Adaptive Aggregation Network for Face Hallucination

Jin Guo, Jun Chen(✉), Zhen Han, Han Liu, Zhongyuan Wang, and Ruimin Hu

National Engineering Research Center for Multimedia Software,
School of Computer Science, Wuhan University, Wuhan 430072, China
gjjggjjg@163.com, hanzhen_1980@163.com, liuhanooo@163.com,
wzy_home@163.com, chenj@whu.edu.cn, hrm1964@gmail.com

Abstract. Face hallucination refers to obtaining a clean face image from a degraded ones. The degraded face is assumed to be related to the clean face through the forward imaging model that account for blurring, sampling and noise. In recent years, many methods have been proposed and improved well progress. These methods usually learn a regression function to reconstruct the entire picture. However, there are huge differences among the optimal learned regression functions in different regions. In other words, the learned regression function needs to process all regions, which makes it difficult to reconstruct a satisfactory picture. As a result, the reconstructed images in some regions are relatively smooth. In order to address the problem, we present a novel face hallucination framework, called Adaptive Aggregation Network (AAN), which uses the aggregation network to guide face hallucination. Our network contains two branches: aggregation branch and generator branch. Specifically, our aggregation branch can explore regression function from low-resolution (LR) to high-resolution (HR) images in different regions, and aggregate the regions by the similarity of the regression function. Then generator module can be used to make a specific hallucination on the selected regions to get a better reconstruction result. After evaluating on datasets, our model was proved to be above the state-of-the-art methods in terms of effectiveness and accuracy.

Keywords: Face hallucination · Adaptive aggregation
Regression function

1 Introduction

Face hallucination, as a representative of low-level vision tasks, is the process of reconstructing a clean face image from the degraded observation. It is not only

Research supported by National Key R&D Program of China (No. 2017YFC0803700), National Nature Science Foundation of China (U1736206, U1611461, 61671332), Natural Science Foundation of Hubei Province (2016CFB573), Hubei Province Technological Innovation Major Project (2016AAA015, 2017AAA123).

R. Hong et al. (Eds.): PCM 2018, LNCS 11165, pp. 190–199, 2018.
https://doi.org/10.1007/978-3-030-00767-6_18

a fundamental problem in face analysis, but can be used as a preprocessor for tasks such as face recognition [1], face alignment [2]. In practical applications, however, the face images captured by surveillance cameras are generally of poor quality and difficult to use directly.

Over the past few decades, many conventional methods [3–6] have been proposed to solve the problem. They assume the correlation between the degraded face image with clean ones, and focus on learning a mapping from degraded images to clean images. However, most methods involve a large number of optimization problems during the reconstruction phase, making it difficult to implement high-performance applications. Due to the complex environment of the degenerative process, the consistency of the hypothesis of degraded face images with clean images is not well. Therefore, the results produced are often unsatisfactory.

Recently, with the development of convolutional neural networks, many methods [7–10] based on deep learning have been used for image reconstruction. Face prior knowledge and spatial structure information are often used as additional information for face hallucination. Despite their high reconstruction quality, most of the methods typically suffer from two major drawbacks. First, there are huge differences in the structure of different regions, making it difficult to generate a mapping that satisfies all regions. Second, in the noisy environment, the prior information and spatial structure of face will be destroyed which makes it difficult to generate satisfactory results.

Therefore, how to reconstruct an HR face image in a noisy environment becomes a difficult problem. Inspired by the recent success of aggregation network in computer vision tasks [11,17–19], we propose the Adaptive Aggregation Network to deal with noise face hallucination. Our network contains two branches: aggregation branch and generator branch. The aggregation branch can cluster face images into two robust regions in a data-driven manner through the similarity of the regression function. Then the generator branch can be used to make a specific face hallucination of the selected regions.

