2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)

RESIDUAL COMPONENT ESTIMATING CNN FOR IMAGE SUPER-RESOLUTION

Xian-Hua Han[1] *and YongQing Sun*[2] *and Yen-Wei Chen*[3]

[1]Graduate School of Science and Technology for Innovation, Yamaguchi University, Japan
[2]NTT, Japan
[3]Ritsumeikan University, Japan

ABSTRACT

With the success of convolutional neural networks (CNNs) for different computer vision applications, CNNs have been widely applied for single image super- resolution (SR). The recent research line for CNN-based image SR mainly concentrates on exploring the pioneering network architectures such as very deep CNN, ResNet, GAN-net, for enhancing performance of the learned high-resolution (HR) image. Although the impressive performance with the recent CNN-based SR work has been achieved, the non-recovered high-frequency (residual) components are unavoidably existed with the current network architectures. This study aims to explore an unified CNN network architecture for learning not only the HR image but also simultaneously the difficultly recovered residual components in the first network. With one existed CNN architecture for image super-resolution, the HR image can be learned while some high-frequency content in the ground-truth image may not be perfectly recovered. For estimating the non-recovered high-frequency content, this study stacks another CNN architecture on the output of the baseline CNN, and construct an end-to-end residual component learning framework for more accurate image SR. Experimental results on benchmark dataset validate that the proposed residual component estimating CNN can overperform the non-stacked CNN architecture, and demonstrates state-of-the-art restoration quality.

Index Terms— Image super-resolution, convolutional neural networks (CNNs), residual component estimating CNN, high frequency component

1. INTRODUCTION

Estimating a highresolution (HR) image from its single low-resolution (LR) counterpart, which called as single image super-resolution (SR) is a highly challenging task in computer vision [1, 3]. Recently, SR received substantial attention within the research community and has a wide range of applications such as satellite imaging [2],medical imaging [**?**], security and surveillance [**?**]. Since many possible HR solutions are existed for a given LR image, the SR problem is widely known to be ill-posed. For recovering an optimal HR image among possible HR images, the state-of-the-art methods mainly constraint the solution space by learning strong prior information, generally called as learning-based SR method. These methods basically exploit the relationship between LR images and HR images via either learning mapping functions from external low- and high-resolution exemplar pairs from training samples, or exploring internal similarities of the same image [4, 5, 6, 7, 8].

In recent years, with the success of convolutional neural networks (CNNs) for different computer vision applications, CNNs are widely used to address the ill-posed inverse problem of Super Resolution (SR), and have demonstrated superiority over other learning paradigms [9, 10, 11, 12]. Motivated by the learning pipeline of the sparse-coding-based SR method (ScSR) [9], Dong et al. [13] proposed an end-to-end nonlinear mapping network between low- and high-resolution images with 3 convolutional layers (SRCNN), which can be explained as three processes: patch extraction and representation, nonlinear mapping and HR image reconstruction. Despite the simple architecture of the pioneer SRCNN model, promising performance can be achieved compared with other learning approaches such as ScSR [9], and for further improving SR performance, different CNN variants have been explored. To take consideration of the self similarity property, deep joint super resolution (DJSR) [14] has been proposed to jointly utilizes both the wealth of external examples and the power of self examples unique to the input. Wang et al. trained end-to-end a cascaded sparse coding network (CSCN) [32] to fully exploit the natural sparsity of images inspired by the idea of the learning iterative shrinkage and thresholding algorithm [5], Kim et al. [15] exploited a very deep CNN architecture based on VGG-network [16], and only learned the lost high-frequency image (residual image) to speed up the training procedure. Ledig et al. [17] combined GAN network for estimating much sharper HR image, and sometime led to unreliable detailed structure. These recent deep learning-based SR methods mainly concentrated more deep and powerful network architectures for learning the effective nonlinear mapping, while some high-frequency components are unavoidable to be non-recovered especially for large expand factor.

