Journal of Real-Time Image Processing (2020) 17:1927–1937
https://doi.org/10.1007/s11554-020-01009-3

SPECIAL ISSUE PAPER

A deep attention-based ensemble network for real-time face hallucination

Dongdong Liu[1] · Jincai Chen[1,2] · Zhenxing Huang[1] · Ni Zeng[3] · Ping Lu[2] · Lin Yang[4] · Haofeng Wang[4] · Jinqiao Kou[4] · Min Wu[4]

Received: 31 March 2020 / Accepted: 8 August 2020 / Published online: 17 August 2020

Abstract
Face hallucination (FH) aims to reconstruct high-resolution faces from low-resolution face inputs, making it significant to other face-related tasks. Different from general super resolution issue, it often requires facial priors other than general extracted features thus leading to fusion of more than one kind of feature. The existing CNN-based FH methods often fuse different features indiscriminately which may introduce noises. Also the latent relations among different features which may be useful are taken into less consideration. To address the above issues, we propose an end-to-end deep ensemble network which aggregates three extraction sub-nets in attention-based manner. In our ensemble strategy, both relations among different features and inter-dependencies among different channels are dug out through the exploitation of spatial attention and channel attention. And for the diversity of extracted features, we aggregate three different sub-nets, which are the basic sub-net for basic features, the auto-encoder sub-net for facial shape priors and the dense residual attention sub-net for fine-grained texture features. Conducted ablation studies and experimental results show that our method achieves effectiveness not only in PSNR (Peak Signal to Noise Ratio) and SSIM (Structural Similarity Index) metrics but more importantly in clearer details within both key facial areas and whole range. Also results show that our method achieves real-time hallucinating faces by generating one image in 0.0237s.

Keywords Face hallucination · Attention mechanism · Residual learning · Ensemble model

1 Introduction

Face hallucination is a domain-specific issue which aims to generate high-resolution (HR) faces from low-resolution (LR) ones. It is critical and fundamental in face analysis and can facilitate several face-related tasks, such as video surveillance [12], face alignment [2, 13], face recognition [27] since most current techniques would degrade when faces are in low resolution.

✉ Jincai Chen
jcchen@hust.edu.cn

Dongdong Liu
ddliu@hust.edu.cn

Zhenxing Huang
huangzx@hust.edu.cn

Ni Zeng
zengni18s@ict.ac.cn

Ping Lu
luping06@hust.edu.cn

Lin Yang
hsjyl@126.com

Haofeng Wang
wanghaofeng@sina.com

Jinqiao Kou
Kjq-smail@163.com

Min Wu
wm_lucy@sina.com

[1] Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China

[2] School of Compter Science and Technology, Huazhong University of Science and Technology, Wuhan, China

[3] Beijing Key Laboratory of Mobile Computing and Pervasive Devices, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

[4] Beijing Institute of Computer Technology and Applications, Beijing, China

Various methods have been proposed for face hallucination (FH). Early methods [1, 4, 18, 19, 21, 28, 32] assume that faces are in controlled setting with small variations and try to learn a mapping function between HR faces and LR faces either by global or part-based models . These methods require faces either to be precisely aligned or to present high similarity in pose and expression with HR reference. Recently Convolutional Neural Networks have gained much reputation in several tasks in computer vision [6, 8, 22, 23] among researchers and different deep networks [3, 5, 15, 16, 25, 26, 31, 33–35, 39] have been proposed to hallucinate faces. Since facial prior is of great importance, various facial prior has been studied and achieve state-of-the-art performance. However, different features extracted from the same original inputs are related with each other, for instance, extracted shape features differ from texture features but they are similar in some positions at the feature maps. The existing methods neglect these inner deeper relations among different extracted features which maybe useful, and often unselectively consider information at every position of features which may introduce noises.

Inspired by the Ensemble Learning which aggregates a set of models to produce better results than a single model, we propose an ensemble network of three extraction sub-nets by using mix of both spatial attention and channel attention. Through spatial attention, the relationships among different features are mined; while through channel attention, the inter-dependencies among channels are exploited. For diversity of extracted features, we utilize three different sub-nets, which are basic sub-net, auto-encoder sub-net and dense residual attention sub-net. We also apply the channel attention in the dense residual attention sub-net to further strength details within interested area.

To be more specific, our contributions of this work are summarized as follows:

- **Constructing an ensemble network.** We propose an ensemble network which integrates multiple different sub-nets to have achieved better results than each of them working independently.
- **Designing an attention-based ensemble strategy.** We exploit the mix of both spatial attention and channel attention in the process of ensemble. For its credit, faces hallucinated by our model could present to be clearer within key facial interested areas.
- **Effectiveness on very low-resolution faces.** Our proposed method can hallucinate face images as low-resolution as 8×8 pixels and achieve higher effectiveness in visual quality than other methods. It can hallucinate faces by at most up-scaling factor of $\times 16$, which is of great significance to the compression of face images and video surveillance in real world.

2 Related work

2.1 Face hallucination

Face hallucination (FH) is a special case of Single Image Super Resolution (SISR) issue, thus details about the facial components such as eyes, noses, mouths and facial contours should be more concentrated on. Various methods have been proposed to extract the most facial prior and use it to assist the hallucinated faces.

Early approaches [1, 4, 18, 19, 21, 28, 32] can be roughly divided into global methods and part-based methods. The global methods aim at learning a holistic model by PCA to hallucinate the entire faces. Since global methods require LR faces to be well aligned and to share similar pose and expressions with the HR references, they are weak in reconstructing local details and usually degrade when faces are not precisely aligned.The part-based methods are then proposed to super-resolve facial parts instead of the whole faces. They try to generate the HR counterparts of LR inputs based on either reference patches or facial components in the training dataset. Since part-based methods require facial components extracted from LR inputs, their performance may deteriorate when the LR faces are tiny or noisy.

