Machine Vision and Applications (2023) 34:106
https://doi.org/10.1007/s00138-023-01460-9

ORIGINAL PAPER

Global attention guided multi-scale network for face image super-resolution

Jinlu Zhang[1,2] · Mingliang Liu[1,2] · Xiaohang Wang[1,2]

Received: 28 November 2022 / Revised: 10 August 2023 / Accepted: 24 August 2023 / Published online: 16 September 2023

Abstract
Face image super-resolution (FSR) is a subtask of image super-resolution that aims to enhance the resolution of facial images. Previous FSR methods have leveraged facial prior information, such as parsing maps and landmarks, to improve their performance. However, these methods have not fully utilized the potential of this prior information, as they typically use the same network structure for different types of facial information. To address this limitation, we propose a new network structure called GAMFSR, which incorporates parsing map prior information and includes a global attention module (GAM) to improve the utilization of the parsing map. Additionally, we developed a multistage super-resolution network for preprocessing, which further improves the prediction accuracy. We conducted ablation studies to investigate the effectiveness of GAM and the impact of different prior information. Our experimental results demonstrate that our approach significantly enhances the guidance of parsing map features and achieves better performance with less prior information.

Keywords Face image super-resolution · Attention mechanism · Facial prior information · Parsing maps

1 Introduction

Due to limitations in data transmission equipment, images are often compressed during transmission, resulting in distortion and blurring. This is particularly problematic for face images, which typically occupy only a small portion of the overall picture, and thus suffer more degradation during transmission. This can negatively impact people's perception and the ability to identify characters in the image. To address this issue, the image super-resolution method (SISR) was developed to generate high-resolution (HR) images from low-resolution (LR) images. However, SISR is a general SR method and is often unable to reconstruct the severely degraded face region into an identifiable image. Thus, a specialized super-resolution method (FSR) is necessary for face images. Image super-resolution is a classic problem in computer vision, with different solutions for various applications such as transportation, medical care and object recognition.

Including from interpolation to deep learning methods, there are many approaches to image super-resolution. The interpolation method is to use the existing pixel points to calculate the unknown pixel value and insert it based on a fixed formula relationship. Such methods are fast and low in memory consumption. However, due to the simplicity of the calculation method, the results lack high-frequency information and are unsatisfactory in terms of quantitative and qualitative. With the development of technology and the update of hardware devices, the method of deep learning has also emerged. For example, SRCNN [2] achieved great success at that time, for the first time proposed to use a three-layer neural network to complete the task of image super-resolution. FSRCNN [3] achieves neural network acceleration by compressing the number of network channels and using a transposed convolution layer instead of preprocessing. VDSR [4] proposes a deep residual block-stacking method to improve the quality of SR results.

This work is supported by natural science foundation of Heilongjiang province of China (No. LH2020F046).

✉ Mingliang Liu
mll_0608@163.com

Jinlu Zhang
jlz_154615@163.com

Xiaohang Wang
xhw_5800@163.com

[1] Department of Automation, Heilongjiang University, Harbin 150080, Heilongjiang, China

[2] Key Laboratory of Information Fusion Estimation and Detection, Heilongjiang University, Harbin 150080, Heilongjiang, China

With the wide range of applications, FSR has become the focus of attention in the field of image super-resolution. FSR is a subdomain of single image super-resolution (SISR). FSR is also a work that aims at searching for a reflection between low-resolution (LR) and high-resolution (HR) images to make images have more details, which is a complex problem for it proposes to use the limited information to search for the best one from several solutions. However, as a subdomain of SISR, FSR is different from SISR tasks cause the resolution of face images always have lower resolution than other images, and the degradation of face images is usually serious to the limitation of physical systems and image transmission conditions; for example, FSR needs to reconstruct a 128×128 image by a scale factor of 8, which SISR tasks always work on the lower factors than it.

