Face Super-Resolution Reconstruction Based on Self-Attention Residual Network

QING-MING LIU\(^1\), RUI-SHENG JIA\(^{1,2}\), CHAO-YUE ZHAO\(^1\), XIAO-YING LIU\(^1\), HONG-MEI SUN\(^{1,2}\), AND XING-LI ZHANG\(^{1,2}\)

\(^1\)College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China
\(^2\)Shandong Province Key Laboratory of Wisdom Mine Information Technology, Shandong University of Science and Technology, Qingdao 266590, China

Corresponding authors: Rui-Sheng Jia (jrs716@163.com) and Hong-Mei Sun (shm0221@163.com)

This work was supported in part by the Natural Science Foundation of Shandong Province, China, under Grant ZR2018MEE008, in part by the National Natural Science Foundation of China under Grant 51904173, and in part by the Project of Shandong Province High Educational Science and Technology Program under Grant J18KA307.

ABSTRACT Aiming at the problems of the face image super-resolution reconstruction method based on convolutional neural network, such as single feature extraction scale, low utilization rate of features and blurred face images texture, a model combining convolutional neural network with self-attention mechanism is proposed. Firstly, the shallow features of the image are extracted by the cascaded 3 × 3 convolutional kernels, and then self-attention mechanism is combined with the residual blocks in depth residual network to extract the deep detail features of faces. Finally, the extracted features are fused globally by skip connections, which provide more high-frequency details for face reconstruction. Experiments on Helen, CelebA face datasets and real-world images showed that the proposed method could make full use of facial feature information, and its peak signal to noise ratio (PSNR) and structural similarity (SSIM) were both higher than the comparison methods with better subjective visual effects.

INDEX TERMS Face super-resolution reconstruction, convolutional neural network, self-attention mechanism, residual network.

I. INTRODUCTION

Face images provide important information for human visual perception and computer analysis, which have been widely studied in recent years. However, in natural scenes, due to the limitation of the distance, face image captured by the cameras is usually small and blurred, which brings challenges to the face recognition. Super-resolution reconstruction is an effective method to improve the resolution of face images, which is of great significance to improve the image quality and enhance the richness of face information.

The existing image super-resolution reconstruction (SR) can be broadly divided into three categories: interpolation-based methods [1], reconstruction-based methods [2], and learning-based methods [3], [4]. The method based on interpolation mainly includes nearest neighbor interpolation, bilinear interpolation, and bicubic interpolation, etc. This method is fast and simple in calculation, but it has poor effect on image detail restoration, and the reconstructed image edge will be jagged. The method based on reconstruction mainly includes the iterative back-projection method [5] and the projections onto convex sets method [6], etc. This method establishes an observation model for the image acquisition process and then realizes super-resolution reconstruction by solving the inverse problem of the observation model. However, due to the limited prior knowledge obtained, more details cannot be recovered for complex images and the reconstruction performance is limited. At present, learning-based methods mainly include methods based on sparse coding [7], neighbor embedding [8] and deep learning [9]. For example, Yang et al. [10], [11] proposed a method based on sparse coding and dictionary learning, which sparsely represented the sample database composed of low-resolution (LR) image blocks and high-resolution (HR) image blocks, and found the over-complete dictionary corresponding to LR and HR image blocks through joint training. However, due to high learning requirements for the over-complete dictionary, its practicability is poor. Similarly, Timofte et al. [12], [13] combined sparse dictionary with neighborhood embedding, and proposed the fixed neighborhood regression method and the anchored neighborhood regression method, which improved the computational efficiency but poorly reconstructed the image details. In recent years, the super-resolution reconstruction methods based on deep learning have become a
research hotspot [14]–[16]. In 2014, Dong et al. [17] firstly proposed super-resolution reconstruction based on convolutional neural network (SRCNN), which utilized the strong feature expression ability of convolutional neural network to improve the accuracy of the reconstructed image. But due to the few convolutional layers, small receptive field and poor generalization ability, the network is difficult to extract the deep features of the image, and the reconstruction performance is limited. In 2016, Shi et al. [18] proposed an efficient sub-pixel convolutional neural network (ESPCN), which rearranged feature maps by the sub-pixel convolutional layer to obtain high-resolution images. In the same year, Kim et al. [19] proposed a very deep convolutional neural network (VDSR) and introduced residual learning to accelerate the training speed of the network. The quality of reconstructed image has been greatly improved, but with the deepening of the network, gradient explosion and network degradation will become more and more obvious. Face super-resolution [20] is a task of image reconstruction in a specific scene. The face reconstruction models should pay more attention to the restoration of facial details. Some scholars have proposed several improved methods based on this point of view. In 2015, Zhou et al. [21] proposed a bi-channel convolutional neural network to improve the problem of the face feature information loss to some extent by cross-layer output of input images. Wang et al. [22] improved the quality of the reconstructed image by introducing additional information (such as texture, edge, etc.) into the deep convolutional network, but the amount of calculation was large. In 2016, Zhu et al. [23] proposed a two-stage iterative method for face super-resolution reconstruction, which is difficult to train and had no obvious improvement effect. In 2018, Sun et al. [24] improved the effect of reconstructing images by increasing the depth of the network, while also increasing the complexity of the model.