Recently, the Convolutional Neural Network (CNN) has manifested great effectiveness in generic image SR problems. Since Dong et al. [7] first propose a spuer resolution convolutional nerual network (SRCNN) to reconstruct effective HR images from LR ones. Many deeper networks have been proposed to hallucinate faces. And utilizing efficient facial prior is of great significance for those existing methods. Among them, various facial prior [5, 15, 16, 26, 31, 33, 34, 39] including parsing maps, semantic face, heatmaps,components and facial landmarks etc. are exploited to facilitate the final well restored faces. Chen et al. [5] propose an end-to-end face super-resolution network to extract parsing maps by Hour-Glass(HG) blocks and use them to enhance the hallucinated faces. Li et al. [15] propose a coarse-to-fine method by constructing a two-branch network, which makes full use of the semantic face prior. Song et al. [26] propose a two-stage method by generating facial components first and then utilize them to super-resolve LR face images. Yu et al. [35] propose a multi-task convolutional neural network with two branches, one of which aims to predict salient regions of facial component heatmaps. However those methods unselectively take in whole area of extracted features and often fuse them in element-wise addition way or concatenation way which may introduce noise. In contrast with them, we propose an ensemble model to mine not only the relations among features but also the inter-dependencies among different channels and then use them to fuse features in selective attention-based technique.

2.2 Attention mechanism

Attention mechanism which simulates the process of human vision has also pulled much focus of researchers, it essentially focuses on learning weighted distributions for features so as to augment some regions and weaken others. Vaswani et al. [30] first propose a transformer network architecture based solely on attention mechanisms which is dispensable with recurrence and convolutions entirely. Inspired by this, Zhang et al. [37] design a self-attention generative adversarial network (SAGAN) which is attention-driven for image generation tasks and achieves clearer details in interested area. Also Zhang et al. [38] propose a residual non-local attention network for high-quality image restoration. Apart from exploiting attention in spatial scale, the channel attention has also been researched. Hu et al. [10] proposed squeeze-and-excitation (SE) block to model channel-wise relationships to obtain significant performance improvement for image classification. Zhang et al. [36] propose a very deep residual channel attention networks (RCAN) which use channel attention to dig out inter-dependencies among channels for image super resolution.

3 Proposed method

3.1 Overall architecture

As is shown in Fig. 1, our network consists of four components: the coarse feature extraction for coarse features, the multiple extraction sub-nets for deep fine-grained features, the ensemble module which aggregates former three sub-nets and the upscale module which reconstructs the final hallucinated faces.

We use two convolutional layers to extract coarse features from the original inputs. Then the extracted coarse features are fed to three different sub-nets to acquire more fine-grained features, which differ from each other. Three sub-nets are the basic sub-net which generates basic features, the auto-encoder sub-net which is opt for shape prior and the the dense residual attention sub-net which extracts more detailed texture features. Then three kinds of features are transferred into the ensemble module to unify them into final ensemble features by exploiting both spatial attention and channel attention. The ensemble features are then fed to upscale module to get the final hallucinated faces, where we use the Pixel Shuffle [24] layers to gradually magnify

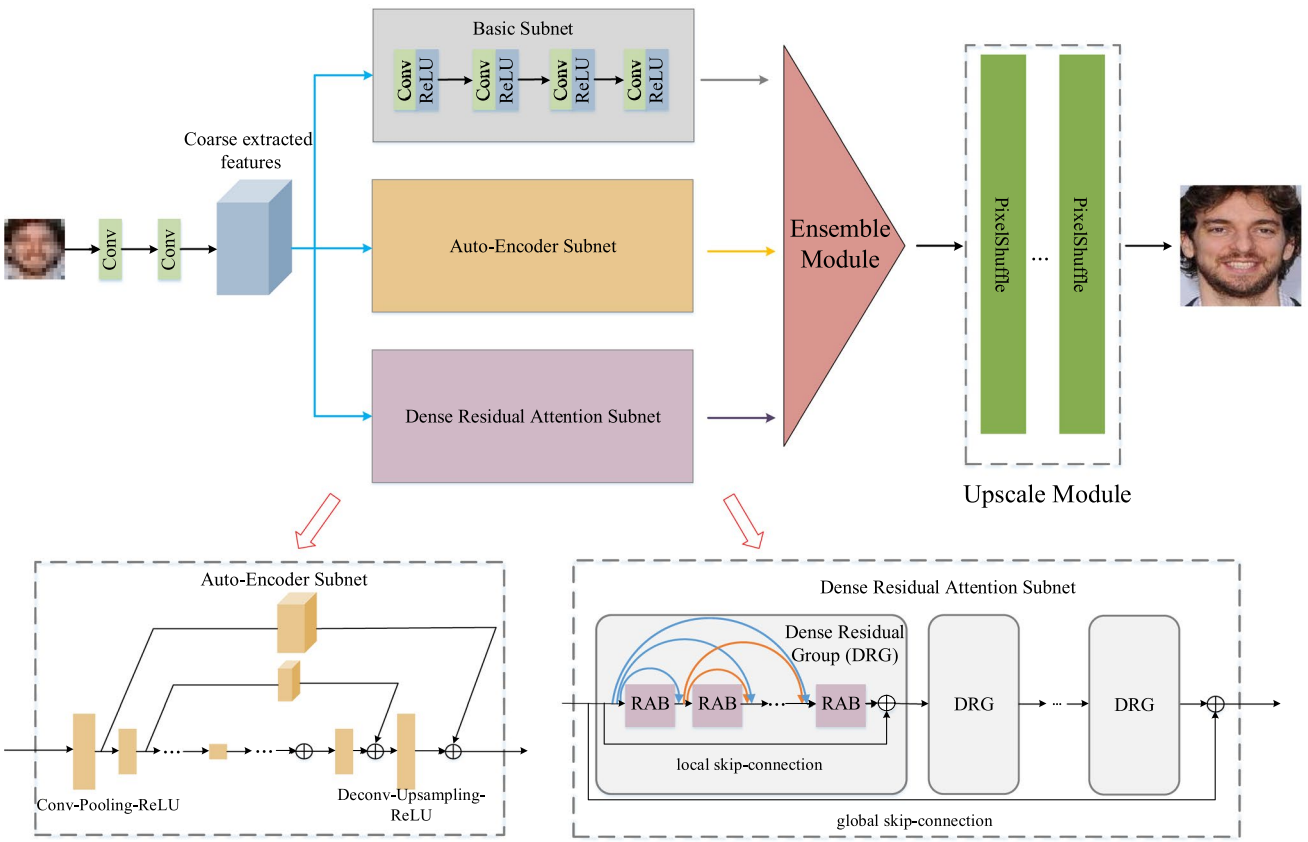

Fig. 1 Overall architecture of our network

features at each turn by two times. The upscale module consists of several pixel shuffle blocks and the number of pixel shuffle blocks depends on the scale needed to be magnified. Since the features are gradually magnified at each turn by two times, the number of pixel shuffle blocks is $\log_2 N$, where N is the scale faces needed to be magnified such as 4, 8 and 16.