Different from SISR, several methods have been proposed for FSR, i.e., general FSR [5–7], prior-guided FSR [8–10], attribute-constrained FSR [11–13], identity-preserving FSR [14–16] and reference FSR [17–19]. One of these methods, prior-guided FSR consider there exists prior knowledge in face images, which can conduct the FSR training work. Previous studies show that the structural information and correspondence prior knowledge hide in face images can help restore more accurate details on HR face images [20]. However, previous studies have several issues that need to be optimized. First, many previous methods, such as SRCNN, VDSR of the SISR method and FSRNet of the FSR method, use the bicubic interpolation algorithm as a preprocessing step, which results in an extra computation and the inaccuracy of the interpolation method also affects the accuracy of the network, resulting in poor network output. Later, based on the SRCNN method, the preprocessing steps of subpixel and transposed convolutional layer instead of interpolation methods are proposed, and these methods have achieved good results under low scaling factors. However, FSR tasks usually work at large scaling factors (e.g., 8x), so these alternative methods are difficult to directly apply to FSR tasks. Second, there is a lot of prior information in the face image. For example, the prior information of parsing map points out the contours and positions of components in the face, and the prior information of landmark marks the entire face in the form of key points. However, the existing FSR methods use similar networks to predict different prior information and cannot make predictions by using the characteristics of prior information in a targeted manner. Moreover, there is information overlap between the prior information. For example, the parsing map and the landmark both record the location information of key components. Predicting too much prior information will bring redundant computation to the network. Third, most of the existing FSR methods are end to end, which is difficult to train at large scaling factors. As a result, the network is made to converge slowly during the process of training work.

To tackle these issues, we propose the GAMFSR network structure, which leverages a convolutional neural network (CNN) and parsing map prior information. Our network comprises three main parts: the base network, estimation network and reconstruction network. In the base network part, we design a lightweight network structure with a cascade of residual blocks to extract features from the input image and then use a transposed convolutional layer to reconstruct images at a low scaling factor. Multiple lightweight networks are stacked in series to form our base network. During the training process, each layer except for the first layer uses the weight of the previous layer as pretraining weight and tunes it as the weight of the current layer. The base network part can output the coarse super-resolution result with the same resolution as the final network output and serve as the input of the prediction network.

In the estimation part, we utilize the residual blocks to extract features from the output of the base network and input the extracted features into a U-Net structure network for multi-scale feature extraction. Furthermore, we propose a global attention mechanism (GAM) and integrate it into the U-Net structure for better utilization of the parsing map prior information. The GAM improves the accuracy of predicting the parsing map with less error. Finally, the output results of the prediction network and the base network part are input to the reconstruction network part to generate the super-resolution image.

We evaluate our proposed method on two popular benchmark datasets: Helen and CelebA-Mask-HQ [21, 22].

The innovations and contributions of our work can be summarized as follows:

1. We design a novel network structure (GAMFSR) for face image super-resolution, which consists of base network, estimation network and reconstruction network. Experiments show that it has excellent performance.
2. We develop a global attention mechanism (GAM) to help the estimation network predicts prior information more effectively. The GAM is applied to the U-Net module of the estimation network to generate parsing maps more accurately.
3. Furthermore, we design a multistage SR network with a special training strategy, which can upsample input images resulting in more high-frequency information.

The remaining sections of the article are summarized as follows: Sect. 2 introduces related work, including SISR methods, FSR methods and attention mechanisms related to SR approaches. Section 3 focuses on the structural details of the GAMFSR network we proposed and the subnetworks in our network are presented. Section 4 presents our experimental results on public datasets and analyzes the effectiveness of our method. Section 5 summarizes the work.

2 Related work

This section will review the development process of the method of single image super-resolution, as well as the method of face image super-resolution. For face image super-resolution methods, we will mainly introduce deep learning-based methods. Finally, we summarize some attention mechanisms that are influential in image super-resolution.