The main contribution of this paper is that we propose an effective model to deal with the face hallucination in noise environments. The noise face hallucination is often difficult to reconstruct a satisfactory result due to the complex degradation process and the destruction of the face prior structure. Compared with other methods, our method not only provides robust face structure information under noise conditions, but turns a complex face hallucination problem into two relatively simple sub-problems. The empirical results show that our designed network surpasses the state-of-the-art methods in terms of effectiveness and efficiency.

2 Related Work

2.1 Face Hallucination and Image Super-Resolution

Face hallucination is a special case of image super resolution, which introduces face prior structure information to reconstruct face images. Early techniques assumed

that the face was in a controlled setting with small variations. Ma et al. [3] utilize face priors information to reconstruct HR face by solving a constrained least squares problem (known as Least Square Representation (LSR)). Yang et al. [4] thought that low-resolution and high-resolution faces have similar sparse priors and reconstruct HR faces through the low-dimensional projections. Later, on the basis of locality and sparseness, Jiang et al. [5,6] proposed a Local Constraint Representation (LcR) method to obtain a better reconstruction face. However, these methods require the face to be landmark detection beforehand which often cannot achieve good results when the images are seriously degraded.

In recent years, many deep learning methods have been used for image hallucination and have achieved great progress. In particular, Dong et al. [7] first proposed a super-resolution convolutional network (SRCNN) for image reconstruction through equally performing sparse coding. Kim et al. [8] proposed a deep convolutional network to achieve better reconstruction performance by skipping connections and learning residuals between HR with LR. Zhou et al [10] proposed a bi-channel convolutional neural network for facial hallucination. They point out the importance of input image, and use full-connected layers to restore HR face. Tuzel et al. [9] added global face information to the network and reconstructed the face image by considering global and local constraints.

2.2 Adaptive Aggregation Network

Adaptive aggregation network developed in recent work and has benefited various tasks [11–13], such as object classification [11] and human pose estimation [12,13]. Since contextual information is important for computer vision problems, most of these works attempt to acquire dense features adaptively by focusing on the top information. Recent proposed Residual Attention Network [11] achieves state-of-the-art results on image classification task. A deep network module capturing top information is used to adaptive aggregation module. The aggregation module is applied to the input image to get important regions and then feed to another deep network module for classification. Chen et al. [12] used a stacked hourglass network structure to fuse information from multiple-context to predict human pose, and benefits from global and local information.

3 Proposed Method

3.1 Problem Formulation

We denote the input noise face image and corresponding clean image as X and Y, respectively. The process of getting the noise LR image from HR image can be modeled as

$$X = DBY + N \tag{1}$$

where D, B and N respectively denote downsampling operator, blur operator and additive noise operator.

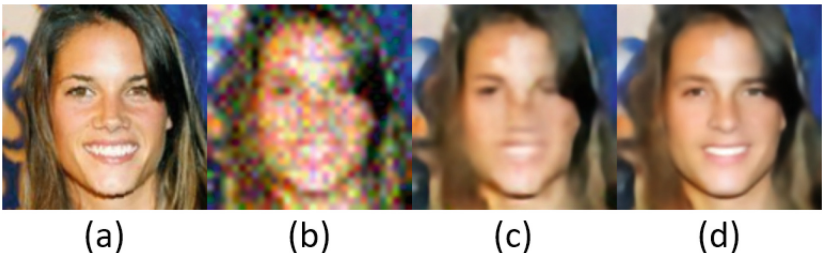

Fig. 1. Examples of experimental results. (a) Target image. (b) LR face image with noise. (c) Result of directly training [8]. (d) Result of our method.

For a given LR image, the face hallucination network F is expected to predict a hallucinated face as similar as the ground truth HR image by minimizing the mean square error (MSE).

$$L = \frac{1}{N_I}||F(X) - Y||_F^2 \tag{2}$$

However, we found that the result obtained by directly training [8] on the image domain (direct network) is not satisfactory. In Fig. 1, we show an example of a hallucinated face image which is used in the training process. In Fig. 1(c) we see that the hallucinated image by directly training has severe smooth in some details. In general, we observe that the learned regression function is performed on the entire picture, which means it need take into account various situations. However, the optimal learned regression function in different regions is different. In other words, the regression function need deal with all regions, which makes it hard to learn well. As a result, the reconstruction results in some areas are relatively smooth.

