J. Meas. Sci. Instrum., 2024, 15(4): 458-465.
DOI: 10.62756/jmsi.1674-8042.2024046
http://jmsi.publish.founderss.cn
jmsi@nuc.edu.cn

Face image super-resolution reconstruction algorithm based on residual attention mechanism

CHE Yali, XU Yan*, XUE Haili, LIU Xuhui

School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

***Corresponding author:** XU Yan (xuyan@mail.lzjtu.cn)

Received: February 15, 2023 **Revised:** March 27, 2023 **Accepted:** May 11, 2023

Abstract: Aiming at the problems such as low reconstruction efficiency, fuzzy texture details, and difficult convergence of reconstruction network face image super-resolution reconstruction algorithms, a new super-resolution reconstruction algorithm with residual concern was proposed. Firstly, to solve the influence of redundant and invalid information about the face image super-resolution reconstruction network, an attention mechanism was introduced into the feature extraction module of the network, which improved the feature utilization rate of the overall network. Secondly, to alleviate the problem of gradient disappearance, the adaptive residual was introduced into the network to make the network model easier to converge during training, and features were supplemented according to the needs during training. The experimental results showed that the proposed algorithm had better reconstruction performance, more facial details, and clearer texture in the reconstructed face image than the comparison algorithm. In objective evaluation, the proposed algorithm's peak signal-to-noise ratio and structural similarity were also better than other algorithms.

Key words: face image; super-resolution reconstruction; residual network; attention mechanism

0 Introduction

Face image super-resolution (SR), also known as face hallucination, refers to the generation of high-resolution (HR) from corresponding low-resolution (LR) face images. This is a fundamental problem in face of analysis, which can significantly facilitate face-related tasks such as face alignment[1], face parsing[2], face recognition[3,4], and 3D face rebuild[5]. The technology has been widely used in many fields, such as intelligent residential districts, face payment, and intelligent transportation. However, the collected face images are blurred and unclear because of uncontrollable environmental factors, such as illumination, distance, and atmospheric disturbance. They cannot recognize the details of the face, so improving the resolution of the face image through the face super-resolution reconstruction technology has become a problem of considerable concern.

After face hallucination was proposed[6], more researchers began studying face SR. Early face image super-resolution reconstruction algorithms are all based on low-level features. The global eigenface method extracted as much feature information as possible to perform feature transformation from low-resolution face images by adopting the way of principal component analysis to achieve face super-resolution reconstruction[7]. Some researchers combined local features and global features to achieve face SR reconstruction. However, the adaptation environment of these traditional algorithms is relatively simple, and it is difficult to deal with blurred images, different image resolutions, and different face pose. Since Dong et al.[8] used shallow convolutional neural networks to achieve image super-resolution reconstruction, the deep learning-based methods based on their outstanding reconstruction capabilities have attracted extensive attention from researchers. At the same time, people began to introduce deep learning into face SR reconstruction, and the traditional face SR algorithms were gradually replaced.

Dong et al.[9] proposed FSRCNN (Fast super-resolution convolutional neural network) based on SRCNN (Super-resolution convolutional neural network). Compared with SRCNN, the network running speed of this algorithm is fast. Still, the accuracy of the reconstruction process is too low because the neural network cannot achieve a balance between convergence speed and convergence. Kim et al.[10]

introduced residual network into image super-resolution reconstruction and proposed a very deep convolutional network for super-resolution (VDSR) model to suppress deep network problems and accelerate convergence. But its model is a simple accumulation of convolutional layers, which cannot effectively extract image information. Zhang et al.[11] proposed a dense residual network that combined a densely connected network and a residual network. It could fully extract information but would increase the number of parameters and the amount of computation. The central part of the above algorithm is a stacked residual block structure. With the deepening of the network, although rich feature information can be extracted, too much redundant information will be extracted at the same time, which makes the network difficult to converge.

To solve these problems, the residual attention face super-resolution reconstruction (RAFSR) was designed, and the attention mechanism was added to the residual network to effectively reduce network redundancy. The adaptive residual network was designed to solve the problem of disappearing gradient, which made it easier to converge in the training of network model, and the features were supplemented according to the needs in the training process. The experimental results showed that the algorithm has a better reconstruction effect compared with the existing algorithms.

