Multi-View Texture Learning for Face Super-Resolution

Yu WANG†, Student Member, Tao LU†(a), Feng YAO†(b), Yuntao WU†, and Yanduo ZHANG†, Nonmembers

SUMMARY In recent years, single face image super-resolution (SR) using deep neural networks have been well developed. However, most of the face images captured by the camera in a real scene are from different views of the same person, and the existing traditional multi-frame image SR requires alignment between images. Due to multi-view face images contain texture information from different views, which can be used as effective prior information, how to use this prior information from multi-views to reconstruct frontal face images is challenging. In order to effectively solve the above problems, we propose a novel face SR network based on multi-view face images, which focus on obtaining more texture information from multi-view face images to help the reconstruction of frontal face images. And in this network, we also propose a texture attention mechanism to transfer high-precision texture compensation information to the frontal face image to obtain better visual effects. We conduct subjective and objective evaluations, and the experimental results show the great potential of using multi-view face images SR. The comparison with other state-of-the-art deep learning SR methods proves that the proposed method has excellent performance.

key words: multi-view face image, texture compensation, face super-resolution, deep learning

1. Introduction

Face super-resolution (SR), which is known as a specific SR algorithm, can reconstruct a high-resolution (HR) image from one or multiple low-resolution (LR) input images. Face SR has been widely used in various face analysis related tasks, such as face recognition [1], [2], face alignment [3]–[5], face parsing [6], [7], pedestrian re-recognition [8], etc. In general, face SR methods include three typical methods: interpolation- [9], reconstruction- [10] and learning-based [11] methods. In other words, we can divide the face SR methods into two categories: traditional and deep learning based methods.

For the traditional learning-based face SR method, learning the mapping relationship between HR and LR from training samples has attracted more and more researchers’ attention. Considering the structural characteristics of face images, Baker and Kanade et al. [12] first proposed a learning-based face SR reconstruction method, which learns the prior distribution of image gradients from local faces. Yang et al. [13] first introduced the sparse coding strategy into the SR algorithm to avoid over-fitting problems. By assuming that the image blocks in the LR and the HR space have similar geometries, Ma et al. [14] used the position block to obtain accurate prior information that represents and reconstructs the image block by the corresponding position. Jiang et al. [15] proposed a local constrained representation (LCR) scheme for face SR. The local regularization was applied to the least-squares solution problem, and the sparsity and locality of the face SR were realized. In addition, in order to improve the SR performance, they also proposed the LCR iterative algorithm [16], which matched the interpolated LR images in HR space and achieved effective results. In addition to these specific domain methods for facial images, there are some general image SR methods that can be directly used for face SR. For example, [17]–[19] used random forests to process large datasets so that different patch patterns can be modeled differently through different decision trees. By stacking multiple decision trees, we can construct LR images hierarchically until the best SR image is obtained.

In recent years, the face SR algorithms based on deep learning has greatly improved the performance of facial reconstruction. Intuitively, the existing learning-based face SR methods can be divided into two kinds: single-input face SR (SFSR) and multi-input face SR (MFSR) methods. Dong et al. [20] firstly introduced Convolutional Neural Network (CNN) into the SR method in an end-to-end manner. Yu et al. [21] proposed a transforming autoencoder network to reconstruct very LR unaligned and noisy face images. Yang et al. [22] proposed a face method based on Generative Adversarial Networks (GAN) to restore reasonable visual output HR face images. Recently, Lu et al. [23] proposed a region-based deep residual network for face hallucination, which utilizes face images to learn further fine structural prior information. Although SFSR provides an end-to-end effective solution for supervised learning. But in actual scenes, the images of the same person are more likely to come from surveillance cameras with different views. If only the SFSR algorithm is used, the face image reconstruction performance will be poor, if the MFSR algorithm is used, more face texture detail information can be used to improve the face reconstruction performance.

On the other hand, the MFSR methods use two or more images as inputs, and these images use additional texture information to assist face SR processing. Generally, the traditional SR method based on multi-frame reconstruction can be directly used for super-resolving face images. However,
the performance of reconstructed SR depends on exploiting a priori information from internal examples (i.e. input LR images). However, it is very challenging to effectively fuse the prior information from multiple frames of LR images. Therefore, using the traditional SR method based on multi-frame reconstruction will limit its reconstruction performance. Jia et al. [24] proposed a general face SR method, which used tensors to model expression and poses changes. Ma et al. [14] used the multi-view face SR method to obtain frontal facial images from non-frontal images. Evgeniya et al. [25] proposed a multi-frame deep face SR, which solved the registration and SR problems in an end-to-end manner. The performance of these methods depends on the performance of the registered sub-network.