3.2 The Network Architecture

In order to solve the problem, we propose an effective model for noise face hallucination. The detailed structure of the network is shown in Fig. 2. It is divided into two branches: aggregation branch and generator branch. According to the similarity of the regression parameters, the aggregation branch can adaptively aggregate face regions into two categories. Then the generator branch can be targeted to recover HR images for selected regions.

We denote the networks input as X. The network can be summarized as

$$L = \frac{1}{N_I}||(G_1(X, \xi_1) - Y)M(X, \Phi) + (G_2(X, \xi_2) - Y)(1 - M(X, \Phi))||_F^2 \tag{3}$$

where M, G represents the output of aggregation branch and generator branch and ξ, Φ denotes the parameters to be learned. The aggregation branch aggregates the face regions into two categories as $M(X, \Phi)$ and $1 - M(X, \Phi)$. Then

each generator branch can be targeted to recover HR images for selected regions. Finally, the reconstructed faces of different generator branches are added to generate the final output.

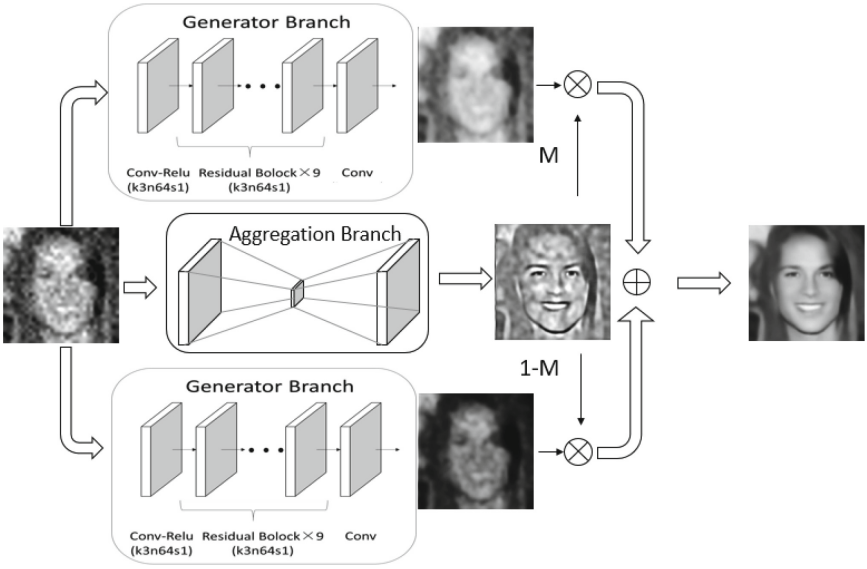

Fig. 2. The detailed structure of network

In aggregation branch, the aggregation network can not only serve as a mask selector during forward inference, but also can guide generator branch gradient update during backward propagation. In the generator branch, the gradient for input image is:

$$\frac{\mathrm{d}(G_1(X,\xi_1)-Y)M(X,\Phi)}{\mathrm{d}\xi_1} = M(X,\Phi)\frac{\mathrm{d}G_1(X,\xi_1)}{\mathrm{d}\xi_1} \tag{4}$$

This property allows the generator module to better reconstruct the selected regions. The aggregation branch can prevent unrelated data to update the parameters of the generator branch.