This study aims to explore an unified CNN network archi-

978-1-7281-5527-2/19/$31.00 ©2019 IEEE
DOI 10.1109/BigMM.2019.00076

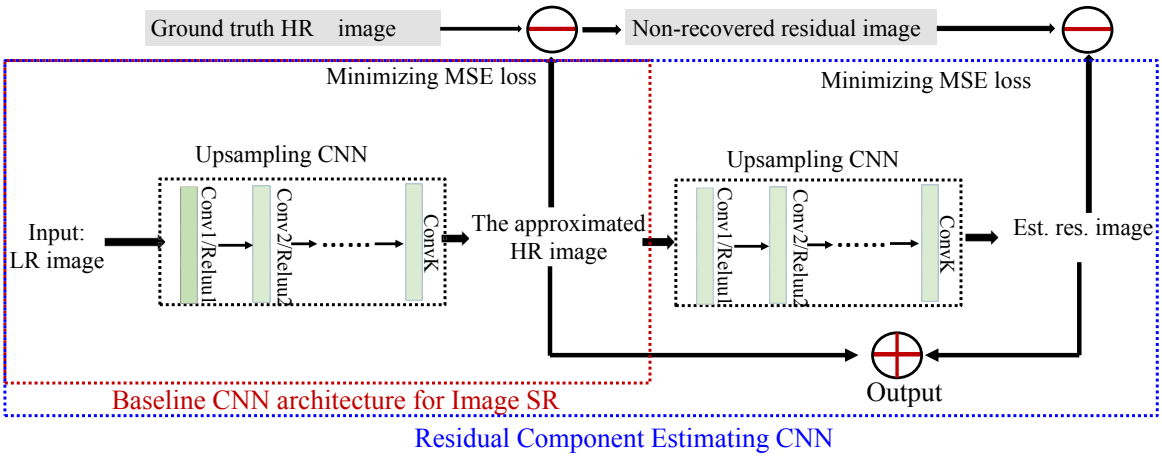


Fig. 1. The schematic concept of the proposed residual component estimating CNN. The parameters (convolutional filter weights) in the proposed CNN framework for image SR are optimized by simultaneously minimizing the reconstruction errors between the estimated HR image/the ground truth pair and the estimated residual component/the corresponding ground truth pair.

tecture for learning not only the HR image but also simultaneously the difficultly recovered residual components in the baseline network. As we know that any designed CNN architecture for image SR can estimate an approximation of the ground-truth HR image from a LR image while some high-frequency content in the ground-truth image may not be perfectly recovered especially for large expand factor. In order to estimate the non-recovered high-frequency content, we propose to stacks another CNN architecture on the output of the baseline CNN, and construct an end-to-end residual component learning framework for more accurate image SR. The proposed residual component learning framework can be constructed on any baseline CNN architecture for image SR, and combines the estimated approximation of the HR image and the residual component as the final predicted HR image. Experimental results on benchmark dataset validate that the proposed residual component estimating CNN can outperform the non-stacked CNN architecture especially for large expand factor, and demonstrates state-of-the-art restoration quality.

2. RESIDUAL COMPONENT ESTIMATING CNN

In this section, we present the technical parts of our proposed residual component estimating CNN for image SR. Specifically, we adopt global residual learning by minimizing the recovering errors not only between the initially estimated HR image and the ground truth image but also simultaneously between the non-recovered high-frequency (residual) component in the initially estimated HR image and the ground truth. Our proposed residual learning strategy is completely different from the residual units of the identity branches in ResNet [18], which adds the identity branches in the middle feature layers for implicitly learning the task-oriented feature excepting the input one, while the proposed residual component estimating CNN stack another CNN architecture on the baseline one for explicitly minimizing the reconstruction error between the non-recovered high-frequency (residual) component in the initially estimated HR image and the ground truth. Let's denote the LR input image as $\boldsymbol{y}$, the HR ground truth image as $\boldsymbol{z}$, the input feature of a residual layer as $\boldsymbol{x}$ and the underlying mapping as $H(\boldsymbol{x})$, the residual feature mapping is defined as $R(\boldsymbol{x}) = H(\boldsymbol{x}) - \boldsymbol{x}$. The objective function to be minimized in CNN based image SR is as:

$$\mathbf{W}_{Opt} = \arg\min_{\mathbf{W}} \|\boldsymbol{z} - F(\boldsymbol{y}, \mathbf{W})\|^2 \tag{1}$$

where $F(\cdot)$ denotes the transformation function from the LR input image $\boldsymbol{y}$ to the HR- ground truth image $\boldsymbol{z}$ with the filter parameters $\mathbf{W}$ of several convolutional layers, and $\hat{\boldsymbol{z}}$ is the estimated HR image via the image SR network.

For a a residual unit structure, the learned output feature is generally formulated as:

$$\hat{\boldsymbol{x}} = RELU(R(\mathbf{W}^R, \boldsymbol{x}) + I(\boldsymbol{x})) \tag{2}$$

where $\hat{\boldsymbol{x}}$ is the output of the residual unit, and $I(\boldsymbol{x})$ is an identity mapping [18] :$I(\boldsymbol{x}) = \boldsymbol{x}$. Function $RELU(R(\mathbf{W}^R, \boldsymbol{x})$ denotes ReLU nonlinear mapping, and $\mathbf{W}^R$ is a set of weights in the residual unit structure, which is optimized by minimizing the objection function of the global network structure. As discussed in [18], ResNet with the residual unit structures, which are implemented with the shortcut connections for identity mapping was mainly designed for solving the degradation problem in much deeper CNN architecture, and has experimental been validated that better performance

can be achieved for image classification problem. However, despite the used shortcut connections in some feature learning layers in the ResNet [18], the global objective function to be optimized is not altered and non-correct estimation can not be refined.

This study proposes to stack CNN architectures to explicitly learn the non-recovered residual component in the initially estimate HR image with the baseline CNN architecture, and constructs an end-to-end network structure for simultaneously learning the required HR image and the difficultly recovered high-frequency (residual) component. With a backbone CNN architecture designed for image SR problem, a HR image $\hat{\boldsymbol{z}}$ can generally be recovered with a LR image $\boldsymbol{y}$ as input, and is formulated as:

$$\hat{\boldsymbol{z}} = F(\boldsymbol{y}, \mathbf{W}_{Opt}) \tag{3}$$

Since the designed CNN network for image SR needs to not only reconstruct low-frequency content but also estimate high-frequency content lost in the input LR image, it is avoidable that some high-frequency content can not be recovered, and thus produce the non-recovered residual image: $\boldsymbol{z}^{Res} = \boldsymbol{z} - \hat{\boldsymbol{z}} = \boldsymbol{z} - F(\boldsymbol{y}, \mathbf{W}_{Opt})$. In order to further learn $\hat{\boldsymbol{z}}^{Res}$, this study overwrites the baseline CNN architecture of the image SR to construct a residual component estimating CNN for learning the non-recovered high-frequency content from the $\hat{\boldsymbol{z}}$. We combine the baseline HR image: $\boldsymbol{z}$ and the residual component: $\boldsymbol{z}^{Res}$ learning procedure to construct an end-to-end residual component estimation network, and the objective function of the residual component learning CNN is formulated as:

$$\begin{aligned} < \mathbf{W}_{Opt}, \mathbf{W}_{Opt}^{Res} > = \underset{\mathbf{W}, \mathbf{W}^{Res}}{\arg\min}\, \omega_1 \|\boldsymbol{z} - F(\boldsymbol{y}, \mathbf{W})\|^2 + \\ \omega_2 \|\boldsymbol{z}^{Res} - F_{Res}(F(\boldsymbol{y}, \mathbf{W}), \mathbf{W}^{Res})\|^2 \end{aligned} \tag{4}$$