In fact, most existing face SR methods only use a single frontal face image as input and then reconstruct the input face image to restore the corresponding HR face image. However, in most practical situations, it is not enough to use only the LR front image texture information. Therefore, it is an attractive idea to restore HR facial images from multi-view reference images. Inspired by the multi-frame face image SR [25] and the parallax attention model [26], we use the view facial texture compensation method to convert SFSR to MFSR. In other words, the multi-view texture compensation method we proposed can solve the problem of recovering the corresponding front HR image of the face from the obtained multi-view face image. Specifically, given a facial image pair (frontal facial image and other multi-view images), first, use residual pooling module (RPM) to generate multi-view facial texture features. Then these facial features are sent to the texture attention module (TAM) to fuse the compensated texture features by calculating the attention map. Finally, the feature map of the target view image is updated through feature fusion to generate the SR results. In summary, our contributions are as follows:

1) We first propose a novel face SR method using multi-view face texture information, which is named multi-view texture learning for face SR (MTL). This method can solve the inadaptability of SFSR in real scenes.
2) We introduce a Texture Attention Module (TAM) with a global receptive field to process multi-view face images with angular differences. This network also proves that the TAM can effectively generate a reliable corresponding global relationship, which can effectively improve the face SR performance. Experimental results prove that our MTL method achieves the state-of-the-art performance in two public datasets. And in real-world actual tests, our model can also achieve advanced reconstruction results.

2. Related Work

In this section, we will briefly review several main work for SFSR, MFSR and attention mechanisms.

2.1 Single-Input Face SR (SFSR)

In recent years, deep-learning based SFSR methods have superior performance in peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) compared to single image SR of traditional algorithms. Since the CNN was introduced into face SR by Dong et al. [20], the deep learning-based approach has dominated the study of SFSR. Kim et al. [27] proposed a SR network with 20 convolutional layers to learn images residual information. Ledig et al. [28], firstly, the generative adversarial network (GAN) is introduced into the image SR to achieve the state-of-the-art performance. Recently, Yang et al. [22] introduced a decision-enhanced generated antagonistic network into the face images SR, which restored the visually reasonable SR face images.

2.2 Multi-Input Face SR (MFSR)

Compared to SFSR that uses a single image as input, the MFSR method takes additional two or more images as input, which can introduce additional texture information to aid the SR process. Evgeniya et al. [25] proposed using a multi-frame face image to train and reconstruct HR face images in an end-to-end manner, this method requires registration between images; Jia et al. [24] proposed a hierarchical tensor spatial representation method using multiple HR face images to achieve the promotion of expression and posture changes; Wang et al. [26] proposed a learning parallax attention SR to reconstruct the stereo images, which needs to combine stereo images response to reconstruct an effective SR results. Our MTL can adaptively transfer face texture information from the multi-view face images to the frontal face images. Unlike existing MFSR methods, we do not need registration between multi-view face images.

2.3 Attention Mechanisms

Attention mechanisms are widely used in computer vision tasks (eg. super-resolution [29], video action recognition [30] and image generation [31]). For the self-attention mechanism, the weighted sum of all positions on the space or time domain is calculated as a response at one location. And, by matrix multiplication, the self-attention mechanism can capture the interaction between any two locations. Therefore, the dependence of the two locations can be modeled without increasing the computational cost of storage. The self-attention mechanism is also widely used in semantic segmentation [32]. It should be noted that since the self-attention mechanism can model the remote dependence of the image, applying these mechanisms directly to the face image SR involves a cost calculation problem.

Inspired by the self-attention mechanism [31]–[33], we designed a texture attention mechanism to model the dependence in the face image compared with the self-attention mechanism. The proposed texture attention mechanism is
more flexible and efficient. In addition, the texture attention mechanism does not collect all similar features for reconstruction, and our network focus on more similar texture features. It turns out that the texture attention mechanism can generate reliable correspondences to improve SR performance.

3. The Proposed Work

In this section, we will introduce the proposed MTL. We take a set of LR face image pairs (a frontal image and two side-face images at different views) as input and reconstruct frontal image through the MTL network. The architecture is shown in Fig. 1.