1 Basic theory

1.1 Residual learning

The main idea of the residual network is to add a skip connection between the input and the output, which can transfer the input directly to the output layer. The network can learn the residual part between the input x and the output directly through this connection[12]. Its basic structural unit is shown in Fig.1. The residual equation is

$$H(x)=F(x)+x, \tag{1}$$

where x is the feature input; $F(x)$ is the residual function between the input and output; and $H(x)$ is the feature output.

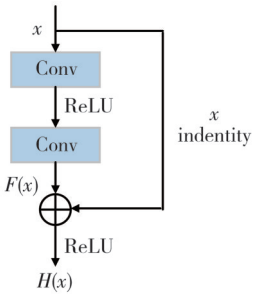


Fig. 1 Residual network structure diagram

1.2 Dual attention module

Convolutional block attention module (CBAM) [13] includes channel attention module (CAM) and spatial attention module (SAM). The CAM extracts global salient feature textures to reduce the impact of redundant information on network model training. The SAM filters the salient features again spatially to enhance the feature learning of images. The structure of CBAM is shown in Fig.2. It can be seen that residual learning can not only solve the problem of performance degradation that occurs when training deep networks by skip connections, but also accelerate the convergence speed of the network. CBAM can suppress the influence of redundant information on the network and enhance the feature learning ability about images.

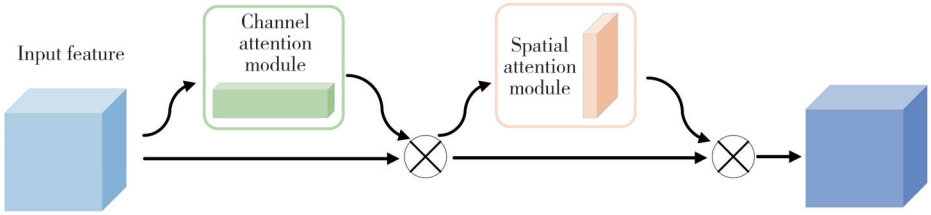


Fig. 2 CBAM structure diagram

2 RAFSR algorithm

In the process of face super-resolution reconstruction, the design of the network model is also crucial. The RAFSR algorithm was proposed combining the attention mechanism and adaptive residual network. Among them, the attention mechanism can make the network pay more attention to the high-frequency information in images during the training process, and enrich the facial details of the reconstructed image[14]. The adaptive residual network can not only solve the problem of gradient disappearing and gradient explosion, but also supplement the features according to the need, so that the network model can converge more easily when training the network.

2.1 Overall network architecture

The overall network framework is shown in Fig. 3,

which is mainly composed of four parts: shallow feature extraction module (SFEM), deep feature extraction module (DFEM), adaptive residual module (ARM), and reconstruction module (RM). The reconstruction process is as follows. First, the face LR image is input into SFEM to obtain shallow feature information, and then the obtained shallow feature information is input into RAM, and then the extracted features in each RAM are fused to obtain deep feature information, and then the feature fusion is used to obtain deep feature information, and finally the shallow feature information extracted from SFEM is input into the adaptive residual network. The output result and the extracted deep feature information are added up and input into the RM to obtain the reconstructed feature map, and finally the high-resolution face image is output.