where $F_{Res}(\cdot)$ denotes the transformation function from the estimated HR-HS image $F(\boldsymbol{y}, \mathbf{W})$ to the residual component $\hat{\boldsymbol{z}}^{Res}$ with the filter parameters $\mathbf{W}^{Res}$ of the convolutional layers, and $\hat{\boldsymbol{z}}^{Res}$ is the estimated residual component via the residual component estimating CNN. $\boldsymbol{z}^{Res}$ represents the non-recovered residual component in the baseline CNN. ω_1 and ω_2 are the weights of the reconstruction errors on the first estimation of $\boldsymbol{z}$ and the residual component estimation of $\boldsymbol{z}^{Res}$. The final estimation of the HR-HS image is the element-wise summation of $\hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{z}}^{Res}$: $\hat{\boldsymbol{z}}^{Final} = \hat{\boldsymbol{z}} + \hat{\boldsymbol{z}}^{Res}$. The schematic concept of the proposed residual component estimating CNN is shown in Fig. 1.

3. EXPERIMENTAL RESULTS

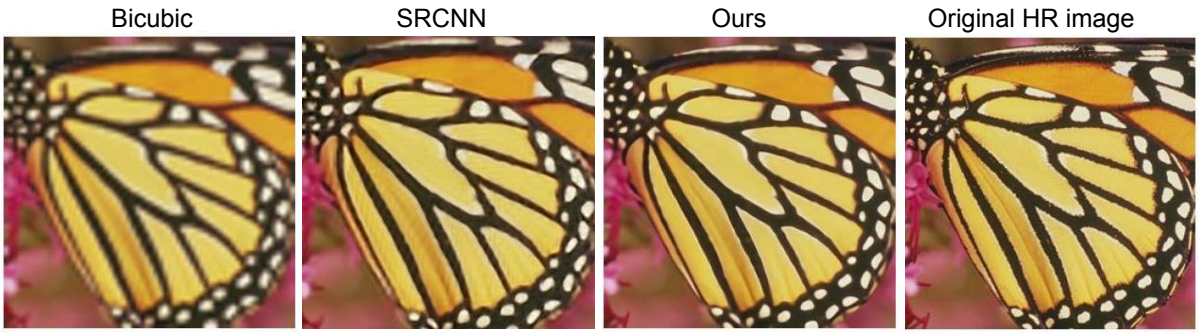
For a fair comparison with the state-of-the-art image SR methods, we use the same training set, test sets, and protocols as in [13]. Specifically, the training set consists of 91 images, and test sets consists of two groups: Set5 (5 images) [**?**] and Set14 (14 images) [12]. The up-scaling factors in our experiments are 3 and 4, and PSNR is used to quantitatively evaluate the performance of the estimated HR image. Following [42], super-resolution is only applied on the luminance channel (Y channel in YCbCr color space), which means the channel number: $c = 1$ in the first-input/last-output layer.

Table 1. The compared average PSNR (dB) of our proposed residual component estimating CNN with the conventional image SR methods including ScSR [11], ANR [8], the baseline SRCNN [13] and VDSR [15] on benchmark Set5, Set14 for upscaling factors 3 and 4.

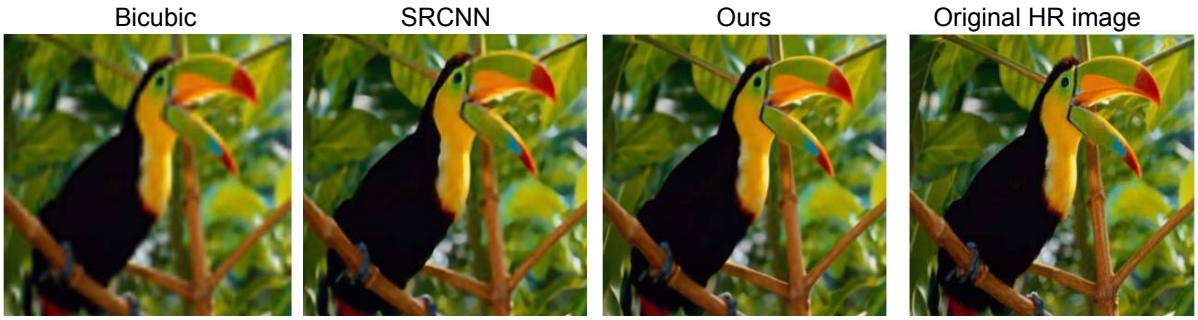
Dataset	Scale	ScSR	ANR	SRCNN	VDSR	Ours
Set5	×3	31.42	31.92	32.75	33.66	33.68
	×4	29.53	29.69	30.48	31.35	31.95
Set14	×3	28.31	28.65	29.28	29.77	29.81
	×4	*	*	27.49	28.01	28.53