3.1 Residual Pooling Module (RPM)

Recent studies have shown that rich facial texture feature information is very effective for improving the performance of facial SR reconstruction. Therefore, the method of multi-scale feature learning can obtain a discriminative representation of facial texture features. Inspired by the Pyramid Dilated Res-U-Net network[34], we propose a residual pooling module to expand the receiving field and extract fine facial texture features from multi-view facial images. Different from the Pyramid Dilated Res-U-Net network, our RPM module cascades 4 RPM groups through residual learning, which more effectively extracts the fine multi-scale facial features.

Specifically, as shown in Fig. 2, our RPM module first connects three 3×3 convolutions with different dilation rates (dilation rates of 1, 4, 8) in parallel, and then connects a 1×1 convolution in series to form a RPM group. Finally, 4 RPM groups are cascaded by residual learning. The RPM module not only expands the receiving filed of texture information, but also effectively extracts multi-scale facial features by combining convolution with different dilation rates. In this network, we alternately cascading a RPM module and a residual block. First, the input facial features are extracted through the RPM module to extract multi-scale facial features, and then the multi-scale facial features are input to the residual block. Repeat this process twice to obtain detailed face multi-view features.

3.2 Texture-Attention Module (TAM)

Inspired by the self-attention mechanism[31], [32], we propose a new attention module structure: Texture Attention Module (TAM). The reasonable use of this structure can efficiently obtain the global correspondence between the front face image and the multi-view image, which can effectively integrate the texture information of the face image pairs, so as to improve the reconstruction performance of the front face image of the target.
Other than a reconstruction loss, we also introduce three kinds of losses, named as: photometric loss, guide loss, and period loss. They help the network to make full use of the most similar correspondence between face images pairs. The overall loss function is expressed as:

\[ L_{\text{overall}} = L_{\text{rec}} + k(L_{\text{pho}} + L_{\text{gui}} + L_{\text{per}}), \]  

where \( k \) is empirically set to 0.005. The face SR performance of our network under the combination of different loss functions will be analyzed in Sect.4.4.4. Next, we will introduce four kinds of loss functions respectively.

(1) Reconstruction Loss

The SR network uses the mean square error (MSE) as loss function to calculate difference between the SR frontal images and the original HR frontal images. The reconstruction loss is expressed as:

\[ L_{\text{rec}} = \sum \| f_x - f_y \|^2, \]  

where \( f_x \) and \( f_y \) represent the SR result and ground-truth of the frontal image, respectively.

(2) Photometric Loss

Due to the variety of lighting conditions in the actual scene, it is challenging to collect high-quality multi-view facial images, so we use an unsupervised method to train our TAM module. Following [35], we introduce the photometric loss using the mean absolute error (MAE) loss. The photometric loss is expressed as:

\[ L_{\text{pho}} = \sum \| I_{m}^{LR} - (P_{m-f} \otimes I_{f}^{LR})\|_1 + \sum \| I_{m}^{LR} - (P_{f-m} \otimes I_{f}^{LR})\|_1, \]  

where \( I_{f}^{LR} \) and \( I_{m}^{LR} \) represent feature maps from a LR face images pair.

(3) Guide Loss

In order to produce accurate and consistent attention in the facial textureless area, guide loss is defined on the attention maps \( P_{f-m} \) and \( P_{m-f} \). The guide loss is expressed as:

\[ L_{\text{gui}} = \sum \sum (\| P(x, y, z) - P(x + 1, y, z)\|_1 + \| P(x, y, z) - P(x, y + 1, z + 1)\|_1), \]  

where \( P \in \{ P_{f-m}, P_{m-f} \} \), \( P(x, y, z) \) represents the contribution of position \((x, z)\) in multi-view face images to position \((x, y)\) in frontal face images. The first and second terms in Eq.(6) are used to achieve vertical and horizontal attention consistency, respectively.

(4) Period Loss

In addition to the photometric loss and the guide loss, we further introduce a period loss to achieve period consistency. Since \( P_{f-m} \) and \( P_{m-f} \) can be regarded as identity matrices, we design a period loss as:
$$L_{\text{per}} = \sum \| P_{f \rightarrow m \rightarrow f} - F \|_1 + \sum \| P_{m \rightarrow f \rightarrow m} - F \|_1,$$

where $F$ is a stack of $H$ identity matrices.

### 4. Experimental Results

In this section, we first introduce datasets and implementation details, and then compare recent SFSR methods and MFSR methods. We further conduct ablation experiments to test our network.

