Residual Learning Based Convolutional Neural Network for Super Resolution

Hwei Jen Lin [1], Yoshimasa Tokuyama [2], and Zi Jun Lin [1]
[1] Tamkang University, Taipei, Taiwan
[2] Tokyo Polytechnic University, Tokyo, Japan
Email: 086204@mail.tku.edu.tw, tokuyama@mega.t-kougei.ac.jp, a0981382251@gmail.com

***Abstract*—Recently, there have been many methods of super resolution proposed in the literature, in which convolutional neural networks have been confirmed to achieve good results. C. Dong *et al.* proposed a convolutional neural network structure (SRCNN) to effectively solve the super resolution problem. J. Kim *et al.* proposed a much deeper convolutional neural network (VDSR) to improve C. Dong *et al.*'s method. However, unlike VDSR proposed by J. Kim *et al.* which trained residue images, SRCNN proposed by C. Dong *et al.* directly trained high-resolution images. Consequently, we surmise the improvement of VDSR is due to not only to the depth of the neural network structure but also the training of residue images. This paper studies and compares the performance of training high-resolution images and training residue images associated with the two neural network structures, SRCNN and VDSR. Some deep CNNs proceed zero padding which pads the input to each convolutional layer with zeros around the border so the feature maps remain the same size. SRCNN proposed by C. Dong *et al.* does not carry out padding, so the size of the resulting high-resolution images is smaller than expected. The study also proposes two revised versions of SRCNN that remain the size the same as the input image.**

***Index Terms*—super resolution, convolutional networks, bicubic interpolation, deep learning, underdetermined inverse problem**

I. Introduction

Super resolution is the process of upscaling and improving the details within an image. Single-image super resolution techniques include not only traditional interpolations but also learning based super resolution algorithms. Most traditional interpolation methods are types of polynomial interpolation, including nearest-neighbor interpolation, bilinear interpolation, and bicubic interpolation, each of which takes a set of neighboring pixels to create the value of a new pixel to achieve super resolution. The advantage of these methods is their simplicity and low computation cost; while the disadvantage is they tend to produce blurry or jagged edges.

Many learning based SR algorithms have been proposed, including example-based methods [1], [2], neighbor embedding [3], [4], and support vector regression [5], [6], sparse representation [7]-[9], and convolutional neural network methods [10]-[14], and these have been confirmed to achieve good results.

C. Dong *et al.* [10] proposed a method of super resolution using a deep convolutional neural network. The network architecture consists of three convolutional layers to learn low-resolution image patches corresponding to high-resolution image patches. J. Kim *et al.* [8] proposed a deeper convolutional neural network for super resolution, consisting of 20 convolutional layers. They claimed their deeper structure outperformed the three-layer convolutional neural network proposed by C. Dong *et al*. However, J. Kim *et al.* trained the corresponding residue image instead of the high-resolution image. Many learning-based super resolution algorithms based on residue image learning have shown better performance than those based on high-resolution image learning [11]-[14].

A residue (image) is the difference between a high-resolution image and a corresponding low-resolution image, which possesses the high frequency information existing in the high-resolution image but not in the low-resolution image. The residue image obtained from the trained network added to the input low-resolution image is a higher resolution image, which is the result of super resolution. We suspected the improvement of VDSR proposed by J. Kim *et al.* might be due to not only the increased number of layers of the network but also the training of residue images. We study and compare the results of the two methods based on the training of residue images and the training of high-resolution images. The computational time required by each of the above versions is also analyzed and compared. C. Dong *et al.*'s SRCNN does not apply zero padding to keep the size of the output image the same as expected. We also propose some methods to solve this problem for SRCNN and maintain the quality of super resolution.

II. Proposed Method

This paper proposes a modified version of SRCNN, proposed by C. Dong *et al.* [10], for super resolution. Different from SRCNN which trains high-resolution images, the proposed version trains residue images. The residue image obtained from the trained network added to the input low-resolution image is the result of super resolution. In addition, we propose solutions to the problem of reducing the size of the output image from SRCNN while maintaining the super resolution quality.

Manuscript received July 23, 2019; revised November 8, 2019.