3.2 Multiple extraction sub-networks

We endeavor to extract as many as efficient and diverse features by feeding the input into three different sub-nets. They are the basic sub-net for basic features, auto-encoder sub-net for facial shape prior and the dense residual attention sub-net for texture features.The basic sub-net only consists of four convolutional layers, which means the extracted features are coarse because of shallow layers. Then the auto-encoder sub-net is applied to extract shape features because the auto-encoder structure is often used to extract semantic segmentation, which is appropriate to preserve the shape priors in faces. And the dense residual attention sub-net is utilized to dig out the texture features because of deep convolutional layers and aided attention maps.

Basic Sub-net. By using only the convolutional layers and ReLU functions, the basic subset is exploited to extract basic global image features $P \in \mathbb{R}^{H\times W\times C}$ from face images. It is composed of several cascaded convolutional layers, each of which is followed by a ReLU function.

Auto-Encoder Sub-net. As is pointed out in [5], the facial shape prior can be well preserved when faces are reduced from high resolution to low resolution. The auto-encoder subset is then applied by us to dig out effective facial shape prior. Diifferent from the basic sub-net above, as is shown in Fig. 1, the auto-encoder sub-net includes the encoder and decoder, each of which consists of k blocks. For block in encoder, features are fed to convolutional layer, max-pooling layer and ReLU function, while in the block of decoder features are fed to deconvolutional layer and ReLU function. To effectively consolidate features and spatial information in different scales,we uses a skip connection between symmetrical layers.

Dense Residual Attention sub-net. The dense residual attention sub-net is utilized to acquire the texture features. As is shown in Fig.1, it consists of m residual groups, each of which cascades n residual attention block(RAB). Inspired by [29], we construct a dense architecture in one residual group, where a residual attention block receives the concatenation of output from all previous ones.

So for the j-th residual attention block in one dense residual group, the input is the concatenated features which have $\frac{j(j-1)}{2}C$ channels, where C is the number of channels of the extracted coarse features. As is shown in Fig. 2, features are firstly fed to a convolutional layer which squeezes the concatenated input into C-channel features, then the output is transferred to two convolutional layers followed by a ReLU layer. They are next fed to a channel attention layer, which exploits the inter-dependencies among feature channels to focus on the most contributory channels. To consolidate the gained information and to prevent from the degradation problem, a skip-connection is applied for every residual block. By applying the dense deep residual learning, more fine-grained texture features are extracted.

3.3 Attention-based ensemble strategy

Inspired by the Ensemble Learning which ensembles different multiple models to get better results than that of each model individually, we design an attention-based strategy to ensemble previous three sub-nets (Fig. 3).

Spatial attention maps. Multiple feature maps from basic sub-net, auto-encoder sub-net and dense residual attention sub-net are obtained respectively, which are all of size $H \times W \times C$. To ensemble those three sub-nets, we first apply spatial attention to calculate relations among three different feature maps and use them to generate spatial attention feature maps. Then the channel attention is utilized in previous spatial attention feature map to further explore which channel is more contributory to the final better results.

More exactly, to generate spatial attention feature maps, these feature maps from three sub-nets are firstly reshaped to features of which the size is $HW \times C$. Then they are fed to three individual 1x1 convolution layers and then more high-frequency features $P, E, R \in \mathbb{R}^{HW\times C}$ are generated.