The DCSCN network proposed in reference [25] enhances the feature extraction capability by cascading multiple $3 \times 3$ convolutional kernels and using skip connection. The PSNR and SSIM of reconstructed images have improved. However, due to the rich features and concentrated high-frequency information of face images, it is difficult for DCSCN to establish the correlation model between facial features. The feature extraction ability of the DCSCN network is limited and the reconstruction effect of face is poor. Inspired by the idea of attention mechanism and non-local neural networks in references [26]–[29], this paper introduces self-attention mechanism into DCSCN and appropriately removes convolutional layers to enhance the feature extraction ability of the network. So the network can learn purposefully, which is more conducive to the accurate reconstruction of face details, and the reconstruction effect is significantly improved.

II. RELATED WORK

The DCSCN proposed by Yamanaka et al. [25] is a deep network structure of cascade convolutions. It used skip connection to improve the utilization rate of feature maps and effectively alleviate the phenomenon of gradient disappearance of the network, making the network training easier. In addition, it also used the parallel convolutional layer (Network in Network) to reconstruct image details, which improved the reconstruction performance of the network. The network structure of DCSCN is shown in Figure 1.

As shown in Figure 1, DCSCN is composed of the feature extraction network and the reconstruction network. Firstly, seven cascaded $3 \times 3$ convolutional kernels constitute the feature extraction network. In order to make full use of local and global image features, the outputs of each convolution layer are input into the reconstruction network by skip connection. Then all the features are connected by concatenation, and a parallel convolutional layer (Network in Network) is used to reconstruct the details of the image. Finally, the residual image output by the up-sampling layer is added to the up-sampling image constructed by bicubic interpolation to reconstruct the original HR image. The focus of the DCSCN is to learn the residuals between the bicubic interpolation of the LR images and the original HR images. Because the feature extraction part of the network adopts 7 cascaded hierarchical structure, which has a huge number of parameters and cannot extract features on face images of rich and concentrated features. So we combine the convolutional

\[ \textbf{FIGURE 1.} \text{DCSCN network structure, which is composed of the feature extraction network and the reconstruction network. The inputs are the low-resolution image, and the outputs are the high-resolution image.} \]
layer with the self-attention mechanism to extract features. In order to reduce network redundancy, the last four layers of the feature extraction network are deleted, and the loss function is targeted improvement.

III. PROPOSED MODEL

The self-attention mechanism [26] can generate detailed information using associations from all the feature positions of the image, designed to capture the global dependency within the image, showing good performance in modeling the global dependency and computational efficiency. Its appearance provides a new idea for face super-resolution reconstruction to obtain the global features to restore texture details. In this paper, the self-attention mechanism is applied to the feature extraction process of DCSCN, named SARCN.

The feature extraction is further subdivided into shallow feature extraction and deep feature extraction. The global features and rich semantic information of low-resolution face images are obtained to restore more high-frequency details of high-resolution images. The proposed network structure is shown in Figure 2.

As shown in Figure 2, the improved network in this paper consists of three parts: shallow feature extraction module, deep feature extraction module and reconstruction module. The shallow feature extraction module is composed of cascaded $3 \times 3$ convolutional kernels in DCSCN, which is used to extract the shallow features. Taking the original LR images as the input can reduce the amount of computation. Then, in order to obtain more global features to further refine obtained feature maps, a self-attention mechanism is designed in the deep feature extraction module. It can explore the global dependency between any two features, which can help to enhance the expression ability of features and restore the texture details of the image. Finally, the feature maps are input into the reconstruction module to realize the up-sampling and added with the bicubic interpolation result of the LR image to obtain the HR face image.