2.1 Single image super-resolution

Due to the photographic environment of life, the face image usually occupies a small proportion in the photograph, and this results in a blurry perception and difficulty for face recognition. To deal with such problem, Baker and Kanade proposed high-scale-factor face image super-resolution in for the first time. Liu et al. [23] implemented a two-stage method which consists of a global parametric model and a local non-parametric model for FSR task. Wang et al. [24] developed a method using eigentransformation to implement the results of emphasize the discrimination between face images. Considering that images at one solution have mappings which similar to linear relationship with their counterparts at other resolutions, Jian et al. [25] proposed a FSR method based on singular value decomposition. However, these traditional methods are difficult to find out the mapping from LR to HR when we need large upsample scale factor.

Dong et al. [2, 3] use deep learning network on the task of SISR for the first time, the method named SRCNN they proposed upsample images as LR using bicubic-interpolated first, and then use neural network learning mapping from LR to HR. Same as SRCNN, Kim et al. [4] proposed VDSR, a 20-layer deep learning network also based on the result of bicubic interpolation [26]. Zhang et al. develop a network structure using residual-in-residual blocks and channel attention mechanism to resolve such issue [27]. Ledig et al. use generative adversarial network (GAN) to reconstruct images [28], and Wang et al. introduce residual-in-residual dense block [29] to optimize [28]. Contrast to the commonly used objective image evaluation indicators, such as PSNR and SSIM [30], the subjective perception of SR images is getting more and more attention. To address with the perceptual quality of images, generative adversarial networks have also been proposed.

2.2 Learning-based face image super-resolution

Recently, many FSR methods using facial prior knowledge based on deep convolutional neural networks have been proposed. Incorporating prior knowledge, these methods can recover more reasonable results and get better performance than SISR methods in the field of face image super-resolution. Song et al. [31] designed a network crop the face image into different parts according to the landmark prior information and learned the mapping between each corresponding part. Yu et al. [32] proposed an end-to-end learning network using landmark, heatmap and parsing map prior information in the intermediate part for FSR task, which achieved good performance. Li et al. [33] developed feedback network to enhance neural network's power and detail representation for image super-resolution. Recently, Ma et al. [34] learn a deep iterative collaboration between two recurrent networks: One network estimates heatmap prior information and helps another one recover face image. And [35–38] design networks which guided by facial component heatmaps.

2.3 Attention mechanism

The attention mechanism has gained widespread use in deep learning in recent years, finding applications in various tasks such as natural language processing and computer image processing. By incorporating attention mechanisms in neural networks, they gain the ability to discriminate feature importance. Attention mechanisms can be divided into spatial attention (SA) and channel attention (CA), and their implementation varies across different tasks. For instance, Zhang et al. proposed the residual channel attention block (RCAB) in RCAN, which combines the residual method and channel attention method, achieving excellent results in SISR. Meanwhile, Hu et al. proposed the squeeze-and-excitation networks (SENet), which use a fully connected layer to predict the importance of each channel and applies it to the corresponding channel. KIM et al. proposed the attention module of the residual attention module (RAM), which effectively combines channel attention and spatial attention, making it better suited for SR problems.

3 Methods

In this section, we demonstrate the design methodology of our GAMFSR. We start with the overall framework of GAMFSR and then introduce the subnetworks of the framework separately.

3.1 Network architecture

The architecture of our network is shown in Fig. 1, which consists of base network, estimation network and reconstruction network.

The input to the network is extremely blurry low-resolution face images, denoted as I^{LR}. Because of the poor resolution of the input image, we first reconstruct the image through the base network for a more accurate result of the estimation network.

