Face Super-resolution Reconstruction Based on Multi-scale Feature Fusion

Shasha Wu[1,a] and Xueyun Chen[1,b]

[1]*College of electrical engineering, Guangxi University, Nanning, Guangxi Province, China*
a. 439450933@qq.com, b. cxy177@163.com

Keywords: Face super-resolution reconstruction, multi-scale, feature fusion, convolutional neural network, image processing.

Abstract: The low resolution and poor recognition of face image in the monitoring environment will lead to the decrease of face recognition accuracy. At present, most super-resolution algorithms suffer from serious image detail losses, due to the low-resolution of input images. In this paper, we propose a super-resolution reconstruction algorithm based on multi-scale feature fusion to alleviate this problem. A feature fusion mapping structure is used to extract features from multi-scale visual fields, and a skip-connection network structure is designed to construct high-resolution face images from them. The experimental results show that the proposed algorithm achieves better super-resolution results: clearer textures, sharper edges, enhanced visual effects, and higher evaluation indexes on FERET face database than the existing mainstream algorithms.

1. Introduction

In the video surveillance field, due to the limitation of camera resolution and the distance between the target and the camera, the obtained face images are often fuzzy and of low resolution, leading to the reduction of recognition rate, making face super-resolution reconstruction particularly significant.

Super-resolution (SR) reconstruction has been a hot topic in the field of image processing. It is a method of using one or more low-resolution (LR) images to predict high- resolution (HR) images in a large variety of fields like remote sensing images, medical images [1], etc. Current super-resolution reconstruction techniques can be divided into three types: interpolation-based, reconstruction-based and learning-based. Bilinear and Bicubic interpolations are classical, simple and intuitive algorithms [2-3]. However, there are some problems in the reconstruction of the images, such as sawtooth artifact and texture blur. Algorithms based on reconstruction [4-5] utilize the signal processing theory to recover images, which suffer from the loss of high-frequency information, such as iterative back-projection method [6] and convex set projection method [7]. This kind of algorithms can reconstruct a clear and high-resolution image, but usually ignore some details. In recent years, learning-based methods [8-10] have attracted much attention. Freeman [11] proposed a super-resolution algorithm based on example learning, which first pointed out that there

are a large number of self-similar blocks in the local spatial neighborhood of the image. Yang et al. [12] used sparse coding theory to construct high- and low-resolution image dictionaries and reconstructed HR images by learning the mapping relationship between dictionaries. With the rapid development of deep learning, convolutional neural networks (CNNs) and random forest [13] have been applied to the super-resolution reconstruction. Dong et al. [14] took the lead in applying the convolutional neural network to image super-resolution, and named the proposed model Super-Resolution Convolutional Neural Network (SRCNN). Deep learning-based approaches have become a hot spot in the field of super-resolution. To improve the speed of image reconstruction, Dong et al. [15] proposed a fast image super-resolution reconstruction (FSRCNN) algorithm. This algorithm directly extracted features from low-resolution images and introduced a deconvolutional layer at the end of the network to reconstruct high-resolution images.

The reconstruction effect based on deep learning is generally better than those based on interpolation. However, it still exists some defects. Although SRCNN realized the first application of CNN to super-resolution reconstruction, but its network structure is relatively simple. Therefore, it's unable to extract deeper features and the reconstruction effect is not perfect. These network models extract the features of the images in a single scale, do not pay attention to the richer details at different scales.

CNN has made great progress in super-resolution reconstruction. To address these problems, we put forward a multi-scale feature fusion convolutional network structure based on CNN, named as MFFCN. The main contributes are listed as:

(1) A multi-scale feature fusion mapping structure (MFF) is proposed to extract features of different scales from the same low-resolution image.

(2) A skip-connection network is presented to extract the deeper features, merge the shallow and deep features, and combine the perceptual loss with texture loss to produce a detailed and high-resolution image.

