Neurocomputing 275 (2018) 267–277

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A two-channel convolutional neural network for image super-resolution

Sumei Li, Ru Fan*, Guoqing Lei, Guanghui Yue, Chunping Hou

School of Electrical and Information Engineering, Tianjin University, Tianjin, China

ARTICLE INFO

Article history:
Received 23 May 2017
Revised 15 August 2017
Accepted 17 August 2017
Available online 13 September 2017

Communicated by Jun Yu

Keywords:
Super-resolution
Convolutional neural networks
Deconvolution
End-to-end training
Multi-scale manner

ABSTRACT

A two-channel convolutional neural network (including one shallow and one deep channel) is proposed for the single image super-resolution (SISR). Most existing methods based on convolution neural networks (CNNs) for super resolution have a shallow channel which easily loses the detailed information. And these methods need preprocessing such as bicubic interpolation enlarging LR images to the size of HR images, which may introduce new noises. Meanwhile, most of them use only one fixed filter during the reconstruction. The proposed algorithm solves the above problems, which is named shallow and deep convolutional networks for image super-resolution (SDSR). First, the proposed method uses two channels: shallow and deep channel. The shallow channel mainly restores the general outline of the image. On the contrast, the deep channel extracts the detailed texture information. Second, the proposed method directly learns an end-to-end mapping between low-resolution (LR) and high-resolution (HR) images, which does not need hand-designed preprocessing. The upsampling of the network by deconvolution is embedded in the two channels, which leads to much more efficient and effective training, reducing the computational complexity of the overall SR operation. Finally, during the last period of reconstruction, the deep channel adopts multi-scale manner, which can extract both the short- and long-scale texture information simultaneously. Our model is evaluated on the different public datasets including images and videos. Experimental results demonstrate that the proposed method outperforms the existing methods in accuracy and visual impression.

1. Introduction

As a ill-posed nature of the underdetermined problem, single image super-resolution (SISR) aims at restoring the high resolution (HR) image with abundant highfrequency details from the low resolution (LR) observation. Deep learning is a new field in the study of machine learning, the motive is to establish neural networks to imitate the human brain mechanism to explain the data, such as images, sound and text. Convolutional neural network (CNN) is a kind of machine learning model of deep learning under the supervision. Recently, CNN has shown excellent performance in various computer vision tasks, such as image classification, object detection, semantic segmentation, and action recognition [1–6]. Recently, CNN also shows excellent performance in the single image super-resolution (SISR) [7–11]. SR is inherently illposed with insufficient knowledge. The ill-posed nature is particularly pronounced for high upscaling factors, for which texture detail in the reconstructed SR images is typically absent. But it is required in some areas to get high resolution images. For example, we can use deep learning technologies to promote the perception and positioning function of the unmanned system.

Traditional SISR methods are based on interpolation, such as bicubic interpolation and Lanczos resampling [12]. Then algorithms based on reconstruction constraint are widely studied, including the iterative back projection (IBP) [13], maximum a posteriori probability (MAP) [14], projections onto convex sets (POCS) [15], which can merge more prior knowledge and be used in varieties motion models. These spatial domain SR methods are time-consuming and need large amount of iterative calculations.

Lately, learning-based methods have been extensively used to model a mapping from LR to HR patches. Neighbor embedding and locally linear embedding (NE+LLE) [16] method interpolates the patch subspace. Sparse coding (SC) [17] method uses a learned compact dictionary based on sparse signal representation. Sparse coding based network (SCN) [18] achieves notable improvement over the generic SC model. The cascade of SCNs (CSCN) [19] also benefits from the end-to-end training of deep network with a specially designed multi-scale cost function. But most of them rely

* Corresponding author .
E-mail addresses: fanru@tju.edu.cn, 503199611@qq.com (R. Fan).

http://dx.doi.org/10.1016/j.neucom.2017.08.041
0925-2312/

on hand-designed features to characterize LR images. The speed of restoring images is slow. Therefore, most of them are with high computational complexity and cannot achieve an end-to-end direct amplification. Recently, a new trend of combining neural network with traditional algorithm is on the rise. Extreme learning machine autoencoder (ELM-AE) [20] adds to a new deep neural network. Autoencoders imbedded into network [21,22] are also used in human pose recovery. [23] presents an AdaBoost-based learning method to learn a non-linear feed-forward artificial neural network with a single hidden layer and a single output neuron.

CNN-based method, as a biologically inspired learning model, has provided a new inspiration and direction for SR problem. Super-Resolution Convolutional Neural Network (SRCNN) [7] proposed by Dong et al. drew considerable attention due to its simple network structure and excellent restoration quality. The authors shortly accelerated the algorithm by reducing network parameters, proposing Fast Super-Resolution Convolutional Neural Networks (FSRCNN) [9]. However, there are still some drawbacks. First, as a pre-processing step, the original LR image is upsampled to the desired size by bicubic interpolation to form the input of the network. Second, Reconstruction of the detailed information is still unsatisfactory. So a two-channel method is proposed to solve the above problems. The shallow channel mainly restores the general outline of the image. On the contrast, the deep channel extracts the detailed texture information. An end-to-end model is established by embedding the upsampling into the two channels, which avoids introducing new errors. The upsampling is carried out by deconvolution. For getting more detailed information, the multi-scale manner is adopted to restore the different scale (including long- and short-scale)texture information in the deep channel. At last, the article combines the deep and shallow channel to get the final HR image. The proposed model is evaluate in the different databases, which shows good robustness, whether on images or highway videos reconstruction, laying a foundation to the unmanned systems.

2. Related work

There have been numerous publications over the last five years using deep learning on the SR areas. Compared with traditional SR methods, which depend on handcrafted features, deep learning may further improve the performance. CNN-based methods, as a biologically inspired learning model, have provided a new inspiration and direction for SR problem.

As in most existing SR methods [24,25]via deep learning, SRCNN is a shallow network, having only three convolution layers. And each layer has fixed filters. It can be seen in Fig. 1(a). Dong et al. attempted to prepare deeper models, but the deeper model performed worse than the shallow one. Finally, they conclude that deeper structure does not always lead to better results. Shallow CNNs [26],[27] is widely used. However, as the deepening of the network, it can fit more complex mapping between LR and HR images. Therefore, a deep network is worth trying.

Before inputing the SRCNN network, images need preprocessing: bicubic interpolation. The interpolation will inevitably introduce new errors, leading to the adverse effects on the feature extraction of CNNs. As a result, they are suboptimal and may even hurt the key information in the original LR images that is crucial for SR. Therefore, upsampling embedded into network is quite necessary.

Motivated by the above observations, we propose to train an combined network, synthesizing the advantages and disadvantages of the shallow and deep networks. As Fig. 1(b) shows. The proposed SDSR combines shallow and deep channels simultaneously, with 3 and 19 layers respectively. Shallow network can restore the rough image. With the deepening of the network, the deep network can restore more details information. Reconstruction performs better. As Fig. 2 shows, the deep network can restore more details, offsetting the defects of the shallow network. Meanwhile, SDSR shows a completely end-to-end CNN structure. Upsampling is embedded into the network by deconvolution layers. Due to the ill-posed nature of image super-resolution restoration, reconstructing a pixel may depend on either short- or long-scale texture information. Therefore, when the size of convolution kernels is same, the network can only extract same range texture information. Inspired by this, we adopt multi-scale manner [28] by using different size of convolution kernels.

