FACE HALLUCINATION BASED ON KEY PARTS ENHANCEMENT

*Ke Li, Bahetiyaer Bare, Bo Yan**

Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, China

Bailan Feng, Chunfeng Yao

Noah's Ark Laboratory, 2012Labs Huawei Technologies Co., Ltd., Beijing, China

ABSTRACT

Face hallucination aims to generate a high resolution face from a low resolution one. Generic super resolution methods can not solve this problem well, because human face has a strong structure. With the rapid development of the deep learning technique, some convolutional neural networks (CNNs) models for face hallucination emerged and achieved state-of-the-art performance. In this paper, we proposed a five-branch network based on five key parts of human face. Each branch of this network aims to generate a high resolution key part. The final high resolution face is the combination of the five branches' output. In addition, we designed a gated enhance unit (GEU) and cascade it to form our network architecture. Experimental results confirm that our method can generate pleasing high resolution faces.

***Index Terms*—** Face hallucination, Convolutional neural networks, Super-resolution

1. INTRODUCTION

Face hallucination is a basic problem in face analysis field, which can improve the performance of face related tasks such as face attributes analysis, face alignment, face recognition and so on. The purpose of face hallucination is to generate a high resolution (HR) face image from a low resolution (LR) one, which is a special case of image super-resolution (SR). Compared with the generic image SR methods [1, 2], the input image of the face hallucination is often fuzzy, so the face hallucination needs more prior knowledge, and the challenges are relatively large.

Early face hallucination methods are proposed under the assumption that the face has a small variation. Wang $et\ al.$ [3] have addressed the mapping between LR and HR faces by feature transformation. Yang $et\ al.$ [4] solved the face hallucination problem by sparse representation, which can recover the high resolution human face accurately from low dimensional projection. Yang $et\ al.$ [5] solved the problem by finding the matching relationship between HR and LR image blocks, and greatly improved the performance of face hallucination technology.

Fig. 1. Proposed method. For an input LR image, we send them to our proposed five-branch network. Each subnetwork generates a HR key part and its mask. Finally, we combine the output of each branch based on output mask and get the HR output.

In recent years, with the wide application of deep learning technology in computer vision field, researchers have solved the problem of face hallucination by deep learning technology and obtained better performance. Zhou $et\ al.$ [6] designed a bi-channel convolutional neural networks (CNNs) to achieve the mapping between LR and HR faces, and achieved good results in the case of large changes in face. In [7], Zhu $et\ al.$ proposed a deep cascaded bi-network for face hallucination. The network also achieved good hallucination results when the input face images were very small. Yu $et\ al.$ [8] proposed a dual network structure consisting of an upsampling network and a discriminative network. With the help of the discriminative network, the performance of the upsampling network can be significantly improved when upsampling small face image at $8\times$ magnification factor. Cao $et\ al.$ [9] proposed a face hallucination model based on human attention, and optimized this model by deep reinforcement learning. This method achieves the best performance on several commonly used datasets.

Inspired by the recent successes of CNNs in computer vision tasks, we propose a five-branch network for face hallucination. Our method is shown in Figure 1. Each subnetwork aims to generate the different HR key part of human face.

*This work was supported in part by NSFC (Grant No.: 61522202; 61772137) and Huawei Technologies Co., Ltd (Contract No.: YBN2017050058).

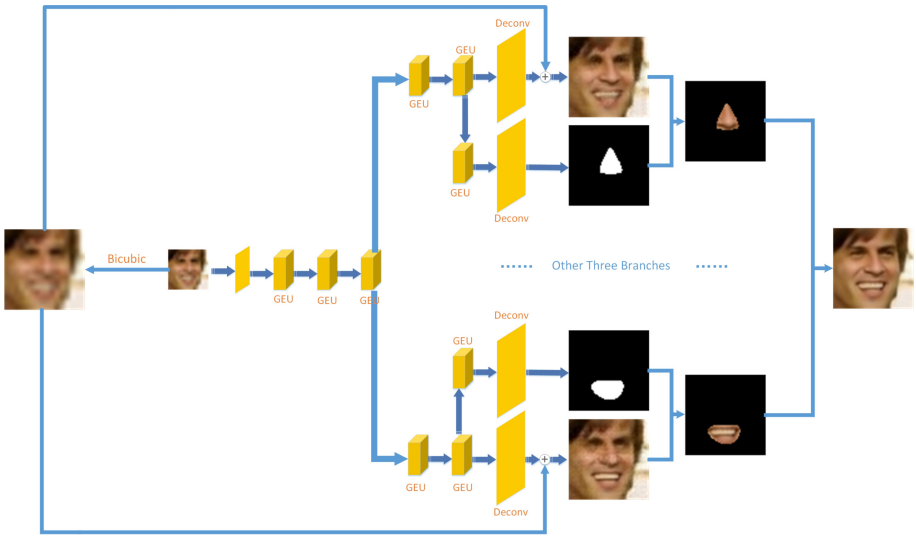

Fig. 2. Our network architecture. Our network has five branches, namely eyes branch, nose branch, mouth branch, eyebrows branch, and the remaining part branch. Each branch outputs the corresponding high resolution key parts