A. SHALLOW FEATURE EXTRACTION MODULE

Shallow feature extraction network is composed of cascaded $3 \times 3$ convolutional kernels, and the number of convolutional kernels in each layer is changed to 64 to reduce the network parameters and computational complexity. Since the self-attention module was added into the network and the feature extraction ability was enhanced. In order to reduce the network redundancy, the last four convolutional layers are deleted, and each layer uses PReLU [30] as the activation function. The network directly takes LR images as the input without the pre-processing process of interpolation and amplification, so the amount of calculation is reduced. In order to make full use of the edge information of the image, the zero-padding operation is performed before each convolution, and the previous three convolutional layers can be expressed as:

$$F_l(Y) = Max[\lambda \times [W_l \times F_{l-1}(Y) + B_l], W_l \times F_{l-1}(Y) + B_l]$$

where $W_l$ is the $l$ convolutional filter which can be expressed as $n \times k \times k \times c$, where $n$ represents the number of convolutional kernels, where is 64; $k$ represents the size of convolutional kernels, where is 3; $c$ represents the image channels. The image normally has three channels called YCbCr. The study [31] has shown that human eyes are more sensitive to the brightness channel (namely Y-channel). It has been proved in the SRCNN paper that only mapping the Y-channel component of the image will not affect the reconstruction quality of the image [17], so here $c$ is taken as 1. $B_l$ is the bias of the $l$ layer, $F_{l-1}(Y)$ is the characteristic map of the previous layer output, and $\lambda$ is the parameter in the activation function PReLU.

B. DEEP FEATURE EXTRACTION MODULE

Deep feature extraction network combines the self-attention mechanism [26] with the residual block in the deep residual network [32], which aims to explore the global dependency between features and extract the deeper feature association of human face. The self-attention mechanism captured the global dependency within the image is similar to non-local neural networks [29] and commonly uses non-local means filter. A plenty of similar approaches have been introduced in the image restoration, such as image denoising,
FIGURE 3. Self-attention residual network. The softmax operation is performed on each row. $H \times W \times C$ means $C$ features with height $H$ and width $W$. The ”⊗” represents matrix multiplication and ”⊕” represents element-wise addition.

demosaicing, compression artifacts reduction [33] and Super-Resolution [34]. For Super-Resolution, self-attention mechanism helps to learn local and non-local information from the hierarchical features. The structure of the self-attention residual network is shown in Figure 3.

First, the shallow feature maps $x$ extracted by the shallow feature extraction module are sent to three $1 \times 1$ convolution layers accompanied by PReLU activation function, respectively, to generate new feature maps $F(x)$, $G(x)$, $H(x)$. Then $F(x)$ performs matrix multiplication with $G(x)$ through a transpose matrix, and applies a softmax layer to calculate the attention features $S(x)$:

$$S_{j,i} = \frac{\exp(F_i, G_j)}{\sum_{i=1}^{N} \exp(F_i, G_j)}$$  \hspace{1cm} (2)

where $N$ is the whole position space, and $S_{j,i}$ represents the influence of position $i$ on position $j$, also called characteristic correlation matrix. The more similar the features of two locations are, the greater the response value is, and the greater the correlation between them. Finally, the attention features $S(x)$ and the feature maps $H(x)$ perform matrix multiplication, and are mapped by a $1 \times 1$ convolutional layer. Then the final attention feature maps $A(x)$ is calculated as:

$$A(x) = S(x) \otimes H(x) + x_i$$  \hspace{1cm} (3)

where “$+x_i$” is referred to as the residual learning [31]. It can be seen that $A(x)$ is the deep feature maps we expect. It represents the features of all locations and therefore has global context information, which can selectively aggregate context information according to self-attention features when fusing with shallow feature maps. In general, the deep feature extraction module can easily capture more global features by learning the relationship between features in all locations to make similar features related to each other. The addition of global features can help restore more texture details.

C. RECONSTRUCTION MODULE

Before entering the reconstruction network, the feature maps extracted by the shallow feature extraction module and the deep feature extraction module need to be concatenated with different scale feature maps by concatenation. The purpose is to comprehensively utilize the local and global features to fuse the upper and lower layers information of the network. Inspired by reference [35], DCSCN proposed a parallel convolutional structure, which is usually composed of one (or more) $1 \times 1$ convolutional layer. The $1 \times 1$ convolutional kernels increase the non-linearity and enhance the robustness of the network while reducing the dimension of the feature maps which enter the reconstruction network, speeding up the calculation, and reducing the information loss. Using the parallel convolutional structure can significantly reduce the number of parameters and computational complexity and improve the reconstruction quality. So we do not improve the reconstruction network.