2. Related Work

2.1.SRCNN

SRCNN first utilizes the convolutional network to image super-resolution reconstruction field. It is a three-layer convolutional network structure based on conventional sparse-coding-based SR methods. First, it uses the Bicubic interpolation algorithm to interpolate the images to get the low-resolution images, then enlarges the images to the same size as the high-resolution images. The fuzzy images got in this way are used as the input of the network. The convolutional kernel of the first convolutional layer is 9×9, the number of channels is 64. This layer is utilized to extract a set of feature maps. The second convolutional layer with the convolutional kernel of 1×1 is used to map these feature maps nonlinearly to high-resolution patch representations. The kernel of the last convolutional layer is 5×5, this layer is used to reconstruct the image to produce the high-resolution image.

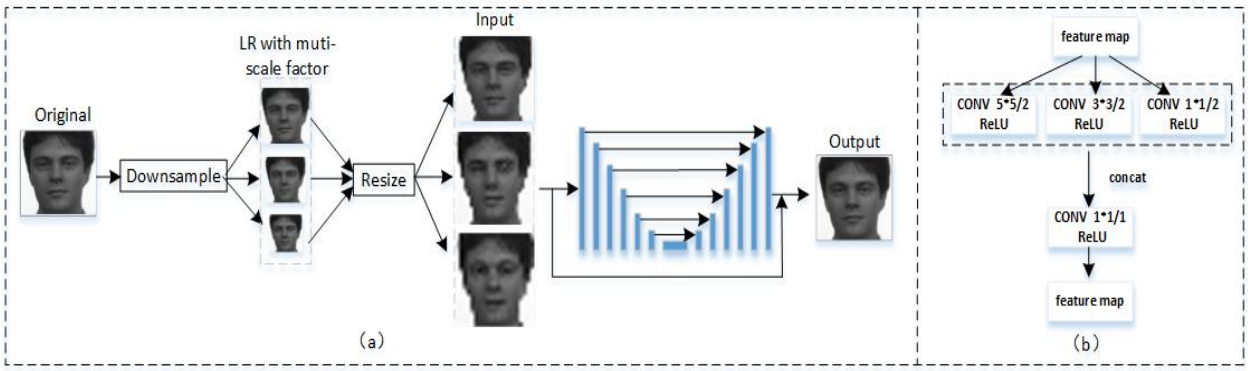

Figure 1: Structure of the network. (a) The multi-scale feature fusion face super-resolution reconstruction network structure (MFFCN). (b) The multi-scale feature fusion mapping module structure (MFF).

2.2.Face Super-resolution Reconstruction Algorithm

Inspired by SRCNN, Nie [16] proposed a face super-resolution reconstruction framework based on CNN. During the training process, First, all the high-resolution images in the training set were down-sampled at f times to get the low-resolution images. Then it puts the low-resolution images into the network, and optimizes the loss function to train the parameters, which calculates the mean square error of input I_i and output I_o. The loss function is shown in equation (1):

$$L = \frac{1}{2} || I_o - I_i ||^2 \tag{1}$$

3. Proposed Method

We propose the framework MFFCN to tackle the face super-resolution reconstruction problem. First, it selects a random down-sampling factor to down-sample the original high-resolution images to get the low-resolution images, then enlarges the low-resolution images to the same size as the original images. It uses the fuzzy images as the input of the network. Second, it extracts features of different scales of the input images by the multi-scale feature fusion module (MFF). The multi-scale features can learn the corresponding relationship between the input images and the original images, so as to ensure the detail clarity of reconstructed images.

3.1.Objective Function

We define the objective function of MFFCN as:

$$\begin{aligned} minV(\Phi(I_{LR}), I_{HR}) = \lambda_1 L_2(\Phi(I_{LR}), I_{HR}) + \lambda_2 L_{part}(\Phi(I_{LR}), I_{HR}) \\ + \lambda_3 TV(\Phi(I_{LR})) \end{aligned} \tag{2}$$

We put the blurred picture I_{LR} into the network Φ, and get the reconstructed image $\Phi(I_{LR})$, calculate the L2 loss, feature loss $L_{part}(\cdot)$ and total variation loss $TV(\cdot)$ between the reconstructed

image $\Phi(I_{LR})$ and the original image I_{HR} to optimize the objective function. The details are as follows.

