SHALLOW AND DEEP CONVOLUTIONAL NETWORKS FOR IMAGE SUPER-RESOLUTION

Ru Fan, Sumei Li, Guoqing Lei, Guanghui Yue

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

ABSTRACT

A shallow and deep convolutional neural network is presented for the single-image super-resolution (SISR). The proposed method doesn't need hand-designed procedures, directly learning an end-to-end mapping between low-resolution (LR) and high-resolution (HR) images. The upsampling of the network by deconvolution leads to much more efficient and effective training, reducing the computational complexity of the overall SR operation. However, most existing methods based on CNNs for super resolution need preprocessing like bicubic interpolating LR images to the size of HR images. This method can restore more details by multi-scale manner, and has strong adaptability whether on images or videos. Our model is evaluated on different datasets, outperforming the existing methods in accuracy and visual impression.

Index Terms— Super-resolution, convolutional neural networks, deconvolution, end-to-end training, multi-scale manner

1. INTRODUCTION

Convolutional neural networks (CNN) have shown excellent performance in various computer vision tasks, such as image classification, object detection, semantic segmentation, and action recognition [1] [2]. Unmanned systems need to get road videos, but the quality of the images is usually poor. At the moment, SISR will play an important role.

Early SISR methods are based on interpolation, such as bicubic interpolation and Lanczos resampling [3]. Then algorithms based on reconstruction constraint is widely studied, including the iterative back projection (IBP) [4] etc., merging more prior knowledge and be used in varieties motion models.

Lately, learning methods are widely used in SR. Neighbor embedding and locally linear embedding (NE+LLE) [5] method interpolate the patch subspace. Sparse coding (SC) [6] methods use a learned compact dictionary based on sparse signal representation. Sparse coding based network (SCN) achieves notable improvement over the generic SC model.The cascade of SCNs (CSCN) [7] can also benefit from the end-to-end training of deep network with a specially designed multi-scale cost function. Above all, most of them rely on hand-designed features to characterize LR images. Reconstruction technology mainly contains three steps of registration, reconstruction and post-processing. The speed of restoring images is slow. Therefore, most of them are with high computational complexity and cant achieve an end-to-end direct amplification.

More recently, Super-Resolution Convolutional Neural Network (SRCNN) [8] proposed by Dong et al. with input of interpolated LR patches 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 needs to be upsampled to the desired size by bicubic interpolation to form the input of the network. Second, the processing speed on large images is still unsatisfactory. Its no use only getting excellent images, we have to consider the application. Inspired by this, the proposed method not only solve the above questions, but also was evaluated on highway and sequence real scene videos. Both show dramatic results.

In our work, We combine shallow and deep networks to improve the accuracy. Deconvolution layers are used to realize upsample. Therefore, the network can learn an end-to-end mapping between the original LR and HR images with no pre-processing. Finally, the proposed model shows excellent performances in highway videos reconstruction, laying a foundation to the unmanned systems.

2. PROPOSED METHOD

In order to get accurate and efficient images, we use a very deep network which totally contains 20 convolution lays and 2 deconvolution layers. Following, we detail how our network works.

Figure 1 shows the architecture of our proposed network. It mainly contains two sections, deep network and shallow network, so we name the proposed model Shallow and Deep Networks for Super-Resolution (SDSR), Deep network is consisted of four steps, feature extraction, mapping, upsampling, and multi-scale reconstruction. The deep lays can make feature map multiple times. Therefore, it can accurately re-

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

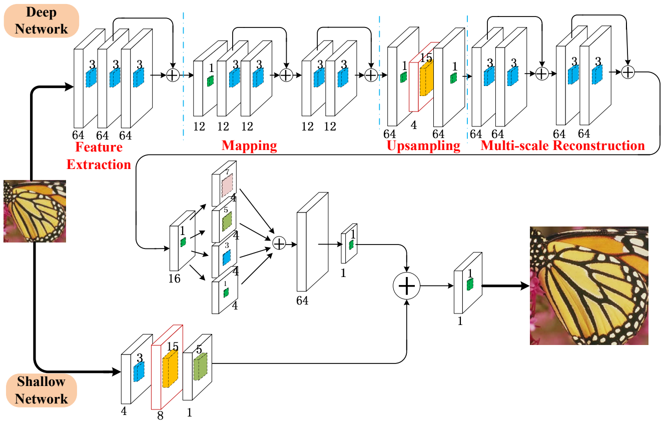

Fig. 1. Our Network Structure: Shallow and Deep Networks for Super Resolution (SDSR). This network mainly contain two sections. First, deep network is consisted of four steps, feature extraction, mapping, upsampling (the red is the deconvolution lay), and multi-scale reconstruction. Second, shallow network is simply upsampling.

store detailed information, like margin. Shallow network is simply upsampling. It can reserve more original images' rough information.

