Multimedia Tools and Applications
https://doi.org/10.1007/s11042-018-6751-5

Image super-resolution based on two-level residual learning CNN

Min Gao[1] · Xian-Hua Han[2] · Jing Li[1] · Hui Ji[1] · Huaxiang Zhang[1] · Jiande Sun[1]

Received: 3 August 2018 / Revised: 20 September 2018 / Accepted: 2 October 2018 /
Published online: 17 October 2018

Abstract
In recent years, CNN has been used for single image super-resolution (SR) with its success of in the field of computer vision. However, in the recovery process, there are always some high-frequency components that cant be recovered from low-resolution images to high-resolution ones by using existing CNN-based methods. In this paper, we propose an image super-resolution method based on CNN, which uses a two-level residual learning network to learn residual components, i.e., high-frequency components. We use the Super-Resolution Convolutional Neural Network (SRCNN) as the network structure in each level so that our proposed method can achieve the high-resolution images with high-frequency components that cant be obtained by the existing methods. In addition, we analyze the proposed method with considering three kinds of residual learning networks, which are different in the structure and superimposed layers of the residual learning network. In the experiments, we investigate the performance of the proposed method with various residual learning networks and the effect of image super-resolution to image captioning task.

Keywords Image super-resolution · Convolutional Neural Networks (CNNs) · Two-level residual learning CNN · Image caption

1 Introduction

Image super-resolution (SR) refers to the recovery of a high-resolution (HR) image from low-resolution (LR) images or image sequences. SR is a long-history topic in the field of computer vision. Image SR reconstruction was first proposed by Harris et al. in 1964 [7].

✉ Xian-Hua Han
hanxhua@yamaguchi-u.ac.jp

✉ Jiande Sun
jiandesun@hotmail.com

[1] School of Information Science and Engineering, Shandong Normal University, Jinan 250014, Shandong Province, China

[2] Graduate School of Science and Technology for Innovation, Yamaguchi University, Yamaguchi, Japan

In many practical applications, people usually expect high-resolution images, though the imaging devices cant obtain the images as high-resolution as they expect. HR images can provide more details which are indispensable in many practical applications. For example, HR medical images are helpful for doctors to make a correct diagnosis [15]. HR satellite images are used to distinguish objects on the earth even from the outside space [18]. HR face images are important data used in daily public surveillance [28]. And the images with higher resolution can usually improve the performance of almost all tasks. There is no end to the pursuit of high resolution in the field of computer vision.

Usually image SR methods can be classified into two kinds, i.e., multiple images based and single image based methods. For single image based SR methods (SISR) [1], the SR problem is widely known to be ill-posed, and this kind of methods can be classified into three categories: interpolation-based, reconstruction-based, and learning-based methods. The interpolation-based methods use traditional interpolation algorithms, e.g., bilinear interpolation or bicubic interpolation to obtain high-resolution images of low-resolution images [26]. This method is simple and fast, but the obtained high-resolution images have the effect of ambiguity. The reconstruction-based method is to obtain high-resolution images by the prior knowledge, such as edge prior [17], gradient profile prior [16], and non-local mean [14]. These methods work well in sharping edges, but the reconstruction effect is poor when the magnification is large. The learning-based methods assume that the low-resolution image is caused by the loss of the high-frequency component of the corresponding high-resolution image, and the loss can be functioned by learning the pair-wised information on the training dataset, which contains the low-resolution images and their corresponding high-resolution images [4].

Recently, some deep learning-based SR studies are carried out. In these methods, low-resolution images can be upgraded to high-resolution images through deep learning model, which can extract the "common features" between them. For instance, SRCNN is one of the popular deep convolutional neural network (CNN) models for image super-resolution. The SRCNN process is mainly divided into three phases: patch extraction and representation, non-linear mapping, and high-resolution reconstruction. Though it achieves good performance, there is still something for SRCNN to be improved. For example, the perceptive field is relatively small during SRCNN feature extraction, and only one layer of convolutional layers is used. That is, the extracted features are local features. However, the SR image usually needs to restore the texture details, so it is difficult to achieve better performance based on only the local features. Given SRCNN, there are several improved methods, such as, RED [13], CSCN [21], VDSR [9] and so on. The SRCNN does not recover well for details, has slow convergence, and requires a fixed scale factor. The VDSR network adopts the deeper residual network to achieve high speed, in which VGG16 is introduced for super-resolution. In addition, the thought of learning residual information is introduced into SR to reduce the "burden" of the network and accelerate the learning rate. The network suitable for different scales is constructed and the reliability of the network is verified by experiments. As a result, VDSR can reach higher image quality than most of the previous works.

In this paper, we propose an image SR method based on two-level residual learning convolutional neural network (TLRLCNN). Here, the TLRLCNN model consists of two successive residual learning convolutional neural networks (RLCNN). In each level of the proposed CNN model, two SRCNNs in series are used to construct the basic network. But, in the first RLCNN, the input of first SRCNN is the low-resolution (LR) image (I_1)and its target is to learn the high-resolution (HR) image (I_2). And the output of the first RLCNN, I_2 , is the input of the second SRCNN, and its target is to learn the residual image (I_3). Hence, the output of the first RLCNN is an image (I_4), which is the combination of the

HR image (I_2) and the residual image (I_3). The second RLCNN has the same structure as the first one. It takes I_2 as its input, and its output is the final HR image. In each RLCNN, we optimize the loss between the ground truth and the estimated HR images and the loss between unrecovered residual components and the estimated high-frequency components simultaneously. The contributions of the study are in two aspects:

1) We propose to learn the residual information via the two-level residual learning CNN, which can learn the unrecovered residual component and restore the high-frequency components as much as possible.
2) We investigate the effect of image resolution to the task of image caption, and demonstrate that image super-resolution is useful to achieve the better image description than that obtained based on images with lower resolution.