3. Proposed method

In order to get accurate and efficient images, a two-channel method is proposed, which mainly contains very deep channel and a shallow channel. The deep channel consists of 19 layers, and the shallow channel includes 3 layers. Fig. 3 shows the architecture of the proposed network. The deep channel conceptually consists of four steps: feature extraction (3 layers), mapping (5 layers), upsampling (3 layers) and multi-scale reconstruction (8 layers). The deep layers with deep hierarchy constantly extract and iterate the underlying characteristics. Therefore, it can accurately restore detailed information, such as margin. The shallow channel is the simply upsampling, which can reserve original images' rough information. In the following, we detail how our network works.

3.1. Feature extraction

The step of feature extraction includes 3 convolution layers, which directly extracts the features from the original LR image Y. Here we set 64 filters with the size of 3×3 for the 3 convolution layers. Convolution layers can be represented as:

$$G_l(Y) = W_l * F_{l-1}(Y), \tag{1}$$

where l represents the lth convolution layer, W_l denotes the filters of the lth layer. G_l is the output feature maps and "$*$" represents the convolution operation. The W_l supports $n_l \times f_l \times f_l$ parameters, where f_l is the kernel size of a filter, n_l is the number of filters. Note that there is no pooling or full-connected layers in SDSR, just convolution and deconvolution layers.

The end of each convolution layer is the activation function. SRCNN adopts Rectified Linear Unit (ReLU, $max(0, x)$) [29] as the activation function. Relu gradient is a constant in most case. Therefore, it can avoid the gradient vanishing problem to some degree. Parametric Rectified Linear Unit (PReLU) [30] not only has the above advantages, but also can make the convergence speed faster. They are different on the coefficient of the negative part. PReLU can be defined as a general activation function:

$$PReLU(x_i) = max(x_i, 0) + a_i min(0, x_i), \tag{2}$$

where x_i is the input signal of the activation function on the ith layer, and a_i is the coefficient of the negative part. The parameter a_i is set to be zero for ReLU, but is learnable for PReLU. PReLU is adopted mainly to avoid the "dead features" [31] caused by zero gradients in ReLU.

Finally, the output of the activation function can be described as:

$$F_l(Y) = PReLU(W_l * F_{l-1}(Y) + B_l), \tag{3}$$

F_l is the final output feature maps, B_l is the biases of the lth layer. For the purpose of confronting with degradation, we use a shortcut connection with identity mapping. The residual network [32] converges much faster.

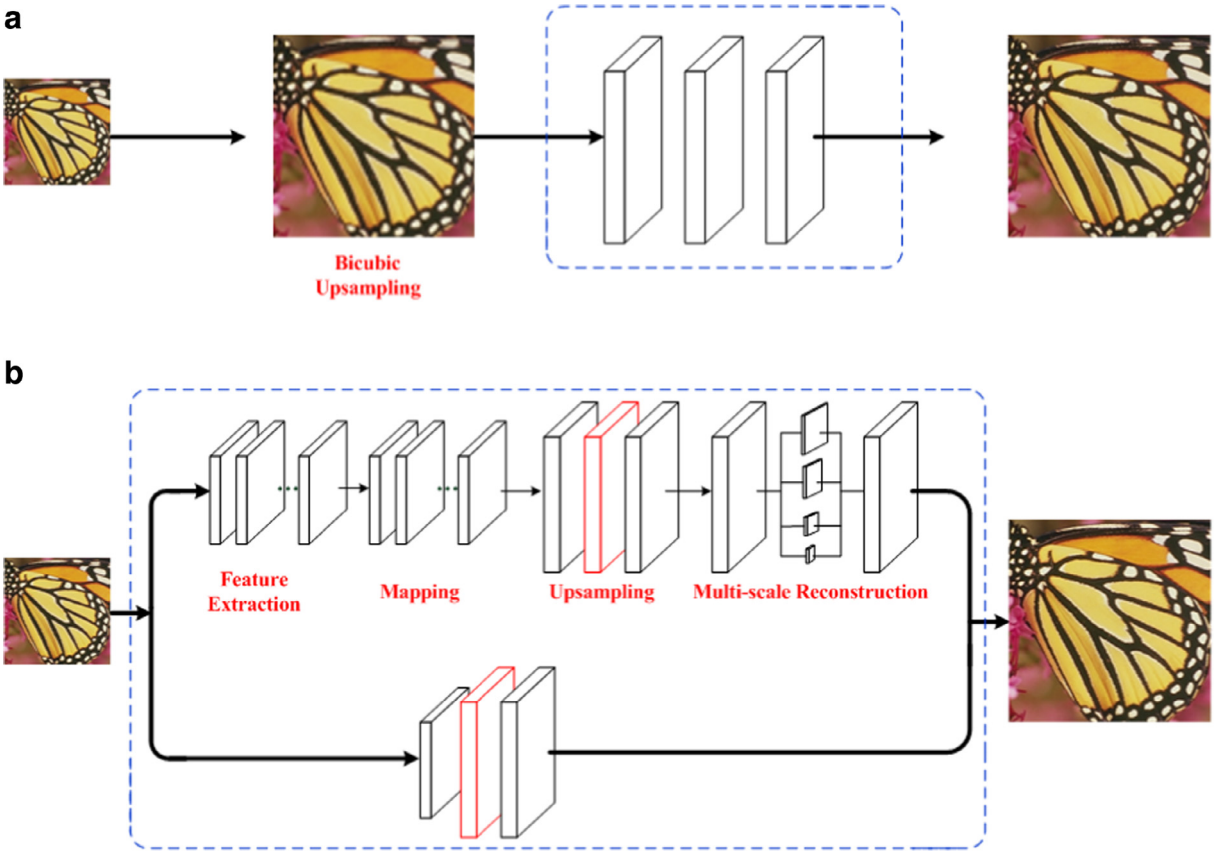


Fig. 1. Overview of SRCNN and our SDSR. (a) Structure of SRCNN: It's a shallow channel. LR images expand to HR images via bicubic interpolating, which is the inputs of the networks. (b) The proposed method SDSR has two channels: shallow and deep channel. Upsampling is embedded into the network, as the red shows. The deep channel integrates all four steps into a completely end-to-end trainable deep model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Fig. 2. Super-resolution results of deep and shallow networks with an upscaling factor 3. Results of the deep network can restore high-frequency details. In contrast, the shallow network can reconstruct the rough information. SDSR combined shallow and deep networks is able to produce perceptually more accurate results.

3.2. Mapping

The step of the mapping consists of 5 convolution layers. The process of convolution is the same as above. First, we map the high-dimensional (64) features into the low-dimensional (12) space using a 1×1 filter. The mapping operation reduces the number of LR feature dimension for the sake of improving computational efficiency. Then we use 4 convolution layers to increase the non-linearity of the model ($12 \times 3 \times 3$ parameters for one layer). Each of the output 12-dimensional vectors is conceptually a representation of a feature map which will be used for upsampling.

