Super-Resolution for Remote Sensing Images via Local-Global-Combined Network

Sen Lei, Zhenwei Shi*, *Member IEEE*, and Zhengxia Zou

Abstract

Super-resolution is an image processing technology that recovers a high-resolution image from a single or sequential low-resolution images. Recently deep Convolutional Neural networks (CNNs) have made a huge breakthrough in many tasks including super-resolution. In this letter, we propose a new single-image super-resolution algorithm named local-global-combined networks (LGCNet) for remote sensing images based on the deep CNNs. Our LGCNet is elaborately designed with its "Multi-Fork" structure to learn multi-level representations of remote sensing images including both local details and global environmental priors. Experimental results on a public remote sensing dataset (UC Merced) demonstrate an overall improvement of both accuracy and visual performance over several state-of-the-art algorithms.

Index Terms

Super-Resolution, Remote Sensing Images, Convolutional Neural Networks (CNNs), Local-Global-Combined Network (LGCNet)

I. Introduction

High-resolution images with rich details are essential for many remote sensing applications such as target detection and recognition. Instead of devoting to physical imaging technology, many researchers aim to recover high-resolution images from low-resolution ones using an image processing technology called super-resolution [1].

There have been many earlier researches on image super-resolution, most of which are designed for multiple images, where a series of low-resolution images (different acquisition time of the same scene) are used to recover the high-resolution image [2]. Some recent researches aim at recovering the high-resolution image from a single low-resolution one by learning mapping functions from low-resolution to high-resolution images, with image priors exploited from a large number of training data [3].

The work was supported by the National Natural Science Foundation of China under the Grants 61671037, the Beijing Natural Science Foundation under the Grant 4152031, the funding project of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University under the Grant BUAA-VR-16ZZ-03, and the Fundamental Research Funds for the Central Universities under Grant YWF-16-BJ-J-30. (*Corresponding author: Zhenwei Shi*).

Sen Lei, Zhenwei Shi (Corresponding author) and Zhengxia Zou are with Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China, (e-mail: senlei@buaa.edu.cn; shizhenwei@buaa.edu.cn; zhengxiazou@buaa.edu.cn).

In the field of remote sensing image processing, both of the single-image and multi-image super-resolution methods have been proposed in recent years. Li $et\ al.$ [4] proposed a multi-images super-resolution method named Hidden Markov Tree with maximum a posterior. For single remote image super-resolution, the Sparsity Prior of natural image statistics is commonly used. Pan $et\ al.$ [5] recovered the high-resolution remote sensing image from a single low-resolution image based on compressive sensing and structural self-similarity. Ponomaryov $et\ al.$ [6] combined discrete wavelet transform and sparse representation to generate the high-resolution image from a single low-resolution image. Li $et\ al.$ [7] explored sparse properties in both spectral and spatial domains for hyperspectral images super-resolution. Although the above approaches have played a catalytic role in the remote sensing image super-resolution filed, their defects are obvious.

First, they are all designed based on low-level features such as dictionary of image edges and contours [8][9], or even raw-pixels [5]. The success of machine learning algorithms generally depends on a right way of how image features are represented [10]. Currently, deep convolutional neural networks (CNNs) has become a popular way to learn high-level feature representation automatically from data and have shown great potential in tasks such as image classification [11] and object detection [12]. The highly complex spatial distribution of remote sensing images indicates higher level abstraction and better data representation are essential for applications such as remote sensing target detection and image super-resolution [13]. In related fields such as natural image super-resolution, some researchers have proposed CNN-based single image super-resolution methods [14][15][22] to learn an end-to-end mapping between the low/high-resolution images and have achieved the state-of-the-art performance.

Second, the ground objects of remote sensing images usually share a wider range of their scales, saying that the object itself ($e.g.$ airplane) and its surrounding environments ($e.g.$ airport) are mutually coupling in the joint distribution of their image patterns, which is highly different from those of natural images. Most of the above methods construct dictionary or learn data priors only in a single object scale while environmental information is neglected.