For preparing samples for CNN training, we extract 33×33 overlapped sub-images from the interpolated LR input image which has the same size with the ground truth HR image. The 91-image dataset can be decomposed into 24,800 sub-images as in SRCNN [13], which are extracted with a stride of 14. Since we aim to evaluate the performance impact of the annexed residual component estimating architecture, we simply take the network structure of SRCNN as the baseline architecture and the the residual component estimating network is same but with added padding in the convolutional layers for ensuring the same spatial size of the output with its input (the output of the baseline CNN). Thus the corresponding ground truth sub-image for a 33×33 LR input is the center patch with size 21×21 in the HR images. The simple Euclidean distance between the estimated output and the ground-truth patches is minimized to learn the residual component estimation CNN parameters. Our network, implemented with Caffe [19], is trained from scratch, using the SGD optimizer. We use a mini-batch size of 128 in training procedure, and train the network for 5000000 iterations with the fixed learning rate 0.0001. Our model parameters are initialized according to Gaussian distribution with standard deviation 0.001. In our experiments, we set ω_1 and ω_2, as 0.2 and 0.8, respectively.

We compare our proposed CNN model with the conventional learning-based image SR methods such as ScSR [11], ANR [8], and the deep learning-based ones: SRCNN [13] and VDSR [15]. The average PSNRs on Set5 and Set14 are shown in Table 1, which manifests our residual component estimation CNN can improve the image recovering performance compared with the baseline SRCNN architecture and the im-



(a) 'Butterfly' test image



(b) 'Bird' test image

Fig. 2. Super-resolution results of 'Butterfly' and 'Bird' images from Set5 dataset with up-scaling factor 4, which manifests our method (the residual component estimation framework can give more favorable visualizing image than the baseline SRCNN network.

proved PSNR values for large upscaling factor (4) are more obvious than the upscaling factor 3. In addition, the residual component estimating framework with the simple CNN architecture (only 3 CNN layers) can over-performer the more complex CNN model such as the VDSR. We believe that this residual-refined architecture can applied to any other complex CNN architecture and is expected to provide much better recovering performance. Fig. 2 shows the recovered HR images with the Bicubic interpolation, the baseline SRCNN and our proposed residual component estimating CNN for the 'Butterfly' and 'Bird' images from Set5 dataset.

4. CONCLUSIONS

This study proposed a novel unified CNN architecture via stacking the baseline CNN structure for learning difficultly recovered residual component in the image SR problem. The proposed CNN framework aims to simultaneously learn the required HR image and the non-recovered residual components in the baseline network. As we know that any designed CNN architecture for image SR problem can estimate the HR image while some high-frequency content existed in the ground-truth image may not be perfectly recovered. For estimating the non-recovered high-frequency content, this study stacked another CNN architecture on the output of the baseline CNN, and constructed an end-to-end residual component learning framework for more accurate image SR, called as residual component estimating CNN. Experimental results on benchmark datasets validated that the proposed residual component estimating CNN can outperform the non-stacked CNN architecture, and demonstrates state-of-the-art restoration quality.

5. REFERENCES

[1] M. Fauvel, Y. Tarabalka, J. Benediktsson, J. Chanusssot, and J. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, 2013.

[2] Y. Tarabalka, J. Chanusssot, and J. Benediktsson, "Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 40, no. 5, pp. 1267–1279, 2010.