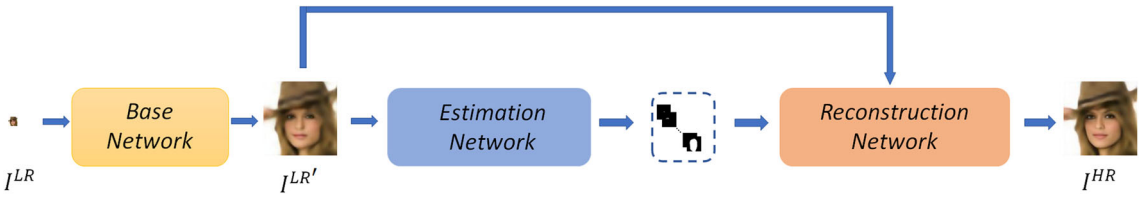

Fig. 1 Structure of the GAMFSR network. The input of the network is LR face images, which are processed by the base network to get the coarse super-resolution images. And the results are input into the estimation network for parsing map prediction. Finally, the estimation results of the estimation network and the coarse SR results of the base network are input into the reconstruction network for HR results reconstruction

$$I^{LR'} = B(I^{LR}), \tag{1}$$

where B is the mapping from LR to HR generated by the base network. $I^{LR'}$ is the SR result of the base network and then sent $I^{LR'}$ to the estimation network to predict parsing map prior information.

$$map^{est} = E(I^{LR'}), \tag{2}$$

where E denotes the mapping of estimation network and map^{est} is the estimation results of the estimation network.

And we integrate the estimation result map^{est} with the results of the base network $I^{LR'}$ by reconstruction network at last. The process can be expressed by:

$$I^{HR} = F(I^{LR'}, map^{est}), \tag{3}$$

where F denotes the mapping which represents by reconstruction network and I^{HR} is the final output of the whole network architecture.

In one layer of the base network, we choose mean square error (MSE) as the cost function. In the stage of the one layer in the base network, the optimization objective can be expressed as:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \|b(Y^i; \theta) - X^i\|_2^2, \tag{4}$$

where Y^i and X^i are the ith LR input and ground-truth image of current layer, and $b(Y^i; \theta)$ is the current layer output for Y^i with parameters θ.

In the estimation network and reconstruction network, we also use MSE as the cost function, and the optimization objective at this stage can be expressed as:

$$\min_{\theta} \frac{1}{2n} \sum_{i=1}^{n} \{\lambda \|E(Z^i; \theta) - P^i\|_2^2 + \|O(Z^i; \theta) - GT^i\|_2^2\}, \tag{5}$$

where θ denotes the parameters, λ is the weight of prior loss, and Z^i, P^i and GT^i are the input, the ground-truth parsing map prior information and image, respectively. $E(Z^i; \theta)$ and $O(Z^i; \theta)$ are the estimation of prior information and the SR results of the whole network, respectively.

3.2 Base network

In most end-to-end SR methods, the upsampling process is usually performed with interpolation, transposed convolutional layers or subpixel convolutions layers. However, since the above methods upsampled by one step under high scaling factors lead to unpleasant results, we progressively predict coarse SR result at each layer in our base network. We need to optimize computation and memory space consumption while achieving high-quality results in the base network. Therefore, we use a lightweight super-resolution network structure as a level in the base network, and each level uses a transposed convolution layer for $\times 2$ upsampling. The base network consists of 3 layers for super-resolving an LR image at a scale factor of 8.

The design of the level in the base network is based on the improvement of FSRCNN. We keep the main framework of FSRCNN and replace the original structure with skip connection in the nonlinear mapping stage. The structure of one layer is shown in Fig. 2.

After each level, the input image is upsampled by a scale factor of 2 with a transposed convolutional layer. We initialize each convolutional layer with a Gaussian distribution with zero mean and standard deviation of 0.001. The output image of each level is fed into the next level until the desired scale factor is reached. The structure of the whole base network is a cascade of CNN networks in the shape of Fig. 2.