3.3. Upsampling

Upsampling operation is an important part at the proposed SDSR system, which aims at increasing the spatial span to the target of HR size. In order to get a good result, we add dimension to 64 with 1×1 filters after the mapping part. Instead of using hand-designed interpolation methods, we adopt the deconvolution layer to achieve upsampling. Fig. 4 shows convolution layers shrink the original images. On the contrary, Fig. 5 reveals the process of a 3×3 patch expanding to 5×5 size.

A convolution is described by k: kernel size, s: stride and p: zero padding and whose input size is i. $i + 2p - k$ is a multiple of s. the convolution has an associated deconvolution described by $i', k' = k, s' = 1$ and $p' = k - p - 1$, where i' is the size of the stretched input obtained by adding $s - 1$ zeros between each input unit. Therefore, the deconvolution layer can achieve upsampling. And its output size is

$$o' = s(i' - 1) + k - 2p. \tag{4}$$

Since we use caffe package [33], when training a set of $f_{sub} \times f_{sub}$-pixel LR sub-images with an upscaling factor n, the

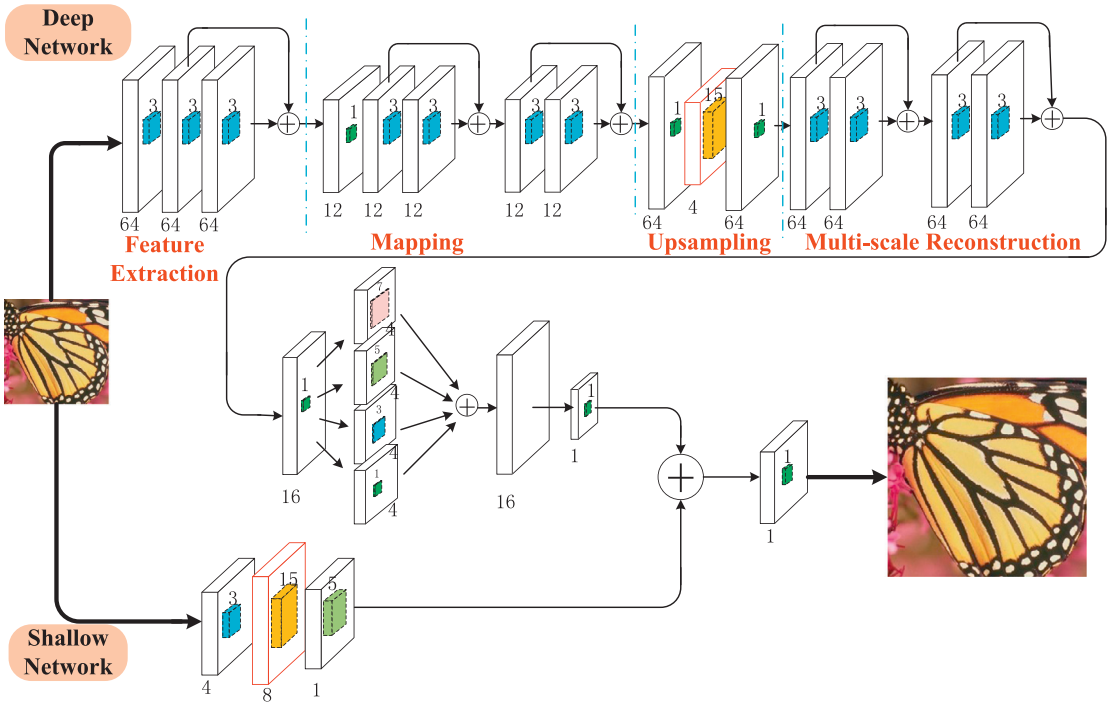


Fig. 3. Our Network Structure: Shallow and Deep Networks for Super Resolution (SDSR). This network mainly contains two channels. The deep channel consists of four steps, feature extraction, mapping, upsampling (the red is the deconvolution lay), and multi-scale reconstruction. The shallow channel is simply upsampling. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

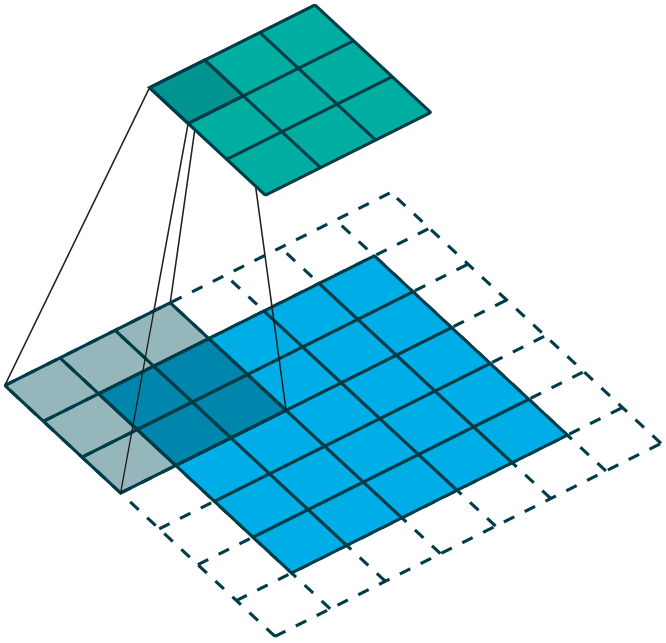


Fig. 4. Convolution process with i=5, k=3, s=2 and p =1.

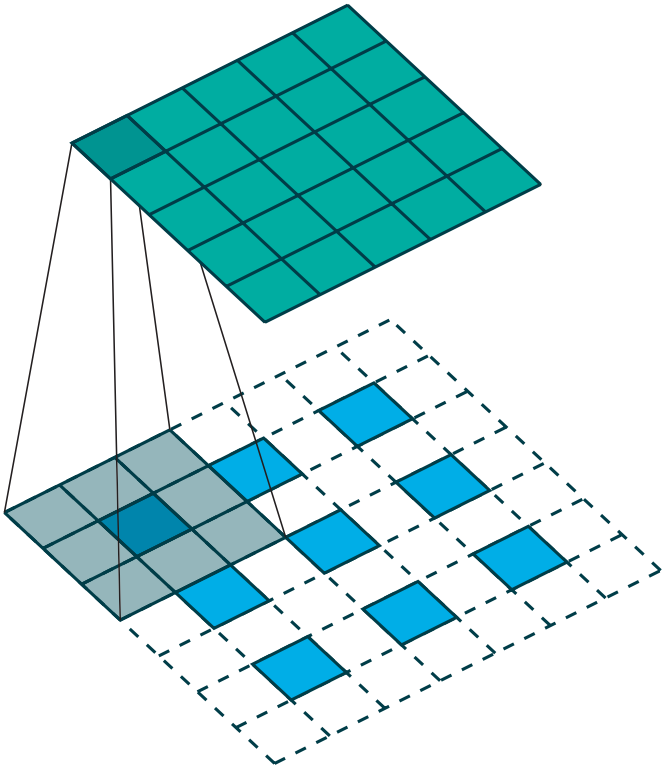


Fig. 5. Deconvolution process with i' =3, k'=3, s'=1 and p' =1.

deconvolution layer can only get a size of $(nf_{sub} - n + 1)^2$ HR images. This method implements the LR images as input of networks.