©2019 Journal of Image and Graphics
doi: 10.18178/joig.7.4.126-129

During the training processing of SRCNN, a training image $\boldsymbol{X}$ is downscaled into a lower-resolution image $\boldsymbol{Y_0}$, which is then upscaled to become the same size as $\boldsymbol{X}$ using bicubic interpolation before it is input into the network. Let $\boldsymbol{Y}$ denote the interpolated image. The purpose of SRCNN is to learn the correspondence F_{SRCNN} between the interpolated image $\boldsymbol{Y}$ and the original high-resolution image $\boldsymbol{X}$; while the purpose of our modified network is to learn the correspondence $F_{LRSRCNN}$ between the interpolated image $\boldsymbol{Y}$ and the residual image $\boldsymbol{R} = \boldsymbol{X} - \boldsymbol{Y}$, so the generated image $F_{LRSRCNN}(\boldsymbol{Y})$ is as close as possible to the difference between the ground truth high-resolution image $\boldsymbol{X}$ and the interpolated image $\boldsymbol{Y}$ ($\boldsymbol{R} = \boldsymbol{X} - \boldsymbol{Y}$), and so the resulting high-resolution image $\boldsymbol{Y} + F_{SRCNN}(\boldsymbol{Y})$ is as close as possible to the ground truth high-resolution image $\boldsymbol{X}$, as shown in Fig. 1. Since our network focuses on the training of residue images, it is called Residual Learning based Super Resolution Convolutional Neural Network (RLSRCNN).

Like SRCNN, the setting for RLSRCNN is $f_1 = 9$, $f_2 = 1$, $f_3 = 5$, $n_1 = 64$, and $n_2 = 32$. In general, the information from $(f_1 + f_3 - 1)^2 = (9 + 5 - 1)^2 = 169$ pixels are utilized to estimate a residue pixel. It uses much more information for reconstruction than existing external example-based approaches, for example, using $(5 + 5 - 1)^2 = 81$ pixels [15]. This is one of the reasons why the SRCNN gives superior performance.

In the training stage, the mean square error is used as the loss function for back propagation, as shown in (1), where $\{\mathbf{r}_i, \mathbf{y}_i\}_{i=1}^N$ are pairs of a residue image patch and a low-resolution image patch for training and $\Theta = \{W_1, W_2, W_3, B_1, B_2, B_3\}$ are network parameters.

RLSRCNN (without padding) and RLSRCNNP (with padding) have their own advantages. The resulting convolution image without carrying out padding has better quality than the corresponding central portion of that with padding, but result in a smaller size. To take advantage of the two strategies, we may combine the result of RLSRCNN and the border of the result of RLSRCNNP or the border obtained by bicubic interpolation, as shown in Fig. 2.

$$L(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\|F(\mathbf{y}_i;\Theta) - \boldsymbol{r}_i\|^2 \quad (1)$$

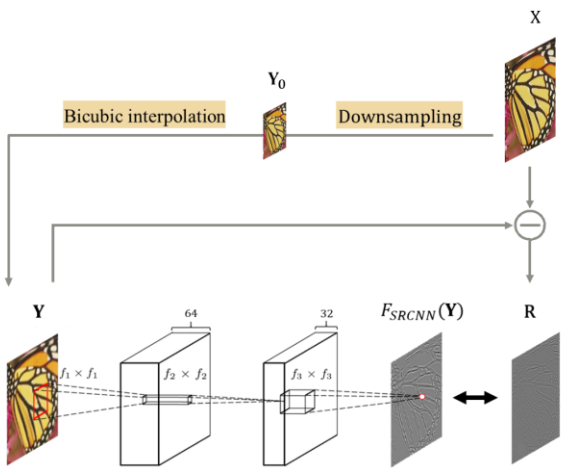

Figure 1. Structure of RLSRCNN

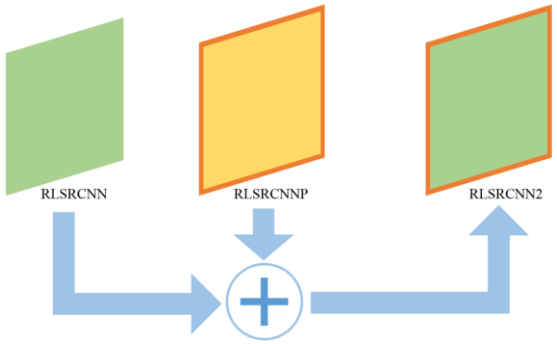

Figure 2. Combination of the result of RLSRCNN and the border of the result of RLSRCNNP

III. Experimental Results

We implemented the proposed algorithms using python programming language on a personal computer equipped with an Intel Core i7-6700 3.4GHz 8CPU and 16 GB memory.