Because the pattern of different key parts is different, using five-branch network to learn each key part individually is better than just using one network to learn a mapping between LR and HR faces. Experimental results show that our method can generate pleasing HR results. The main contributions of our can be summarized as:

1) We propose five-branch CNNs model for face hallucination based on five key parts of human face, namely eyes, nose, eyebrows, mouth, and the remaining part. Each subnetwork is responsible for generate an HR key part.
2) For each subnetwork, we design a gated enhance unit (GEU) and cascaded it to form our network architecture. The main idea of designing a gated block is to connect the input of gated block and output of gated block with trainable weight values.
3) We introduce a trick for training a five-branch network and prove the superiority of the proposed model with experimental results.

The rest of the paper is organized as follows. In Section 2, we describe the proposed five-branch network. Then we present experimental results and discuss the results of proposed deep CNNs model in Section 3. Finally, we draw conclusions in Section 4.

2. OUR METHOD

In this section, we first introduce our proposed method. Then, we introduce proposed five-branch network architecture.

2.1. proposed method

Our method is shown in Figure 1, for an input LR image, we generate an HR image through our novel method. Firstly, we send the input LR image to the proposed five-branch network. Each branch is responsible for generating a specific key part of human face. These five key parts are: eyes branch, nose branch, mouth branch, eyebrows branch, and the remaining part branch. Then, each branch outputs the HR key part and its mask. Based on generated masks, our model combines the each branch's ouput and outputs the HR face image.

Different from the methods in [7, 9], our model separate the face image to five key parts and learn the mapping between LR and HR for each key part separately. Theoretically, it is better than just training a CNNs to learn the mapping between LR and HR faces directly. In addition, different from region based enhancement methods, our method can find the five key parts correctly. Thus the predicted HR face image is better than existing methods.

2.2. Network Architecture

Our network architecture is demonstrated in Figure 2, which consists of several GEUs, several deconvolution layers, and several skip connections [10]. The network architecture of each branch is similar, so we just draw two branches' architecture for easy to view. For easy to train, each branch learns a residual image. Therefore, the output of each branch is the combination of bicubic interpolated image and residual image. Based on mask, we can get the HR key part image. In the end, we combined each HR key part image and get the final HR face. We discuss each layer as follows.

We proposed a GEU in order to enhance the hallucinating results. In Figure 3, we demonstrated the architecture of the proposed GEU. As we can see from this figure, the GEU learns a weight value g to combine the input signal x and the output of the third convolution layer y . Thus the output of

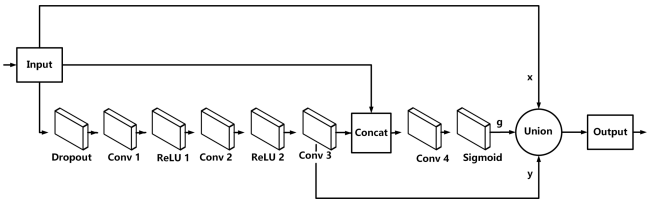

Fig. 3. GEU. We cascade GEUs to form our deep CNNs model. GEU learns a weight value g for combing the input signal x and output y of the third convolution layer.

the GEU is the weighted combination of x and y. We define the output of the union layer as:

$$Output = g \times y + (1 - g) \times x. \quad (1)$$

The convolutional layers of our model are used to extract local features and the deconvolution layers in our model are used to upsample feature maps. In our proposed model, ReLU is employed as the activation function of deconvolution layers and convolutional layers. In order to fasten the converging speed of our model, we add several standard dropout layers. By preventing complex co-adaptations on training data, dropout layer can reduce overfitting problem in neural network training. The skip connection in our model adds the learned residual image and the input upsampled LR key part of human face. Our proposed model has several concatenate layers to fuse the feature maps, which are connected to it. We set the loss function of our model as mean squared error (MSE) followed by previous works [11, 2]. We train our five-branch network by minimizing the loss function for N training samples as follows:

$$L(w_0, w_b, w_{bm}) = \frac{1}{N}\sum_{i=1}^{N}[\sum_{b=1}^{5}(L_{bm} \times ||F(I; w_0, w_b) - L_f||^2 + a||P(I; w_0, w_b, w_{bm}) - L_{bm}||^2)] \quad (2)$$

where I and L_f denote the LR face image and ground truth HR face image respectively; and $F(I; w_0, w_b)$ represents the predicted HR face image, which is generated from our model with branch-shared weights w_0 and independent weights w_b; and $P(I; w_0, w_b, w_{bm})$ and L_{bm} represent predicted the b^{th} mask image and its label respectively; and w_{bm} denotes the independent weights, which are used to predict the b^{th} mask. In order to learn the HR face image and mask image together, we set a weight value a to balance them. In our experiment, we set a to 0.01. The loss function is minimized using stochastic gradient descent.