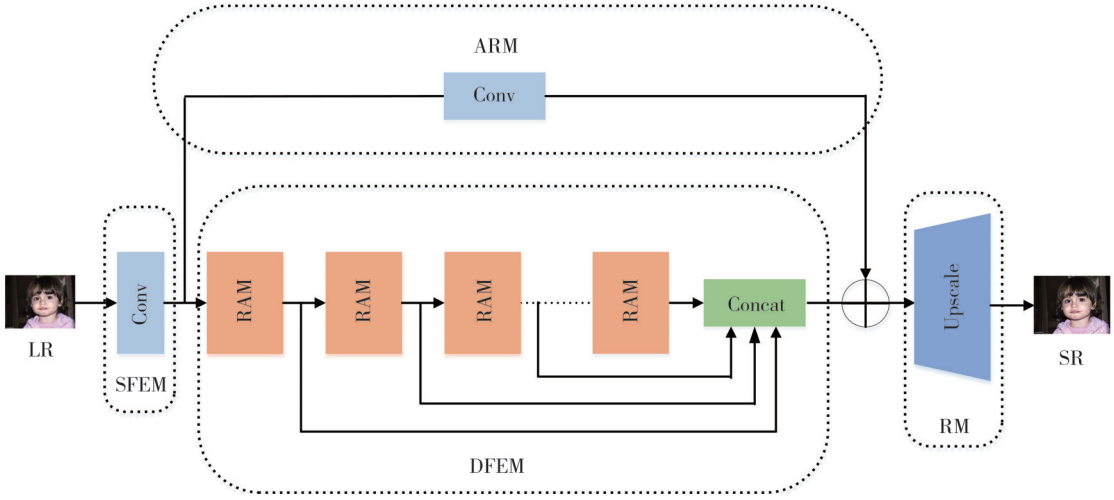


Fig. 3 RAFSR structure diagram

2.1.1 Shallow feature extraction module (SFEM)

SFEM mainly extracts the shallow features of low-resolution face images. A convolution kernel of 3×3 is used for extracting, and the extracted shallow features only have the facial contour information of the image. The extracted information is used as the input for the next module. The convolution calculation formula is

$$F_0 = C_{\mathrm{LR}}\left(W_{\mathrm{LR}}^{3\times 3}, I_{\mathrm{LR}}\right), \tag{2}$$

where F_0 is the shallow feature of the extracted face image; $C_{\mathrm{LR}}(\cdot)$ is the convolution function; I_{LR} is the low-resolution face image; and $W_{\mathrm{LR}}^{3\times 3}$ is the convolution kernel of 3×3.

2.1.2 Deep feature extraction module (DFEM)

DFEM consists of 8 residual attention modules (RAM). Its main function in the face image super-resolution reconstruction network is to extract the high-frequency information of the face image, which largely determines the quality of the reconstructed image.

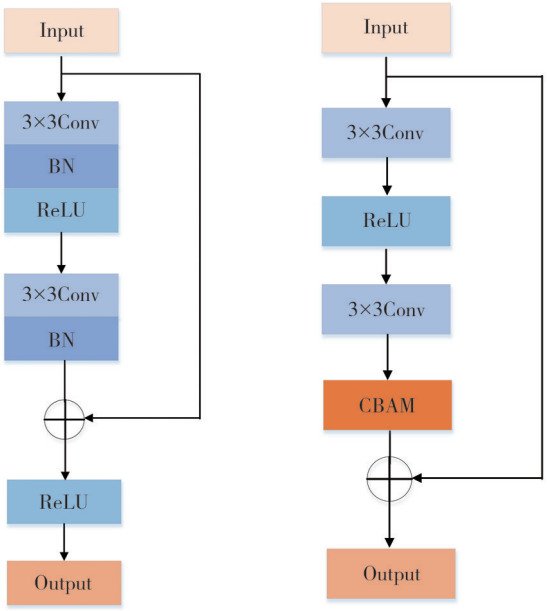
1) Residual attention module

The super-resolution reconstruction algorithm of face image based on convolutional neural network firstly extracts shallow feature information from low resolution face image, then obtains high frequency information through depth feature extraction module, and finally obtains high resolution face image from high frequency information through reconstruction module. However, since the depth of the network model with excellent generalization performance is generally deeper in the reconstruction process, it will inevitably introduce too much redundancy and invalid information. In order to diminish the impact of this information on the network, the channel attention (CA) mechanism is usually introduced into the reconstructed network[15]. The CA mechanism performs global average pooling on the input features and gives each feature an equal weight.

However, this makes the network process high-frequency and low-frequency information equally, which not only causes the network to lack the learning ability of essential features, but also goes against the core idea of image super resolution, that is, to recover as much high-frequency detail information in the image as possible. In order to solve this problem, some researchers have also introduced spatial attention[16] into the reconstruction network. Although this approach assigns more weight to high-frequency information[17], it increases the parameters and computational complexity.

Therefore, the CBAM module was introduced, which combined the channel and spatial attention modules. Given an intermediate feature map, the attention weights can be sequentially inferred along the two dimensions of space and channel and then multiplied with the original feature map to adaptively adjust the features so as to make the network focus on helpful information and suppress redundant and invalid

information. This can not only save parameters and computational power but also recover more high-frequency details. Secondly, CBAM is introduced into the residual block. As shown in Fig. 4, the output of RAM convolution is used as the input of CBAM. To make full use of it, CBAM is added to each RAM block to strengthen the extraction of information, which is more conducive to recovering face image details. In addition, skip connections are added to the RAM to improve the circulation of information and solve the problems of gradient disappearance and network degradation caused by excessive network depth.