In addition, the excellent recovery of the generator branches in turn causes the aggregation branches to cluster more similar structures into the region. Assuming that the parameters of our generator branch are fixed, the loss function of our network is:

$$\arg\min_{\Phi} ||G_1(X,\xi_1)M(X,\Phi) + G_2(X,\xi_2)(1-M(X,\Phi)) - Y||_F^2 \tag{5}$$

In order to minimize the loss, our aggregation branch can get greater weight for better reconstruction regions, which means the aggregation branch cluster more similar structures into categories. Through the similar process of alternating minimization, our network is constantly optimized to generate better results.

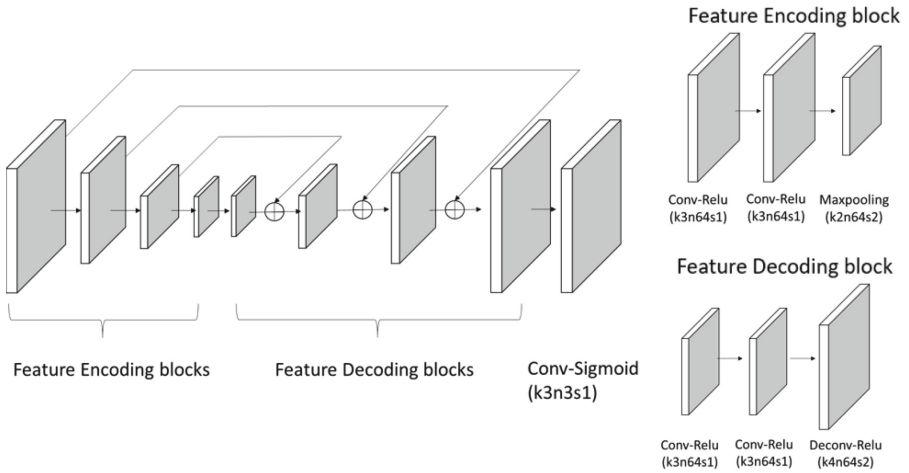

Fig. 3. The detailed structure of aggregation branch

Aggregation Branch. Our aggregation branch adopts a similar Hour-Glass structure, which used to human pose estimation [12,13], to cluster face regions. The detailed structure of the aggregation network is shown in Fig. 3. The network consists of multiple Feature Encoding blocks and Feature Decoding blocks. Each pair of feature encoding block and feature decoding block brings the feature representation into a new spatial scale, so that the whole network can process information on different scales. To effectively consolidate and preserve spatial information in different scales, the hourglass block uses a skip connection mechanism between symmetrical layers. Specifically, the feature information of input is quickly collected through multiple feature encoding blocks, and the feature decoding blocks amplifies the feature information to the same scale as the input. Finally a sigmoid layer normalizes the output range to [0, 1] to get a mask. Compared with CNN, the Hour-Glass structure can obtain a wider range of input information with less computational cost.

Generator Branch. For the generator branch, it be served to generate face images and can be adapted to any state-of-the-art network structures. Considering the success of residual network [14] in computer vision tasks, we choose the residual block as our network basic unit to hallucination face images. The generator branch consists of a cascade of multiple residual blocks. Each residual block contains two convolutional layers, then the input data is passed through a skip connection for element-wise sum with the output of the last convolutional layer.

4 Experiments

4.1 Dataset

We evaluated extensive experiments in Celebrity Face Attributes (CelebA) dataset [1]. The CalebA dataset contains 202,599 face images with 10,177 celebrity identities which is a very common dataset for face-related training. In our experiment, we first aligned the images with Mtcnn method [15] and crop the center image patches with size of 128×128 as the HR face images to be processed. Then we generate LR images by applied blur operation, down-sampling operation, and noise adding operation on the HR image. We set fixed Gaussian blur kernel b = 1.0, down-sampling factor 4 and we consider three noise levels $\sigma = 5$, 15 and 25. We select 22k faces from the dataset, of which 20k face images are trained and the rest are used for testing.