2 Two-level residual learning CNN

We propose a residual learning convolutional neural network (RLCNN) model in order to obtain as much as possible image details. Figure 1 shows the framework of our proposed RLCNN model. In our proposed model, two RLCNNs are in series as shown in Fig. 1, which is helpful to restore more high frequency components, and obtain the HR images with more details. In the first RLCNN, the input of first SRCNN is the low-resolution (LR) image (I_1) and its target is to learn the high-resolution (HR) image (I_2). And the output of the first RLCNN, I_2 , is the input of the second SRCNN, and its target is to learn the residual image (I_3). Hence, the output of the first RLCNN is an image (I_4), which is the combination of the HR image (I_2) and the residual image (I_3). The second RLCNN has the same structure as the first one. It takes I_4 as its input, and its output is the final HR image. In each RLCNN, we optimize the loss between the ground truth and the estimated HR images and the loss between unrecovered residual components and the estimated high-frequency

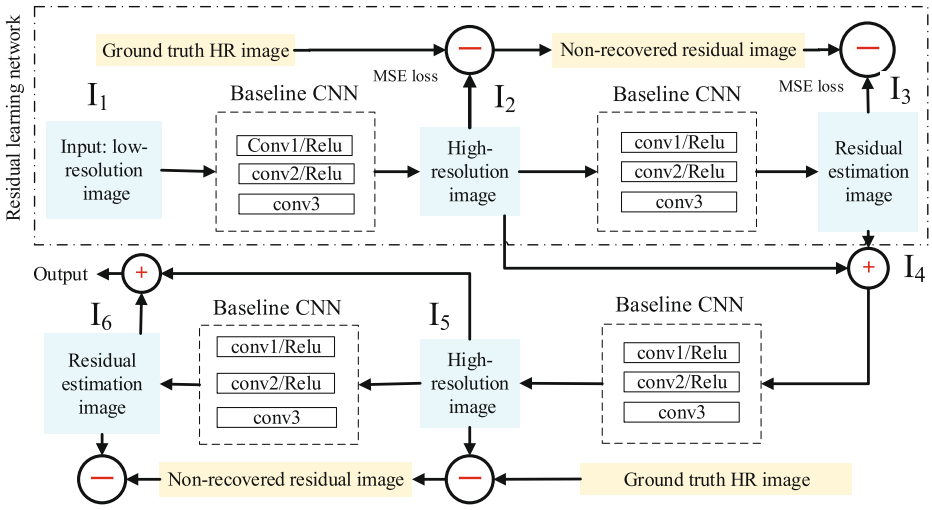

Fig. 1 We stack another RLCNN on an RLCNN to perform TLRLCNN on the image. The TLRLCNN learn twice for the residual components. In the two RLCNNs, the unrecovered residual image minus the residual estimated image is the loss of the residual

components simultaneously. In order to emphasize the high-frequency components, we give a higher weight to the loss between the unrecovered residual component and the estimated high frequency component.

2.1 First-level residual learning convolutional neural

We use y to denote LR images, g to denote the HR ground truth image, and x to denote the input features of the first-level network. The minimized objective function of the first-level CNN network is expressed as:

$$\mathrm{G}' = \min_{W} \|g - F(\mathrm{y}, W)\|^2 \tag{1}$$

Where $F(\cdot)$ denotes the transfer function from the LR input image y to the HR image. W denotes the filter parameter of several convolution layers on the input LR images, G′ is the unrecovered HR component of the low-resolution image after the first SRCNN. The feature output of first SRCNN structure is formulated as:

$$\mathrm{x}' = \mathrm{ReLU}(\mathrm{R}(\mathrm{W}^R, \mathrm{x}) + \mathrm{i}(\mathrm{x})) \tag{2}$$

Where x' is the output of the first-level network, W^R is a set of weights in the first-level network structure, and it is used to minimize of the objective function in the global network structure. Function ReLU(R(w, x) denotes ReLU nonlinear mapping function, $i(x)$ is an identity mapping function as in [24], i.e, $i(x)$= x. The HR image recovered from the LR image through the first-level SRCNN is:

$$I_2 = F(y, W) \tag{3}$$

It can be seen from (1)-(3) that the RLCNN can not only reconstruct the low-frequency components, but also learn the high-frequency components that have not been learned after reconstruction. The recovered high-frequency components are formulated as:

$$g^{res1} = g - F(y, W) \tag{4}$$

In order to learn the high-frequency components as many as possible, another RLCNN is stacked in addition to the first residual learning CNN.

$$\begin{aligned} G'' &= \min_{W, W_{res1}} (\left\| g^{res1} - F_{res1}(F(y, W), W^{res1}) \right\|^2) \\ &= \min_{W, W_{res1}} (\left\| g^{res1} - F_{res1}(I_2, W^{res1}) \right\|^2) \end{aligned} \tag{5}$$