(a) Original residual block (b) RAM

Fig. 4 Original residual block and improved RAM

2) Feature fusion module

In face image super-resolution reconstruction, the dependency between different depth feature maps is often ignored, which leads to a weakened recovery performance of the reconstruction network for texture details, and different feature maps carry different receptive field information. The obtained features all have a certain impact on the reconstruction performance. At the same time, the high-frequency information of the image is easily lost when it propagates in the deep network, which leads to the problem of blurred texture details in the reconstructed HR image. To make full use of the different information carried by each module in the network, a feature fusion module is added to fuse the feature information extracted from each layer of the network, strengthen feature reuse, reduce the loss of high-frequency information, and make the reconstructed face image. Texture details are richer. As shown in Fig. 3, the output of 8 RAM is allowed to concatenate feature information according to dimension for concatenation, that is feature fusion. If the dimensions of the two input features x and y are p and q, the output feature dimension is $p+q$.

2.1.3 Adaptive residual module

In order to extract features further, the residual idea was introduced to improve the network quality. As shown in Fig.3, an adaptive residual network is added to the model[18,19], in which skip connections can not only solve the problem of the gradient disappearance caused by excessive network depth, but also make it easier to converge network model training. Unlike the traditional residual network, the adaptive residual in this paper adds a 1×1 convolution to the skip connection path. After the input image undergoes this convolution, the image feature information will be extracted. After the weights are updated in transmission, the needed features can also be obtained further to supplement features for the overall network. Therefore, the richness of the output features can be enhanced, and the gradient disappearance can be alleviated by the adaptive residual network. In essence, adaptive residual networks still function as residual networks and can increase the diversity of feature extraction.

2.1.4 Reconstruction module

The traditional reconstruction algorithm uses the method of bicubic interpolation[20] to sample the low-resolution image to the same resolution as the high-resolution image in the super-resolution reconstruction process and then learns the mapping relationship between the low-resolution image and high-resolution image after obtaining the same resolution. The traditional algorithm both increases the computational complexity and adds redundant information in the network, affecting image reconstruction quality. Most of the latest algorithms train the network using an unamplified, low-resolution image as input, and then up-sample the end of the network to zoom the image to the same resolution as the HR image.

The reconstruction module consists of convolution and subpixel convolution. Its convolution calculation formula is

$$I_{\mathrm{HR}}=H_{\mathrm{US}}(F_{\mathrm{DF}}), \tag{3}$$

where I_{HR} represents the HR image; F_{DF} represents the extracted deep features; and $H_{\mathrm{US}}(\bullet)$ represents the upsampling operation.

2.2 Loss function

The image pixel loss is calculated by the L1 paradigm, namely the difference between the

reconstructed super-resolution image (SR) and the original high-resolution image (HR). The function expression is

$$L=\frac{1}{r^2HW}\sum_{x=1}^{rW}\sum_{y=1}^{rH}\left\| I_{x,y}^{\mathrm{HR}}-G_{\theta_G}(I_{\mathrm{LR}})_{x,y}\right\|, \tag{4}$$

where W and H are the image sizes; $G_{\theta_G}(I^{\mathrm{LR}})$ is the reconstructed image; and I^{HR} is the original high-resolution image.

3 Experiment

3.1 Dataset and training settings

This paper conducted experiments on two public datasets: CelebA[21] and Helen[22]. CelebA is a large-scale face dataset that contains 202 599 face images of 10 177 different celebrities, and each image is labeled accordingly. The Helen dataset contains 2 330 face images, and each image has its corresponding parsing map.

The first 13 000 face images from the CelebA dataset and the first 2 000 face images from the Helen dataset are selected as the joint dataset for training the network. The remaining images in the two datasets are selected randomly as the test set. To minimize the loss function and better update the parameters, the Adam optimizer[23] is used to adaptively optimize the network parameters, where $\varepsilon=1\times10^{-8}$, $\beta_1=0.9$, $\beta_2=0.99$. The initial learning rate is 0.000 1, and the number of iterations and epochs are set to 200.

The experimental hardware is Intel(R) Core(TM) i7 11700K CPU@3.80GHz, 32GB memory, and two NVIDIA GeForce 6GB 2080Ti graphics cards. Install Anaconda3 under the Windows10 operating system, then build the deep learning framework Pytorch in Python3.7.