3.4. Multi-scale reconstruction

Considering that HR image restoration usually relies on both short- and long-scale texture information, we propose to perform HR reconstruction with multi-scale convolutions to extract multi-scale texture information. Multi-scale method has been widely studied in human vision problems, which can aggregate local information effectively.

Multi-scale reconstruction part is made up of 10 trainable layers. The first 4 layers are 3×3 convolution layers with 64 filters to extract high-dimension features, whose function is similar to

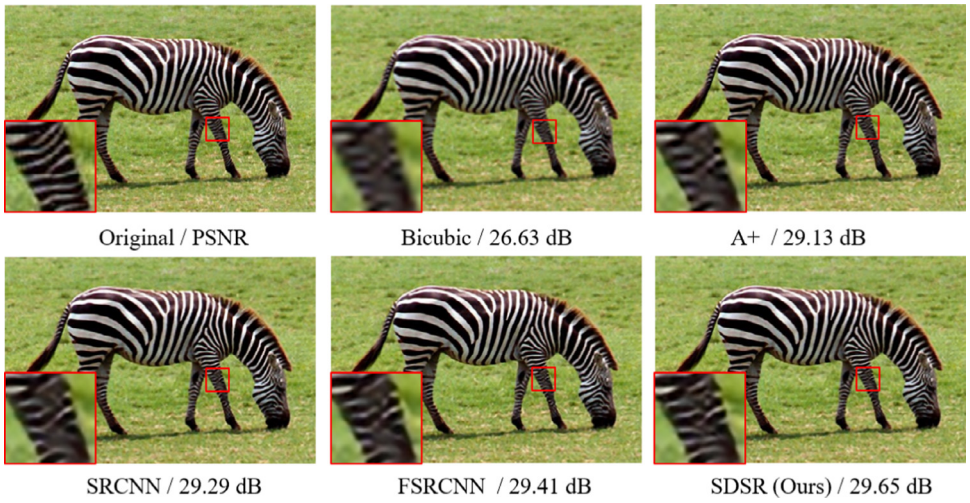


Fig. 6. Super-resolution results of "zebra" (Set14) with an upscaling factor 3. SDSR recovers sharp lines.

the beginning of the feature extraction part. Every two convolution layers form a block, where the input is added to the output of the block through a shortcut connection with identity mapping. Then the 5th layer is the dimension reduction layer with 1×1 kernel size and 16 channels. Multi-scale convolution process is as below.

$$F_{li}(Y) = PReLU(W_{li} * F_{i-1}(Y) + B_l). \quad (5)$$

The subsequent multi-scale convolution layer consists of 4 kinds of filters with 7×7, 5×5, 3×3 and 1×1 kernel sizes, respectively. They are represented as W_{li}, i=1,2,3,4. The 4 filters are parallel. Each of them output 4 feature maps, then they are concatenated into 16 feature maps, such that features in different scales can be extracted. Finally, there is a 1×1 convolution layer, which serves as a weighted combination of multi-scale texture features.

Multi-scale reconstruction can restore short- and long-scale texture information. Compared with single convolution, multi-scale convolution provides a new way of thinking in CNNs.

3.5. Combining shallow and deep networks

The shallow network has only 3 layers, so it just restores rough images, which lacks of high-frequency details. On the contrary, the deep network can more accurately restore the high-frequency content of the HR images. Finally, we concatenate the shallow with the deep network, then input a convolution layer. Therefore, image quality is greatly improved.

During the period of training, we use Mean Squared Error (MSE) as the loss function: $\frac{1}{2}\|y - f(x)\|^2$, where y is the ground truth images, $f(x)$ is the output of our SDSR network.

4. Experiments and results

4.1. Datasets for tTraining and tTesting

Training dataset The 91-image dataset proposed in [17] is widely used as the training set in learning based SR methods [7–9,34]. As big data generally get better results, we also use General-100 dataset [9] that contains 100 bmp-format images with no compression, which are very suitable for the SR training. Therefore, the original data is totally 191 images. To make the dataset more efficient, we augment the original images with two steps. 1) Scaling: each image is scaled by the factor of 0.9, 0.8, 0.7, 0.6. 2) Rotation: each image is rotated by the degree of 90, 180, 270. Thus the final training dataset is 20 times of the original data. That is, the number of training images are total $191 \times 20 = 3820$ images.

Testing dataset For benchmark, the Set5 [35] (5 images), Set14 [36] (14 images) and BSD100 [37] (100 images) are used to evaluate the performance of upscaling factors $\times 2$, $\times 3$ and $\times 4$. PSNR and SSIM metrics are utilized for quantitative evaluation.

For luminance is the most important influence factor for human visual, we only restore the luminance channel in YCrCb space. In order to get a better display, the Cr and Cb channels are upsampled by bicubic interpolation.

Training We use batches of size 64, momentum of 0.9 and learning rate of 0.00001. This net applies stochastic gradient descent. All the filters in convolution layers are randomly initialized from a Gaussian distribution with zero mean and standard deviation $\sqrt{2/n}$ (and 0 for biases). Training is on GPU Titan X Pascal.

4.2. Experimental results

In this section, we compare the performance of our models with the state-of-the-art methods on above image datasets and real scene videos.

4.2.1. Image super-resolution results

Figs. 6–9 show the restored images and particular regions of images are enlarged to see the difference clearly. It is obvious that the proposed method performs better in visual perception. From the quantitative analysis, the Peak Signal to Noise Ratio (PSNR) of our SDSR reachs to 33.97 dB, which is much higher than Bicubic Interpolation, Adjusted Anchored Neighbourhood Regression method (A+) [38], SRCNN [7] and FSRCNN [9]. The results of the compared methods are either obtained using the publicly available codes or provided by the authors. It is obvious that our method performs better in details.In order to create intuitive feelings of the differences among various methods, we draw Fig. 10. The output of our SDSR network has less error pixels, especially around the marginal areas, for the deep network part can extract more detail information. As Fig. 11 shows, the result of a generative adversarial network (GAN) for image super resolution (SRGAN) [39] has more detailed information , even more than the original image. Therefore, it has a lower-level PSNR. But it is more suitable for human perception for the improved loss. SDSR is a little smooth at the texture detail, it has a higher-level PSNR.