3.3 Estimation network and global attention module

Estimation Network In the context of face image super-resolution, prior information such as color and texture is essential. Additionally, face images have unique prior infor-

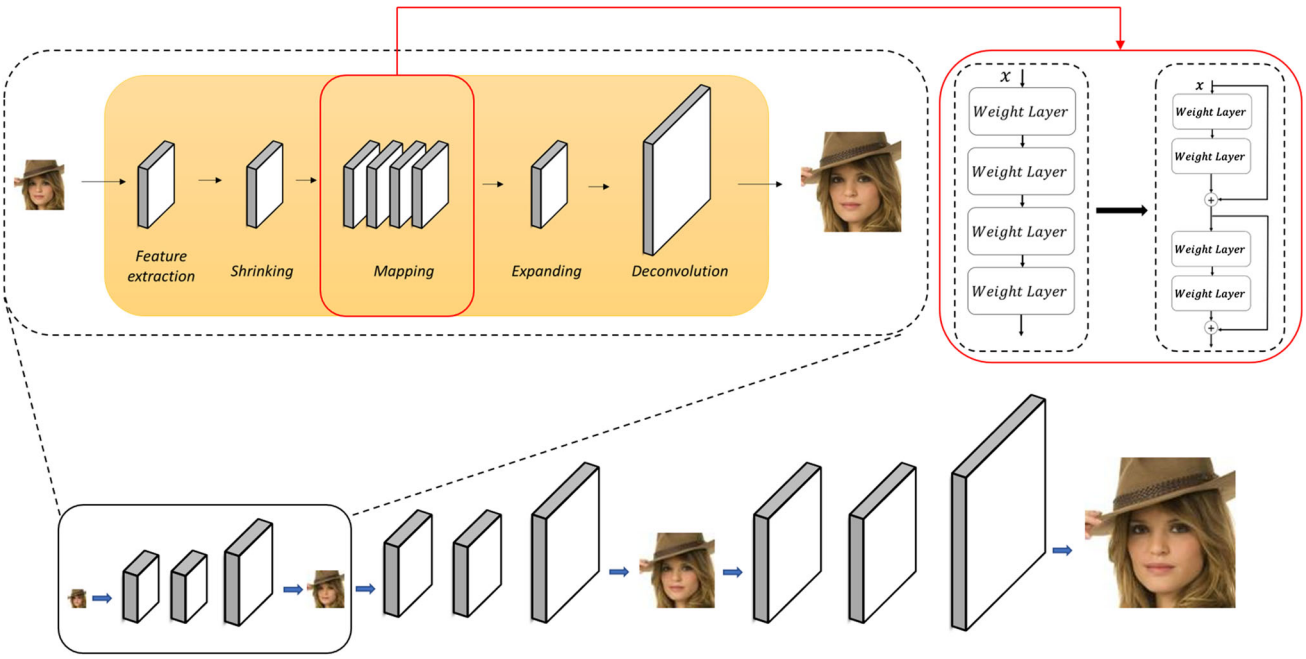

Fig. 2 The network structure of the base network is shown in the figure. Super-resolution with a high scale factor is achieved by stacking multiple levels. For example, to achieve a super-resolution task with a scale factor of 8, three levels need to be cascaded. By using the skip connection structure, not only the convergence of the network can be accelerated, but also the information transfer from LR to HR can be better maintained. The structure in the red block shows how the skip connection is equipped on the mapping layer

mation in the form of the shape and position of facial components, which can be obtained through parsing map prior information. To leverage this information, we designed an estimation network that predicts the parsing map prior information based on the reconstruction results of the base network. This network searches for the mapping relationship between the reconstruction results and the parsing map to better integrate these two sets of information. By combining the reconstruction results and the predicted parsing map results, further optimization can be performed to achieve even better super-resolution images.

The network structure of the estimation network is shown in Fig. 3. In the first part of the estimation network, feature extraction is the first of all. Considering that the input image of the estimation network has a higher resolution than the input image of the base network, in order to extract more feature information, the filter size is set to be 9 for feature extraction. This can cover almost all the feature information extracted by the filter of size 5 used in the base network, which can work with less information loss. The part of the feature extraction can be denoted as $Conv(9, 64, 3) - Relu$.