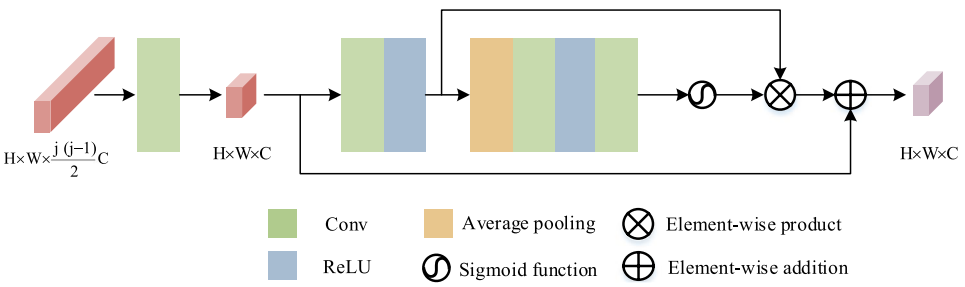

Fig. 2 The j-th residual attention block(RAB) in one residual group

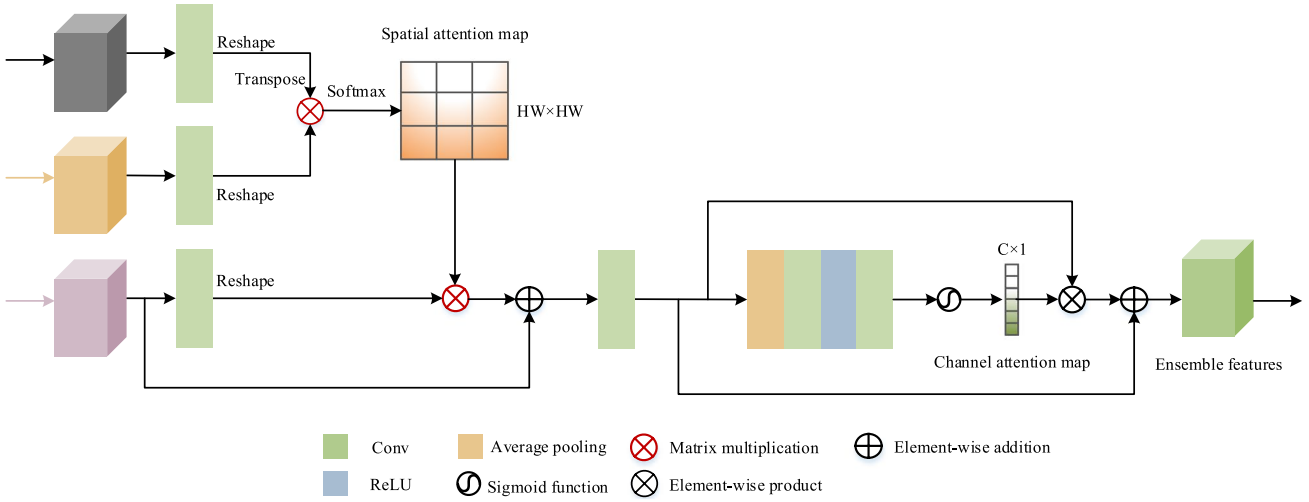

Fig. 3 Our ensemble strategy

We firstly conduct the matrix multiplication for feature $P \in \mathbb{R}^{HW\times C}$ from basic sub-net and the transposed feature $E^{\mathrm{T}} \in \mathbb{R}^{C\times HW}$ from auto-encoder sub-net, then a softmax layer is applied to get the attention map $G \in \mathbb{R}^{HW\times HW}$. It demonstrates the specific relations between features at every position. The relations ρ between the i-th position of basic features p_i and j-th position of facial prior e_j can be formulated as below:

$$\rho(p_i, e_j) = \frac{exp(p_i^{\mathrm{T}} W_u^{\mathrm{T}} W_v e_j)}{\sum_{j=1}^{N} exp(p_i^{\mathrm{T}} W_u^{\mathrm{T}} W_v e_j)}, \tag{1}$$

where W_u, W_v are the learned weight matrix. Then the final blended spatial attention feature maps $\beta \in \mathbb{R}^{HW\times C}$ are generated by conducting matrix multiplication for attention maps G and feature $R \in \mathbb{R}^{HW\times C}$ from the dense residual sub-net. The attention feature map β aggregates the map in each one of the whole C channels:

$$\beta = [\beta_1, \beta_2, ..., \beta_C], \tag{2}$$

thereinto the attention map in the j-th channel can be represented as followed:

$$\beta_j = W_g \sum_{i=1}^{N} \rho(p_i, e_j) W_h r_i \tag{3}$$

where W_g is also the learned weight matrix.

Channel attention maps. Having acquired the spatial attention feature map β, we endeavor to further produce the channel attention feature map by exploiting the inter-channel relationships of features. We firstly reshape β into $\beta_0 \in \mathbb{R}^{H\times W\times C}$, then the adaptive average pooling layer is applied to it to get the channel-wise statistic $z \in \mathbb{R}^C$. Then it is transferred to a convolutional layer, ReLU layer and another convolutional layer,and then the sigmoid function is applied to it to generate the final channel attention map $z \in \mathbb{R}^C$, which can be formulated as below:

$$z = \delta(W_1(\sigma(W_0(AvgPool(\beta))))), \tag{4}$$

where $AvgPool(.)$ denotes the average pooling layer; W_0, W_1 are the learned weight matrix; and $\sigma(.)$, $\delta(.)$ denotes the sigmoid function and ReLU function respectively. Then an element-wise product is produced by multiplying the channel attention map z and spatial attention feature map β to generate the final ensemble features.

3.4 Loss function

Given a training set $\{x^{(i)}, y^{(i)}\}_{i=1}^{N}$, N is the number of training face images, $y^{(i)}$ is the ground-truth high-resolution(HR) image corresponding to its low-resolution image pair $x^{(i)}$. To update and learn weights and parameters above, we adopt MSE loss between original ground-truth HR face images and hallucinated face images generated by our end-to-end network. The loss can be formulated as below:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left\| F(x^{(i)}) - y^{(i)} \right\|^2 \tag{5}$$

, where θ denotes the parameters, F represents the learned mapping function between low-resolution inputs and final hallucinated face images by our network.

4 Experiments

In this section, implementations details, ablation analysis, and the comparisons between our model with other methods will be elaborated.

4.1 Implementation details

The datasets and training setup will be described as followed.

Datasets. Both our training dataset and test dataset are from the CelebFaces Attributes Dataset (CelebA), which is a large-scale face attributes dataset with 202,599 faces, 10,177 identities, and 5 landmark annotations per image [20]. We randomly pick up 20100 images from the CelebA dataset to form our dataset. Then we split it into training dataset which contains 20000 images and test dataset which includes 100 images.

Training Setup. We coarsely crop the training images according to their face regions and resize them to 128× 128 pixels as the high resolution(HR) face images. The HR images are then downsampled to 32×32 pixels, 16×16 pixels, and 8×8 pixels as the corresponding low resolution(LR) ones, respectively, by bicubic interpolation. The number of residual group in the dense residual attention sub-net is set as 4, and the number of residual attention block within one residual group is also set to 4. We use the ADAM optimizer to minimize the loss function during the training process. And we initiate the learning rate as 0.00025 following the setup in [5] and halve it every 200 epochs. The whole network is implemented through PyTorch with a TITAN GPU.

4.2 Ablation analysis

Three sets of ablation studies are conducted to demonstrate the significance of ensemble, the efficiency of our ensemble strategy and the rationality of our designed three sub-nets respectively.

Why ensemble is needed? Trying to figure out significance to construct the ensemble of multiple models, we conduct a set of study by only using our three sub-nets respectively as the main body to form three independent networks. Then we retrains these three models using the same training dataset, eventually PSNR and SSIM metrics are calculated. As is shown in Fig. 4, our ensemble network of three different sub-nets performs better in both PSNR and SSIM metrics than each one of them working independently.