3.1.1. L2 Loss

The L2 loss is a measure of the loss in pixels between the super-resolution image generated by the multi-scale convolutional network Φ and the original image I_{HR}. By solving the least square error between the generated picture $\Phi(I_{LR})$ and the original picture to optimize the pixel loss. It is used to ensure the consistency of the generated image and the original image on the basic features of the face. L_2 loss is shown in equation (3):

$$L_2(\Phi(I_{LR}), I_{HR}) = \| I_{HR} - \Phi(I_{LR}) \|^2 \tag{3}$$

3.1.2. Feature Loss

To pursue more realistic details, we propose a feature loss that combines perceptual loss with texture loss. Compared with the simple loss between pixels, the perceptual loss hopes that the features between the two images are more similar. Texture loss seeks to effectively reconstruct texture from the aspects of high-level global information and low-level detail information. We feed original image and reconstructed image into the feature loss network φ to extract their features, and calculate the perceptual loss and the texture loss. In this paper, the feature loss network uses the pre-trained network VGG19 [17] on the ImageNet database. We input the image into VGG19, and use the layer of the VGG19 which unused activation function before the fifth pooling layer to extract features. Perceptual loss is shown in equation (4):

$$L_{per}(\Phi(I_{LR}), I_{HR}) = \| \varphi(I_{HR}) - \varphi(\Phi(I_{LR})) \| \tag{4}$$

We put the reconstructed image and the original image into the loss network φ, minimize the absolute deviation of the feature values between them, make them more semantically similar.

Texture loss is shown in equation (5):

$$L_{text}(\Phi(I_{LR}), I_{HR}) = \| g(\varphi(I_{HR})) - g(\varphi(\Phi(I_{LR}))) \| \tag{5}$$

We use the original image as the target texture image, the output image is generated iteratively by matching statistics extracted from loss network to the target textures. $g(\cdot)$ represents the Gam matrix of the feature.

Overall feature loss is shown in equation (6):

$$L_{part}(\Phi(I_{LR}), I_{HR}) = L_{per}(\Phi(I_{LR}), I_{HR}) + L_{text}(\Phi(I_{LR}), I_{HR}) \tag{6}$$

3.1.3. Total Variation Loss

The total variation of the image contaminated by noise is significantly larger than the total variation of the noiseless image. Limiting the total variation can control the noise of the image during the generation process. We reduce the noise in the reconstructed image by minimizing the difference in pixel values between two adjacent points in the image. The total variation loss is as:

$$TV(\Phi(I_{LR})) = \sum_{i,j}((\Phi(I_{LR})_{i,j+1} - \Phi(I_{LR})_{ij})^2 + (\Phi(I_{LR})_{i+1,j} - \Phi(I_{LR})_{ij})^2) \quad (7)$$

3.2.Multi-scale Feature Fusion Network

We present a multi-scale feature fusion face super-resolution reconstruction network structure (MFFCN). The network structure is illustrated in Figure 1.

First, it randomly selects the down-sampling factor to down-sample the original face images to obtain low-resolution images, and then enlarges the low-resolution image to the same size as the original face images. The resulting blurred images are used as the input of the network. In the entire network structure, the multi-scale feature fusion mapping module (MFF) is used to replace the original convolutional layer. We add skip connections, following the general shape of a "U-Net" [18]. Specifically, we add skip connections between each layer i and layer n−i, where n is the total number of layers. Each skip connection simply concatenates all channels at layer i with those at layer n−i. The feature map is directly superimposed and passed to the next layer. This framework can fuse deep and shallow features. Finally, the input images are directly superimposed with the output of the network to fuse the main features of the input images.

Most existing algorithms almost only use smaller or larger receptive fields and learn from one dimension. The structural information of the images is usually with different scales. The single scale-feature extraction is not enough to completely restore the high-frequency texture area of the images. Motivated by this fact, we propose a multi-scale feature fusion mapping module (MFF) to learn the features from different receptive fields. The MFF shown in Fig.1 is composed of three convolutional kernels of three different scales of 1×1, 3×3 and 5×5. It is used to extract features from multi-scale visual fields. Then it uses cascading to merge and reorganize these feature maps containing multi-scale information. Finally, it puts the feature map into a convolutional layer with the kernel which size is 1×1 for feature mapping to generate a new feature map.