After feature extraction and non-linear mapping through the above network, it is also necessary to up-sample the acquired feature maps to achieve the final image reconstruction. Bicubic interpolation or bilinear interpolation, and transposed convolution can be used. However, the bicubic or bilinear interpolation kernels are fixed, which cannot restore the details well for face images. Although the transposed convolution can learn up-sampling kernels of the input feature maps, it need zero-padding in the learning process, resulting in distortion of images edge. In 2016, reference [18] proposed a new up-sampling method, sub-pixel convolution, which can reduce the influence of transposed convolution with zero-padding and reduce the amount of computation. Therefore, we introduce sub-pixel convolutional layer at the end of the network to realize up-sampling.
The essence of the sub-pixel convolution is that the image size can be changed by adding a phase shift layer after the conventional convolutional layer, which is utilized in the process of up-sampling. The interpolation function is implicitly included in the previous convolution layers and can be automatically learned, as shown in Figure 4. The sub-pixel convolution includes two steps of convolution and rearrangement. For a network with the scale factor of $k$, the number of convolution channels of the sub-pixel convolutional layer should be set to $r^2$ ($r = k$). Then, $r^2$ channels of each pixel of feature maps are rearranged into $r \times r$ region in a specific order, corresponding to $r \times r$ sub-block in the high-resolution image, so that feature maps of size $r^2 \times H \times W$ are rearranged into a $1 \times rH \times rW$ high resolution image. The previous convolutions are carried out on the low-resolution images, and the arrangement does not need convolution operation, so the efficiency will be very high.

In the reconstruction process, sub-pixel convolution performs up-sampling operation on the feature maps and combines the final output features to generate a residual image. Because of the great similarity between the input LR image and the original HR image, the pixel information of the residual image mostly approaches to 0. So the final HR image can be obtained by adding the generated residual image and the input LR image after up-sampling by bicubic interpolation. Therefore, the focus of the improved CNN model is to learn the residual between the bicubic interpolation of the LR images and the HR original images. The input LR images, the original HR images and the generated residual images are used to complete the reconstruction of HR images, which further improves the quality of the image reconstruction. The output of the network is:

$$ y = R( F_l( Y ) + B_l ) + U( Y ) $$  \hspace{1cm} (4)

where $R$ and $U$ represent reconstruction results and bicubic interpolation results respectively.

### D. LOSS FUNCTION

In the field of super-resolution reconstruction, Mean Square Error (MSE) is currently adopted by most methods for network training as the loss function. In view of the complex features of face images and the concentration of high-frequency information, we change the loss function into Mean Absolute Error (MAE) function to estimate the error between the reconstructed HR image $\hat{I}$ and ground truth HR image $I$, which is expressed as:

$$ \text{Loss} = \frac{1}{n} \sum_{i=1}^{N} \| I_i - \hat{I}_i \| $$  \hspace{1cm} (5)

### IV. EXPERIMENTS

In this paper, Helen [36] dataset, CelebA [37] dataset, CMU + MIT [38] face dataset and WHU-SCF [39] (Wuhan University - Surveillance Camera Face) dataset are used to train and evaluate our improved model. The experimental environment is TensorFlow and Caffe deep learning framework based on Win10. The hardware configuration is Intel (R) core (TM) i7-8750h @ 2.20GHz CPU, 8GB RAM. The graphics card is NVIDIA GeForce GTX 1070 GPU. CUDA 9.0 and cudnn 6.1 are used for GPU acceleration.

#### A. TRAINING DATASET

There are 350 front face images in Helen [36] dataset to be used as the training set. Each face image in the training set is different in expression, posture, scale, illumination and other aspects. Since the proposed network has a large number of parameters that need to be trained, it is necessary to enhance the training set to better match the proposed model. In this paper, 350 face images in the training set were firstly rotated clockwise by 0°, 90°, 180° and 270°, respectively, and then mirrored to obtain a total of 2,800 images.

Since high-frequency detail information will be lost after interpolation and down-sampling of HR images, the LR images after interpolation can be taken as the input of the network. The detail information between the high-resolution and low-resolution images can be obtained through the improved network for reconstruction. In this paper, the original HR images are down-sampled by factor $k$ ($k = 2, 3, 4$) using bicubic interpolation to generate the corresponding LR images. The LR training images are cropped into a set of sub-blocks of $l_{sub} \times l_{sub}$, and the corresponding HR images are cropped into sub-blocks of $kl_{sub} \times kl_{sub}$. There, we choose $l_{sub} = 32$ and clipping step size of $kl_{sub}$ to obtain about 100,000 groups of LR / HR blocks. Some images before clipping are shown in Figure 5.