How is our ensemble strategy? To testify our ensemble strategy, we compare metrics when three kinds of features from our three sub-nets are fused in our ensemble strategy, element-wise addition and concatenation respectively. We keep the the three extraction sub-nets and upscale module as the same, while replacing our ensemble module by element-wise addition and concatenation respectively, and retrain these two different models. As is shown in Fig.4, our method outperforms other two traditional fusion methods by at least 0.31 dB in PSNR metrics.

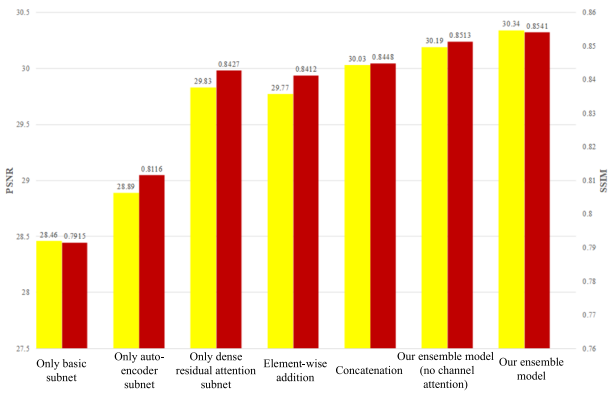

Fig. 4 Ablation study of ensemble, yellow bars denote PSNR results and red bars denote SSIM results

Meanwhile, we also test the effects of the used attention by removing the latter channel attention from our model, and retrain this model. As is shown in Fig.4, our ensemble of multi-sub-nets by using mix of spatial attention and channel attention performs better than by just using the spatial attention.

Effects of three extraction sub-nets. To validate the effectiveness of the three sub-nets designed by us, we make a group of ablation studies by removing the sub-net in turn. To keep the whole structure, the removed sub-net is replaced by a direct path with skip-connection and all of other parts are kept same. As is shown in Table 1, the model which has all three sub-nets performs the best, the one model which removes the basic sub-net works out as the second best, followed by the one which removes the auto-encoder sub-net and the one which removes the dense residual attention sub-net. This, to some extent, proves the importance of features from each sub-net. However all of them are indispensable for the final hallucinated faces because the ensemble of our all three sub-nets performs way better than other comparative models.

4.3 Comparisons

We make comparisons between our model and other methods to testify the efficiency of our proposed network. Among all those image super resolution algorithms which have been researched, we pick up some representative methods including the general SISR algorithm such as SRCNN [7], EDSR [17], SRResNet [14] and the specific facial prior aided algorithm like TDAE [34] and FSRNET [5]. The facial prior

Table 1 Ablation analysis of three sub-nets by removing each sub-net in turn

Basic sub-net	×	✓	✓	✓
Auto-encoder sub-net	✓	×	✓	✓
Dense residual attention sub-net	✓	✓	×	✓
PSNR/SSIM	29.74/0.8368	29.27/0.8261	26.93/0.7868	**30.34/0.8541**

The bold value indicates the best performance in PSNR/SSIM metrics

Table 2 Comparisons on quantitative metrics

	Methods	PSNR/SSIM		
		×4	×8	×16
General SR method	Bicubic	26.92/0.7369	23.41/0.5936	20.41/0.4941
	SRCNN	28.12/0.7862	24.48/0.6568	21.84/0.4941
	EDSR	*29.69/0.8366*	*26.68/0.7519*	23.50/0.6520
	SRResNet	29.27/0.8258	25.62/0.7196	22.86/0.6331
Facial prior aided method	TDAE	28.94/0.8155	25.30/0.7084	21.87/0.5813
	FSRNET	*29.92*/0.8424	26.21/0.7441	*23.72/0.6674*
	Ours	**30.34/0.8541**	**27.29/0.7799**	**24.57/0.7015**

The bold/italics font indicates the best/second best performance respectively

aided method also focus on exploiting original facial priors other than the whole image quality which the general SISR methods center on. We retain these models by feeding them with the same training dataset and also test them with the same test dataset.

As is shown in Table 2, among all those methods which are testified, our method outperforms other methods by gaining 0.42 dB, 0.61 dB and 0.85 dB over the second best algorithm when faces are hallucinated by 4, 8 and 16 times respectively.

Results of whole range. Apart from quantitative comparison in PSNR/SSIM metrics, large differences also appear in the qualitative vision effects of the generated results by different methods. Fig. 5 shows that our method can rehabilitate good-quality faces at each scale factor (×4, ×8, ×16) when the original LR inputs are very tiny. It is also shown in Fig. 6 that the reconstructed face by our method appears to be closer to both what the face originally looks like and what human beings expect to see than other testified methods.

Results of specific facial components. To further investigate the effectiveness of restoring details of specific facial components, we compare typical regions in hallucinated faces produced by our model and other methods. We pick up five characteristics of a face, which are facial contour, eyes, eyebrows, nose and mouth. As is shown in Fig. 8, the edge of whole eyes are more kept by our model than other methods. And details inner eyes, such as the eyeballs, are restored to be clearer by our method. It can be concluded from Figs. 7, 8, 9, 10, 11 that both the shape and texture within considered typical regions on the face are reconstructed by our model to be clearer and more similar to its original look than other methods.

