 93. [CrossRef]
41. Yu, Y.; Tang, S.; Aizawa, K.; Aizawa, A. Category-based deep CCA for fine-grained venue discovery from multimodal data. *Front. IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *99*, 1–9. [CrossRef]

42. Yu, Y.; Tang, S.; Raposo, F.; Chen, L. Deep cross-modal correlation learning for audio and lyrics in music retrieval. *Front. ACM Trans. Multimed. Comput. Commun. Appl.* **2019**, *15*, 20. [CrossRef]
43. Achanta, R.; Hemami, S.; Estrada, F.; Susstrunk, S. Frequency-tuned salient region detection. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, USA, 20–26 June 2009; pp. 1597–1604.
44. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
45. Huang, J.; Singh, A.; Ahuja, N. Single image super-resolution from transformed self-exemplars. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5197–5206.
46. 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]
47. Zhang, L.; Zhang, L.; Mou, X.; Zhang, D. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Trans. Image Process.* **2011**, *20*, 2378–2386. [CrossRef] [PubMed]
48. Sheikh, H.R.; Bovik, A.C. Image information and visual quality. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; Volume 3.
49. Lu, T.; Wang, J.; Zhou, H.; Jiang, J.; Ma, J.; Wang, Z. Rectangular-Normalized Superpixel Entropy Index for Image Quality Assessment. *Entropy* **2018**, *20*, 947. [CrossRef]
50. 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, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
51. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.
52. Wang, S.; Yue, B.; Liang, X.; Jiao, L. How Does the Low-Rank Matrix Decomposition Help Internal and External Learnings for Super-Resolution. *IEEE Trans. Image Process.* **2018**, *27*, 1086. [CrossRef] [PubMed]