Where $F_{res1}(\cdot)$ is the transfer function from the high-resolution image to the learning residual image, W_{res1} is the filter parameter of the convolutional layer, and g^{res1} is the residual component that has not been recovered from the learning residual convolutional neural network (RLCNN).The resulting residual estimate image is represented as:

$$I_3 = F_{res1}(F(y, W), W^{res1}) \tag{6}$$

Thus, the objective function of the first residual learning neural network (RLCNN) is:

$$\begin{aligned} < G', G'' > = \min_{W, W^{res1}} & (w_1 \|g - F(y, W)\|^2 \\ & + w_2 \left\| g^{res1} - F_{res1}(F(y, W), W^{res1}) \right\|^2) \end{aligned} \tag{7}$$

2.2 Second-level residual learning convolutional neural network

In the RLCNN framework, the input of the second-level network is the output of the first-level network, and the second-level network learns the residual components that the first-level network unrecovered explicitly. The second RLCNN is exactly the same as the first-level RLCNN. The output of the first-level RLCNN is used as the input of the second-level RLCNN, and it is formulated as:

$$I_4 = I_2 + I_3 \tag{8}$$

The input of the second SRCNN in the second-level RLCNN is

$$I_5 = F_H(I_4, W^H) \tag{9}$$

Where $F_H(\cdot)$ is the transfer function from learning residual image to second HR image, and W_H is the filter parameter of the convolutional layer.

So the objective function for the second HR image is formulated as:

$$\begin{aligned} \mathrm{G}''' &= \min_{W_H}(\left\| g - F_H(F_{res}(I_4, W^{res}), W^H) \right\|^2) \\ &= \min_{W_H}(\left\| g - F_H(I_5, W^H) \right\|^2) \end{aligned} \tag{10}$$

The unrecovered residual component of the second-level RLCNN is formulated as:

$$g^{res2} = g - F_H(I_5, W^H) \tag{11}$$

The HR image of the final-level SRCNN output is formulated as:

$$I_6 = F_{res2}(I_5, W^{res2}) \tag{12}$$

The objective functions of second-level RLCNN is:

$$\begin{aligned} < G''', G'''' > = \min_{W^H, W^{res2}} (& w_3 \left\| g - F_H(I_5, W^H) \right\|^2 \\ & + w_4 \left\| g^{res2} - F_{res2}(I_6, W^{res2}) \right\|^2) \end{aligned} \tag{13}$$

So the final objective function is:

$$\begin{aligned} < G', G'', G''', G'''' > = \min_{W, W^{res1}, W^H, W^{res2}} (& w_1 \| g - I_2 \|^2 \\ & + w_2 \left\| g^{res1} - F_{res1}(I_2, W^{res1}) \right\|^2 \\ & + w_3 \left\| g - F_H(I_5, W^H) \right\|^2 \\ & + w_4 \left\| g^{res2} - F_{res2}(I_6, W^{res2}) \right\|^2 \end{aligned} \tag{14}$$

Where w_1, w_2, w_3 and w_4 are the weights of the reconstruction errors on the estimation of g, the residual component estimation of g^{res1} and the second estimation of g and the residual component estimation of g^{res2} respectively.The final output is the sum of the outputs of the two SRCNN:

$$I_{out} = I_5 + I_6 \tag{15}$$

3 Variation of the proposed method

3.1 Residual learning CNN (RLCNN)

As we describe in Section 2, the residual learning CNN (RLCNN) (Fig. 2) learns the residual component once after learning the HR image, and the final super-resolution image is obtained by adding the learned HR image to the learned residual image.

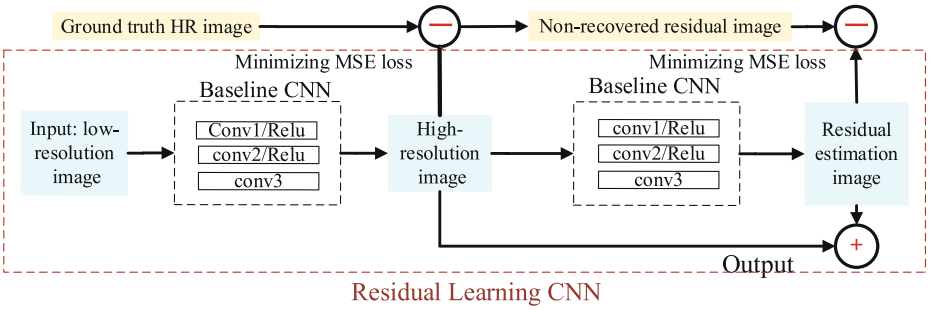

Fig. 2 Residual learning CNN(RLCNN)

3.2 Residual learning CNN+

Compared with RLCNN, the residual learning CNN+ (RLCNN+) (Fig. 3) learns the residual components twice after learning the HR image, the final super-resolution image is obtained by combining the learned HR image with the two residual images.

4 Experiment

4.1 Dataset

In image SR task, the training dataset consists of 91 images. Each image is divided into sub-images with the size of 33× 33, and the 91 images are divided into 24,800 sub-images, which are extracted from the original image with a stride of 14. There is a test dataset: set5 [1], which consists of 5 images. We compare the proposed model with its two variations on this dataset. In addition, we apply the image super-resolution in image caption to demonstrate the usability of the proposed model. In image caption, we use the MSCOCO dataset in the training and test of image caption. We select 82,783 images from the MSCOCO dataset to train the image caption model.