This section compares the proposed RLSRCNN method with SRCNN VDSR in terms of SR quality and the time cost. The 91 images, of size between 78×78 and 508×508, used in C. Dong *et al.*'s experiments [10] and dataset Set14 [9] are used for training and testing, respectively. Besides, the FERET face image dataset [16] consisting of 130 images of size 100×120 are also used, of which 100 images were for training and 30 images were for testing.

For comparison of different methods, the experiments are two-fold. Since C. Dong *et al.*'s SRCNN does not carry out padding, in the first set of experiments, RLSRCNN and SRCNN are compared. For comparison with J. Kim *et al.*'s VDSR structure, which carries out padding, we modify SRCNN and RLSRCNN into versions using padding, called SRCNNP and RLSRCNNP, respectively. In the second set of experiments, VDSR, SRCNNP, RLSRCNNP, and RLSRCNN2 are compared.

TABLE I. PSNR/SSIM Comparison of SR Methods without Padding

datasets	Set14		FERET	
	PSNR	SSIM	PSNR	SSIM
Bicubic	26.31	0.6181	31.64	0.8727
SRCNN	27.41	0.6524	32.66	0.8813
RLSRCNN	**27.68**	**0.6644**	**32.91**	**0.891**

TABLE II. PSNR/SSIM Comparison of SR Methods with Padding

datasets	Set14		FERET	
	PSNR	SSIM	PSNR	SSIM
Bicubic	26.07	0.617	30.94	0.864
VDSR	26.83	0.645	31.64	**0.881**
SRCNNP	26.92	0.628	31.45	0.871
RLSRCNNP	27.12	0.633	32.06	**0.881**
RLSRCNN2	**27.32**	**0.66**	**32.16**	**0.881**

©2019 Journal of Image and Graphics

The comparisons are shown in Table I and Table II, in which the central portions of the results of bicubic interpolation serve as a baseline for comparison. The tables show the results of RLSRCNN are better than SRCNN. Fig. 3 and Fig. 4 show some results of RLSRCNN and SRCNN on Set14 and FERET.

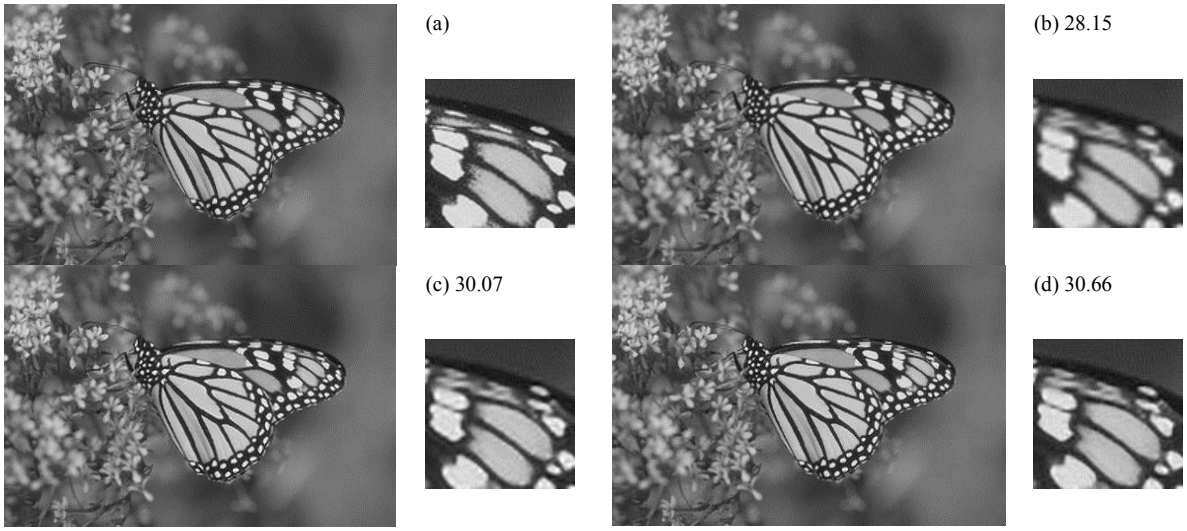

Figure 3. Comparison of SR methods: (a) ground truth, (b) bicubic, (c) SRCNN, (d) RLSRCNN