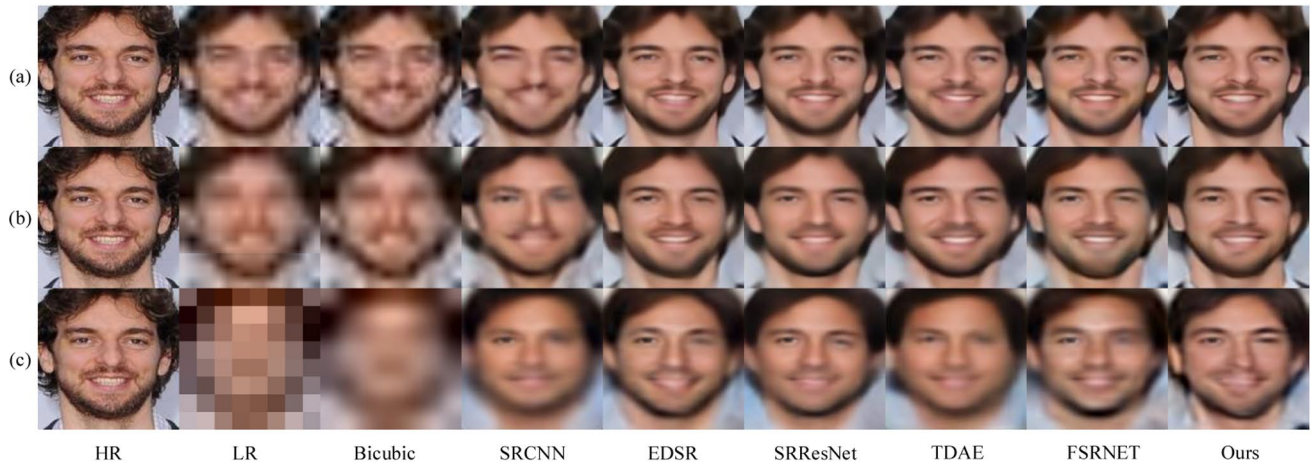

Fig. 5 Hallucinated faces of different methods on different scale factor. **a** shows the results when faces are hallucinated by 4 times, **b** shows the results when faces are hallucinated by 8 times, **c** shows the results when faces are hallucinated by 16 times

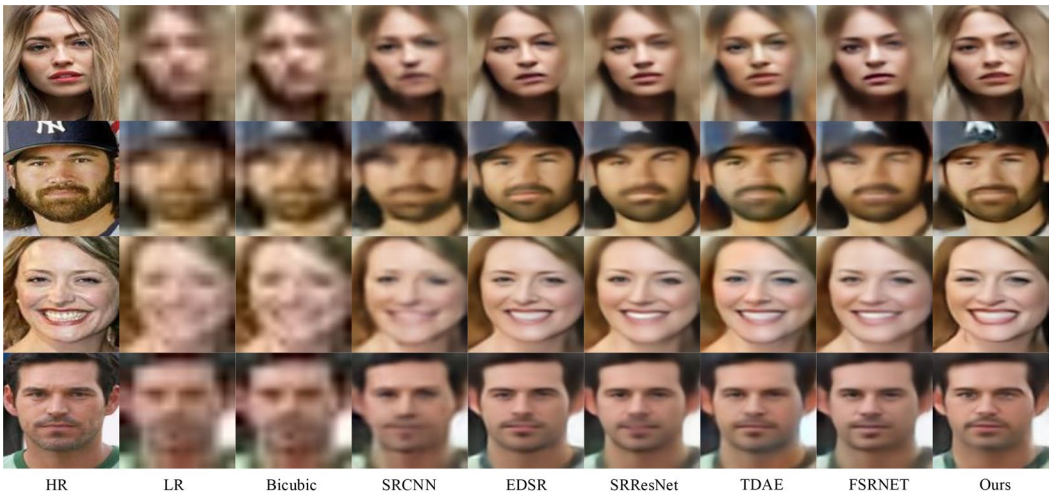

Fig. 6 Visualization of results on scale factor of 8

Table 3 Comparisons in average running time

Methods	SRCNN	ESDR	SRResNet	FSRNet	TDAE	Ours
PSNR(dB)	28.12	29.69	29.27	29.92	28.94	30.34
Average running time(s)	0.0031	0.0133	0.0243	0.0181	0.0226	0.0237

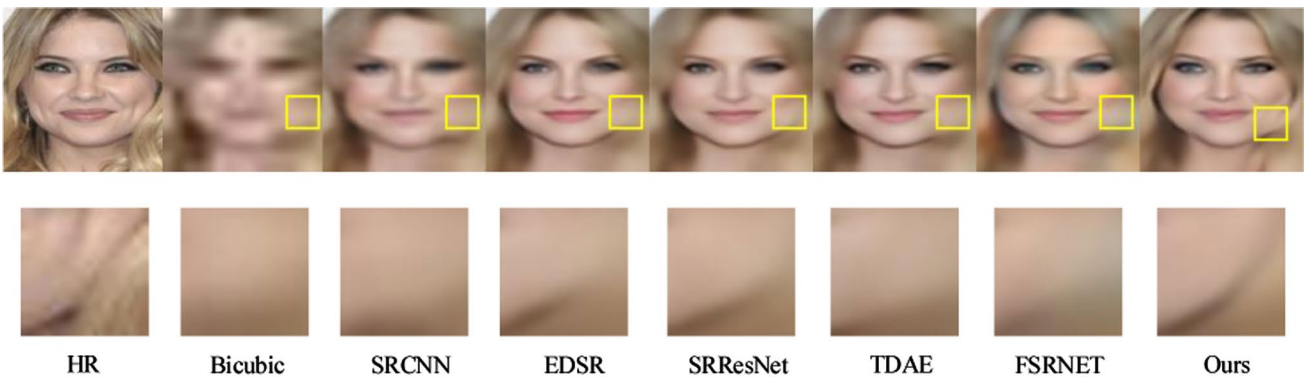

Fig. 7 Comparisons in details of facial contour

4.4 Discussion and future work

Ablation study and experimental results above demonstrate the effects of our ensemble model in both PSNR and SSIM metrics and details within reconstructed face images. On the other hand, our method may have larger time complexity due to the abundant multiplication operations. As is shown in Table 3, we can ignore SRCNN which consumes the least time but performs badly, then our method leverages the performance in PSNR by at least 0.42dB than other methods at the cost of consuming more time. Even though, our method can still be considered as effective because the improvements in performance is relatively worthwhile compared with no more than 0.011s increased running time.

We have acquired better results with the help of our ensemble network. However the work can be improved by introducing more attention-related module such as the spatial transformer network [11]. Apart from that, designing the model in the manner of Generative Adversarial Network [9] is also worthy to be researched and is where we attempt to go further.