4.2 Training

In image super-resolution experiments, the size of the convolution filter is 9, 1, 5, and the stride is 1. The first two output features are 64 and 32 respectively. We set upscaling factors to 3, 4. We use PSNR to evaluate the quality of the obtained HR images. We use the SRCNN method as our baseline. In order to ensure that the output image size is consistent with the

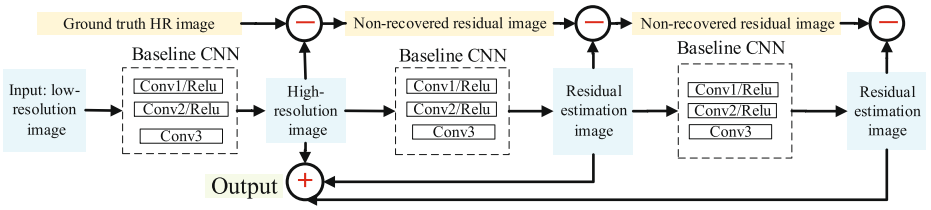

Fig. 3 Residual Learning CNN+ (RLCNN+)

Table 1 The PSNR of the transfer method on the dataset set5

Scale	Bicubic interpolation	SRCNN	RLCNN	RLCNN+	TLRLCNN
×2	32.45dB	32.92dB	33.45dB	33.27dB	33.29dB
×3	27.79dB	27.90dB	28.25dB	28.86dB	29.69dB
×4	25.90dB	26.44dB	26.45dB	26.59dB	27.83dB

labels, the padding mode in the first baseline CNN is set to 'VALID', and the padding mode in the remaining three baseline CNN layers is set to 'SAME'. The padding is 'SAME' in the convolutional layers for ensuring the same spatial size of the output with its input. Therefore, the center of the 33 × 33 input LR sub-image is taken as the ground truth HR image with the size of 21 × 21. The Euclidean distance between the predicted output and the ground truth patch is adopted to measure the similarity. Our network uses the tensorflow framework. We set the batch as 128, the fixed learning rate as 0.0001, and the color channel dimension as 1. We set w_1 , w_2 , w_3 , and w_4 as 0.2, 0.8, 0.2 and 0.8 respectively in training. The SGD optimizer is used in the optimization. Our model parameters are initialized according to Gaussian distribution with standard deviation 0.001.

4.3 Super-resolution experiments

Table 1 shows quantitative comparisons for ×3 and ×4 SR. In Tables 2, 3 and 4, we show the PSNR value after 'butterfly' and 'bird' images processing. Comparing the three transformed models shows that the more residual layers, the larger the PSNR value. When compared with the RLCNN and RLCNN+, our TLRLCNN achieves the best average results on Set5. Specifically, for the scaling factor ×3, our TLRLCNN performs the best on Set5.

In Figs. 4 and 5 we show visual comparisons on scale ×2, ×3. For image "butterfly" and "bird", we observe that RLCNN and RLCNN+ would produce noticeable artifacts and produce blurred edges. In contrast, our TLRLCNN can recover clearer edges, more close to the ground truth. These results further indicate the benefits of learning residual component.

In Fig. 6, we show visual comparisons on scale ×4. For image "butterfly" and "bird", we can see that the edges of low resolution images are blurry. And we observe that our TLRLCNN can recover clearer edges, more close to the ground truth. These results also further indicate the benefits of learning residual component.

4.4 Image caption experiments

Image caption comes from the basic idea of language translation, and it tries to find a natural language statement to describe the content of an image automatically. Image captioning helps visually impaired people to understand visual content [20]. Image captioning can be used for multimedia search, video content query and visual understanding of chat bots. And Image captioning can also be used for surveillance event detection[2, 12]. In

Table 2 The PSNR of "butterfly" and "bird" (Set5) with scale factor × 2

Scale×2	Bicubic interpolation	SRCNN	RLCNN	RLCNN+	TLRLCNN
butterfly	27.07dB	28.73dB	29.55dB	29.22dB	29.42dB
bird	36.34dB	36.66dB	37.43dB	37.19dB	36.96dB

Table 3 The PSNR of "butterfly" and "bird" (Set5) with scale factor × 3

Scale×3	Bicubic interpolation	SRCNN	RLCNN	RLCNN+	TLRLCNN
butterfly	22.11dB	22.56dB	23.38dB	23.18dB	23.33dB
bird	30.64dB	30.34dB	30.68dB	30.73dB	29.82dB

recent years, most image captioning methods are based on the encoder-decoder model [3, 10, 22, 25]. The encoder is usually a convolutional neural network, and the features of the final fully connected layer or convolution layer are used as image features. The decoder is usually a recurrent neural network, which is mainly used for image description generation. Vinyals et al. propose an encoder-decoder framework, in which image features are extracted by CNN, and then the target language is generated by LSTM. The objective function is to maximize the maximum likelihood estimation of the target description [19]. Fang et al. use multi-instance learning to train visual detectors to extract the words contained in an image, and then learn a statistical model for generating descriptions [5]. Inspired by the recent development of attention mechanisms in machine translation, Xu et al. propose a method of combining spatial attention mechanisms in the convolutional features of images, and then input the context information into the encoder-decoder framework [23]. Xu et al. use three kinds of semantic information to guide the generation of words at each moment [8]. Zhou et al. use the text-conditional method, and combine with the image features, which can finally be used for the generation of the current word according to a specific region of the image [27]. Lu et al. propose the concept of visual sentinel, and it can determine whether the generated word is adaptive and whether image features or text features are used [11]. However, in image captioning task, there are very few researches discussing the effect caused by image quality on performance, for example, the image resolution. In this paper, we demonstrate that low-resolution images can degrade the accuracy of image caption, and our proposed method is helpful to improve the performance of image caption (Figs. 7 and 8). Firstly, we reconstruct the LR image into the HR image by using the proposed TLRLCNN (Fig. 9). Then we use the obtained HR image for image caption. After the super-resolution method, the resolution of the image is improved. There is an example of the process of changing the resolution of an image. When scale is 3, the resolution of original image 'butterfly' is 255×255, we use bilinear cubic interpolation to downsample the original image. Therefore, the resolution of image 'butterfly' becomes 85×85. After upsampling, the resolution of the image becomes 255×255. In Figs. 7, and 8, the puppy on the ship or the airplane above the sea are easier to identify, improved resolution makes it easier to identify small objects, which is of great benefit to the image captioning.