4.2 Parameter Settings

In our aggregation network, the number of feature encoding blocks and feature decoding blocks is 4, and each residual block consists of 2 convolution layers with kernels size of 3×3. In addition to the feature map of input layer and each branch's output layer is 3, the feature map of other layer is 64. For implementation, we train our model with the Tensorflow platform. The model is trained using the Adam optimization algorithm with an initial learning rate of 1e−3. We total train 50 epochs, and the later 20 epochs with learning rate of 1e−4. Training our network on celebA dataset takes about 6 h on 1 Titan X GPU.

Table 1. Quantitative comparison under Gaussian noise.

Method	$\sigma = 5$			$\sigma = 15$			$\sigma = 25$		
	PSNR	SSIM	FSIM	PSNR	SSIM	FSIM	PSNR	SSIM	FSIM
Bicubic	25.0573	0.7538	0.8421	22.9231	0.6607	0.8192	20.5045	0.5423	0.7781
BCCNN	25.5556	0.7640	0.8520	23.1615	0.6672	0.8209	21.7181	0.5946	0.7952
GLN	25.9653	0.7990	0.8785	25.2131	0.7548	0.8532	23.6237	0.7108	0.8353
SRCNN	27.4359	0.8144	0.8857	25.4672	0.7534	0.8532	23.9494	0.6902	0.8265
VDSR	28.1647	0.8389	0.8994	26.0401	0.7767	0.8659	24.4137	0.7061	0.8341
Ours	**28.6401**	**0.8533**	**0.9056**	**26.4691**	**0.7991**	**0.8762**	**24.9496**	**0.7394**	**0.8476**

4.3 Comparisons

We compare our approach with two types of methods: general image super-resolution methods and face hallucination approaches. For general image SR methods, we compare with SRCNN [7] and VDSR [8]. For face hallucination methods, we choose GLN [9] and BCCNN [10] as the contrast methods. Then we use the widely used PSNR (peak signal to noise ratio), SSIM (structural similarity) and FSIM (feature similarity) [16] to evaluate the reconstructed face.

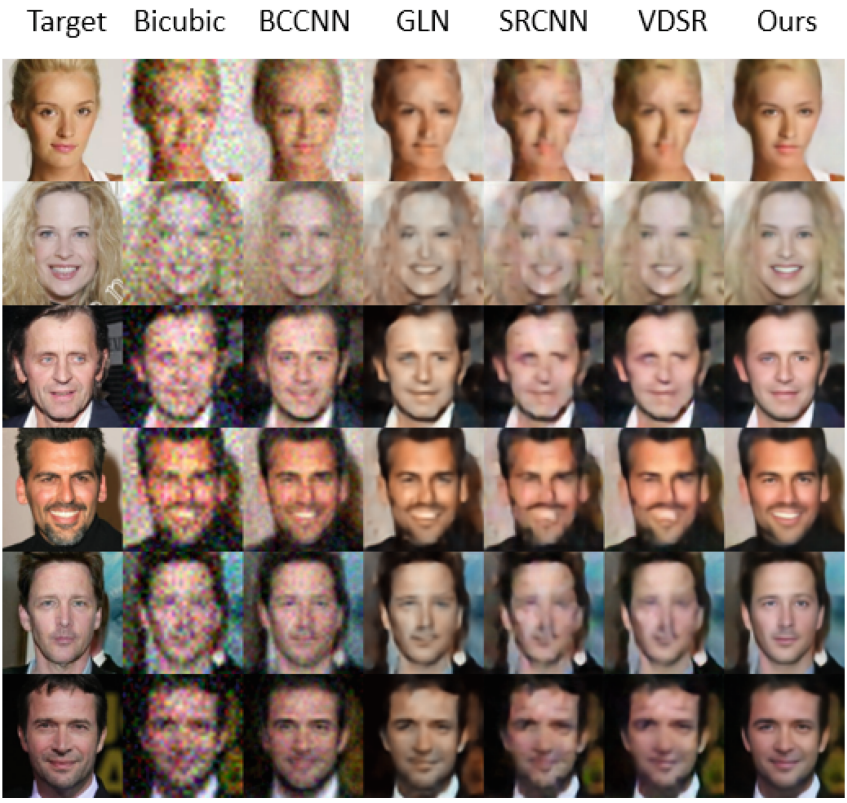

Fig. 4. Comparison of the hallucinated HR images with noise level 25.