#### 4.1 Datasets

We conduct extensive experiments using two available datasets: FEI face dataset [36] and Youtube Faces (YTF) dataset [37]. Samples of face datasets are shown in Fig. 4. The FEI dataset consists of 200 people with 14 images per 200 people and a total of 2800 images. All images are colored and photographed in an upright front position on a white homogeneous background with a contour rotation of up to approximately 180 degrees. The angle difference between each face image is about 10 degrees, and the original size of each image is $640 \times 480$ pixels. We select 180 people as the training set and each person select three images, one is a frontal face image, and two are face images from different views. Then select another 20 people as the testing set, and also select three face images per person. Finally, the selected image is downsampled to the corresponding HR and LR by MATLAB to form a training image pair.

#### 4.2 Implementation Details

We only focus on 4× upsampling. In the training phase, we first trim the $32 \times 24$ pixel image block with a step size of 2 from these LR images via MATLAB. During this time, the corresponding HR image is also cut. As with most image SR methods, we increase the data by randomly flipping these image blocks horizontally and vertically. It should be noted that we did not do any pre-alignment operations. We use four evaluation indicators: Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM) [40], Visual Information Fidelity (VIF) [41] and Natural Image Quality Evaluator (NIQE) [42] to test face SR reconstruction performance. For all the test results, we tailored the same size boundaries to achieve a fair comparison.

Our face SR model is implemented using the Pytorch framework on a PC with single NVIDIA GTX 1080Ti GPU. We used the Adam [43] method to optimize all models and $\beta_1 = 0.9$, $\beta_2 = 0.999$. The initial learning rate is set to 0.0002 and is reduced by half every 30 epochs. Since more epochs do not provide a significant improvement, the training phase FEI face dataset and YTF dataset stopped after 116 and 160 epochs respectively.

#### 4.3 Experimental Results

As far as we know, the MFSR method is a new topic in the field of computers, and there is no relevant open source code. Therefore, we compare the latest single-input face SR methods and multi-input generic image SR methods, including LCGE [38], SRCNN [20], EDGAN [22], TDAE [21], PRDRN [23], PASSRent [26], and SRNTT [39] methods. For a fair comparison, we retrain and test the comparison methods using the same face datasets.

Table 1 shows the quantitative performance of the FEI [36] face dataset under the 4x metrics of PSNR, SSIM, VIF and NIQE. The results show that our algorithm performs the best in PSNR and SSIM evaluation indicators, and performs the second in VIF and NIQE evaluation indicators. In particular, MTL is at least 0.14 dB higher than these SFSR algorithms. Because of TAM can get more reliable
Table 1 Comparison of PSNR, SSIM, VIF and NIQE results of different algorithms on the FEI dataset [36] and YTF dataset [37]. (red color represent the first performance and blue color represent the second performance)

| Dataset | Algorithm | Bicubic | LCGE [38] | SRCNN [20] | EDGAN [22] | TDAE [21] | PRDRN [23] | PASSRnet [26] | SRNTT [39] | Ours |
|---------|-----------|---------|-----------|------------|------------|-----------|------------|-------------|-------------|------|
| FEI [36]| PSNR/dB   | 33.30   | 36.53     | 36.76      | 37.90      | 34.71     | 37.14      | 37.01       | 36.10       | 38.04|
|         | SSIM      | 0.9318  | 0.9537    | 0.9497     | 0.9555     | 0.9530    | 0.9570     | 0.9543      | 0.9564      | 0.9627|
|         | VIF       | 0.5293  | 0.6095    | 0.6051     | 0.6859     | 0.5019    | 0.6454     | 0.6154      | 0.6285      | 0.6695|
|         | NIQE      | 10.7924 | 9.2048    | 11.0833    | 9.3552     | 8.1761    | 9.9523     | 9.0283      | 9.8074      | 8.7816|
| YTF [37]| PSNR/dB   | 29.06   | 30.50     | 30.95      | 30.99      | 28.59     | 30.75      | 30.95       | 29.91       | 31.25|
|         | SSIM      | 0.8538  | 0.8758    | 0.8853     | 0.8873     | 0.8269    | 0.8870     | 0.8920      | 0.8931      | 0.8936|
|         | VIF       | 0.4451  | 0.4723    | 0.4922     | 0.5207     | 0.4000    | 0.5036     | 0.5183      | 0.5161      | 0.5234|
|         | NIQE      | 10.2943 | 8.3517    | 9.3430     | 7.4168     | 8.4838    | 9.1493     | 8.5057      | 8.6019      | 8.3383|

Fig. 5 Subjective comparison of our method with other algorithms on FEI datasets. (a) Bicubic, (b) LCGE, (c) SRCNN, (d) EDGAN, (e) TDAE, (f) PRDRN, (g) PASSRnet, (h) SRNTT, (i) ours, (j) HR.
texture information. On the YTF [37] dataset indicate similar quantitative performance in Table 1. Our method is also shown to favor better numerical scores than SR methods.