In image caption, we train image captioning models with MSCOCO dataset. There are 82873 images in the training data. We use ResNet [6] to extract the feature of images. Deep Residual Network (ResNet) is a stack of Residual Units. ResNet mainly solves the problem of degraded network effects when the network is too deep. The solution is to define a local structure called bottleneck, which passes the input directly to the output as part

Table 4 The PSNR of "butterfly" and "bird" (Set5)with scale factor × 4

Scale×4	Bicubic interpolation	SRCNN	RLCNN	RLCNN+	TLRLCNN
butterfly	20.12dB	21.12dB	21.18dB	20.21dB	21.37dB
bird	27.83dB	28.21dB	28.00dB	28.02dB	29.42dB

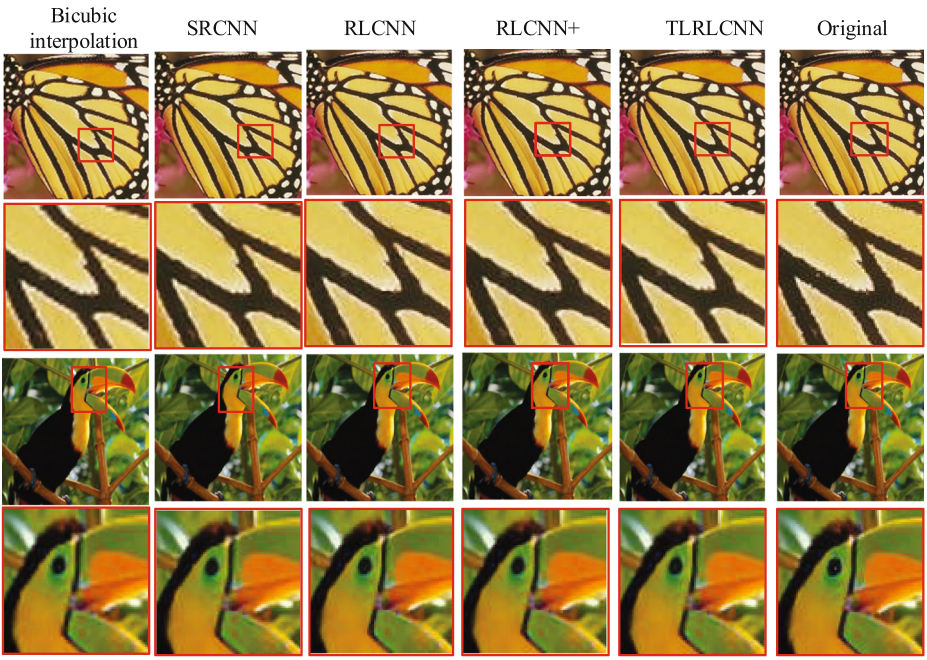

Fig. 4 Super-resolution results of "butterfly" and "bird" (Set5) with scale factor × 2

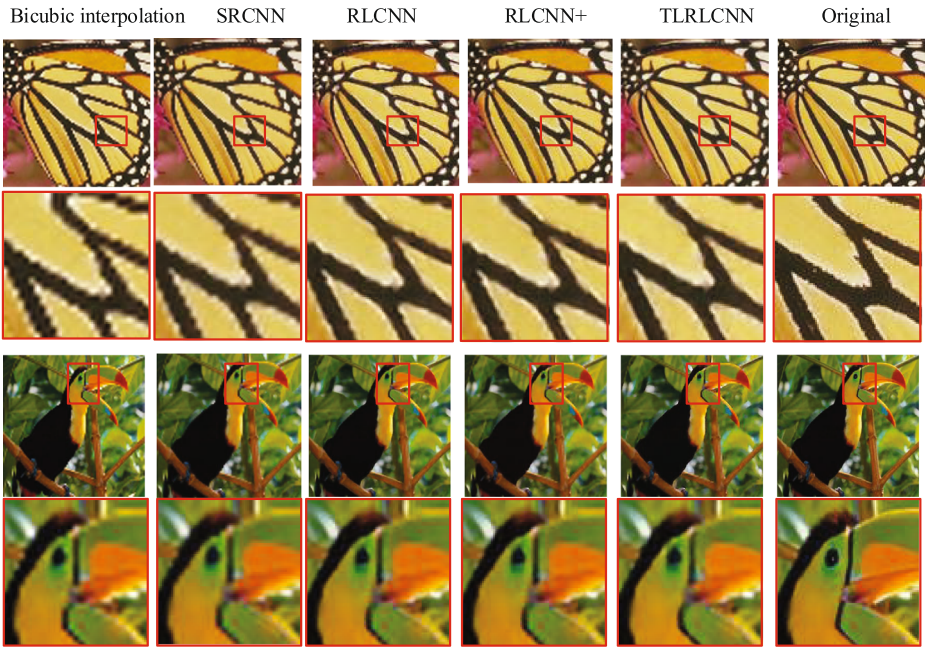

Fig. 5 Super-resolution results of "butterfly" and "bird" (Set5) with scale factor × 3