3.2 Result analysis

The proposed algorithm is compared with the Bicubic, SRCNN, VDSR, and SRGAN[24] algorithms. To ensure the fairness and accuracy of the experimental results in the subjective and objective evaluation standards, the scale factor of each model in the experiment is all set to 4 and trained with the same training set.

Fig. 5 shows the reconstruction results of 5 super-resolution reconstruction models on face images.

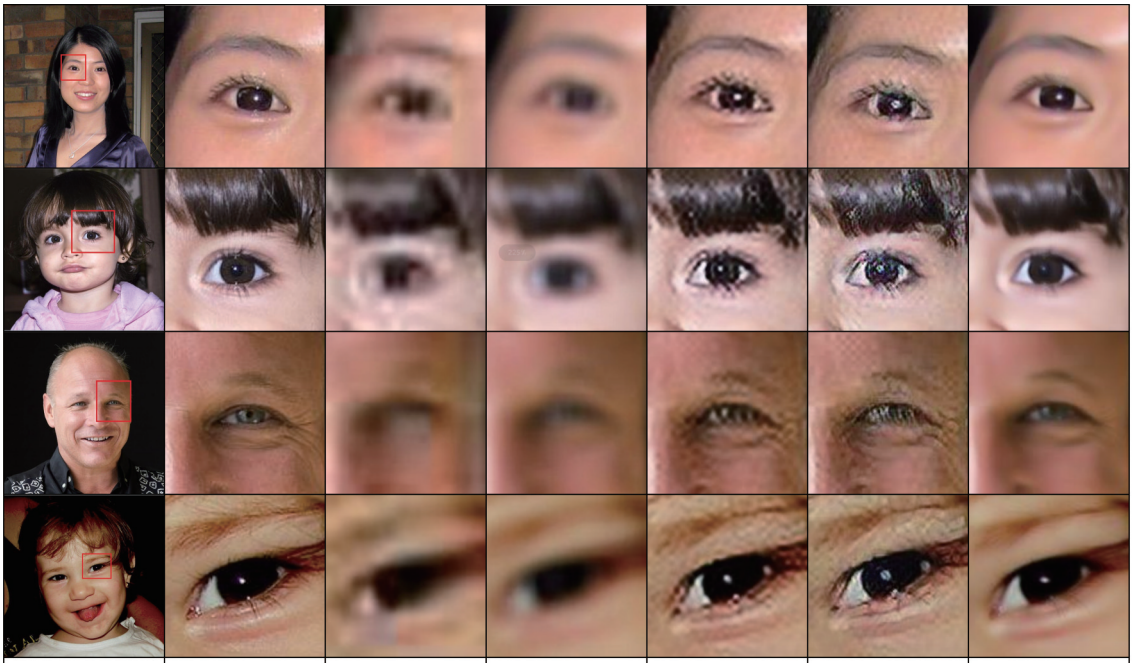


Fig. 5 Comparison of subjective effects of reconstructed images with different algorithms

Among them, the original represents the original high-resolution image. The Bicubic algorithm is a traditional interpolation method, and the reconstructed face image has the worst sensory effect. It can only restore the general outline of the face, and the facial features are relatively blurred. Compared with the Bicubic algorithm, the reconstruction effect of the SRCNN algorithm has been enhanced, but due to its small number of network layers, the extracted feature information is limited, and therefore the recovered detailed information is also relatively small. The VDSR algorithm has a deep network in which a residual

network is added. Its reconstructed face image has a better effect, but some details are still lost. The SRGAN algorithm has a better effect in reconstructing the images, but some textures are too smooth and the color contrast is relatively low. On the whole, the proposed algorithm can restore a good overall effect. As can be seen from Fig. 6, the reconstruction effect of eyes, eyelashes, eyebrows, and other areas in the image is also very good. Therefore the proposed algorithm has a better effect on some edges or details.