Table 1 shows the quantitative comparison of the state-of-the-art methods measured by average PSNR and SSIM. As the result shows, our method outperforms most cases, especially as

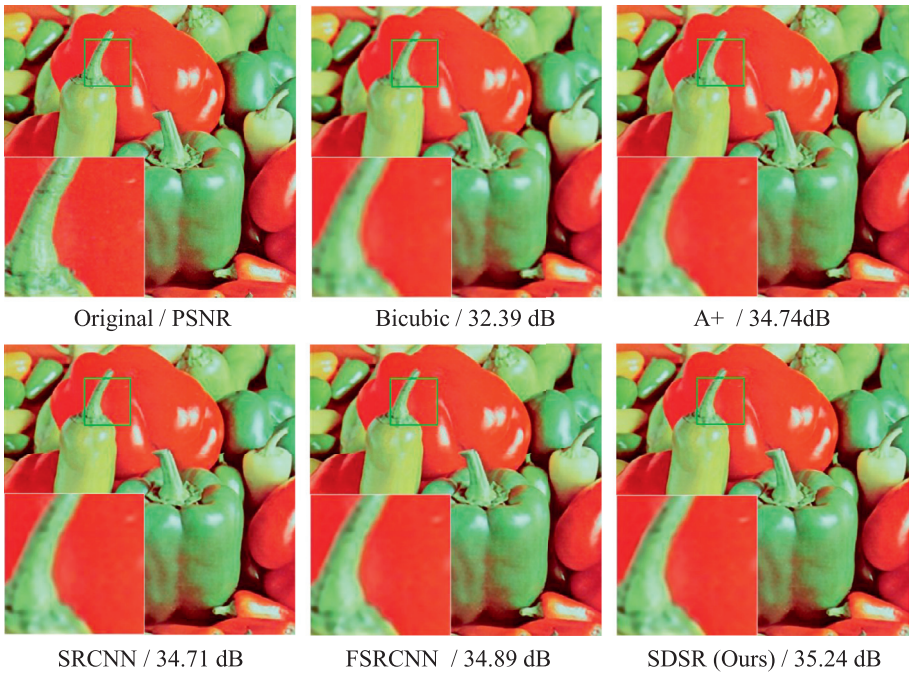


Fig. 7. Super-resolution results of "pepper" (Set14) with an upscaling factor 3.

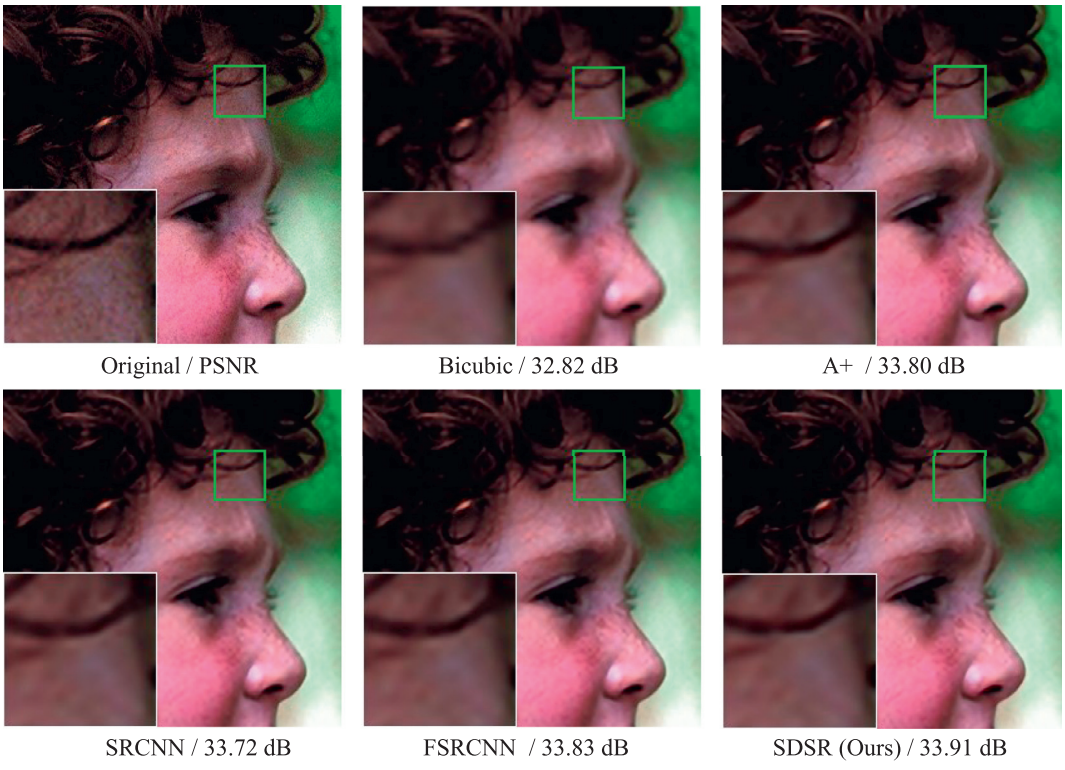


Fig. 8. Super-resolution results of " head_GT " (Set5) with an upscaling factor 3.

the increasing number of images. Therefore, our model has better generalization ability. It can be used in various complicated environments. The intuitive comparison can be seen in Fig. 12. The state-of-the-art SR methods including: SC - sparse coding-based method of Yang et al. [17], NE+LLE - neighbour embedding + locally linear embedding method [16], ANR - Anchored Neighbourhood Regression method [40], A+ - Adjusted Anchored Neighbourhood Regression method [38], and KK - the method described in [41]. It is worth pointing out that SDSR surpasses the existing state-of-the-art methods' lines at the very beginning of the learning stage, which indicates that SDSR is convergent faster.

Computational complexity A low-complexity metric is more favored and welcome in many practical applications. Therefore, we also conduct an experiment to demonstrate the time complexity of the proposed method. For this purpose, we record the run time of the proposed method as well as the completing methods. All

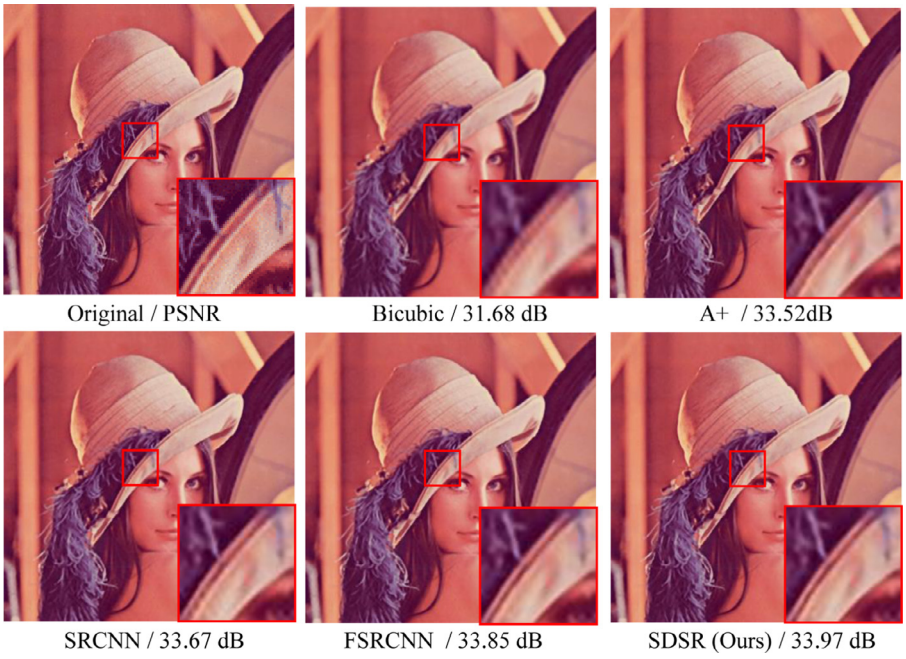


Fig. 9. Super-resolution results of "lenna" (Set14) with an upscaling factor 3. SDSR recovers sharp lines.