Fig. 8 Comparisons in details of eyes

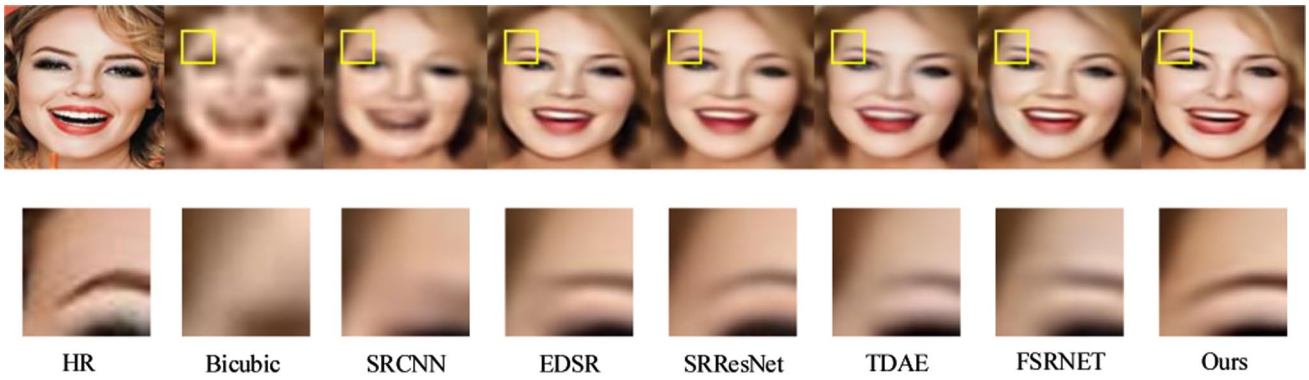

Fig. 9 Comparisons in details of eyebrows

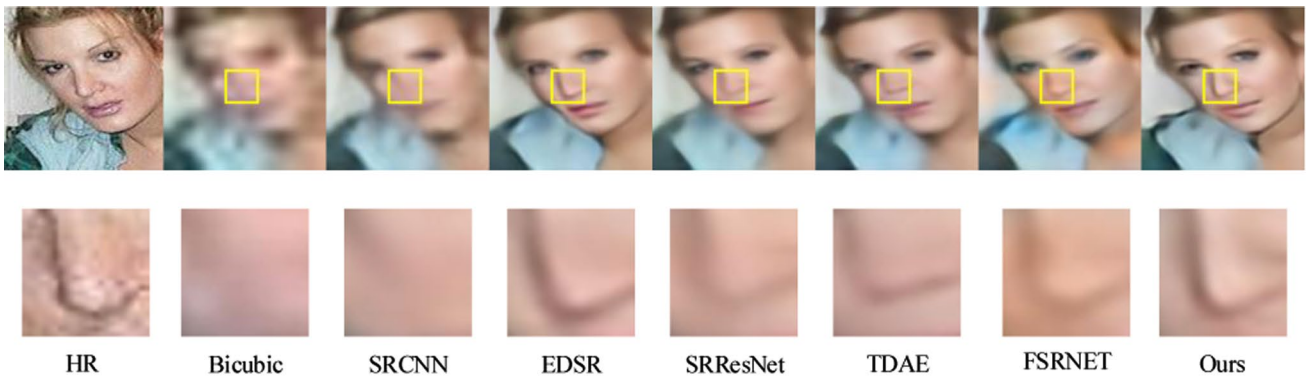

Fig. 10 Comparisons in details of nose

5 Conclusion

In this paper, we propose a deep ensemble network which aggregates three extraction sub-nets by using mix of both spatial attention and channel attention to hallucinate faces. We design three different sub-nets, each of which performs its own function, to extract diverse features. They are the basic sub-net for basic features, auto-encoder sub-net for shape prior and dense residual attention sub-net for more fine-grained texture features. Thereinto, the channel attention is applied in the dense residual attention sub-net to further consolidate details in interested areas. Most importantly, the exploitation of both spatial attention which reflects relations among different features and channel attention which presents dependencies among channels

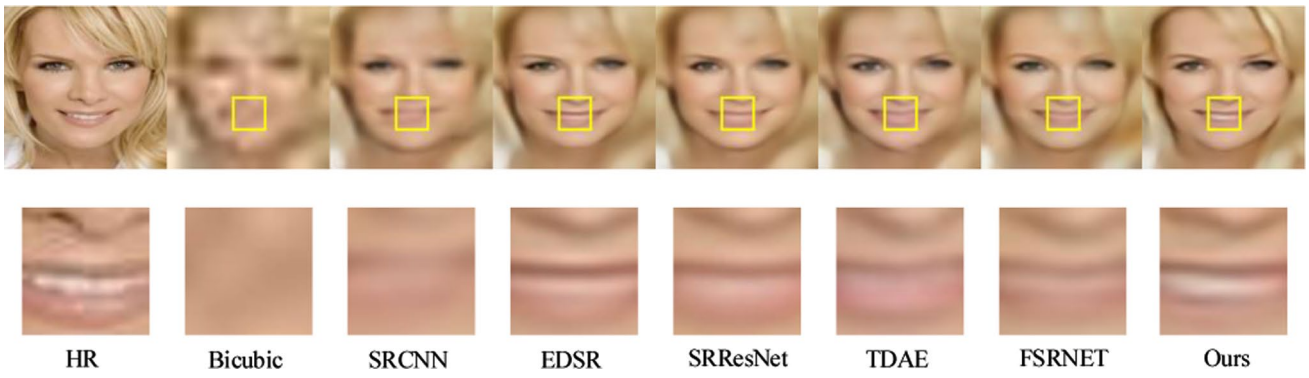

Fig. 11 Comparisons in details of mouth

allows our model to focus details within key facial components. And experimental results demonstrate the efficiency of our method in not only quantitative PSNR/SSIM metrics but also details within key facial areas and whole range.

Acknowledgements This work was supported by the National Natural Science Foundation of China under the Grant No.61672246, No.61272068, and the Fundamental Research Funds for the Central Universities, HUST:2016YXMS018.