Mapping is an important part of the estimation network, which largely affects the prediction results of the parsing map. Therefore, a complex network structure is required to preserve the mapping relationship between input and output at this part. Since the feature extraction method used in the previous part has obtained a wide range of feature information, the accuracy should be considered in the mapping process. The structure, depth and width of the network all have an impact on the accuracy of the results, considering the computational of the network and the effect of the parsing map prior information on the super-resolution results. Based on the results of the experiment, we adopt the structure of stacking residual blocks to learn the mapping information, which consists of $Conv(3, 128, 128) - self - attentionlayer$.

To integrate local feature information with the global one, we utilize a U-Net structure for achieving multi-scale feature fusion. Firstly, the encoding operation is applied to process the input features, which can enhance the receptive field of the network, capture spatial location information of the parsing map more precisely and observe a broader range of features. Then, the decoding operation is performed, and the corresponding layers in the encoding and decoding processes are connected to achieve multi-scale feature fusion. Finally, based on the result of U-Net network, the parsing map prior information is reconstructed by a reconstruction layer, and this layer can be represented as $Conv(1, 11, 128) - Relu$.

Global Attention Module As shown in Fig. 4, the parsing map prior information divides the face image into 11 parts, and we aim at enhancing the recovery ability of facial components guided by the parsing map prior information. By observing the parsing maps, it can be seen that the pars-

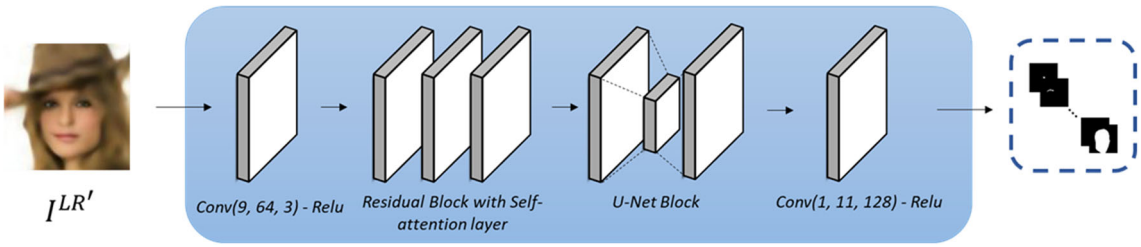

Fig. 3 Network structure of the estimation network.The network consists of feature extraction, mapping and multi-scale information fusion parts. The feature extraction module is a convolutional layer of $Conv(9, 64, 3) - Relu$, the mapping structure is a cascade of residual blocks and self-attention layers, the U-Net structure is used in the multi-scale information fusion, and the global attention module is inserted into it. Finally, the parsing map prior information is reconstructed by a layer of $Conv(1, 11, 128) - Relu$

Fig. 4 Parsing map prior information. It demonstrates the shape and position information of the facial components, which can be used to verify the accuracy of the reconstruction results, and can also optimize the super-resolution network. **a** Ground truth **b** from the top left, left eye, right eye, upper lip, lower lip, mouth, left ear, right ear, nose, left eyebrow, right eyebrow and skin, respectively

ing maps are a priori information that enhances the contrast of key features and other irrelevant information. Therefore, we believe that enabling the network to distinguish the input feature maps and enhance the expression of key features in the input data can effectively improve the network's ability to predict the parsing maps. In light of this, we propose a global attention module (GAM) used on the U-Net part of the estimation network to improve the network's ability to reconstruct the components of the parsing maps.

To capture the information across spatial and channels, average-pooling is commonly used in attention mechanism. And max-pooling is usually used to capture the distinctive features. We consider that the features extracted by max-pooling can conduct the network learning the features of parsing maps more efficiently. Thus, we design an attention mechanism to enhance the distinctive features. Given a tensor from a layer of the network as input,

$$B = Avg(input), \tag{6}$$

$$F_1 = input - B, \tag{7}$$

where $Avg()$ is the operation of whole map average pool. B, F_1 and input denote baseline, result of comparison and input feature maps, respectively.