3.3.Algorithm

Algorithm 1: MFFCN. Default value λ_1=10, λ_2=1, λ_3 =0.0001, η=0.0001, m=10.
Require: The batch size m. original face dataset S_{HR} . The maximum number of iterations t_{max} . Parameter to be optimized ω. learning rate η .

1. for $t < t_{max}$ do
2. Randomly select m sample pictures $\{I_{HR}^1, I_{HR}^2 \cdots I_{HR}^m\}$ from the original picture data set S_{HR}, randomly select a factor to down-sample the original images and enlarge them to the same size as the original images to obtain the input image samples $\{I_{LR}^1, I_{LR}^2 \cdots I_{LR}^m\}$
3. Take I_{LR} as input, get m reconstruction images $\{I_{con}^1, I_{con}^2 \cdots I_{con}^m\}$, $I_{con}^i = \Phi(I_{LR}^i)$
4. Update parameter ω to optimize objective function V

$$Minmize\{\nabla V(\omega) = \nabla_\omega \frac{1}{m}\sum_i^m [\lambda_1 L_2(\Phi(I_{LR}), I_{HR}) + \lambda_2 L_{part}(\Phi(I_{LR}), I_{HR}) + \lambda_3 TV(\Phi(I_{LR}))]\}$$

$$\omega \leftarrow \omega - \eta \nabla V(\omega)$$

5. end for

4. Experiment

In this section, we use experiments to validate the effectiveness of the proposed method. We use mini-batch SGD and apply the Adam solver [19], with a learning rate of 0.0001, and momentum parameters $\beta 1 = 0.5, \beta 2 = 0.999$.

We evaluate our model on the FERET face database, which contains more than 10,000 photos of more than 1,000 people, and each person includes photos of different expressions, lighting, poses and age. In this paper, we select 3122 photos of frontal faces in the data set, and normalize them to the size of 128×128. The training set and the test set are separated according to the ratio of 8: 2.

To evaluate the performance of super-resolution reconstruction, visualization comparison and objective indicators are used as the quality evaluation indicators of face image reconstruction.

4.1. Visualization Comparison with Other Models

The proposed algorithm is compared with the existing algorithms in terms of visualization, shown in Figure 2. (a), (b), and (c) respectively show the results of face super-resolution reconstruction in the test set of different algorithms under different sampling factors.

In Figure 2, from left to right are the original images, the input images, the images reconstructed by Bicubic, the images reconstructed by SRCNN, and the images reconstructed by MFFCN. From the comparison diagram, we observe that the images reconstructed by MFFCN are sharper in contour than which reconstructed by Bicubic. Compared with SRCNN, the images have clearer textures. The edges of them are more distinct, especially in the areas of the eyes and lips. As the down-sampling factor increases, the images reconstructed by the Bicubic become ever more blurred. The details around the eyes in the images reconstructed by SRCNN are gradually blurred, especially the eyelids. The images reconstructed by MFFCN do not change obviously with the increase of the sampling factor. They have consistently maintained clear textures and sharp edges.