The qualitative results of our model are shown in Fig. 5 and 6, where Fig. 5 is the test result on the FEI dataset, and Fig. 6 is the test result on the YTF dataset. It should be noted that the algorithm (b) to the algorithm (g) is only an input frontal image for testing, and algorithm (g) and algorithm (h) is to input two test images and reconstruct the frontal image. From the enlarged area, we can see that the qualitative results of our proposed MTL algorithm on both datasets can achieve better visual performance and restore satisfactory structural information. It can be seen from Table 1 that in the full reference quality evaluation method VIF and the non-reference quality evaluation method NIQE (the smaller the NIQE value, the better the image quality), our performance has also achieved excellent results. The proposed MTL integrates multi-view facial attention maps to compensate for the texture information of different perspectives. This method not only retains the spatial structure of the face components, but also more effectively recovers the facial high-frequency details. In short, the proposed face SR network produces fine details and obtains good quantitative values.

4.4 Ablation Study

In this section, we present ablation experiments to demonstrate the validity of our model, including the structure of...
Table 2 Comparison of PSNR, SSIM, VIF and Params. results of different models on the FEI datasets. (TFSR refers to the network structure of two branches.)

| Model                  | PSNR | SSIM | VIF   | Params. |
|------------------------|------|------|-------|---------|
| MTL with SFSR          | 37.67| 0.9612| 0.6594| 1.55M   |
| MTL with TFSR          | 37.85| 0.9620| 0.6681| 1.66M   |
| MTL without TAM        | 37.92| 0.9625| 0.6685| 1.57M   |
| MTL without RPM        | 37.95| 0.9625| 0.6689| 1.67M   |
| MTL                    | 38.04| 0.9627| 0.6695| 1.67M   |

4.4.1 Network Architecture

Single Input VS Multiple Input: Multiple input images provide face texture information from different perspectives compared to single input images. In order to verify the effectiveness of multi-view texture information on face SR performance, we design two sets of comparative experiments. First, compare the network structure of the three branches to the network structure of the two branches, and then compare the network structure of the single branch. The experimental results on the FEI dataset [36] is listed in Table 2.

Compared with the original network, with two-branch network training and single-branch network training, the indicators decreased by 0.19 dB and 0.37 dB respectively in terms of PSNR. However, from the experimental results, even without texture information from multi-view face images, our single-branch network performance can perform better than other single-image face SR methods. When the input is a multi-view face image, our network structure performance is significantly improved.

4.4.2 Effectiveness of RPM

In our original network, the RPM module is used to extract the texture features of the face image. In order to verify the validity of this module, the RPMs in our network structure are deleted and then replaced with ordinary convolutions. From the comparison results in the Table 2, we can see that the SR performance of the original network benefits from the proposed module. If the RPM is removed and replaced with a normal convolution, the PSNR value will be reduced from 38.04 dB to 37.92 dB. This is because the long distance space between local features in the face image and its relationship in the multi-view image hinders the ordinary CNN from effectively integrating these features.

4.4.3 Effectiveness of TAM

TAM was introduced to integrate texture information from multi-view face images. To prove the validity of this module, we removed the TAM and directly stacked the RPM output. It can be observed from Table 2 that if the TAM is removed, the objective indicators will be decreased from 38.04 dB to 37.92 dB. This is because the long distance space between local features in the face image and its relationship in the multi-view image hinders the ordinary CNN from effectively integrating these features.

4.4.4 Effectiveness of Losses

In order to verify the effectiveness of introducing loss into the network, we retrain the MTL network with different loss combinations. Table 3 shows the comparison results of different loss combinations. It can be seen from Table 3 that if the MTL network is trained only by reconstruction loss, the PSNR value of our MTL network will be reduced from 38.04 dB to 37.79 dB. This is because the large convolution receptive field allows our network to use context information for face images over a wide range. Therefore, a more accurate correspondence can be obtained to improve the SR performance, thereby obtaining a good face reconstruction image.