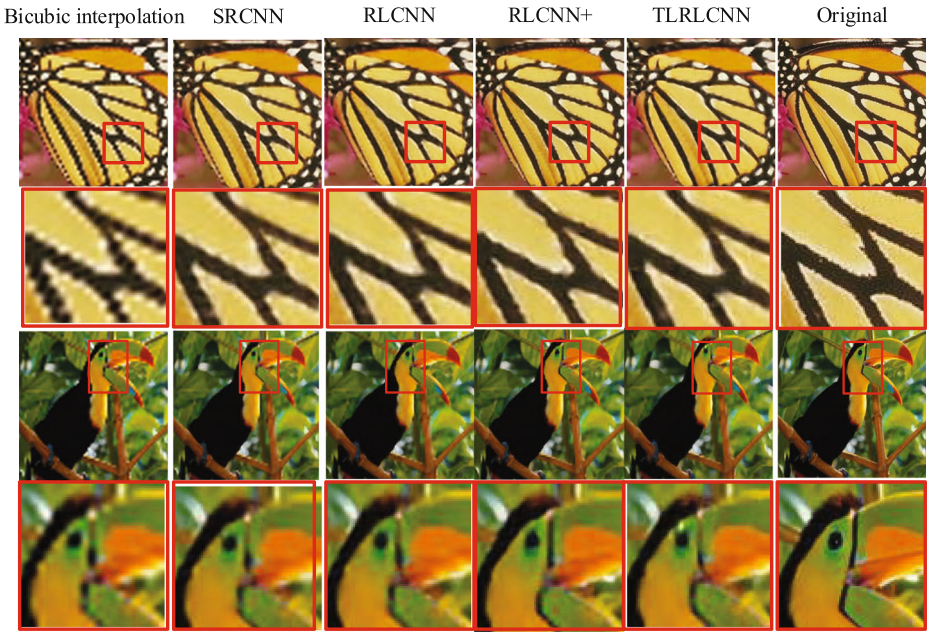

Fig. 6 Super-resolution results of bird (Set5) with scale factor ×4

of the network output, thus transforming the original network fit map $F(x)$ into a fitting $F(x) - x$. Each image is represented by a 1536-dimensional vector. When we train on MSCOCO dataset, we use the following settings. All the LSTM units are set to be 1,024. We set the learning rate ρ =0.001, learning rate decay factor β = 0.99, encoder max sequence length is L_1=30, and decode max sentence length is L2=20, empirically. We adopt Tensorflow framework to carry out our experiments. The experiment results are listed in Tables 5 and 6. We performed experiments when the image upscaling factor was 3 or 4, and the image description evaluation scores are as shown in Tables 5 and 6.