(1) Feature extraction: SDSR extracts features on the original LR image without bicubic interpolation, aiming at learning an end-to-end mapping. Therefore, we extract features directly on the original LR image Y with three convolution layers. SRCNN adopts ReLU as the activation function, while we use Parametric Rectified Linear Unit (PReLU) [10] in the new networks. They are different on the coefficient of the negative part. PReLU can be defined as a general activation function:

$$PReLU(x_i) = max(x_i, 0) + a_i min(0, x_i), \qquad (1)$$

where x_i is the input signal of the activation function on the i-th channel, 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" [11] caused by zero gradients in ReLU, and is set behind each convolution layer apart from the last convolution layer. The convolution layers can be represented as:

$$F_l(Y) = PReLU(W_l * F_{l-1}(Y) + B_l), \qquad (2)$$

where W_l and B_l denote the filters and biases of the l-th convolution layer respectively, F_l is the output feature maps and " $*$ " represents the convolution operation. The W_l contains n_l filters of support $n_{l-1} \times f_l \times f_l$, where f_l is the spatial support of a filter, n_l is the number of filters, and n_0 is the number of channels in the input image. Here we set 64 filters with the size of 3×3 for 3 convolution layers during the period of feature extraction. Note that there is no pooling or full-connected layers in SDSR, just convolution and deconvolution layers. For the purpose of confronting with degradation [12], we use a shortcut connection with identity mapping. The residual network converges much faster.

(2) Mapping: First, we map the high-dimensional (64) features into the low-dimensional (12) space using $12 \times 1 \times 1$ layer to save the computational cost. Then we use 4 convolution layers to increase the non-linearity of the model ($12 \times 3 \times 3 \times 12$ parameters for one layer).

(3) Upsampling: Upsampling operation is very important part at the end-to-end trainable system. It aims at increasing the spatial span to the target of HR size. The mapping operation reduces the number of LR feature dimension for the sake of improving computational efficiency. Our upsampling takes place at high dimension. Therefore, 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. Since we use caffe package [13], when training a set of $f_{sub} \times f_{sub}$-pixel LR sub-images with an upscaling factor n, the deconvolution layer can only get a size of $(nf_{sub} - n + 1)^2$ HR images.

(4) Multi-scale reconstruction: Considering that HR image restoration usually relies on both short- and long-range context information, we propose to perform HR reconstruction with multi-scale convolutions to extract multi-context information. Multi-scale inference 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 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 5-th layer is the dimension reduction layer with 1×1 kernel size and 16 channels. The subsequent multi-scale convolution layer is consisted of 4 convolution layers with 7×7, 5×5, 3×3 and 1×1 kernel sizes, respectively. The 4 convolution layers 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, that serves as a weighted combination of multi-context features.

(5) Combining Shallow and Deep Networks:

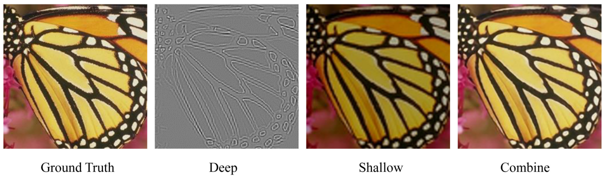

Fig. 2. Super-resolution results of deep network, shallow network and combined network.

The shallow network has only three layers, so it just restores rough images, which lacks of high-frequency details. In contrary, the deep network can more accurately render 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.

3. EXPERIMENTS AND RESULTS

3.1. Datasets for Training and Testing

Training dataset The 91-image dataset proposed in [6] is widely used as the training set in learning based SR methods [8] [14] [15] [9]. As big data generally get better results, we also use General-100 dataset [9] that contains 100 bmp-format images with no compression, thus 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. Thus the final training dataset is 20 times of the original data. That is, totally 191 × 20 = 3820 images.

Testing dataset For benchmark, the Set5 [16] (5 images), Set14 [17] (14 images) and BSD100 [18] (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. Number of iterations can be seen in figure 5. This net applys 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). During the period of training, we use Mean Squared Error (MSE) as the loss function: $1/2\|y - f(x)\|^2$, where y is the ground truth images, $f(x)$ is the output of our SDSR network. Training is on GPU Titan X Pascal.

3.2. 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.

3.2.1. Image super-resolution results

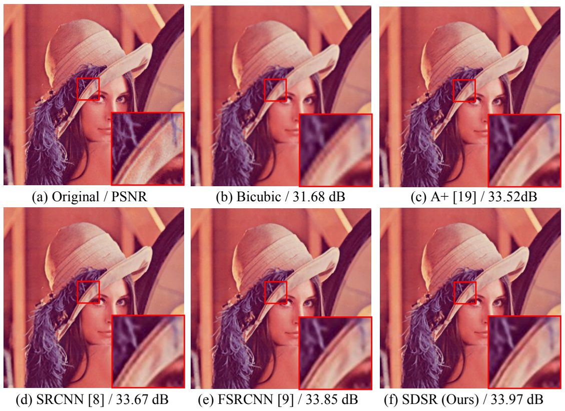

Fig. 3. Super-resolution results of "lenna" (Set14) with an upscaling factor 3. SDSR recovers sharp lines.