Table 3 Comparative results achieved on FEI datasets by our MTL trained with different losses.

| Model       | $L_{rec}$ | $L_{phi}$ | $L_{psr}$ | PSNR | SSIM |
|-------------|-----------|-----------|-----------|------|------|
| MTL         | ✓         |           |           | 37.79| 0.9614|
| MTL         | ✓ ✓       |           |           | 37.80| 0.9616|
| MTL         | ✓ ✓ ✓     |           |           | 37.88| 0.9624|
| MTL         | ✓ ✓ ✓ ✓   | ✓ ✓ ✓ ✓   | ✓ ✓ ✓ ✓   | 38.04| 0.9627|

Tables 3 and 4 show the comparison results of different multi-view combinations (red color represent the first performance and blue color represent the second performance).

| Frontal/PSNR Right-18° Right-36° Right-54° Right-72° Right-90° |
|---------------------------------------------------------------|
| Left-18°           | 37.12dB   | 37.06dB   | 37.09dB   | 37.10dB | 37.02dB |
| Left-36°           | 37.10dB   | 37.06dB   | 37.08dB   | 37.09dB | 37.06dB |
| Left-54°           | 37.10dB   | 37.09dB   | 37.08dB   | 37.11dB | 37.04dB |
| Left-72°           | 37.02dB   | 37.06dB   | 37.03dB   | 37.07dB | 37.00dB |
| Left-90°           | 36.96dB   | 36.93dB   | 36.91dB   | 37.02dB | 36.92dB |
ture features to provide accurate correspondence, and cannot make better use of the texture information of multi-view face images. In addition, if photometric loss, guide loss and period loss are added, the face SR reconstruction performance will gradually improve. This is because these losses encourage our TAM to get reliable and more consistent correspondence. Overall, our MTL achieves the best performance when it is trained with all losses.

5. Performance Comparison of Different Multi-View Training Sets Combinations

In this section, we design a set of comparative experiments in which the multi-view face images between about 18° and 90° in the FEI face dataset [36] were combined differently and tested by multi-view images with an amplitude of about 18° each. The test result image is shown in Fig. 7. In terms of the image quality evaluation index PSNR, it can be seen from Table 4 that the texture information of the multi-view image with a small angle can better improve the image quality. And the multi-view image with 90° left and right can also have a good effect on the reconstruction of the front image. It is proved that our model can better reconstruct the front image by using the texture information of multi-view image.

6. Experimental Results in Real Scenarios

In this section, in order to prove the applicability of our model, we conduct a set of test experiments in actual scenarios. We collect a set of multi-view datasets of real scenes from the research team members through the camera. Specifically, we place multiple sets of cameras to photograph each team member from different perspectives, and calculated the distance between the camera and the face to simulate HR-LR face image pairs. We use the LR multi-view images obtained by shooting as the test set, and perform SR reconstruction on the images. The test result is shown in Fig. 8. It can be seen from the partially enlarged detail information in Fig. 8 that the test results obtained by our model recover better on the structural information of the face, such as the eyes, nose, and mouth areas. The experimental results show that the model also has good applicability and effectiveness in real scenarios.
7. Conclusions

In this paper, we propose a multi-view face SR network, which is called MTL. Our face SR network learns the texture information of multi-view face images through texture compensation. It can generate clearer and more realistic face texture information of multi-view face images through texture compensation. The experimental results on FEI face dataset [36] and YTF dataset [37] show that our method outperforms state-of-the-art face SR methods. In future work, we guess that researchers can focus on facial SR models at larger magnifications (e.g. 8x, 16x) based on a multi-input image method to obtain better face SR images. It is also possible to use face image with different backgrounds as reference images, and then reconstruction target face image, which will make the face image reconstruction become more interesting. These face SR algorithms based on multiple inputs are indeed a good direction for future face SR research.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (62072350, U1903214, 62071339), Hubei Technology Innovation Project (2019AA045), the Central Government Guides Local Science and Technology Development Special Projects (2018ZYD059), the High value Intellectual Property Cultivation Project of Hubei Province, the Enterprise Technology Innovation Project of Wuhan (202001602011971). Scientific Research Foundation of Wuhan Institute of Technology (CX2019244).