In this letter, we propose a novel image super-resolution method named Local-Global-Combined Networks (LGCNet) by leveraging the multi-level data representation ability of deep learning for remote sensing images. In a typical CNN model, the neurons of lower convolutional layers share small size of receptive fields and focus more on local details, while those in higher layers, bigger receptive fields are accumulated which covers a larger area of data. Our LGCNet is elaborately designed with its “multi-fork” structure to learn multi-scale representations of remote sensing data including both local details ($e.g.$ edge and contours of an object) and global priors ($e.g.$ environmental type).

The rest parts of this letter are organized as follows. Section II gives the implementation details of the proposed method. Experimental results are described in Section III. Some conclusions are drawn in Section IV.

II. Methodology

Given a single low-resolution remote sensing image $\mathbf{X}$, all we need to do is learning a mapping from $\mathbf{X}$ to its original high-resolution image Y. [1]

[1] we first upscale X to the desired size by a fixed factor (e.g. 2, 3 and 4), using bicubic interpolation.

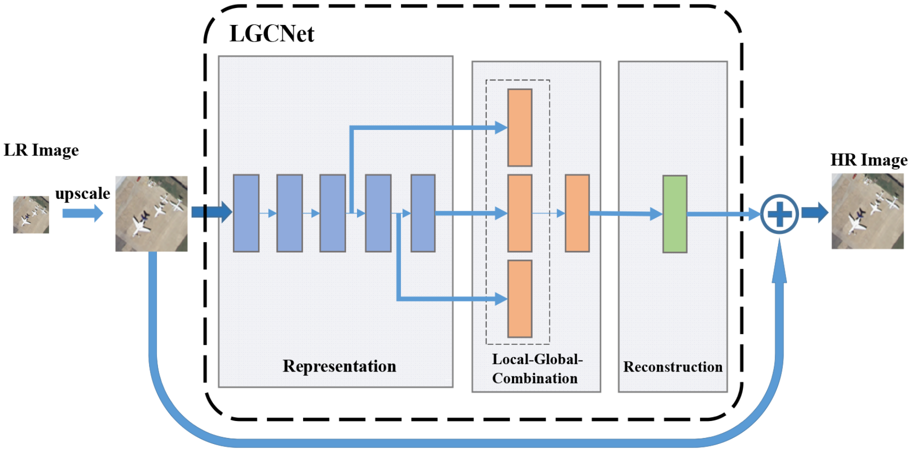


Fig. 1. Flowchart of the proposed super-resolution method for remote sensing images.

A. Convolutional Neural Networks for Super-Resolution

Convolution, nonlinear mapping and pooling are three main components of CNNs. Through these operations, CNNs can adaptively transform the input image space into an effective feature space for a specific task via supervised training. Considering that in an image super-resolution task, a low-resolution image would further lose detailed information after pooling causing a worse reconstructed result, in our model, only convolution and nonlinear mapping operations are used.

Let us represent the size of inputs $\mathbf{X}$ as $H \times W \times C$, where C denotes the channel numbers of remote sensing images. For a network consisting of L convolutional layers, outputs after convolution and nonlinear mapping can be computed as

$$f_1(\mathbf{X}; W_1, b_1) = \sigma(W_1 * \mathbf{X} + b_1) \tag{1}$$

$$f_l(\mathbf{X}; W_l, b_l) = \sigma(W_l * f_{l-1}(\mathbf{X}) + b_l) \tag{2}$$

where W_l, b_l, $l \in (1, ..., L)$ are network weights and bias respectively to be learned. W_l is a tensor with size of $k_l \times k_l \times n_{l-1} \times n_l$, in which k_l denotes the kernel size at layer l, and n_l denotes the number of feature maps at the same layer ($n_0 = C$). b_l is a vector whose size equals n_l. The nonlinear function σ is an element-wise operation and rectified linear function ($max(0, x)$) is mostly adopted nowadays, which makes CNNs converge much faster than traditional saturating nonlinearities [11].

B. Local-Global-Combined Network

The flowchart of the proposed method is illustrated in Fig. 1, in which the part enclosed by a bold dashed box illustrates our proposed LGCNet. When a network goes deep, learning residuals can make the network converge faster and obtain a better minimum and performance [15][16][17]. Therefore, we design LGCNet to reconstruct high-frequency information (residuals)

$$Res(\mathbf{Y}) = \mathbf{Y} - \mathbf{X} = f(\mathbf{X}; W, b) \tag{3}$$

Multi-level information could show great potential for image super-resolution tasks, especially for that of remote sensing images. Deep CNNs with numerous convolution layers are hierarchical models and naturally give multi-level representations of input images, where in low layers the representations focus on local details ($e.g.$ edge and contours of an object) and in higher layers the representations involve more global prior ($e.g.$ environmental type). LGCNet makes full use of the local and global representations and consists of three main parts which are carefully described as follows.