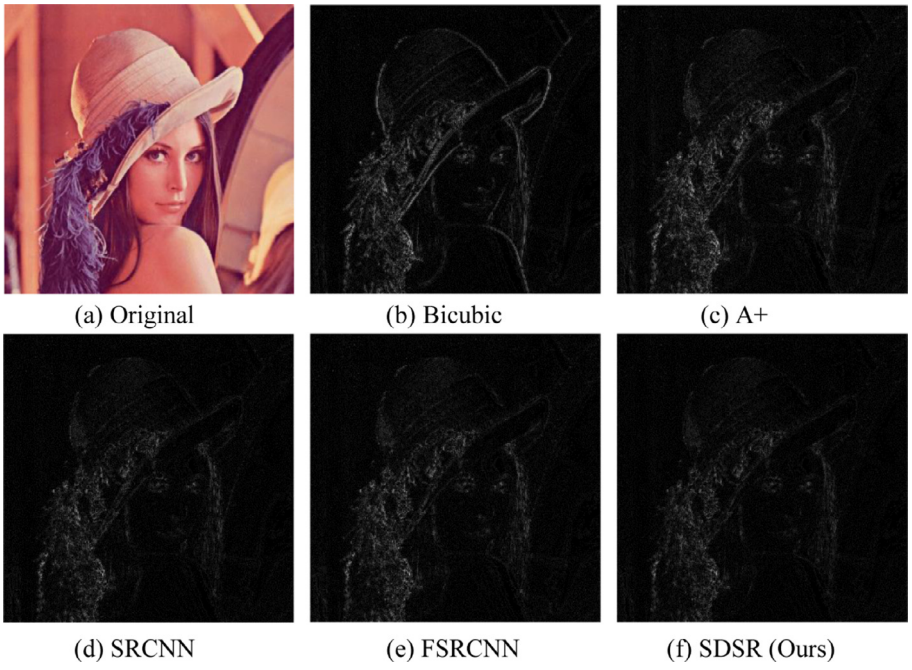


Fig. 10. Error maps for "lenna" (Set14) an upscaling factor 3. Black pixels are good and white pixels are bad pixels. Our SDSR with less bad pixels.

algorithms are implemented in MATLAB 2014b and executed on a 2.6 GHz processor with 8 GB RAM, Windows 7 Pro 64-bit laptop. For each method, all the images in Set5 and Set14 database are evaluated and total time consumed is recorded by the MATLAB functions. Then, the average processing time per image (in second) is given in Table 2. As Table 2 clearly shows, the proposed SDSR has moderate computational complexity. The main costs of the proposed method are spent on image feature extraction. To be specific, the deep channel with large quantities of parameters is the main cost. SRCNN need input a large image at the beginning of the network, which slow down the speed of the network. The inputs of FSRCNN are small images. And it has less layers and parameters which affect the precision of it to a certain extent. Therefore, FSRCNN has a high processing speed. Precision and speed need to be balanced considered at any time according to the application scenario.

4.2.2. Highway real scene videos super-resolution results

As for the practicability of our SDSR, we test it in highway real scene videos[1] with 352 × 288 in size and 5 s in length. Fig. 13 shows the representative 6 frames from the highway video. The original frames are in the left corner. The big high-resolution images are the restored results of SDSR with an upscaling factor of 3. It can still restore abundant details.

[1] Highway of YUV Videos Sequences: http://www.trace.eas.asu.edu/yuv/.

Fig. 11. Super-resolution results of "comic" (Set14) with an upscaling factor 4. SRGAN optimized for a loss more sensitive to human perception. SDSR has a higher-level PSNR.

Table 1
Average PSNR/SSIM for upscaling factors ×2, ×3 and ×4 on datasets Set5, Set14 and BSD100. The best performances are in bold.

Datasets	Set5			Set14			BSD100		
Upscaling	×2	×3	×4	×2	×3	×4	×2	×3	×4
Bicubic	33.66 (0.9299)	30.39 (0.8682)	28.42 (0.8104)	30.24 (0.8687)	27.55 (0.7736)	26.00 (0.7019)	29.56 (0.8431)	27.21 (0.7385)	25.96 (0.6675)
A+ [38]	36.54 (0.9544)	32.58 (0.9088)	30.28 (0.8603)	32.28 (0.9056)	29.13 (0.8188)	27.32 (0.7491)	31.21 (0.8863)	28.29 (0.7835)	26.82 (0.7087)
SRCNN [7]	36.66 (0.9542)	32.75 (0.9090)	30.49 (0.8628)	32.45 (0.9067)	29.30 (0.8215)	27.50 (0.7513)	31.36 (0.8879)	28.41 (0.7863)	26.90 (0.7103)
FSRCNN [9]	37.00 (0.9558)	33.16 (0.9140)	30.71 (0.8657)	32.63 (0.9088)	29.43 (0.8242)	27.59 (0.7535)	31.50 (0.8906)	28.52 (0.7893)	26.96 (0.7128)
CSCN [19]	37.00 (0.9557)	33.18 (0.9153)	30.94 (0.8755)	**32.65 (0.9081)**	29.41 (0.8234)	27.71 (0.7592)	31.46 (0.8891)	28.41 (0.7863)	26.90 (0.7167)
SRGAN [39]	–	–	29.40 (0.8472)	–	–	26.02 (0.7397)	–	–	25.16 (0.6688)
SDSR(Ours)	**37.07 (0.9564)**	**33.42 (0.9181)**	**31.01 (0.8744)**	32.64 (0.9093)	**29.47 (0.8288)**	**27.73 (0.7614)**	**31.52 (0.8911)**	**28.65 (0.7933)**	**27.10 (0.7186)**

Table 2
Time complexities of SR methods on Set5 and Set14 database (in seconds).

Datasets	Bicubic	SRCNN	FSRCNN	SDSR
butterfly	0.0082	1.2706	0.0638	0.7832
baby	0.0205	1.4412	0.0769	0.2028
bird	0.0106	1.3947	0.0676	0.1745
head	0.0117	1.3606	0.0592	0.1716
woman	0.0100	1.3011	0.0585	0.1656
Set14	0.5056	92.34	1.6883	6.5165

Table 3
Results on HD videos from Ultra Video Group datasets with an upscaling factor of 3.

Datasets	Bicubic	SRCNN [7]	ESPCN [10]	Our SDSR
Bosphorus	39.38	41.07	41.25	**41.96**
ReadySetGo	34.64	37.33	37.37	**38.42**
Beauty	39.77	40.46	40.54	**41.73**
YachtRide	34.51	36.07	36.18	**36.91**
ShakeNDry	38.79	40.26	40.47	**41.08**
HoneyBee	40.97	42.66	42.89	**44.32**
Jockey	41.86	43.62	43.73	**44.81**
Average	38.56	40.21	40.35	**41.32**

In addition, we also use the Ultra Video Group datasets[2], containing 7 videos of 1920 × 1080 in size and 5 s in length to evaluate the proposed method. Table 3 shows the mean PSNR for different models. Best results for each category are shown in bold. There is significant difference between the PSNR of the proposed method and ESPCN [10]. SDSR can dramatically improve the quality of videos, and has strong flexibility both in images and videos. Therefore, SDSR has broad application prospects in unmanned systems.

[2] Ultra Video Group Test Sequences:c http://www.ultravideo.cs.tut.fi/.