References

[1] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.4690–4699, 2019.
[2] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.815–823, 2015.
[3] Z. Zhang, P. Luo, C.C. Loy, and X. Tang, “Learning deep representation for face alignment with auxiliary attributes,” IEEE Trans. Pattern Anal. Mach. Intell., vol.38, no.5, pp.918–930, 2016.
[4] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Process. Lett., vol.23, no.10, pp.1499–1503, 2016.
[5] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou, “Joint multi-view face alignment in the wild,” IEEE Transactions on Image Processing, vol.28, no.7, pp.3636–3648, 2019.
[6] Y. Li, S. Liu, J. Yang, and M.-H. Yang, “Generative face completion,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.5892–5900, 2017.
[7] P. Luo, X. Wang, and X. Tang, “Hierarchical face parsing via deep learning,” 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp.2480–2487, 2012.
[8] K. Nasrollahi and T.B. Moeslund, “Super-resolution: a comprehensive survey,” Mach. Vision. Appl., vol.25, no.6, pp.1423–1468, 2014.
[9] L. Zhang and X. Wu, “An edge-guided image interpolation algorithm via directional filtering and data fusion,” IEEE Trans. Image Process., vol.15, no.8, pp.2226–2238, 2006.
[10] T.R. Y., “Multiple frame image restoration and registration,” Computer Vision and Image Processing, 1984.
[11] K. Zeng, T. Lu, X. Liang, K. Liu, H. Chen, and Y. Zhang, “Face super-resolution via bilayer contextual representation,” Signal Processing-image Communication, vol.75, pp.147–157, 2019.
[12] S. Baker and T. Kanade, “Hallucinating faces,” Proc. 4th IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), pp.83–88, 2000.
[13] J. Yang, J. Wright, T.S. Huang, and Y. Ma, “Image super-resolution via sparse representation,” IEEE Trans. Image Process., vol.19, no.11, pp.2861–2873, Nov. 2010.
[14] X. Ma, H. Huang, S. Wang, and C. Qi, “A simple approach to multiview face hallucination,” IEEE Signal Process. Lett., vol.17, no.6, pp.579–582, June 2010.
[15] J. Jiang, R. Hu, Z. Han, Z. Wang, T. Lu, and J. Chen, “Locality-constraint iterative neighbor embedding for face hallucination,” 2013 IEEE International Conference on Multimedia and Expo (ICME), pp.1–6, July 2013.
[16] J. Jiang, R. Hu, Z. Wang, Z. Han, and J. Ma, “Facial image hallucination through coupled-layer neighbor embedding,” IEEE Trans. Circuits Syst. Video Technol., vol.26, no.9, pp.1674–1684, Sept. 2016.
[17] J.-J. Huang and W.-C. Siu, “Learning hierarchical decision trees for single-image super-resolution,” IEEE Trans. Circuits Syst. Video Technol., vol.27, no.5, pp.937–950, May 2017.
[18] Z.-S. Liu, W.-C. Siu, and J.-J. Huang, “Image super-resolution via weighted random forest,” 2017 IEEE International Conference on Industrial Technology (ICIT), pp.1019–1023, March 2017.
[19] Z.-S. Liu, W.-C. Siu, and Y.-L. Chan, “Fast image super-resolution via randomized multi-split forests,” 2017 IEEE International Symposium on Circuits and Systems (ISCAS), pp.1–4, May 2017.
[20] C. Dong, C.C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” Computer Vision – ECCV 2014, Lecture Notes in Computer Science, vol.8692, pp.184–199, Springer International Publishing, Cham, 2014.
[21] Y. Yu and F. Porikli, “Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autencoders,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.5367–5375, July 2017.
[22] X. Yang, T. Lu, J. Wang, Y. Zhang, Y. Wu, Z. Wang, and Z. Xiong, “Enhanced discriminative generative adversarial network for face super-resolution,” Pacific Rim Conference on Multimedia, vol.11165, pp.441–452, Springer, 2018.
[23] T. Lu, X. Hao, Y. Zhang, K. Liu, and Z. Xiong, “Parallel region-based deep residual networks for face hallucination,” IEEE Access, vol.7, pp.81266–81278, 2019.
[24] K. Jia and S. Gong, “Generalized face super-resolution,” IEEE Trans. Image Process., vol.17, no.6, pp.873–886, June 2008.
[25] E. Ustinova and V.S. Lempitsky, “Deep multi-frame face hallucination for face identification,” CoRR, vol.abs/1709.03196, 2017.
[26] L. Yang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo, “Learning parallax attention for stereo image super-resolution,” CoRR, vol.abs/1903.05784, 2019.
[27] J. Kim, J.K. Lee, and K.M. Lee, “Accurate image super-resolution using very deep convolutional networks,” CoRR, vol.abs/1511.04587, 2015.
[28] C. Ledig, L. Theis, F. Huszar, J. Caballero, A.P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” CoRR, vol.abs/1609.04802, 2016.
[29] Y. Liu, Y. Wang, N. Li, X. Cheng, Y. Zhang, Y. Huang, and G. Lu, “An attention-based approach for single image super-resolution,” CoRR, vol.abs/1807.06779, 2018.
[30] X. Wang, R.B. Girshick, A. Gupta, and K. He, “Non-local neural networks,” CoRR, vol.abs/1711.07971, 2017.
[31] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” arXiv preprint arXiv:1805.08318, 2018.