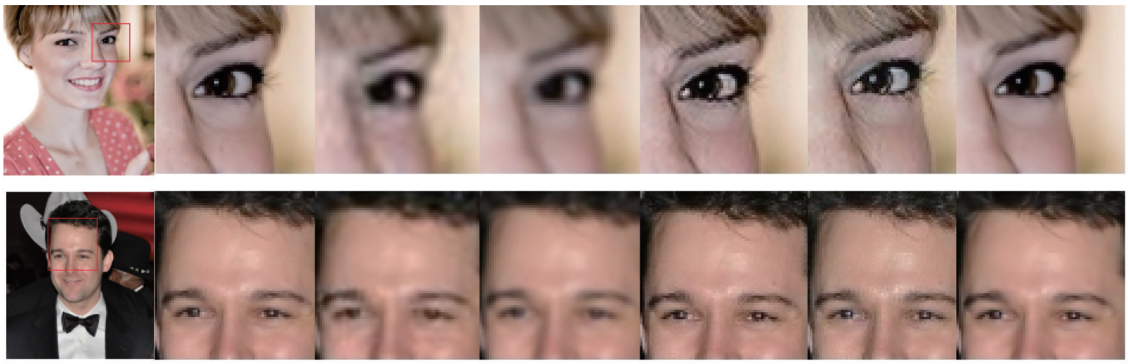


Fig. 6 Comparison of local details of subjective effects of reconstructed images with different algorithms

3.3 Quantitative analysis

The most common evaluation metrics, namely peak signal to noise ratio (PSNR) [25] and structural similarity (SSIM) [25], are used to evaluate the reconstructed images. Among them, PSNR is an error-sensitive image quality evaluation. The higher the value is, the better the quality of the reconstructed image. The value range of SSIM is [0, 1], which compares the structural similarity of two images. The closer it is to 1, the higher the similarity between the two images, and the better the reconstruction effect.

100 images in the Helen dataset are randomly selected for testing (except the images in the training set). The results of different algorithms are shown in Table 1 when the scale factor is 4. It can be seen that the proposed algorithm has a higher PSNR value and SSIM value.

Table 1 Comparison of PSNR and SSIM of different SR algorithms

Algorithm	PSNR/dB	SSIM
Bicubic	25.50	0.769
SRCNN	26.60	0.798
VDSR	29.69	0.837
SRGAN	29.26	0.775
RAFSR	**32.83**	**0.880**

3.4 Ablation experiment

To verify the effectiveness of the CBAM module and the adaptive residual module, 50 images in the Helen dataset were randomly selected as the test set (except the images in the training set) to design ablation experiments. The model with the two modules removed from the whole network is regarded as the original model (OM). The model with only the CBAM module is OM+CBAM. The model with only the adaptive residual module is OM+ARM, and the model with both the adaptive residual module and the CBAM module is the RAFSR algorithms. The ablation results are shown in Table 2.

Table 2 Comparison of PSNR and SSIM of different algorithms in ablation experiments

Algorithm	PSNR/dB	SSIM
OM	32.26	0.829
OM+ARM	32.39	0.857
OM+CBAM	32.60	0.869
RAFSR	**33.29**	**0.874**

It can be seen that the PSNR value and SSIM value tested by the network model are higher than the network with missing modules. Therefore, the CBAM module and the adaptive residual module are highly effective in this algorithm. Secondly, to better realize face image reconstruction, a comparative experiment was also carried out on selecting the number of residual blocks. The results are shown in Table 3. Since the PSNR value and SSIM value are both high when the number of residual blocks is 8, the number of residual blocks in this model is 8.

Table 3 Comparison of PSNR and SSIM with different numbers of residual blocks

Number of residual blocks	PSNR/dB	SSIM
4	32.34	0.858
8	**33.29**	**0.874**
16	32.89	0.871

4 Conclusions

A face image super-resolution network model was

designed based on the residual attention mechanism. Among them, the attention mechanism was introduced into the residual block, which could solve the problem of redundant information and improve the efficiency of the reconstruction network. The added adaptive residual network could also solve the problem of disappearing gradient, which made it easier to converge in the training of network model and supplement the features according to the needs in the training process. The experimental results showed that compared with Bicubic, SRCNN, VDSR, and SRGAN, the proposed algorithm had obvious improvement in the overall visual effect and an excellent performance in the objective evaluation of PSNR and SSIM values, which proved the effectiveness of the proposed method in face image resolution reconstruction.

Acknowledgement

This work was supported by National Natural Science Foundation of China (No.62063014)

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] JOURABLOO A, LIU X. Pose-invariant 3D face alignment//International Conference On Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 3694-3702.