Representation. The first part utilizes L convolutional layers, where each layer is followed by the nonlinear mapping, to adaptively transform inputs into effective feature space and obtain different level representations. Since large convolutional filter size would make the network redundant and slow, we set the filter size k_l and the number of feature maps n_l in each layer relatively small: $k_l = 3$ and $n_l = 32$.

Local-Global-Combination. This part is the core of the multi-scale learning. Local-global-combination is mainly implemented through a "multi-fork" structure by concatenating convolutional results of different layers. One convolutional layer is further applied to merge these combined representation for final reconstruction. To obtain richer representation of the merged layer, we set the filter size and the number of feature maps relatively large, where $k = 5$ and $n = 64$. In this way, the concatenated representation f_c is defined as

$$f_c = [f_i, f_j, f_k, ...] \tag{4}$$

wheref_i, f_j, f_k are different level representations. Then the overall local-global-combined representation f_{lgc} can be computed as follows:

$$f_{lgc} = \sigma(W_{lgc} * f_c + b_{lgc}) \tag{5}$$

Reconstruction. We directly utilize one convolutional layer in this final part of LGCNet to recover residuals (high-frequency components) from the aforementioned local-global-combined representation

$$\mathbf{R} = W_f * f_{lgc} + b_f \tag{6}$$

and the final high-resolution image $\hat{\mathbf{Y}}$ can be further obtained by adding its low-resolution component

$$\hat{\mathbf{Y}} = \mathbf{X} + \mathbf{R} \tag{7}$$

For LGCNet, we set $L = 5$ to make a fast investigation and verification for the proposed idea. For each convolutional layer, in order to assure the output feature maps has the same size as the inputs, the padding is utilized with size of 1 for $k = 3$ and 2 for $k = 3$. Table I presents the detailed configurations, in which the local-global-combination part is determined by experiments in the subsection of Local-Global-Combination Analysis.

We use mean square error as loss function to train the proposed network

$$\frac{1}{2N} \sum_{i=1}^{N} ||\mathbf{Y}_i - \hat{\mathbf{Y}}_i||^2 \tag{8}$$

where N is the total number of training samples.

TABLE I

THE DETAILED CONFIGURATIONS OF LGCNET

Three Main Parts	Configurations
Representation	conv1: $32 \times 3 \times 3$, stride=1, pad=1
	conv2: $32 \times 3 \times 3$, stride=1, pad=1
	conv3: $32 \times 3 \times 3$, stride=1, pad=1
	conv4: $32 \times 3 \times 3$, stride=1, pad=1
	conv5: $32 \times 3 \times 3$, stride=1, pad=1
Local-Global-Combination	concat: conv3+conv4+conv5
	conv6: $64 \times 5 \times 5$, stride=1, pad=2
Reconstruction	conv7: $3 \times 3 \times 3$, stride=1, pad=1

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. *Dataset and Similarity Metrics*

Since there are no open data sets for super-resolution of remote sensing images, we choose UC Merced data set [18], which is a classical scene classification dataset with relatively high spatial resolution (0.3m/pixel), to evaluate our method. UC Merced dataset contains 21 classes of ground features in total with 100 images per class. We split half images (50 images per class) for training and the others for test. Further we randomly select 20% of the training samples as validation set for model selection and the other 80% for training. All the images are firstly to be down-sampled as low-resolution images with the original images acting as high-resolution reference images. In this letter, two classical evaluation criteria, PSNR [dB] (Peak Singal-to-Noise Ratio) and SSIM (Structural Similarity Index Measure) [19] are chosen to measure the performance of several different super-resolution methods. As images in this data set are RGB images, PSNR and SSIM are computed by averaging similarities among these three channels.

Furthermore, real data is used to test the robustness of our proposed method. The three visible bands of the GaoFen-2(GF-2) multispectral image (3.2 m/pixel) are extracted and stacked into a pseudo RGB image for experiments. Since there are no corresponding high-resolution images for reference, the results are displayed and compared with other methods qualitatively.