4.2.3. Numbers of layers, multi-scale and ways of combine analysis

In order to verify the deeper network can restore more detail information, we short the deep network part of the SDSR model to only 10 layers and 3 layer, which are named small SDSR and smaller SDSR respectively. As Fig. 14 presents that smaller SDSR performs worst, small SDSR is second, our SDSR shows best. Results in Table 4 show that the performance can be further improved by increasing layers of the network. As with the deepening of network, it can fit different complex mappings. Meanwhile, large networks entail more computational overhead, which leads

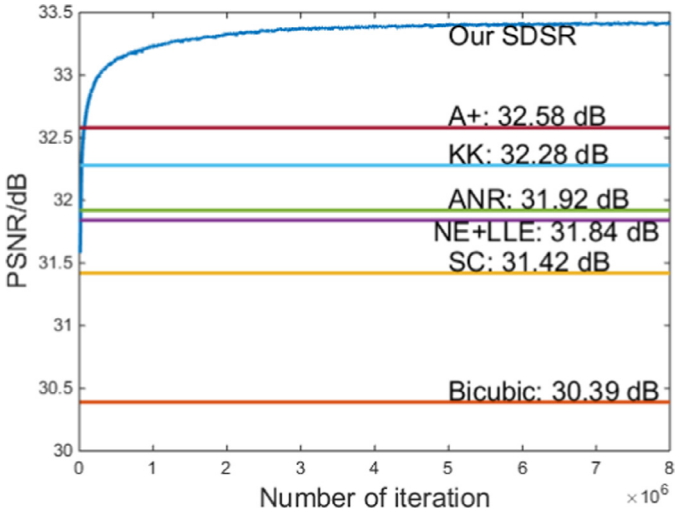


Fig. 12. The test convergence curve of our SDSR and results of other methods on the Set5 dataset.

Fig. 13. The result of highway real scene videos reconstruction.

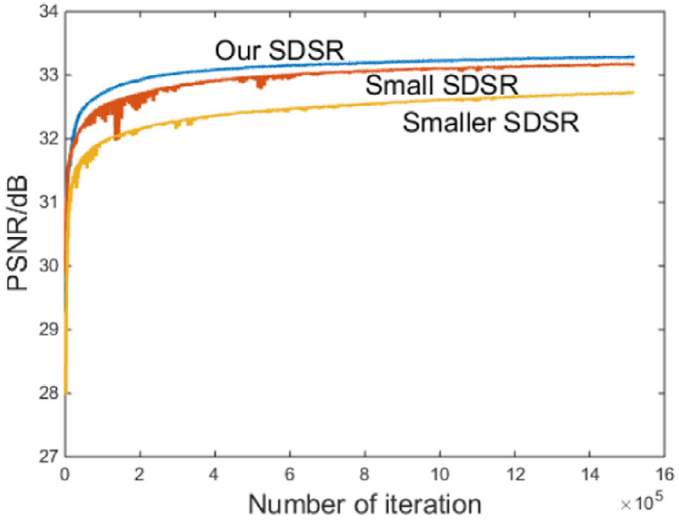


Fig. 14. Small SDSR and SDSR on Set5 with an upscaling factor 3.

Table 4
Average PSNR (dB) and reconstruction time of different models on datasets with an upscaling factor 3.

Model	Smaller SDSR	Small SDSR	SDSR
Set5	32.73	33.17	**33.42**
Set14	29.13	29.34	**29.47**
BSD100	28.38	28.55	**28.65**
BSD100 Times	**9.45 s**	11.60 s	17.71 s

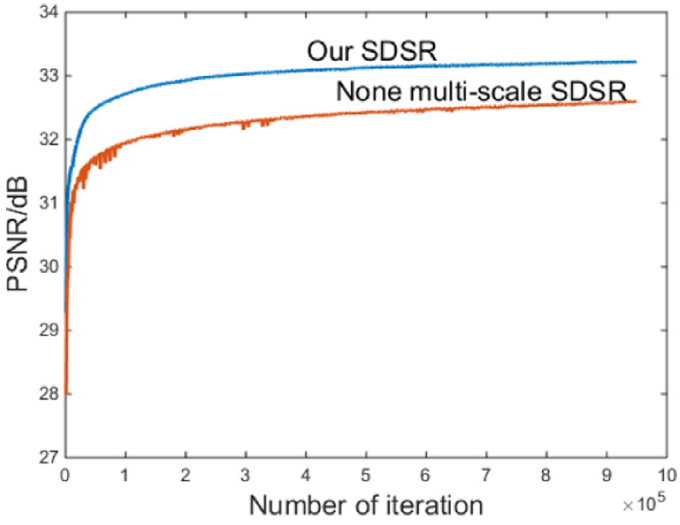


Fig. 15. None multi-scale SDSR and SDSR on Set5 with an upscaling factor 3.

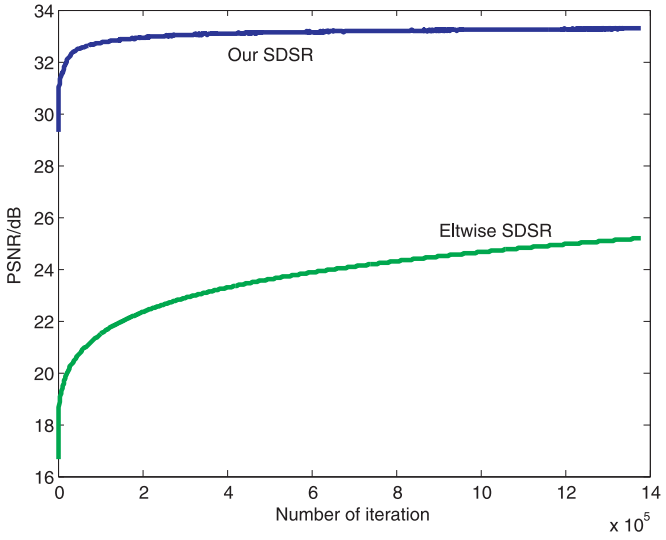


Fig. 16. Different ways of combine shallow with deep channel networks..

to a longer recovery time. In the practical application the pros and cons should be weighted.

To investigate the contribution of the multi-scale module to the final results, we set a model which is none multi-scale in the bottom of the network. As can be seen in Fig. 15, once there is none multi-scale procedure, reconstruction will become quite bad, which indicates that multi-scale plays a important role in the reconstruction. Since the large size convolution kernel has a big receptive filed, on the contrary, small size convolution kernel has a small receptive filed, both of them coexist and complement each other. Therefore, multi-scale method can extract more detailed information.

SDSR combines shallow with deep channel networks using concatenation, which just large the dimension of the feature maps. Then, we try other ways of combination. Concatenation is substituted by eltwise, which add the deep output to the shallow output in pixel. As we can see in Fig. 16, the result of eltwise SDSR is very bad, for the pixel-level sum has many uncertainty factors. Simple sum may import unnecessary information causing network convergence slowly.

5. Conclusion

In this paper, we have presented a super-resolution method using shallow and deep networks, which directly extracts features from the original LR images, and learns to upscale the resolution in the latent feature space. For the reconstruction with multi-scale manner, it can restore more details both on images and videos. As to lay a good foundation for self-driving technology. Experimental results both visually and objectively support that our SDSR method outperforms state-of-the-art SR algorithms.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (NSFC), one is the key international (regional) cooperation research projects -the study of 3D image/video coding, content processing, key techniques of quality evaluation under grant number 61520106002. Another is the general project -based on the motion and depth perception of stereoscopic visual comfort study (61471262). We gratefully acknowledge the support of NSFC.