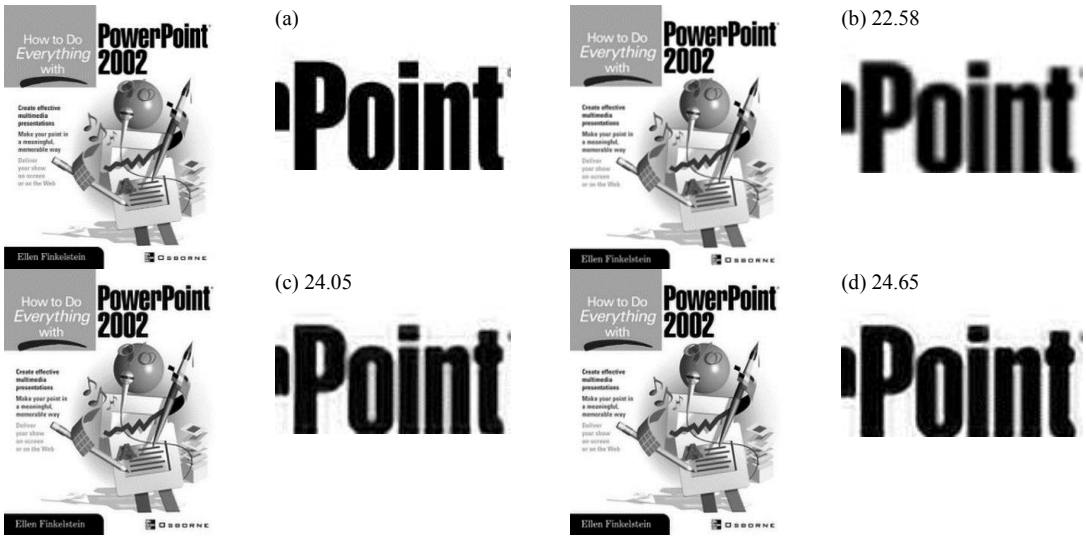

Figure 4. Comparison of SR methods: (a) ground truth, (b) bicubic, (c) SRCNN, (d) RLSRCNN

TABLE III. RLSRCNN2 ON SET14 AND FERET

datasets	Set14		FERET	
	PSNR	SSIM	PSNR	SSIM
Bicubic	26.07	0.617	30.94	0.864
VDSR	26.83	0.645	31.64	**0.881**
SRCNNP	26.92	0.628	31.45	0.871
RLSRCNNP	27.12	0.633	32.06	**0.881**
RLSRCNN2	**27.32**	**0.66**	**32.16**	**0.881**

The comparison of the four structures carries out padding, including VDSR, SRCNNP, RLSRCNNP, and RLSRCNN2, on Set14 and FERET are shown in Table III. Table IV and Table V compare the results of training over a short time period and along time period, respectively. From these tables, we can see that all methods improve their performance after longer time training. When training time reaches about 2350 seconds, VDSR outperforms all the other methods.

TABLE IV. COMPARISON OF SRCNNP, VDSR, RLSRCNNP, AND COMPARISON OF SRCNNP, VDSR, RLSRCNNP, AND RLSRCNN2 UNDER A SHORT TRAINING TIME PERIOD: (A) BICUBIC, (B) VDSR (RESIDUE LEARNING), (C) VDSR (HR LEARNING), (D) SRCNNP, (E) RLSRCNNP, (F) RLSRCNN2

datasets	91pic		FERET	
methods	Average PSNR	Training time (sec)	Average PSNR	Training time (sec)
(A)	26.31	-	30.94	-
(B)	26.86	186.58	31.64	41.4
(C)	25.6	186.86	20.5	39.64
(D)	26.92	182.09	31.45	41.5
(E)	27.12	182.37	32.07	40.64
(F)	**27.32**	182.37 + 178.86	**32.16**	40.64 + 40.95

TABLE V. COMPARISON OF SRCNNP, VDSR, RLSRCNNP, AND RLSRCNN2 UNDER A LONG TRAINING TIME PERIOD: (A) BICUBIC, (B) VDSR (RESIDUE LEARNING), (C) VDSR (HR LEARNING), (D) SRCNNP, (E) RLSRCNNP, (F) RLSRCNN2

datasets	91pic		FERET	
methods	Average PSNR	Training time (sec)	Average PSNR	Training time (sec)
(A)	26.31	-	30.94	-
(B)	**27.49**	2327.71	**32.7**	376.52
(C)	27.35	2341.04	31.49	372.77
(D)	27.29	2328.91	32.1	373.02
(E)	27.34	2327.4	32.32	375.98
(F)	27.38	2327.4 + 2298.3	32.38	375.98 + 375.84