#### B. TRAINING SETUP

We used the Adam optimization method (parameter setting: $\beta_1 = 0.9, \beta_2 = 0.999$) instead of the stochastic gradient

---

**FIGURE 5.** Sample faces from training dataset, all of which are uncropped HR face images.
FIGURE 6. Visual results comparison of different methods and proposed model on Test1-man with scale factor of 3. Column (a) Bicubic interpolation; (b) SRCNN; (c) FSRCNN; (d) ESPCN; (e) VDSR; (f) DCSCN; (g) Ours; (h) Ground-truth HR.

descent method (SGD) to minimize loss. Adam uses first-order moment estimation and second-order moment estimation of the gradient to adjust the learning rate of each parameter dynamically. The advantage is that after bias correction, the learning rate of each iteration has a certain range so that the parameters are relatively stable. The learning rate $lr$ was initialized to 0.002; the learning rate attenuation factor was set to 0.005; the momentum was set to 0.9. When the loss does not decrease after $5 \times 10^4$ iterations of training, the learning rate is decreased by a factor of 2 and training will be finished if the learning rate goes lower than $2 \times 10^{-5}$. At the end of the training, iteration is about $3.5 \times 10^5$ times. The network convolutional layers used MSRA method proposed by He et al. [40] for parameter initialization.

C. EXPERIMENTAL RESULTS AND ANALYSIS

We mainly studied the reconstruction effect of images on the factor of 2, 3 and 4. Experiments were carried out on Helen [36] and CelebA [37] testing sets, and compared with bicubic interpolation, SRCNN [17], FSRCNN [41], ESPCN [18], VDSR [19], DCSCN [25] from two aspects of subjective visual evaluation and objective quantitative standard. Two widely used objective quantification included peak signal-to-noise ratio (PSNR) and structural similarity.
index (SSIM), were used to evaluate the reconstruction results. Test1 is 35 test images randomly selected from Helen dataset; Test2 and Test3 are 20 and 50 test images randomly selected from CelebA dataset.

Figure 6 and Figure 7 have shown the qualitative evaluation results of the above methods on Test1-man and Test3-woman respectively. Because human eyes are the region with high signal energy and the most important feature of human face, we have partially enlarged the eyes region of human face images, which can visually evaluate the quality of face images reconstruction with various SISR methods. Figure 6 showed that for the reconstruction of human eyeballs and wrinkles, except VDSR and DCSCN, other methods are seriously blurred, and even leading to distortion of wrinkles. However, the face reconstruction image of the proposed method has clear eyeballs and obvious wrinkles, which is greatly improved compared with other methods. Similarly, as shown in Figure 7, the proposed method perfectly reconstructed human eyeballs, eyebrows and eyelashes, while images reconstructed by other methods have obvious ringing effect and blurred edges. Compared with above methods, the proposed method can restore more details of face images, which shows that the self-attention module can extract features of face images better, resulting in sharper edges and better visual effects.

Table 1 provides the quantitative evaluation results of the proposed method and Bicubic, SRCNN, FSRCNN, ESPCN, VDSR and DCSCN at different scales ($\times 2$, $\times 3$, $\times 4$) on different testing sets (Test1, Test2, Test3). The best performances are shown in bold. The results of all methods are retrained and tested by our training sets in the author’s published source code. From table 1, it can be seen that the proposed method can handle different scales better, and the PSNR and SSIM of reconstructed images have basically
achieved the highest value, slightly higher than VDSR and DCSCN, but significantly higher than other comparison methods. Compared with the bicubic method, the average PSNR and SSIM of the proposed method are increased by 4.420 dB and 0.047, respectively, when the scale factor is 3.

### D. EXPERIMENTS ON REAL-WORLD IMAGES

In previous experiments, test LR faces are generated from HR faces smoothly by fixing down-sampling. It does not consider the causes of real-world LR image formation. Hence, in order to further examine the strengths of the proposed model, experiments are carried out on two real-world datasets: CMU+MIT [38] dataset and WHU-SCF [39] dataset.

The CMU+MIT dataset consists many pictures of real-life situations, few sample images are shown in Figure 8. Three faces from each picture were shown in Figure 8. We cut out the captured faces, which are the LR images of the real scene, and directly input them into the corresponding model for reconstruction. The result was shown in Figure 9, with a scale factor of 3.

As shown in Figure 9, the first column on the left is the input LR images, each column in the middle is the reconstruction results of different methods, and the last column is the results of the proposed model. It can be stated that images obtained by bicubic interpolation are generally fuzzy, with few high-frequency details, and it is difficult to distinguish facial features. Images obtained by SRCNN have more noise points, and results from FSRCNN to DCSCN have become more and more clear, but the edges of eyebrows, eyes and other parts are not sharp enough. The results of the proposed method are sharper, although there is a little noise, but less fuzzy, with better details. In addition, we have only LR images to be super-resolved, without the corresponding HR ground truth. Therefore, we need to introduce quantitative reference-free image quality evaluation indexes [42]. We use natural image quality evaluator (NIQE) for assessment.