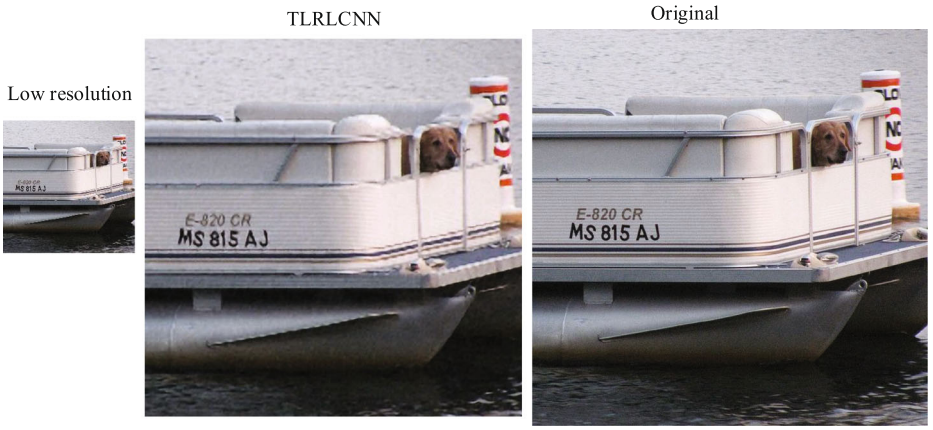

Fig. 7 Super-resolution results of mscoco data

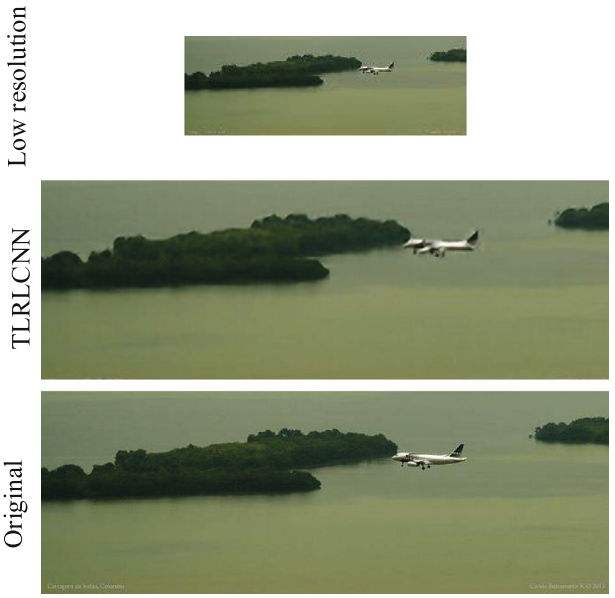

Fig. 8 Super-resolution results of mscoco data

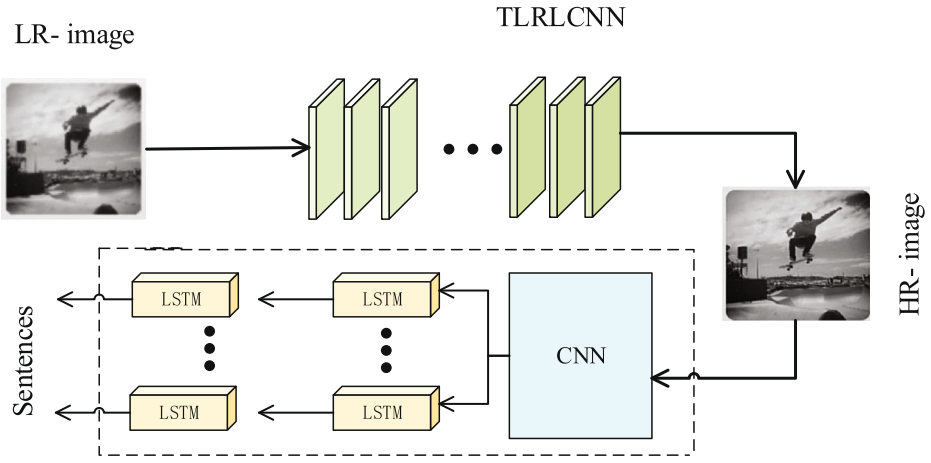

Fig. 9 The framework of image caption based on image super-resolution. The low-resolution image is used as an input to the TLRLCNN to generate a high-resolution image, and the high-resolution image is used as an input to the image captioning model to generate the image description text

Table 5 The evaluation results of the image captioning when scale=3

image	Belu_1	Belu_2	Belu_3	Belu_4	METEOR	ROUGE_L	CIDEr
Scale×3	63.5491	44.5380	30.3013	20.5778	19.8356	46.4477	69.3667
TLRLCNN	64.5950	45.4788	31.01832	21.3820	21.0205	46.6551	69.4927
Original image	66.1544	47.4316	32.4797	22.1064	21.4389	48.9014	77.3624

Table 6 The evaluation results of the image captioning when scale=4

image	Belu_1	Belu_2	Belu_3	METEOR	ROUGE_L
Scale×4	61.3772	41.8184	27.7436	18.9319	45.0850
TLRLCNN	61.6240	42.1719	27.7979	19.0570	45.3490
Original image	66.1544	47.4316	32.4797	21.4389	48.9014

Tables 5 and 6, show quantitative comparisons for×3, ×4 image captioning. It can be seen from the tables that the image captioning produced by our TLRLCNN is more accurate than the low resolution image. These results further indicate that resolution has an inevitable effect on image captioning. Our TLRLCNN can be applied to image captioning for higher image captioning accuracy. Our method has a certain improvement in all evaluation criteria, which proves that the high-resolution images produced by our method have an impact on image captioning.

5 Conclusion

In this paper, we propose a new network structure that stacks two residual learning networks to learn high-frequency residual components that are not recoverable by traditional image super-resolution methods. The proposed structure not only obtains the HR images just like the traditional methods, but also obtains the additional residual components. Experimental results show that the proposed TLRLCNN overperforms its unstacked structure, i.e., RLCNN. Though SRCNN is used as the basic structure in the proposed model, the proposed model does not depend on a fixed network structure. Given the experimental results, we can see that it is promising to construct such CNN-based model for image super-resolution to improve the quality of the obtained HR image via learning more residual information.

Acknowledgments This work is supported by Natural Science Foundation for Distinguished Young Scholars of Shandong Province (JQ201718), Key Research and Development Foundation of Shandong Province (2016GGX101009), the Natural Science Foundation of China (U1736122) and Shandong Provincial. Key Research and Development Plan (2017CXGC1504). And we gratefully acknowledge the support of NVIDIA Corporation with the donation of the TITAN X GPU used for this research. The contact author is Jiande Sun (jiandesun@hotmail.com).

References

1. Bevilacqua M, Roumy A, Guillemot C, Alberi-Morel ML (2012) Low-complexity single-image super-resolution based on nonnegative neighbor embedding
2. Chang X, Yu Y-L, Yang Y, Xing EP (2017) Semantic pooling for complex event analysis in untrimmed videos. IEEE Trans Pattern Anal Machine Intell 39(8):1617–1632
3. Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua T-S (2017) Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning[C]. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 6298–6306.
4. Dong C, Loy CC, He K, Tang X (2014) Learning a deep convolutional network for image super-resolution. In: European conference on computer vision. Springer, pp 184–199