IV. CONCLUSIONS

This paper proposes a super resolution method with the so-called RLSRCNN structure, which is a modified version of SRCNN proposed by C. Dong *et al.* Unlike SRCNN which directly trains a target high-resolution image, RLSRCNN trains a residue image. The experimental results show training residue images obtains better results than training target high-resolution images. Our experimental results also show the much deeper structure VDSR proposed by J. Kim *et al.* outperforms SRCNN due to not only a long training period but also the training of residue images.

In addition to the training of residue image for improving SR results, we suggest combining the result of RLSRCNN and the border of the result of RLSRCNNP or the border obtained by bicubic interpolation, to produce a SR result of good quality while preserving the desired size.

ACKNOWLEDGMENT

We thank the Ministry of Science under Grant no. MOST-108-2922-I-032-003 and Technology and Ministry of Education, Taiwan for funding this study.

REFERENCES

[1] W. Freeman, T. Jones, and E. Pasztor, "Example-based super resolution," *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 55-65, 2002.
[2] K. I. Kim and Y. Kwon, "Example-based learning for single-image super resolution," in *Proc. DAGM Symposium*, Munich, 2008.
[3] H. Chang, D. Y. Yeung, and Y. Xiong, "Super resolution through neighbor embedding," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, June-July 2004, vol. 1, pp. I-275-I-282.
[4] X. Gao, K. Zhang, D. Tao, and X. Li, "Image super resolution with sparse neighbor embedding," *IEEE Transactions on Image Processing*, vol. 21, pp. 3194-3205, March 2012.
[5] T. C. Ho and B. Zeng, "Super resolution images by support vector regression on edge pixels," in *Proc. IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, Nov.-Dec. 2007, pp. 674-677.
[6] K. S. Ni and T. Q. Nguyen, "Image super resolution using support vector regression," *IEEE Transactions on Image Processing*, vol. 16, pp. 1596-1610, June 2007.
[7] A. B. Rao and J. V. Rao, "Super resolution of quality images through sparse representation," in *Proc. the 48th Annual Convention of Computer Society of India - vol. II. Advances in Intelligent Systems and Computing*, 2014.
[8] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, pp. 2861-2873, 2010.
[9] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. International Conference on Curves and Surfaces*, 2012, pp. 711-730.
[10] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 295-307, June 2015.
[11] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super resolution using very deep convolutional networks," *Computer Vision and Pattern Recognition*, arXiv: 1511.04587v2 (cs.CV), November 2015.
[12] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super resolution," *Computer Vision and Pattern Recognition*, arXiv:1707.02921v1 (cs.CV), July 2017.
[13] Y. Tai, J. Yang, and X. Liu, "Image super resolution via deep recursive residual network," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp. 2790-2798.
[14] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super resolution," *Computer Vision and Pattern Recognition*, arXiv:1802.08797v2 (cs.CV), Feb. 2018.
[15] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," in *Proc. IEEE International Conference on Computer Vision*, 1999, vol. 2, pp. 25-47.
[16] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1090-1104, Oct. 2000.

Hwei Jen Lin received the B.S. degree in Applied Mathematics from National Chiao Tung University, Hsinchu, Taiwan in 1981 and the M.S. and the Ph.D. degrees in Mathematics from Northeastern University, Boston, U.S.A. in 1983 and 1989. She is currently a professor at the Department of Computer Science and Information Engineering of Tamkang University. Her current research interests include pattern recognition, image processing, computational intelligence, and deep learning.

Yoshimasa Tokuyama is a professor of Department of Media and Image Technology, Faculty of Engineering, Tokyo Polytechnic University. He received his MS in Mechanical Engineering in 1986 and doctor degree in Computer Graphics in 2000 from The University of Tokyo. He was a member of the 3D CAD project at RICOH's Software Division from 1986 to 2002. His areas of research interest include computer graphics, image processing, virtual reality, augmented reality, game, haptic interface, and their applications.

Zi Jun Lin is currently a Master student in the Department of Computer Science and Information Engineering of Tamkang University. Her research interests focus on computer vision, pattern recognition, and deep learning.