### TABLE 1. Average PSNR(dB)/SSIM on Test1, Test2 and Test3 compared with different SR methods.

| Dataset | Scale | Bicubic PSNR/SSIM | SRCNN PSNR/SSIM | FSRCNN PSNR/SSIM | ESPCN PSNR/SSIM | VDSR PSNR/SSIM | DCSCN PSNR/SSIM | Ours PSNR/SSIM |
|---------|-------|-------------------|-----------------|------------------|----------------|----------------|----------------|---------------|
| Test1   | 2     | 38.229/0.920      | 39.893/0.931    | 40.061/0.931     | 40.263/0.932   | 40.382/0.947   | 40.053/0.954   | 40.526/0.960   |
|         | 3     | 34.788/0.906      | 36.801/0.909    | 37.056/0.900     | 37.384/0.902   | 37.490/0.919   | 37.495/0.929   | 37.703/0.922   |
|         | 4     | 33.261/0.871      | 34.823/0.880    | 35.100/0.882     | 35.256/0.892   | 35.442/0.897   | 35.639/0.893   | 35.930/0.902   |
| Test2   | 2     | 34.280/0.935      | 37.001/0.937    | 37.106/0.937     | 37.673/0.947   | 37.905/0.949   | 38.324/0.967   | 38.876/0.974   |
|         | 3     | 31.169/0.881      | 34.906/0.910    | 35.332/0.912     | 35.962/0.920   | 35.988/0.928   | 36.179/0.941   | 36.166/0.942   |
|         | 4     | 29.022/0.827      | 31.308/0.871    | 31.756/0.873     | 31.892/0.875   | 32.207/0.901   | 32.236/0.899   | 32.617/0.904   |
| Test3   | 2     | 35.302/0.939      | 38.066/0.950    | 38.107/0.950     | 38.642/0.960   | 38.805/0.961   | 39.486/0.975   | 40.124/0.980   |
|         | 3     | 31.948/0.904      | 35.849/0.824    | 36.321/0.926     | 36.925/0.931   | 37.179/0.934   | 37.189/0.940   | 37.321/0.956   |
|         | 4     | 29.683/0.858      | 32.067/0.904    | 32.539/0.916     | 32.672/0.916   | 33.005/0.921   | 33.055/0.919   | 33.697/0.926   |
The comparison results are shown in Table 2. Note that the larger the NIQE is, the worse the image quality is. It can be stated that the proposed method achieved the highest score. It has demonstrated that the feature extraction ability is stronger, the model can learn more edge information, further showed the effectiveness of self-attention module for feature extraction.

Further, in order to examine the effectiveness of the proposed model on surveillance images, we also carried out experiments on the WHU-SCF dataset. WHU-SCF dataset has videos taken by Wuhan University surveillance cameras in different environments. Few frames have been selected from a video locally captured by the surveillance camera under various circumstances. The camera is close to the human face, and the video has been shot under low illumination. The representative video frames which have been selected from the video are given in Figure 10. Faces extracted from the frames were used as LR test faces, and LR reconstruction results are also shown in the right of Figure 10. After the reconstruction of our model, the definition of LR
faces is improved, and the facial features can be basically distinguished, but the reconstructed images also contain a small amount of noise and blur, which is the shortcoming of our model and the direction of our next improvement.

V. MODEL ANALYSIS

A. ANALYSIS ON SELF-ATTENTION MODULE

In order to explore the influence of the self-attention module on training, we compared the proposed network with the original network without the self-attention module. The experimental results are shown in Figure 11.

It can be seen from Figure 11 that the network using self-attention had a faster convergence rate than the network without self-attention, and its PSNR was higher about 0.2dB, which had fully proved that using self-attention module can effectively improve feature extraction capability and reconstruction performance.

B. ANALYSIS ON DIFFERENT LAYERS OF SHALLOW FEATURE EXTRACTION NETWORK

Researches have shown that the network depth is an important factor to influence the effect of super-resolution reconstruction. Increasing network depth can extract more features, but as the depth increases, the gradient disappearance or dispersion of the network will also become more and more obvious. And the overall training difficulty will increase, so it is necessary to strike a balance between the reconstruction accuracy and training difficulty. Since the feature extraction capability of the model is greatly enhanced by the addition of the self-attention module, the number of layers of the shallow feature extraction can be appropriately reduced. We had trained and tested the network with shallow feature extraction depths of 1 to 7. The experimental results were shown in Table 3, and the scale factor is 3.

| Depth | PSNR(dB) | SSIM | Time/s |
|-------|----------|------|--------|
| 1     | 37.498   | 0.918| 28.2   |
| 2     | 37.679   | 0.920| 28.8   |
| 3     | 37.703   | 0.922| 29.8   |
| 4     | 37.704   | 0.921| 30.5   |
| 5     | 37.703   | 0.921| 31.4   |
| 6     | 37.703   | 0.922| 33.2   |
| 7     | 37.704   | 0.922| 33.7   |
As we can see from Table 3 that the average PSNR and SSIM values of the reconstruction model increased with the increase of the depths of shallow feature extraction network, and model with the depth of 7 obtained best reconstruction performance. Considering the reconstruction speed and calculation amount, since PSNR of the model with the depth of 3 model is very close to the model with the depth of 7, and its PSNR value tends to be stable. Therefore, we selected the model with depth of 3 as the shallow feature extraction network.