B. *Implementation Details*

In the training phase, we extract 41×41 sub-images from low-resolution images $\mathbf{X}$ and its corresponding reference image $\mathbf{Y}$ to form the training-sample pairs. The total number of these pairs is around 140k and training uses mini-batch size of 128. Learn rate is initially set to 0.1 to obtain a fast convergence. The training for LGCNet is iterated for 80 epoches in total and the learning rate decreases by a factor of 10 after the 40th epoch. Meanwhile, in order to prevent gradient explosion, we clip gradients by its L2 norm which is often used in training recurrent networks [20]. Specifically, the gradient $\mathbf{g}$ is replaced by $\frac{\mathbf{g}\times t}{||\mathbf{g}||_2}$ before parameter update when $||\mathbf{g}||_2$ is above the threshold t. Momentum and weight decay is set to 0.9 and 0.0001 as most deep learning tasks do. All these experiments are carried on an Inter i7 CPU 4.0 GHz with 32G RAM and Nvida Titan Z, and the Caffe package [21] is utilized to implement our proposed method.

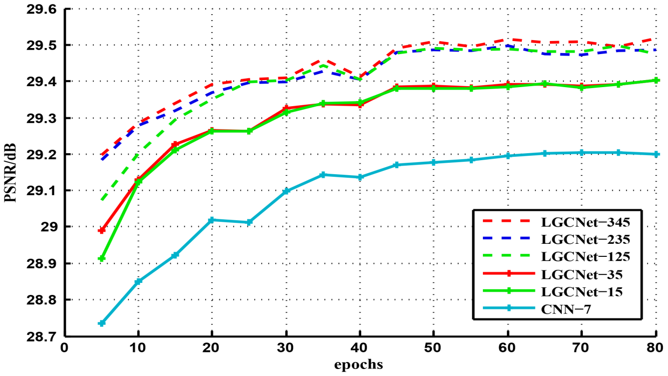


Fig. 2. The experiment results (mean PSNR) of validation set with different training epochs. All models are trained with upscale factor of 3 with the same training configuration.

C. Local-Global-Combination Analysis

The most important property of LGCNet is that it combines different level representations of deep CNNs model which involve relatively local details and global environmental prior to obtain a better super-resolution consequence. To verify if it really be helpful for this task, we design a set of experiments. Firstly, we use a network consisting of 7 convolutional layers (CNN-7) to be a benchmark, which only utilizes global and high level representation to learn residuals. Then we combine the fifth convolutional layer and different lower ones to import into the following concatenated layer, where one or two layers are selected. For fairness, all these models are designed to recover the remote sensing images for the upscaling factor of 3 with the same training configuration.

Fig. 2 shows the experiment results measured by the mean PSNR of validation set with training epochs proceeding. The models, designed with different strategies, are denoted with the corresponding names. Take LGCNet-345 for example, it denotes that this model combines the representations of the third, fourth and fifth layer. As we expected, layer combination gives better super-resolution results for remote sensing images with more layers combined, where more local and global representations are incorporated. The performance of LGCNet-345 is slightly better than other two three-layer-combination models, hence we take this model as the final LGCNet architecture and Table I presents its detailed configurations.