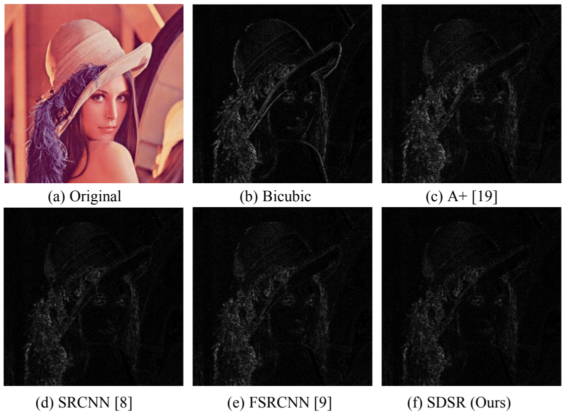

Fig. 4. Error maps for "lenna" (Set14) an upscaling factor 3. Black pixels are good and white pixels are bad pixels. Our SDSR with less bad pixels.

Figure 3 shows super restored images and particular regions of images are enlarged to better see the difference. 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 reaches to 33.97 dB, which is much higher than Bicubic Interpolation, Adjusted Anchored Neighbourhood Regression method (A+) [19], SR-CNN [8] and FSRCNN [9]. The results of the compared methods are either obtained using the publicly available codes

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+ [19]	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[8]	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[7]	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)
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)**

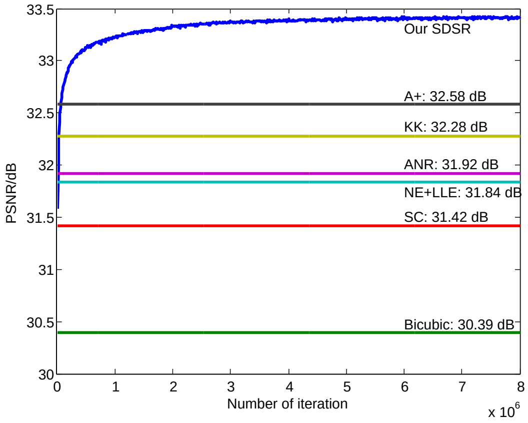

Fig. 5. The test convergence curve of our SDSR and results of other methods on the Set5 dataset.

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 figure 4. 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.

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 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 Figure 5. The state-of-the-art SR methods including: SC - sparse coding-based method of Yang et al.[6], NE+LLE - neighbour embedding + locally linear embedding method [5], ANR - Anchored Neighbourhood Regression method [20], A+ - Adjusted Anchored Neighbourhood Regression method [19], and KK - the method described in [21]. It is worth pointing out that SRCNN surpasses the existing state-of-the-art methods' lines at the very beginning of the learning stage, which indicates that SDSR is fast convergence. That should owe to the applying of residual learning.

3.2.2. *Highway real scene videos super-resolution results*

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

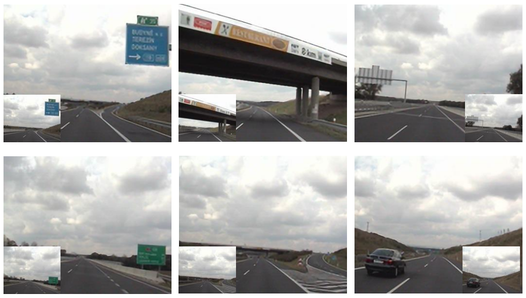

Fig. 6. The result of highway real scene videos reconstruction

Table. 2. Results on HD videos from Ultra Video Group datasets with an upscaling factor of 3.

Datasets	Bicubic	SRCNN	ESPCN	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 seconds in length to evaluate the proposed method. Table 2 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 [22]. 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.

4. CONCLUSION

In this paper, we have presented a super-resolution method using shallow and deep networks. 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.

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

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

5. REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 2, pp. 1097–1105, 2012.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Computer Science*, pp. 580–587, 2014.

[3] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of Applied Meteorology*, vol. 18, no. 8, pp. 1016–1022, 1979.

[4] M. Irani and S. Peleg, "Motion analysis for image enhancement: Resolution, occlusion, and transparency," *Journal of Visual Communication and Image Representation*, vol. 4, no. 4, pp. 324–335, 1993.

[5] H. Chang, D. Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. IEEE, 2004, vol. 1, pp. I–I.

[6] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[7] D. Liu, Z. Wang, B. Wen, J. Yang, W. Han, and T. S. Huang, "Robust single image super-resolution via deep networks with sparse prior," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3194–3207, 2016.

[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*. Springer, 2014, pp. 184–199.

[9] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision*. Springer, 2016, pp. 391–407.

[10] K. He, X. Zhang, S. Ren, and 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.

[11] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[14] C. Dong, Chen C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[15] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," *arXiv preprint arXiv:1511.04587*, 2015.

[16] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[17] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[18] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. IEEE, 2001, vol. 2, pp. 416–423.

[19] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 111–126.

[20] R. Timofte, V. De Smet, and 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.

[21] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 6, pp. 1127–1133, 2010.

[22] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and 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.