[3] X. Xu, Z. Wu, J. Li, A. Plaza, and Z. Wei, "Anomaly detection in hyperspectral images based on low-rank and sparse representation," *IEEE Trans. Geoscience and Remote Sensing*, vol. 54, no. 4, pp. 1990–2000, 2016.

[4] Y. Zhou, H. Chang, K. Barner, P. Spellman, and B. Parvin, "Classification of histology sections via multispectral convolutional sparse coding," *CVPR*, pp. 3081–3088, 2014.

[5] J.M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. M. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, 2013.

[6] P.S. Chavez, S.C. Sides, and J.A. Anderson, "Comparison of three different methods to merge multiresolution and multispectral data: Landsat tm and spot panchromatic," *Photogramm. Eng. Rem. S.*, vol. 30, no. 7, pp. 1779–1804, 1991.

[7] R. Haydn, G.W. Dalke, J. Henkel, and J.E. Bare, "Application of the ihs color transform to the processing of multisensor data and image enhancement," *Int. Symp on Remote Sens. of Env.*, 1982.

[8] B. Aiazzi, S. Baronti, F. Lotti, and M. selva, "A comparison between global and context-adaptive pansharpening of multispectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 302–306, 2009.

[9] M. Cetin and N. Musaoglu, "Merfing hyperspectral and panchromatic image data: Qualitative and quantitative analysis," *Int. J. Remote Sens.*, vol. 30, no. 7, pp. 1779–1804, 2009.

[10] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization for hyperspectral and multispectral data fusion," *IEEE Trans Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, 2012.

[11] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J.Y. Toureret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Trans Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3658–3668, 2015.

[12] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," *ICCV*, pp. 3586–3595, 2015.

[13] W.S. Dong, F.Z. Fu, G.M. Shi, X. Cao, J.J. Wu, G.Y. Li, and X. Li, "Hyperspectral image super-resolution via non-negative strutured sparse representation," *IEEE Transaction on Image Processing*, vol. 25, no. 3, pp. 2337–2352, 2016.

[14] X.-H. Han, B.X. Shi, and Y.Q. Zheng, "Self-similarity constrained sparse representation for hyperspectral image super-resolution," *IEEE Transaction on Image Processing*, vol. 27, no. 11, pp. 5625–5637, 2018.

[15] C. Dong, C. C. Loy, K.M. He, and X.O. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 38, no. 2, pp. 295–307, 2015.

[16] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] Aitor Alvarez-Gila, Joost van de Weijer, and Estibaliz Garrote, "Adversarial networks for spatial context-aware spectral image reconstruction from rgb," *IEEE International Conference on Computer Vision Workshop (ICCVW 2017)*, 2017.

[18] S. Galliani, C. Lanaras, D. Marmanis, E. Baltsavias, and K. Schindler, "Learned spectral super-resolution," *arXiv preprint arXiv:1703.09470*, 2017.

[19] Y.S. Li, J. Hua, X. Zhao, W.Y. Xie, and J.J. Li, "Hyperspectral image super-resolution using deep convolutional neural network," *Neurocomputing*, vol. 266, pp. 29–41, 2017.

[20] X.-H. Han, B.X. Shi, and Y.Q. Zheng, "Ssf-cnn: Spatial and spectral fusion with cnn for hyperspectral image super-resolution," *ICIP*, pp. 2506–2510, 2018.

[21] K.M. He, X.Y. Zhang, S.Q. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[22] A. Robles-Kelly, "Single image spectral reconstruction for multimedia applications," *Proc. of ACM Multimedia Conference (MM)*, pp. 251–260, 2015.

[23] F. Yasuma, D. Iso, and S. Nayar, "Generalized assorted pixel camera: Post-capture control of resolution, dynamic range and spectrum," *IEEE Transaction on Image Processing*, vol. 19, no. 9, pp. 2241–2253, 2010.

[24] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," *CVPR*, pp. 193–200, 2011.