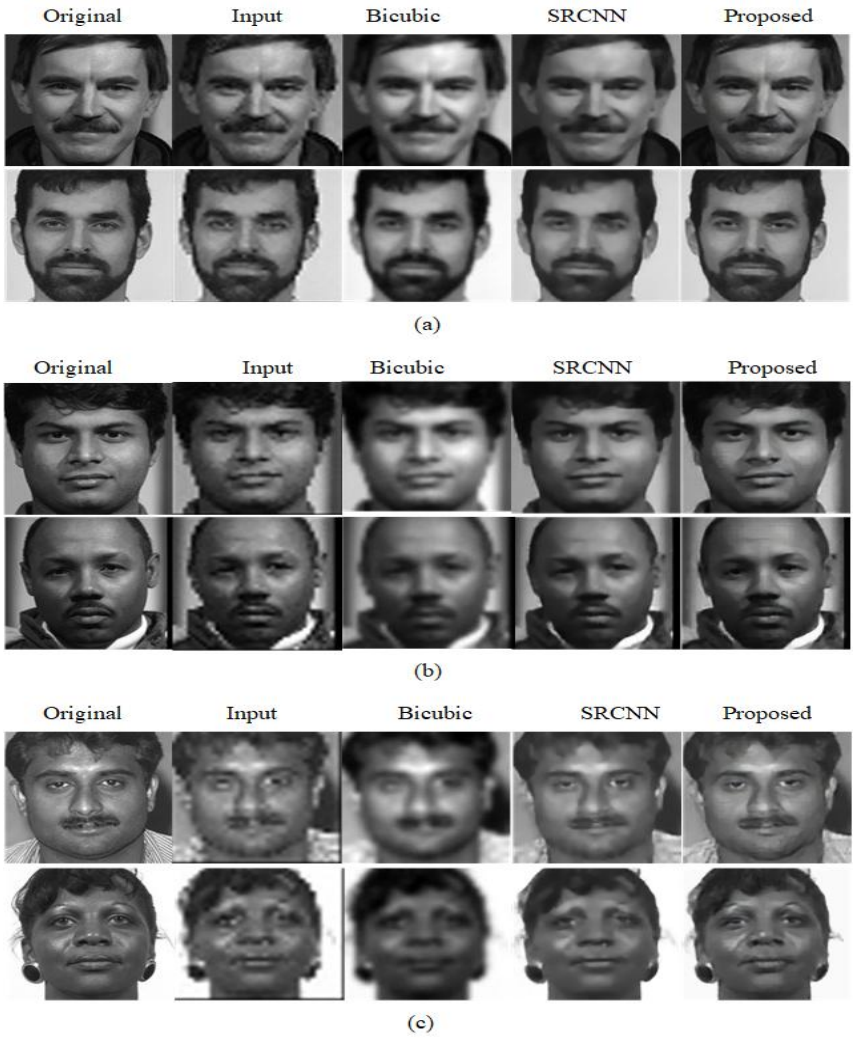

Figure 2: Comparison of images reconstructed results. (a) The down-sampling factor equals 2. (b) The down-sampling factor equals 3. (c) The down-sampling factor equals 4.

4.2.Quantitative Comparison

We measure the reconstruction effect from two objective evaluation indicators: structural similarity (SSIM) and peak signal-to-noise ratio (PSNR).

4.2.1. SSIM

SSIM can evaluate the similarity of two pictures effectively and objectively. It measures in terms of brightness, contrast and structure, which is more consistent with the visual recognition and perception characteristics of human eyes. Range of SSIM is between 0 and 1. The larger the SSIM value, the higher the similarity of the two pictures. When the two images are the same, the SSIM value is 1. The SSIM of the reconstructed image I and the original image I_{HR} can be obtained according to equation (8), where μ_I and $\mu_{I_{HR}}$ are the average of the images I and I_{HR}. σ_I^2 and $\sigma_{I_{HR}}^2$ are the variance of I and I_{HR}, $\sigma_{II_{HR}}$ is the covariance of I and I_{HR}. $c_1=(k_1L)^2$, $c_2=(k_2L)^2$, they are constants. L is the dynamic range of image pixels, $k_1=0.01$, $k_2=0.03$.

$$\mathrm{SSIM}(I, I_{HR}) = \frac{(2\mu_I\mu_{I_{HR}} + c_1)(2\sigma_{II_{HR}} + c_2)}{(\mu_I^2 + \mu_{I_{HR}}^2 + \mathrm{c}_1)(\sigma_I^2 + \sigma_{I_{HR}}^2 + c_2)} \tag{8}$$

Table 1: Comparison of SSIM values of various algorithms.

Factor SSIM	Algorithms		
	Bicubic	SRCNN	MFFCN
2	0.9205	0.9644	0.9736
3	0.8047	0.9316	0.9484
4	0.7860	0.9023	0.9234

Table 1 shows the SSIM values between the original images and the reconstructed images of various algorithms. As is observed from the table, compared with other algorithms, the SSIM values of this method are the largest. When the down-sampling factor is 4, compared to the traditional Bicubic algorithm, the SSIM value of the algorithm set out in the present is 0.13 higher than the Bicubic. Compared with SRCNN, it is 0.02 higher than it. It indicates that the images generated by MFFCN are more realistic and have the highest similarity with the original images.