[32] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.3146–3154, 2019.

[33] A. Vyasani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, pp.5998–6008, 2017.

[34] Q. Zhang, Z. Cui, X. Niu, S. Geng, and Y. Qiao, “Image segmentation with pyramid dilated convolution based on resnet and u-net,” International Conference on Neural Information Processing, vol.10635, pp.364–372, Springer, 2017.

[35] C. Godard, O.M. Aodha, and G.J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.270–279, 2017.

[36] C.E. Thomaz and G.A. Giraldi, “A new ranking method for principal components analysis and its application to face image analysis,” Image and Vision Computing, vol.28, no.6, pp.902–913, 2010.

[37] L. Wolf, T. Hassner, and I. Maoz, “Face recognition in unconstrained videos with matched background similarity,” CVPR 2011, pp.529–534, 2011.

[38] Y. Song, J. Zhang, S. He, L. Bao, and Q. Yang, “Learning to hallucinate face images via component generation and enhancement,” arXiv preprint arXiv:1708.00223, 2017.

[39] Z. Yang, Z. Wang, Z. Lin, and H. Qi, “Image super-resolution by neural texture transfer,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.7982–7991, 2019.

[40] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE Trans. Image Process., vol.13, no.4, pp.600–621, 2004.

[41] H.R. Sheikh and A.C. Bovik, “Image information and visual quality,” IEEE Trans. Image Process., vol.15, no.2, pp.430–444, 2006.

[42] A. Mittal, R. Soundararajan, and A.C. Bovik, “Making a “completely blind” image quality analyzer,” IEEE Signal Process. Lett., vol.20, no.3, pp.209–212, March 2013.

[43] D.P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

Yu Wang received the B.S. degree from the College of Post and Telecommunication, Wuhan Institute of Technology, Wuhan, China, in 2018. He is currently studying computer vision, image processing and machine learning for the M.S. degree in the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China.

Tao Lu received the B.S. and M.S. degrees from the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China, in 2003 and 2008, respectively, and the Ph.D. degree from the National Engineering Research Center for Multimedia Software, Wuhan University, in 2013. He is currently an Associate Professor with the School of Computer Science and Engineering, Wuhan Institute of Technology, and also a Research Member with the Hubei Provincial Key Laboratory of Intelligent Robot. He held a postdoctoral position with the Department of Electrical and Computer Engineering, Texas A&M University, from 2015 to 2017. His research interests include image/video processing, computer vision, and artificial intelligence.

Feng Yao received the M.S. degrees from the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China. He is currently an Associate Professor with the School of Computer Science and Engineering, Wuhan Institute of Technology. His research interests include computer application technology and computing intelligence.

Yuntao Wu received the Ph.D. degree in information and communication engineering from the National Key Lab. for Radar Signal Processing at Xidian University in December 2003. From April 2004 to September 2006, he was a post doctoral researcher at Institute of Acoustics, Chinese Academy of Sciences, Beijing, China. During Oct. 2006–Feb. 2008, He worked as a senior research fellow at the City University of Hong Kong. He was a Visiting Researcher in the Faculty of Engineering, Bar-Ilan University from April 2013 to March 2014, Israel. Currently, He is a full-time professor in Wuhan Institute of Technology. At the same time, He is also a Chutian Scholar Project in Hubei Province distinguish professor at the same university. He has published over 80 journal and conference papers, and he is a senior member of the Chinese Institute of Electronic Engineers (CIE), an associate editor of Multidimensional Systems and Signal Processing (an international journal of Springer publisher). His research interests include signal detection, parameter estimation in array signal processing and source localization for wireless sensor networks, biomimetic signal analysis, etc.

Yanduo Zhang received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 1999. He was a Visiting Professor with the Center for Vision, Cognition, Learning, and Art, University of California at Los Angeles, Los Angeles, from 2012 to 2013. He is currently a Professor with the Wuhan Institute of Technology, Wuhan, China. He is also the Vice President with the Wuhan Institute of Technology and the Director of the Hubei Province Key Laboratory of Intelligent Robot. He has coauthored two books and published more than 100 research papers. His research interests include computer vision, robot control, and intelligent computing. He is a member of the ACM.