3. EXPERIMENT

In this section, we firstly introduce train and test datasets. Then, we introduce training details. Finally, we demonstrate experimental results and discuss the priority of our model.

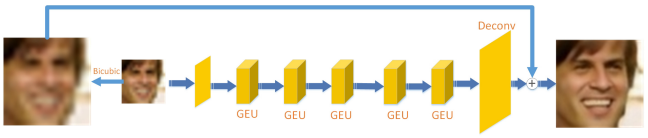

Fig. 4. Pre-trained network. For initialize parameters of our model, we pre-train a network to learn the mapping between LR and HR faces.

3.1. Datasets

We used CelebA [12] dataset to train our model. This dataset contains 202,599 face images with 5 landmarks and 40 binary attributes per image. We used all images of CelebA dataset to form our train set and used images in LFW [13] dataset to form the test set. We used landmarks information of CelebA dataset to crop five key parts for each image of train set. These five key part images are the training images for each branch of our proposed network. Each branch is trained to learn HR key part of human face.

3.2. Training Details

In the process of running experiment, we found that if we directly train each branch separately, the network can not converge. So, we introduce a trick for training our five-branch network. Firstly, we pre-train a network (shown in Figure 4) in order to learn the mapping between LR and HR face images. Then, we use parameters of this network to initialize each branch of our five-branch network except for mask generation part, which is randomly initialized. Finally, we train each branch of our network together and it can converge.

Training parameters were set as follows. We initialized the learning rate to 0.0001 and decreased it by 10 for every 15 epoches. In order to prevent the large gradient change, we used adjustable gradient clipping [2] to limit the parameters' gradient in $[-\theta/\gamma, \theta/\gamma]$, where γ denotes current learning rate and θ is set to 0.005. We set the batch size, momentum factor, and decay factor as 32, 0.9, and 0.0005 respectively. Finally, we train our network for 370,000 iterations. We employed Tensorflow [14] to implement our model with GTX1070 GPU. It took 24 hours to train our proposed model.

3.3. Experimental results

We tested our proposed model on the images from LFW dataset [13]. We design the following experiment to test our model. Firstly, we crop the face regions from test set. Then, we downsample the cropped face region images to 16×16 size. In our experiment, we upsample the face images to 64×64 with our trained $4\times$ upsample model and report the qualitative results.

We compared our method with bicubic interpolation, SRCNN [1], and VDSR [2]. We generate compared images us-