4.2.2. PSNR

PSNR is another objective indicator for evaluating image quality. It directly compares the pixel differences between the two pictures. The larger the value, the better the image quality. Assuming that both the sizes of two images I and I_{HR} are m×n, the mean square error (MSE) between them is:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - I_{HR}(i,j)]^2 \tag{9}$$

Calculate PSNR through MSE, the formula is as equation (10):

$$PSNR = 10 \cdot \log_{10}(\frac{255^2}{MSE}) \tag{10}$$

Table 2: Comparison of PSNR values of various algorithms.

Factor PSNR (dB)	Algorithms		
	Bicubic	SRCNN	MFFCN
2	32.69	37.98	39.13
3	27.42	35.89	36.37
4	26.41	32.68	34.45

From Table 2, we can see that the algorithm we proposed has the largest PSNR value and the quality of the image is the best. When the down-sampling factor is 4, compared with the traditional Bicubic algorithm, our algorithm has a value of 7.74dB higher than it. It is 1.77dB higher than the approach of SRCNN.

5. Conclusions

This paper presents MFFCN for face super-resolution reconstruction. The proposed framework supports images of low-resolutions at different scales. A skip-connection network structure is proposed to extract deep and shallow features from the multiple convolutional layers, and reconstruct the super-resolution face images. The feature loss term is utilized in training to get clear textures. The experiments on the FERET database demonstrate its advantages over existent methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61661006.

References

[1] Sung Cheol Park, Min Kyu Park and Moon Gi Kang, "Super-resolution image reconstruction: a technical overview," in IEEE Signal Processing Magazine, vol. 20, no. 3, pp. 21-36, May 2003.

[2] Xin Li and M. T. Orchard, "New edge directed interpolation," Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101), Vancouver, BC, Canada, 2000, pp. 311-314 vol.2.

[3] M. Unser, A. Aldroubi and M. Eden, "Fast B-spline transforms for continuous image representation and interpolation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 3, pp. 277-285, March 1991.

[4] R. R. Schultz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," in IEEE Transactions on Image Processing, vol. 5, no. 6, pp. 996-1011, June 1996.

[5] M. Irani, S. Peleg," Improving resolution by image registration," Cvgip-Graphical Models Image Process., 53 (1991), pp. 231-239.

[6] Jian Sun, Zongben Xu and Heung-Yeung Shum, "Image super-resolution using gradient profile prior," 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1-8.

[7] Stark, Henry, and P. Oskoui. "High-resolution image recovery from image-plane arrays, using convex projections." Journal of the Optical Society of America A Optics & Image Science 6.11(1989), pp. 1715-1726.

[8] R. Timofte, V. De and L. V. Gool, "Anchored Neighborhood Regression for Fast Example-Based Super-Resolution," 2013 IEEE International Conference on Computer Vision, Sydney, NSW, 2013, pp. 1920-1927, doi: 10.1109/ICCV.2013.241.

[9] Hong Chang, Dit-Yan Yeung and Yimin Xiong, "Super-resolution through neighbor embedding," Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., Washington, DC, USA, 2004, doi: 10.1109/CVPR.2004.1315043.

[10] J. Huang, A. Singh and N. Ahuja, "Single image super-resolution from transformed self-exemplars," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 5197-5206, doi: 10.1109/CVPR.2015.7299156.

[11] W. T. Freeman, T. R. Jones and E. C. Pasztor, "Example-based super-resolution," in IEEE Computer Graphics and Applications, vol. 22, no. 2, pp. 56-65, March-April 2002, doi: 10.1109/38.988747.

[12] Jianchao Yang, J. Wright, T. Huang and Yi Ma, "Image super-resolution as sparse representation of raw image patches," 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1-8, doi: 10.1109/CVPR.2008.4587647.

[13] S. Schulter, C. Leistner and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 3791-3799

[14] C. Dong, C. C. Loy, K. He and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 295-307, 1 Feb. 2016

[15] Dong, Chao, C. C. Loy, and X. Tang. "Accelerating the Super-Resolution Convolutional Neural Network." (2016).

[16] H. Nie, Y. Lu and J. Ikram, "Face Hallucination via Convolution Neural Network," 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), San Jose, CA, 2016, pp. 485-489.

[17] Simonyan, Karen, and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." Computer ence (2014).

[18] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015.

[19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. ICLR, 2015.