C. ANALYSIS ON DIFFERENT LOSS FUNCTION

In the process of network training, the choice of loss function will affect the speed of training and the performance of the model. In view of the features of face images, such as strong correlation and high-frequency information concentration, we choose the Mean Absolute Error (MAE), which is more stable to deal with outliers, instead of the Mean Square Error (MSE). In order to prove the superiority of the used loss function, a series of comparative experiments were carried out on MSE and MAE.

Average PSNR and SSIM of different test sets after training and testing with MSE and MAE loss functions are shown in Table 4. The results have shown that both PSNR and SSIM are the best when using MAE, which further proved the effectiveness of the loss function MAE.

VI. CONCLUSION

In order to solve the problems of weak feature expression, low feature utilization and network redundancy in the DCSRN network, we improve the network from the perspective of feature extraction. By adding the self-attention residual module, simplifying convolutional layers to reduce network redundancy, and improving the loss function, we have improved the network reconstruction performance. Experimental results on standard face datasets and surveillance video frames have shown that the self-attention module can effectively capture the global dependencies between features and reconstruct more high-frequency information. The proposed method has some advantages compared with other face super-resolution reconstruction methods. Both subjective reconstruction effect and objective evaluation criteria of the method are improved, and the reconstructed face images have higher quality.

REFERENCES

[1] M. Batz, A. Eichenseer, J. Seiler, M. Jonscher, and A. Kaup, “Hybrid super-resolution combining example-based single-image and interpolation-based multi-image reconstruction approaches,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2015, pp. 56–62.
[2] K. I. Kim and Y. Kwon, “Single–image super–resolution using sparse regression and natural image prior,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 6, pp. 1127–1133, Jun. 2010.
[3] H. Chang, D.-Y. Yeung, and Y. Xiong, “Super–resolution through neighbor embedding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Washington, DC, USA, Jun. 2004, pp. 275–282.
[4] K. Ni, S. Kumar, N. Vasconcelos, and T. Nguyen, “Single image super-resolution based on support vector regression,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., Aug. 2006, p. 2.
[5] M. Irani and S. Peleg, “Improving resolution by image registration,” CVGIP, Graph. Models Image Process., vol. 53, no. 3, pp. 231–239, May 1991.
[6] H. Stark and P. Oskoui, “High–resolution image recovery from image–plane arrays, using convex projections,” J. Opt. Soc. Amer. A, Opt. Image Sci., vol. 6, no. 11, p. 1715, Nov. 1989.
[7] J. Jiang, R. Hu, Z. Wang, and Z. Han, “Noise robust face hallucination via locality–constrained representation,” IEEE Trans. Multimedia, vol. 16, no. 5, pp. 1268–1281, Aug. 2014.
[8] J. Jiang, R. Hu, Z. Wang, and Z. Han, “Face super–resolution via multilayer locality–constrained iterative neighbor embedding and intermediate dictionary learning,” IEEE Trans. Image Process., vol. 23, no. 10, pp. 4220–4231, Oct. 2014.
[9] Q. Qi, R.-S. Jia, Q.-C. Mao, H.-M. Sun, and L.-Q. Zuo, “Face detection method based on cascaded convolutional networks,” IEEE Access, vol. 7, pp. 110740–110748, 2019.
[10] J. Yang, J. Wright, T. Huang, and Y. Ma, “Image super–resolution as sparse representation of raw image patches,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Anchorage, AK, USA, Jun. 2008, pp. 1–8.
[11] 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.
[12] R. Timofte, V. De, and L. V. Gool, “Anchored neighborhood regression for fast example–based super–resolution,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Sydney, NSW, Australia, Dec. 2013, pp. 1920–1927.
[13] R. Timofte, V. De, and L. V. Gool, “A+: Adjusted anchored neighbor–hood regression for fast super–resolution,” in Proc. IEEE Int. Conf. Asian Conf. Comput. Vis. (ACCV), Singapore, Apr. 2015, pp. 111–126.
[14] Y. Tian, R. S. Jia, S. H. Xu, R. Hua, and M. D. Deng, “Super–resolution reconstruction of remote sensing images based on convolutional neural network,” J. Appl. Remote Sens., vol. 13, no. 4, pp. 1–12, 2019.
[15] K. Jiang, Z. Wang, P. Yi, J. Jiang, J. Xiao, and Y. Yao, “Deep distillation recursive network for remote sensing imagery super–resolution,” Remote Sens., vol. 10, no. 11, p. 1700, Oct. 2018.
[16] Q.-C. Mao, H.-M. Sun, Y.-B. Liu, and R.-S. Jia, “Mini–YOLOv3: Real–time object detector for embedded applications,” IEEE Access, vol. 7, pp. 133529–133538, 2019.
[17] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super–resolution,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland, Springer, 2014, pp. 184–199.
[18] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real–time single image and video super–resolution using an efficient sub–pixel convolutional neural network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1874–1883.
[19] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super–resolution using very deep convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pat- tern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 1646–1654.
[20] J. Shi, X. Liu, Y. Zong, C. Qi, and G. Zhao, “Hallucinating face image by regularization models in high–resolution feature space,” IEEE Trans. Image Process., vol. 27, no. 6, pp. 2980–2995, Jun. 2018.
[21] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, “Learning face hallucination in the wild,” in Proc. 29th AAAI Conf. Artif. Intell. (AAAI), vol. 7, Mar. 2015, pp. 3871–3877.
[22] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, “Deep networks for image super–resolution with sparse prior,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 370–378.
[23] S. Z. Zhu, S. Liu, C. C. Loy, and X. Tang, “Deep cascaded bi–network for face hallucination,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Cham, Switzerland: Springer, 2016, pp. 614–630.
[24] Y. T. Sun, H. Song, K. Zhang, and F. Yan, “Face super–resolution via very deep convolutional neural network,” J. Comput. Appl., vol. 38, no. 4, pp. 1141–1145, 2018.
[25] J. Yamanaka, S. Kuwashima, and T. Kurita, “Fast and accurate image super resolution by deep CNN with skip connection and network in network,” in Proc. Int. Conf. Neural Inf. Process. Cham, Switzerland: Springer, 2017, pp. 217–225.