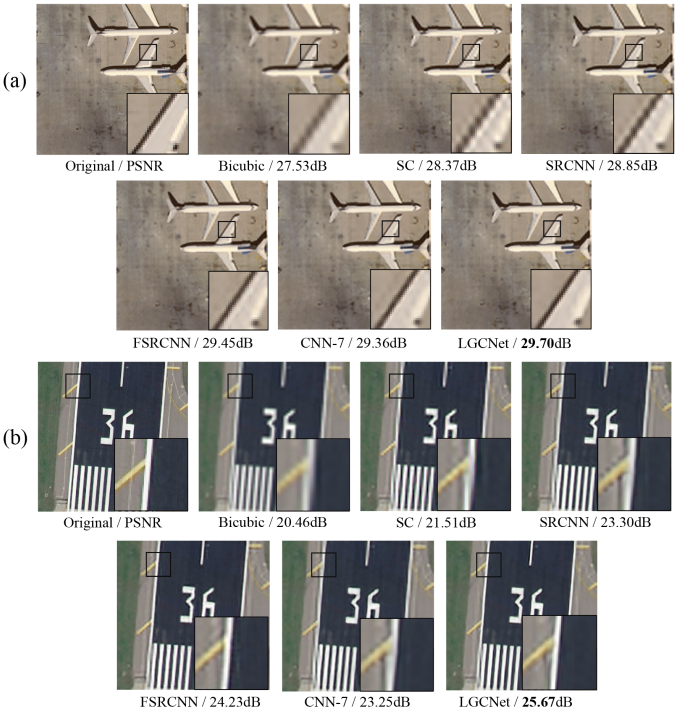
D. Results Comparison and Analyses

Here, we further evaluate the performance of LGCNet on the test set with different upscaling factors, comparing with some other methods including the classic bicubic interpolation, Sparse Coding (SC)[8], CNN-based SRCNN [14] and FSRCNN [22] (state-of-the-arts) and our baseline model CNN-7. Since test images have three channels and it makes no sense, in the context of remote sensing, to turn original channels into YCbCr as it does in SC, SRCNN and FSRCNN, we slightly adjust these three methods to take three-channel images as inputs for fair and convincing comparison. SRCNN and FSRCNN is retrained under our experimental dataset to obtain their best performance for a fair comparison.

TABLE II

MEAN PSNR (dB) AND SSIM OVER ALL THE TEST DATA SET

scale	Bicubic	SC[8]	SRCNN[14]	FSRCNN[22]	CNN-7(ours)	LGCNet(ours)
	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM
2	30.76 / 0.8789	32.77 / 0.9166	32.84 / 0.9152	33.18 / 0.9196	33.15 / 0.9191	**33.48** / **0.9235**
3	27.46 / 0.7631	28.26 / 0.7971	28.66 / 0.8038	29.09 / 0.8167	29.02 / 0.8155	**29.28** / **0.8238**
4	25.65 / 0.6725	26.51 / 0.7152	26.78 / 0.7219	26.93 / 0.7267	26.86 / 0.7264	**27.02** / **0.7333**



(a)

Original / PSNR

Bicubic / 27.53dB

SC / 28.37dB

SRCNN / 28.85dB

FSRCNN / 29.45dB

CNN-7 / 29.36dB

LGCNet / **29.70**dB

(b)

Original / PSNR

Bicubic / 20.46dB

SC / 21.51dB

SRCNN / 23.30dB

FSRCNN / 24.23dB

CNN-7 / 23.25dB

LGCNet / **25.67**dB

Fig. 3. The super-resolution results: (a)“airplane” image (upscaling factor = 3); (b) “runway” image (upscaling factor = 4).

Table II presents the ultimate mean PSNR and SSIM over all the test images of these six methods for three upscaling factors (2, 3, and 4). Among these methods, LGCNet has the best performance with the highest PSNR and SSIM. Fig. 3 illustrates some super-resolution results of these methods. The high-resolution remote sensing images recovered by the LGCNet have clearer edges and more distinct contours.

Table III gives the detailed reconstruction results (upscaling factor = 3) for each class of ground feature, which indicates that our model has achieved an overall improvement for all of the 21 classes [2] over other methods including the state-of-the-arts. Among these classes, “harbor” images (class11) have the lowest PSNR of 23.63 dB (still better than other methods). It should be noticed that some classes such as baseball-diamond (class3), beach (class4) and

[2] All these 21 classes: 1-agricultural,2-airplane, 3-baseballdiamond, 4-beach, 5-buildings, 6-chaparral,7-denseresidential, 8-forest, 9-freeway, 10-golfcourse, 11-harbor, 12-intersection, 13-mediumresidential, 14-mobilehomepark, 15-overpass, 16-parkinglot, 17-river, 18-runway, 19-sparseresidential, 20-storagetanks, 21-tenniscourt