5. Fang H, Gupta S, Iandola F, Srivastava RK, Deng L, Dollár P, Gao J, He X, Mitchell M, Platt JC et al (2015) From captions to visual concepts and back. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1473–1482
6. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
7. Irani M, Peleg S (1991) Improving resolution by image registration. CVGIP: Graph Models Image Process 53(3):231–239
8. Jia X, Gavves E, Fernando B, Tuytelaars T (2016) Guiding long-short term memory for image caption generation[C]. In: IEEE international conference on computer vision. IEEE, pp. 2407–2415
9. Kim J, Kwon Lee J, Lee K (2016) Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1646–1654
10. Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S, Choi Y, Berg AC, Berg TL (2013) Babytalk: Understanding and generating simple image descriptions. IEEE Trans Pattern Anal Machine Intell 35(12): 2891–2903
11. Lu J, Xiong C, Parikh D, Socher R (2017) Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol 6, p 2
12. Ma Z, Chang X, Yang Y, Sebe N, Hauptmann AG (2017) The many shades of negativity. IEEE Trans Multimedia 19(7):1558–1568
13. Mao X, Shen C, Yang Y-B (2016) Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In: Advances in neural information processing systems, pp 2802–2810
14. Rousseau F (2010) A non-local approach for image super-resolution using intermodality priors. Med Image Anal 14(4):594–605
15. Shi W, Caballero J, Ledig C, Zhuang X, Bai W, Bhatia K, de Marvao AMSM, Dawes T, ORegan D, Rueckert D (2013) Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 9–16
16. Sun J, Xu Z, Shum H-Y (2008) Image super-resolution using gradient profile prior. In: IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008. IEEE, pp 1–8
17. Tai Y-W, Liu S, Brown MS, Lin S (2010) Super resolution using edge prior and single image detail synthesis. In: 2010 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2400–2407
18. Thornton MW, Atkinson PM, Holland D (2006) Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. Int J Remote Sens 27(3):473–491
19. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3156–3164
20. Wang S, Li X, Yao L, Sheng QZ, Long G et al (2017) Learning multiple diagnosis codes for icu patients with local disease correlation mining. ACM Trans Knowl Discovery Data (TKDD) 11(3):31
21. Wang Z, Liu D, Yang J, Han W, Huang T (2015) Deep networks for image super-resolution with sparse prior. In: Proceedings of the IEEE international conference on computer vision, pp 370–378
22. Wu Q, Shen C, Liu L, Dick A, van den Hengel A (2016) What value do explicit high level concepts have in vision to language problems? In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 203–212
23. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp 2048–2057
24. Yang J, Lin Z, Cohen S (2013) Fast image super-resolution based on in-place example regression. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1059–1066
25. Yao T, Pan Y, Li Y, Qiu Z, Mei T (2017) Boosting image captioning with attributes. In: IEEE International conference on computer vision, ICCV, pp 22–29
26. Zhou F, Yang W, Liao Q (2012) Single image super-resolution using incoherent sub-dictionaries learning. IEEE Trans Consumer Electron, 58(3)
27. Zhou L, Xu C, Koch P, Corso JJ (2017) Watch what you just said: Image captioning with text-conditional attention[C]. In: Proceedings of the on thematic workshops of ACM multimedia 2017. ACM, pp. 305–313
28. Zou WW, Yuen PC (2012) Very low resolution face recognition problem. IEEE Trans Image Process 21(1):327–340

Min Gao received the bachelor degrees in: communication engineering from the ShanDong Normal University of School of Information Science and Engineering, JiNan, in 2016. She is currently working toward the master degree in communication and information systems at the Shandong Normal University. Her research interests include computer vision, machine learning, and signal processing. She is a student member of the CCF.

Xian-Hua Han received a B.E. degree from ChongQing University, ChongQing, China, a M.E. degree from ShanDONG University, JiNan, China, a D.E. degree in 2005, from the University of Ryukyus, Okinawa, Japan. From Apr. 2007 to Mar. 2013, she was a post-doctoral fellow and an associate professor with the College of Information Science and Engineering, Ritsumeikan University, Japan. From Apr. 2016 to Feb. 2017, she was a senior researcher at the Artificial Intelligence Researcher Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan. She is now an associate Professor at Graduate School of Science and Technology for Innovation, Yamaguchi University, Japan. Her current research interests include image processing and analysis, feature extraction, machine learning, computer vision, and pattern recognition. She is a member of the IEEE, IEICE.

Jing Li is a Ph.D. student in Shandong Normal University, Jinan, China. She is also a lecturer with School of Mechanical and Electrical Engineering, Shandong Management University. Her research interests include machine learning, multimedia processing and retrieval, etc.

Hui Ji received the B.S degree in communication engineering from Shandong University, China, in 2007. He received the Ph.D. degree in Electronics and Telecommunication from INSA-Rennes, France, in 2015. Since Dec.2015 he has been a lecture of school if Information Science and Engineering (ISE), Shandong Normal University. His current research interests include wireless communications, Multiple- Input-Multiple-Output, image processing and Big data analysis.

Huaxiang Zhang is currently a professor with the School of Information Science and Engineering & the Institute of Data Science and Technology, Shandong Normal University, China. He received his Ph.D. from Shanghai Jiaotong University in 2004, and worked as an associated professor with the Department of Computer Science, Shandong Normal University from 2004 to 2005. He has authored over 160 journal and conference papers and has been granted 9 invention patents. His current research interests include machine learning, pattern recognition, evolutionary computation, cross-media retrieval, web information processing, etc.

Jiande Sun received the Ph.D. degree in communication and information system from Shandong University, Jinan, China, in 2000 and 2005, respectively. From September 2008 to August 2009, he was a Visiting Researcher with the Institute of Telecommunications System, Technical University of Berlin, Berlin, Germany. From October 2010 to December 2012, he was a Post-Doctoral Researcher with the Institute of Digital Media, Peking University, Beijing, China, and with the State Key Laboratory of Digital-Media Technology, Hisense Group, respectively. From July 2014 to August 2015, he was a DAAD Visiting Researcher with Technical University of Berlin and University of Konstanz, Germany. From October 2015 to November 2016, he was a Visiting Researcher with the Language Technology Institute, School of Computer Science, Carnegie Mellon University, USA. He is currently a Professor with the School of Information Science and Engineering, Shandong Normal University. He has published more than 60 journal and conference papers. He is the co-author of two books. His current research interests include multimedia content analysis, video hashing, gaze tracking, image/video watermarking, 2D to 3D conversion, and so on.