[2] LI Y J, LIU S F, YANG J M, et al. Generative face completion//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 5892-5900.

[3] JIN Y G, BOUGANIS C S. Robust multi-image based blind face hallucination//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 5252-5260.

[4] YANG J, LUO L, QIAN J J, et al. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(1): 156-171.

[5] ROTH J, TONG Y Y, LIU X M. Adaptive 3D face reconstruction from unconstrained photo collections//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 4197-4206.

[6] BAKER S, KANADE T. Hallucinating faces//Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), March 28-30, 2000, Grenoble, France. New York: IEEE, 2002: 83-88.

[7] WANG X G, TANG X O. Hallucinating face by eigentransformation. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2005, 35(3): 425-434.

[8] DONG C, LOY C C, HE K M, et al. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 295-307.

[9] DONG C, LOY C C, TANG X O. Accelerating the super-resolution convolutional neural network//European Conference on Computer Vision, October 11-14, 2016, Cham: Springer, 2016: 391-407.

[10] KIM J, LEE J K, LEE K M. Accurate image super-resolution using very deep convolutional networks//Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 1646-1654.

[11] ZHANG Y L, TIAN Y P, KONG Y, et al. Residual dense network for image super-resolution//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 2472-2481.

[12] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[13] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module//European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. Cham: Springer, 2018: 3-19.

[14] LIU Y, DONG Z L, LIM K P, et al. A densely connected face super-resolution network based on attention mechanism//2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), November 9-13, 2020, Kristiansand, Norway. New York: IEEE, 2020: 148-152.

[15] HU J, SHEN L, SUN G. Squeeze-and-excitation networks//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 7132-7141.

[16] ZHU X Z, CHENG D Z, ZHANG Z, et al. An empirical study of spatial attention mechanisms in deep networks//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE, 2019: 6687-6696.

[17] CHEN C F, GONG D H, WANG H, et al. Learning spatial attention for face super-resolution. IEEE Transactions on Image Processing, 2020, 30: 1219-1231.

[18] LIU S, GANG R P, LI C H, et al. Adaptive deep residual network for single image super-resolution. Computational Visual Media, 2019, 5(4): 391-401.

[19] WANG H F, XU Y, WEI Y M, et al. Image super-

resolution reconstruction based on parallel convolution and residual network. Journal of Computer Applications, 2022, 42(5): 1570-1576.

[20] KEYS R. Cubic convolution interpolation for digital image processing. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1981, 29(6): 1153-1160.

[21] LIU Z W, LUO P, WANG X G, et al. Deep learning face attributes in the wild//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 3730-3738.

[22] LE V, BRANDT J, LIN Z, et al. Interactive facial feature localization//European Conference on Computer Vision, October 7-13, 2012, Florence, Italy, Heidelberg: Springer, 2012: 679-692.

[23] PEREYRA M, SCHNITER P, CHOUZENOUX É, et al. A survey of stochastic simulation and optimization methods in signal processing. IEEE Journal of Selected Topics in Signal Processing, 2016, 10(2): 224-241.

[24] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic single image super-resolution using a generative adversarial network//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 105-114.

[25] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2004, 13(4): 600-612.

基于残差注意力机制的人脸图像超分辨率重建算法

车亚丽，徐　岩*，薛海丽，刘旭辉

兰州交通大学 电子与信息工程学院，甘肃 兰州 730070

摘　要：针对人脸超分辨率重建算法中存在的重建效率不高、重建图像纹理细节模糊和重建网络不易收敛等问题，本文提出了一种残差注意力人脸超分辨率重建算法。首先，该算法为了解决冗余信息和无效信息对重建效果造成的影响，在网络的特征提取模块中引入了注意力机制，提高了整体网络的特征利用率；其次，为了缓解梯度消失等问题，在网络中引入自适应残差，让网络模型训练起来更易收敛，并在训练时根据所需进行特征补充。实验结果表明，所提算法与对比算法相比，重建性能更好且重建出的人脸图像面部细节更多，纹理更清晰。客观评价也表明，所提算法的峰值信噪比和结构相似性均优于其他算法。

关键词：人脸图像；超分辨率重建；残差网络；注意力机制

引用格式：CHE Yali, XU Yan, XUE Haili, et al. Face image super-resolution reconstruction algorithm based on residual attention mechanism. Journal of Measurement Science and Instrumentation, 2024, 15(4): 458-465.