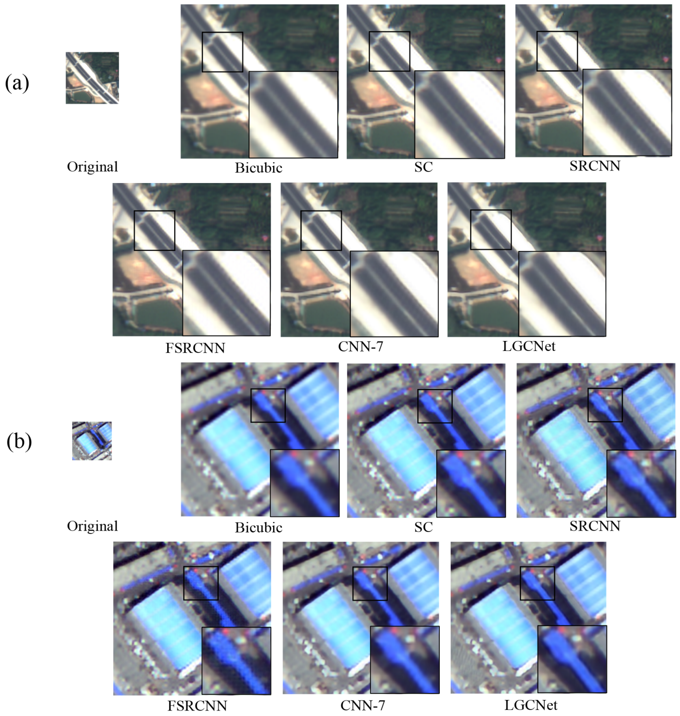


Fig. 4. The super-resolution results of real data: (a)upscaling factor = 3; (b) upscaling factor = 4.

golf-course (class10) may share relatively high PSNR, this is because images of these classes are much smoother than those of other classes thus essentially may be not suitable for evaluating super-resolution tasks and can be excluded. Nevertheless, we still take this complete dataset as a fair judgment. Since local details and environmental priors are essential in all the ground features, our LGCNet with local-global-combinations outperforms other methods in each class.

Fig. 4 illustrates some super-resolution results of the GF-2 satellite data. Even though the resolution of these images (3.2 m/pixel) is different from the training sets, which are 0.9 m/pixel (0.3×3) and 1.2 m/pixel (0.3×4) for upscaling factor 3 and 4 respectively, the LGCNet still obtains better results with fewer jaggies and ringing artifacts. These results indicate our model is more robust than other methods.

E. Evaluations of Depth

In order to explore the influence of architecture depth, We extent our model with five more layers (totally10 layer in the representation part), which combines the 4th, 7th and 10th layer and called LGCNet+. Moreover, we implement VDSR [15] (state-of-the-art) as a comparison, which is an end-to-end deep model with 20 layers. Table IV shows the results on UC Merced test data and the inference time is tested with the Nvida Titan Z (in GPU mode). It can be found that LGCNet+ acquires better super-resolution results than LGCNet, because of deeper representations. Although VDSR is deeper and owns more parameters, LGCNet+ still obtains a little better quality than VDSR with a large speed improvement, which proves the effectiveness of the local-global-combination. LGCNet is a lighter model with the faster speed for super-resolution.

TABLE III
Mean PSNR (dB) of Each Class for Upscaling factor 3

class	Bicubic	SC	SRCNN	FSRCNN	CNN-7	LGCNet
		[8]	[14]	[22]	(ours)	(ours)
1	26.86	27.23	27.47	27.61	27.59	**27.66**
2	26.71	27.67	28.24	28.98	28.81	**29.12**
3	33.33	34.06	34.33	34.64	34.59	**34.72**
4	36.14	36.87	37.00	37.21	37.22	**37.37**
5	25.09	26.11	26.84	27.50	27.39	**27.81**
6	25.21	25.82	26.11	26.21	26.22	**26.39**
7	25.76	26.75	27.41	28.02	27.89	**28.25**
8	27.53	28.09	28.24	28.35	28.35	**28.44**
9	27.36	28.28	28.69	29.27	29.16	**29.52**
10	35.21	35.92	36.15	36.43	36.39	**36.51**
11	21.25	22.11	22.82	23.29	23.32	**23.63**
12	26.48	27.20	27.67	28.06	27.99	**28.29**
13	25.68	26.54	27.06	27.58	27.48	**27.76**
14	22.25	23.25	23.89	24.34	24.30	**24.59**
15	24.59	25.30	25.65	26.53	26.19	**26.58**
16	21.75	22.59	23.11	23.34	23.37	**23.69**
17	28.12	28.71	28.89	29.07	29.03	**29.12**
18	29.30	30.25	30.61	31.01	30.93	**31.15**
19	28.34	29.33	29.40	30.23	29.94	**30.53**
20	29.97	30.86	31.33	31.92	31.87	**32.17**
21	29.75	30.62	30.98	31.34	31.32	**31.58**