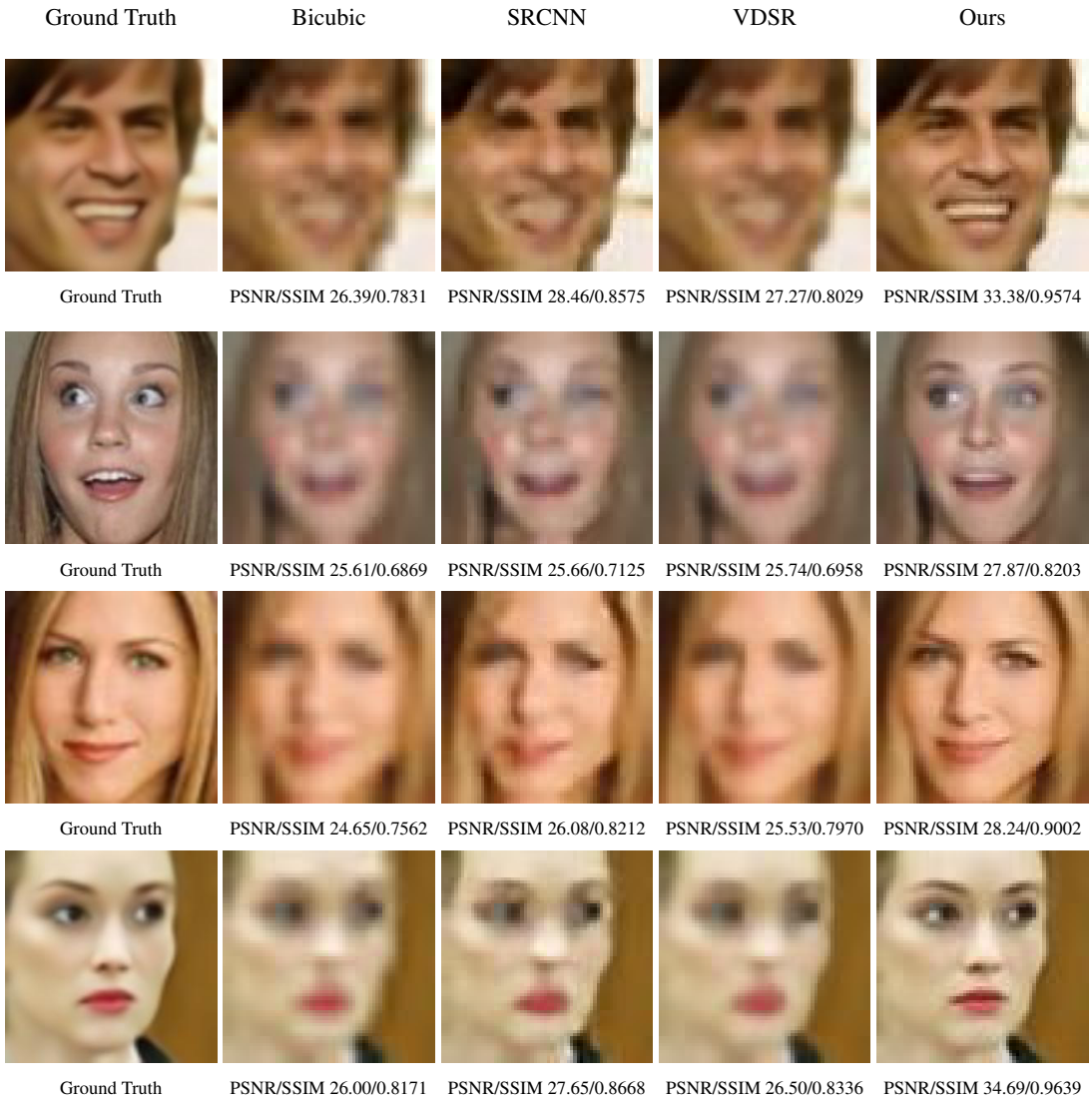

Fig. 5. Visual results for $\times 4$ on test dataset. From left to right: ground truth, bicubic interpolation, SRCNN [1], VDSR [2], and Ours.

ing publicly available code of SRCNN and VDSR. Because SRCNN model has two version, we choose the version, which is trained on imagenet dataset [15]. Visual comparisons between our model and other methods are demonstrated in Figure 5. We also calculated PSNR and SSIM [16] values between generated HR results and ground truth images, and put these values under each image. As we can observe from this figure, on contrary to serious blurred image that are generated from other methods, our method can generate pleasing results with higher SSIM and PSNR values. The results of our proposed method are more similar to ground truth images. This is because our network model is designed based on key parts enhancement, so each branch of our model can learn different pattern for each key part of human face. Therefore, its performance is better than one-branch network, which learns a direct mapping between LR and HR face images. More importantly, our proposed GEU can help to achieve better results than other methods. Overall, our proposed five-branch network enhanced each key part obviously and achieved state-of-the-art performance.

4. CONCLUSION

In this paper, we proposed a five-branch network for face hallucination. This five-branch network aims to generate a high resolution image of human face's five key parts. Thus the final high resolution image is obtained by combining each branch's output. Moreover, we proposed a gated enhance unit to further improve the performance of face hallucination. Besides that, we introduce a trick for training a five-branch network. Experimental results show that our proposed model can generate pleasing results in comparison with state-of-the-arts.

5. REFERENCES

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

[2] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu 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.

[3] Xiaogang Wang and Xiaoou Tang, "Hallucinating face by eigentransformation," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 3, pp. 425–434, 2005.

[4] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[5] Chih-Yuan Yang, Sifei Liu, and Ming-Hsuan Yang, "Structured face hallucination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1099–1106.

[6] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin, "Learning face hallucination in the wild.," in *AAAI*, 2015, pp. 3871–3877.

[7] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang, "Deep cascaded bi-network for face hallucination," in *European Conference on Computer Vision*. Springer, 2016, pp. 614–630.

[8] Xin Yu and Fatih Porikli, "Face hallucination with tiny unaligned images by transformative discriminative neural networks.," in *AAAI*, 2017, pp. 4327–4333.

[9] Qingxing Cao, Liang Lin, Yukai Shi, Xiaodan Liang, and Guanbin Li, "Attention-aware face hallucination via deep reinforcement learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 690–698.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[11] Chao Dong, Chen Change Loy, and Xiaoou Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision*. Springer, 2016, pp. 391–407.

[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.

[13] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep., Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[14] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., "Tensorflow: A system for large-scale machine learning.," in *OSDI*, 2016, vol. 16, pp. 265–283.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[16] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.