[26] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” 2018, arXiv:1805.08318. [Online]. Available: https://arxiv.org/abs/1805.08318

[27] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Washington, DC, USA, Jul. 2017, pp. 3156–3164.

[28] Y. L. Zhang, K. P. Li, K. Li, L. C. Li, B. N. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Sep. 2018, pp. 286–301.

[29] X. L. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 7794–7803.

[30] C. Xu, T. Liu, D. Tao, and C. Xu, “Local rademacher complexity for multi-label learning,” IEEE Trans. Image Process., vol. 25, no. 3, pp. 1495–1507, Mar. 2016.

[31] J.-B. Huang, A. Singh, and N. Ahuja, “Single image super-resolution from transformed self-exemplars,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 5197–5206.

[32] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[33] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, “Residual non-local attention networks for image restoration,” 2019, arXiv:1903.10082. [Online]. Available: https://arxiv.org/abs/1903.10082

[34] Q.-M. Liu et al., “Face Super-Resolution Reconstruction Based on Self-Attention Residual Network,” IEEE Access, vol. 8, pp. 40194–40203, Oct. 2020.

RUI-SHENG JIA is currently a Full Professor with the College of Computer Science and Engineering, Shandong University of Science and Technology, China. He has 32 first-author publications and has 25 coauthor publications. His research interests include artificial intelligence, big data processing, information fusion, microseismic monitoring, and inversion.

CHAO-YUE ZHAO was born in Shandong, China, in 1996. She received the B.S. degree from Shandong Women’s University, China, in 2018. She is currently pursuing the M.S. degree with the Shandong University of Science and Technology. Her research interests include image processing and deep learning.

HONG-MEI SUN received the B.S. and M.S. degrees in computer science from the Shandong University of Science and Technology, China, in 1995 and 2005, respectively. She is currently a Lecturer with the College of Computer Science and Engineering, Shandong University of Science and Technology, and the Leader of the Key Research and Development Projects of Shandong Province, China. She has four first-author publications and has five coauthor publications. Her research interests include micro seismic monitoring technology and software engineering.

XING-LI ZHANG received the master’s degree in computer software and theory from the Taiyuan University of Technology, in 2010. She is currently a Lecturer with the Shandong University of Science and Technology and the Leader of the National Natural Science Foundation of China. Her main research interests are microseismic signal analysis and processing, and deep learning algorithm.

Q.-M. Liu et al.: Face Super-Resolution Reconstruction Based on Self-Attention Residual Network

VOLUME 8, 2020