References

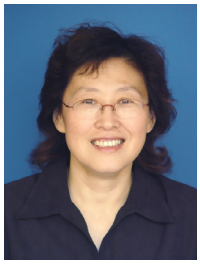
[1] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2) (2012) 1097–1105.
[2] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition,, 2014, pp. 580–587.
[3] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Simultaneous detection and segmentation, in: European Conference on Computer Vision, Springer, 2014, pp. 297–312.
[4] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 221–231.
[5] P. Tang, H. Wang, S. Kwong, G-ms2f: googlenet based multi-stage feature fusion of deep cnn for scene recognition, Neurocomputing 225 (2017) 188–197.
[6] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, F.E. Alsaadi, A survey of deep neural network architectures and their applications, Neurocomputing 234 (2017) 11–26.
[7] C. Dong, C.C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in: European Conference on Computer Vision, Springer, 2014, pp. 184–199.
[8] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2016) 295–307.
[9] C. Dong, C.C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in: European Conference on Computer Vision, Springer, 2016, pp. 391–407.
[10] W. Shi, J. Caballero, F. Huszár, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, Z. Wang, 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, 2016, pp. 1874–1883.
[11] Y. Liang, J. Wang, S. Zhou, Y. Gong, N. Zheng, Incorporating image priors with deep convolutional neural networks for image super-resolution, Neurocomputing 194 (2016) 340–347.
[12] C.E. Duchon, Lanczos filtering in one and two dimensions, J. Appl. Meteorol. 18 (8) (1979) 1016–1022.
[13] M. Irani, S. Peleg, Motion analysis for image enhancement: resolution, occlusion, and transparency, J. Vis. Commun. Image Represent. 4 (4) (1993) 324–335.
[14] R.R. Schultz, R.L. Stevenson, Extraction of high-resolution frames from video sequences, IEEE Trans. image Process. 5 (6) (1996) 996–1011.
[15] H.H. Bauschke, J.M. Borwein, On projection algorithms for solving convex feasibility problems, SIAM Rev. 38 (3) (1996) 367–426.
[16] H. Chang, D.Y. Yeung, Y. Xiong, Super-resolution through neighbor embedding, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., volume 1, IEEE, 2004. I–I
[17] J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (11) (2010) 2861–2873.
[18] Z. Wang, D. Liu, J. Yang, W. Han, T. Huang, Deep networks for image super-resolution with sparse prior, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 370–378.
[19] D. Liu, Z. Wang, B. Wen, J. Yang, W. Han, T.S. Huang, Robust single image super-resolution via deep networks with sparse prior, IEEE Trans. Image Process. 25 (7) (2016) 3194–3207.
[20] K. Sun, J. Zhang, C. Zhang, J. Hu, Generalized extreme learning machine autoencoder and a new deep neural network, Neurocomputing 230 (2017) 374–381.
[21] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, Multimodal deep autoencoder for human pose recovery, IEEE Trans. Image Process. 24 (12) (2015) 5659–5670.
[22] C. Hong, J. Yu, X. Chen, Image-based 3d human pose recovery with locality sensitive sparse retrieval, in: IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2013, IEEE, 2013, pp. 2103–2108.
[23] M.M. Baig, M.M. Awais, E.-S.M. El-Alfy, Adaboost-based artificial neural network learning, Neurocomputing 248 (2017) 120–126.
[24] Y. Romano, J. Isidoro, P. Milanfar, RAISR: Rapid and accurate image super resolution, IEEE Trans. Comput. Imaging 3 (1) (2017) 110–125.
[25] G. Duan, W. Hu, J. Wang, Research on the natural image super-resolution reconstruction algorithm based on compressive perception theory and deep learning model, Neurocomputing 208 (2016) 117–126.
[26] C. Dong, Y. Deng, C. Change Loy, X. Tang, Compression artifacts reduction by a deep convolutional network, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 576–584.
[27] K. Yu, C. Dong, C.C. Loy, X. Tang, Deep convolution networks for compression artifacts reduction, arXiv preprint arXiv:1608.02778 (2016).
[28] D. Eigen, C. Puhrsch, R. Fergus, Depth map prediction from a single image using a multi-scale deep network, in: Advances in Neural Information Processing Systems, 2014, pp. 2366–2374.
[29] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 807–814.
[30] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.
[31] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.
[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[33] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, 2014, pp. 675–678.
[34] J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1646–1654.
[35] M. Bevilacqua, A. Roumy, C. Guillemot, M.L. Alberi-Morel, Low-complexity single-image super-resolution based on nonnegative neighbor embedding (2012).
[36] R. Zeyde, M. Elad, M. Protter, On single image scale-up using sparse-representations, in: International Conference on Curves and Surfaces, Springer, 2010, pp. 711–730.
[37] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Proceedings of the Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001., volume 2, IEEE, 2001, pp. 416–423.
[38] R. Timofte, V. De Smet, L. Van Gool, A+: adjusted anchored neighborhood regression for fast super-resolution, in: Asian Conference on Computer Vision, Springer, 2014, pp. 111–126.
[39] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, arXiv preprint arXiv:1609.04802 (2016).
[40] R. Timofte, V. De Smet, L. Van Gool, Anchored neighborhood regression for fast example-based super-resolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1920–1927.
[41] K.I. Kim, Y. Kwon, Single-image super-resolution using sparse regression and natural image prior, IEEE Trans. Pattern Anal. Mach. Intell. 32 (6) (2010) 1127–1133.

Sumei Li received her Ph. D. degree from the Nankai University, Tianjin, China, in 2004. She jointed Tianjin University, China, in 2006, where she is currently an associate professor in School of electrical automation and information engineering. Her research interests are the area of (3D) digital image processing, visual quality evaluation, pattern recognition and neural network.

Ru Fan received the B.S. degree in communication engineering from Hebei University, Baoding, China, in 2015, and she is currently working towards the master's degree at the school of Electrical and Information Engineering, Tianjin University, Tianjin, China. His research interests include image processing, stereo matching and deep learning

Guoqing Lei graduated from Changchun University of Science and Technology in 2016, now studying at the University of Tianjin for a master's degree. Her research interests are the area of image reconstruction and neural network.

Guanghui Yue received the B.S. degree in communication engineering from Tianjin University, Tianjin, China, in 2014, and he is currently working towards the Ph.D. degree at the school of Electrical and Information Engineering, Tianjin University, Tianjin, China. His research interests include bioelectrical signal processing, image quality assessment and 3-D image visual discomfort prediction.



Chunping Hou received the M.Eng. and Ph.D. degrees, both in electronic engineering, from Tianjin University, Tianjin, China, in 1986 and 1998, respectively. Since 1986, she has been the faculty of the School of Electronic and Information Engineering, Tianjin University, where she is currently a Full Professor and the Director of the Broadband Wireless Communications and 3D Imaging Institute. Her current research interests include 3D image processing, 3D display, wireless communication, and the design and applications of communication systems.