4.4 Results

We use test sets to generate several types of LR surfaces with different noise level. Figure 4 shows the performance of our model and comparisons with other methods. It has been observed that Zhou's BCCNN [10] cannot eliminate noise. The final result of BCCNN is partly from the input noise face image, so the network does not converge well to obtain better results. Tuzel's GLN [9], like BCCNN, introduced the structural information of the face. However, in the noisy environment, the prior structural information of the face is not stable enough to converge well to obtain satisfactory results. Dong's SRCNN [7] also cannot remove the noise because there are only 3 convolutional layers and the parameters of the network are too small to produce satisfactory results. Kim's VDSR [8] makes the face clean and has more facial detail compare with SRCNN. However, this method uses the same regression function for all regions, resulting in poor reconstruction of some facial regions. Obviously, our method produces better results which not only removes the noise but preserves more face features information.

Table 1 shows the results of comparing our model with other state-of-the-art methods at different noise level. In terms of PSNR, SSIM and FSIM indicators, our model is much better than all comparison methods.

5 Conclusion

In this paper, we present a novel face hallucination framework which uses adaptive aggregation network to guide noise face hallucination. Our network contains two branches: aggregation branch and generator branch. Specifically, our aggregation branch can explore mapping relationships from LR to HR images in different regions, and aggregate the regions by the similarity of the mapping. Then generator branch can be used to make a specific face hallucination of the selected regions to get a better reconstruction result. The experimental results show that our model achieves state-of-the-art performance in noise face hallucination.

References

1. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: IEEE International Conference on Computer Vision, pp. 3730–3738 (2015)
2. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. IEEE Trans. Pattern Anal. Mach. Intell. **38**(5), 918–930 (2016)
3. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**(11), 2861–2873 (2010)
4. Yang, C.Y., Liu, S., Yang, M.H.: Structured face hallucination. In: Computer Vision and Pattern Recognition, pp. 1099–1106(2013)
5. Jiang, J., Hu, R., Han, Z., Lu, T., Huang, K.: Position-patch based face hallucination via locality-constrained representation. In: IEEE International Conference, pp. 212–217 (2012)
6. Jiang, J., Hu, R., Wang, Z., Han, Z.: Noise robust face hallucination via locality-constrained representation. IEEE Trans. Multimed. **16**(5), 1268–1281 (2014)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 184–199. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_13
8. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654 (2016)
9. Tuzel, O., Taguchi, Y., Hershey, J.R.: Global-local face upsampling network. arXiv preprint arXiv:1603.07235 (2016)
10. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Learning face hallucination in the wild. In: AAAI, pp. 3871–3877 (2015)
11. Wang, F., et al.: Residual attention network for image classification. arXiv preprint arXiv:1704.06904 (2017)
12. Chen, Y., Shen, C., Wei, X.S., Liu, L., Yang, J.: Adversarial posenet: a structure-aware convolutional network for human pose estimation. CoRR, abs/1705.00389, 2 (2017)

13. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
15. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
16. Zhang, L., Zhang, L., Mou, X., Zhang, D.: FSIM: a feature similarity index for image quality assessment. IEEE Trans. Image Process. **20**(8), 2378–2386 (2011)
17. Sun, Y., Liang, D., Wang, X., Tang, X.: Deepid3: face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2015)
18. Caicedo, J.C., Lazebnik, S.: Active object localization with deep reinforcement learning. In: IEEE International Conference on Computer Vision, pp. 2488–2496 (2015)
19. Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D.: DRAW: a recurrent neural network for image generation. arXiv preprint arXiv:1502.04623 (2015)