TABLE IV
Mean PSNR (dB), SSIM and time (s) over all the test data set

scale	VDSR[15]	LGCNet(ours)	LGCNet+(ours)
	PSNR/SSIM/time	PSNR/SSIM/time	PSNR/SSIM/time
2	33.47/0.9234/0.119	33.48/0.9235/**0.063**	**33.53/0.9242**/0.070
3	29.34/**0.8263**/0.118	29.28/0.8238/**0.061**	**29.35**/0.8251/0.069
4	27.11/0.7360/0.120	27.02/0.7333/**0.061**	**27.13/0.7375**/0.073

IV. Conclusion

We designed a novel network named local-global-combined network (LGCNet) to make full use of the representations of deep convolutional neural networks (CNNs) for the super-resolution of remote sensing images. The LGCNet focuses on the reconstruction of residuals between low-resolution and corresponding high-resolution image pairs by learning multi-level representation of ground objects and environmental priors. Experimental results show that the fusion of different layers gives more accurate reconstruction results. Our method obtains an overall improvements on overall improvement (for all the 21 classes) of both accuracy and visual performance over several state-of-the-art algorithms. Moreover, experiments on real data verify the robustness of our LGCNet and more layers adopted in representation part contribute to quality improvements with a lower speed.

References

[1] L. Yue, H. Shen, J. Li, *et al*. "Image super-resolution: The techniques, applications, and future," *Signal Processing*, vol. 128, pp. 389-408, 2016.

[2] S. Borman and R. L. Stevenson, "Super-Resolution from Image Sequences-A Review," *Midwest Symposium on Circuits and Systems*, pp. 374-378, 1998.

[3] C. Y. Yang, C. Ma and M. H. Yang, "Single-image super-resolution: A benchmark," *Proc. Eur. Conf. Comput. Vis.*, pp. 372-386, 2014.

[4] F. Li, X. Jia, D. Fraser and A. Lambert, "Super resolution for remote sensing images based on a universal hidden Markov tree model," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1270-1278, 2010.

[5] Z. Pan, J. Yu, H. Huang, *et al*. "Super-resolution based on compressive sensing and structural self-similarity for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4864-4876, 2013.

[6] H. Chavez-Roman and V. Ponomaryov, "Super resolution image generation using wavelet domain interpolation with edge extraction via a sparse representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1777-1781, 2014.

[7] J. Li, Q. Yuan, H. Shen, *et al*. "Hyperspectral Image Super-Resolution by Spectral Mixture Analysis and Spatial-Spectral Group Sparsity," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 9, pp. 1250-1254, 2016.

[8] J. Yang, J. Wright. T. S. Huang, *et al*, "Image super-resolution via sparse representation," *IEEE Trans. on Image Processing*, vol. 19, no. 11, pp. 2861-2873, 2010.

[9] W. Dong, L. Zhang, G. Shi, *et al*, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. on Image Processing*, vol. 20, no. 7, pp. 1838-1857, 2011.

[10] Y. Bengio, A. Courville, P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798-1828, Aug. 2013.

[11] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proc. Advances in Neural Information Processing Systems*, 2012, pp. 1097-1105.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580-587.

[13] Z. Zou and Z. Shi, "Ship Detection in Spaceborne Optical Image with SVD Networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5832-5845, Oct. 2016.

[14] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295-307, Feb. 2016.

[15] J. Kim, J. K. Lee, K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 1646-1654

[16] K. He, X. Zhang, S. Ren, *et al*, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 770-778

[17] T. Hui, C. C. Loy, X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *European Conference on Computer Vision*, 2016

[18] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM GIS*, 2010, pp. 270-279.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, *et al*, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.

[20] R. Pascanu, T. Mikolov, and Y. Bengio., "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, 2013.

[21] Y. Jia, E. Shelhamer, J. Donahue, *et al*, Caffe: Convolutional architecture for fast feature embedding, in *Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 675C678.

[22] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision*, 2016