References

1. BAKER, S.: Hallucinating faces. In: IEEE international conference on automatic face and gesture recognition, pp 83–88 (2000)
2. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. Int. J. Comput. Vis. **107**(2), 177–190 (2014)
3. Cao, Q., Lin, L., Shi, Y., Liang, X., Li, G.: Attention-aware face hallucination via deep reinforcement learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 690–698 (2017)
4. Chang, H., Yeung, DY., Xiong, Y.: Super-resolution through neighbor embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), IEEE, vol 1, pp I–I (2004)
5. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: end-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2492–2501 (2018)
6. Chen, Z., Wang, R., Zhang, Z., Wang, H., Xu, L.: Background-foreground interaction for moving object detection in dynamic scenes. Inf. Sci. **483**(5), 65–81 (2019)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)
8. Gao, L., Li, X., Song, J., Shen, H.T.: Hierarchical lstms with adaptive attention for visual captioning. IEEE Trans. Pattern Anal. Mach. Intell. **42**, 1112–1131 (2019)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. **27**, 2672–2680 (2014)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 7132–7141 (2018)
11. Jaderberg, M., Simonyan, K., Zisserman, A., kavukcuoglu, k: Spatial transformer networks. Adv. Neural Inf. Process. Syst. **28**, 2017–2025 (2015)
12. Jiang, H., Deng, W., Shen, Z.: Surveillance video processing using compressive sensing. Inverse Probl. Imaging **6**(2), 201–214 (2012)
13. Jourabloo, A., Ye, M., Liu, X., Ren, L.: Pose-invariant face alignment with a single cnn. In: Proceedings of international conference on computer vision(ICCV), pp 3200–3209 (2017)
14. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4681–4690 (2017)
15. Li, M., Sun, Y., Zhang, Z., Yu, J.: A coarse-to-fine face hallucination method by exploiting facial prior knowledge. In: 2018 25th IEEE international conference on image processing (ICIP), IEEE, pp 61–65 (2018a)
16. Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., Yang, R.: Learning warped guidance for blind face restoration. In: Proceedings of the European conference on computer vision (ECCV), pp 272–289 (2018b)
17. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) workshops, pp 136–144 (2017)
18. Liu, C., Shum, HY., Zhang, CS.: A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Citeseer, pp 192–198 (2001)
19. Liu, C., Shum, H.Y., Freeman, W.T.: Face hallucination: theory and practice. Int. J. Comput. Vis. **75**(1), 115–134 (2007)
20. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of international conference on computer vision(ICCV), pp 3730–3738 (2015)
21. Ma, X., Zhang, J., Qi, C.: Hallucinating face by position-patch. Pattern Recognit. **43**(6), 2224–2236 (2010)
22. Shamsolmoali, P., Zareapoor, M., Wang, R., Jain, D.K., Yang, J.: G-GANISR: gradual generative adversarial network for image super resolution. Neurocomputing **366**, 140–153 (2019a)
23. Shamsolmoali, P., Zareapoor, M., Wang, R., Zhou, H., Yang, J.: A novel deep structure u-net for sea-land segmentation in remote sensing images. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sensing **12**(9), 3219–3232 (2019b)

24. Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, AP., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1874–1883 (2016)
25. Shi, Y., Guanbin, L., Cao, Q., Wang, K., Lin, L.: Face hallucination by attentive sequence optimization with reinforcement learning. IEEE transactions on pattern analysis and machine intelligence (2019)
26. Song, Y., Zhang, J., He, S., Bao, L., Yang, Q.: Learning to hallucinate face images via component generation and enhancement. In: 26th international joint conference on artificial intelligence (IJCAI 2017), International joint conferences on artificial intelligence, pp 4537–4543 (2017)
27. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1701–1708 (2014)
28. Tappen, MF., Liu, C.: A bayesian approach to alignment-based image hallucination. In: Proceedings of the European conference on computer vision (ECCV), Springer, pp 236–249 (2012)
29. Tong, T., Li, G., Liu, X., Gao, Q.: Image super-resolution using dense skip connections. In: Proceedings of international conference on computer vision(ICCV), pp 4799–4807 (2017)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, AN., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Adv. Neural Inf. Process. Syst. pp 5998–6008 (2017)
31. Xin, J., Wang, N., Gao, X., Li, J.: Residual attribute attention network for face image super-resolution. Proceedings of the AAAI conference on artificial intelligence **33**, 9054–9061 (2019)
32. Yang, CY., Liu, S., Yang, MH.: Structured face hallucination. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1099–1106 (2013)
33. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: Proceedings of the European conference on computer vision (ECCV), Springer, pp 318–333 (2016)
34. Yu, X., Porikli, F.: Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 3760–3768 (2017)
35. Yu, X., Fernando, B., Ghanem, B., Porikli, F., Hartley, R.: Face super-resolution guided by facial component heatmaps. In: Proceedings of the European conference on computer vision (ECCV), pp 217–233 (2018)
36. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV), pp 286–301 (2018)
37. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International conference on machine learning, pp 7354–7363 (2019a)
38. Zhang, Y., Li, K., Li, K., Zhong, B., Fu, Y.: Residual non-local attention networks for image restoration. arXiv preprint arXiv :190310082 (2019b)
39. Zhu, S., Liu, S., Loy, CC., Tang, X.: Deep cascaded bi-network for face hallucination. In: Proceedings of the European conference on computer vision (ECCV), Springer, pp 614–630 (2016)

Dongdong Liu is current a master student in Huazhong University of Science and Technology and is supervised by Prof. Jincai Chen, his research interests include single image super resolution, face hallucination etc.

Jincai Chen is current a professor in Huazhong University of Science and Technology, his research interests include image super resolution, video abnormaly detection etc.

Zhenxing Huang is current a phd. student in Huazhong University of Science and Technology and is supervised by Prof. Jincai Chen, his research interests include single image super resolution, medical image processing etc.

Ni Zeng is current a master student in Institute of Computing Technology, Chinese Academy of Sciences, her research interests include ubiquitous computing and sign language synthesis etc.

Ping Lu is current a professor in Huazhong University of Science and Technology, her research interests include image super resolution, video abnormaly detection etc.

Lin Yang is currently working in Beijing Institute of Computer Technology and Applications.

Haofeng Wang is currently working in Beijing Institute of Computer Technology and Applications.

Jinqiao Kou is currently working in Beijing Institute of Computer Technology and Applications.

Min Wu is currently working in Beijing Institute of Computer Technology and Applications.