Then we send the comparison result into the residual block and activate the output.

$$Output = sigmoid(res(F_1)), \tag{8}$$

where $res()$ denotes the mapping of residual block and $sigmoid()$ denotes the activation function of sigmoid. Output is the spatial attention map we need.

Given a tensor from a layer of the network, we design a channel global attention block for learning channel attention vector. In detail, we max-pool the whole feature map to obtain the most representative features in spatial to represent the channel features and average pool the channel features to find out the baseline between channels. Finally, we compare the channel features with the baseline.

$$Out = sigmoid(res(max(input) - avg(input))), \tag{9}$$

where input denotes the input feature maps, and $max()$ and $avg()$ are max-pool and average pool operation individually.

Figure 5 illustrates how the U-Net module works, and U-Net module [39] first downsamples the input, then upsamples and connects the output of low layers to the output of deep layers which can capture the features from different scales. We add our attention blocks which consist of spatial and channel attention mechanism after every layer in the U-Net module.

3.4 Reconstruction network structure

In the reconstruction network, we aim at optimizing the super-resolution result of the base network using the parsing map prior information predicted by the estimation network to obtain a final super-resolution image. The network architec-

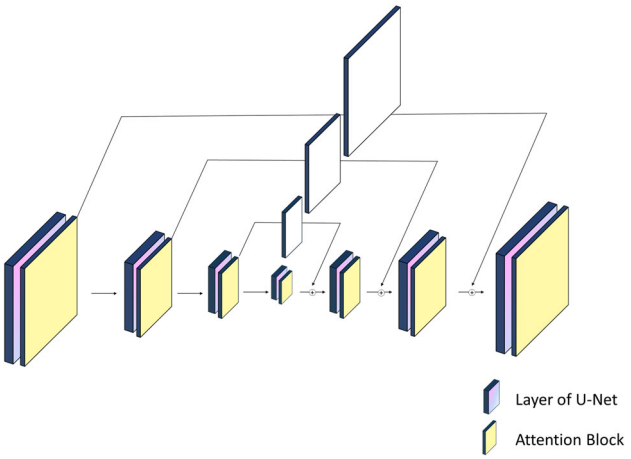

Fig. 5 U-Net structure armed with our global attention module, which can help reconstruct more accurate details from different scales. GAM is set after each layer of U-Net module

ture is shown in Fig 6, which consists of preprocessing layer, mapping layer and reconstruction layer.

In preprocessing layer, feature extraction needs to be performed on the input data. Different from the method of extracting input image features in the estimation network, we do not need to extract the correlation information between the target pixel and surrounding pixels. We tend to use parsing map prior information to refine the input coarse SR image. Therefore, we extract features from the parsing map output of the estimation network and the coarse super-resolution output of the base network using the convolutional layers of $Conv(1, 32, 11)$ and $Conv(1, 32, 3)$, respectively, and then concatenate the data. The process of the layer of preprocessing can be expressed as:

$$M = Conv(map^{est}) \otimes Conv(I^{LR'}), \tag{10}$$

where $Conv()$ denotes the convolution layer with 1*1 filters, $\otimes$ denotes concatenated features and M denotes the output of preprocessing layer.

The preprocessed data will be sent into the mapping layer for processing, and the high-frequency information in the $I^{LR'}$ will be optimized based on the results of the estimation network. Finally, the result of optimizing the high-frequency information is reconstructed by the reconstruction layer $Conv(3, 3, 64)$ to obtain the final SR image result.

$$I_{SR} = R(E_1(M)), \tag{11}$$

where E_1 and R denote the mapping generated by mapping layer and reconstruction layer, respectively. I^{SR} is the SR result of the network.

4 Experiments

In this section, we evaluate and compare the performance of bicubic, SISR methods, FSR method and our methods on the face image super-resolution dataset. We will introduce this section in the following turn, the preparation for training, the study environment and result evaluation methods, and the comparison of the results of different methods. Table 1 shows the comparison of different methods.

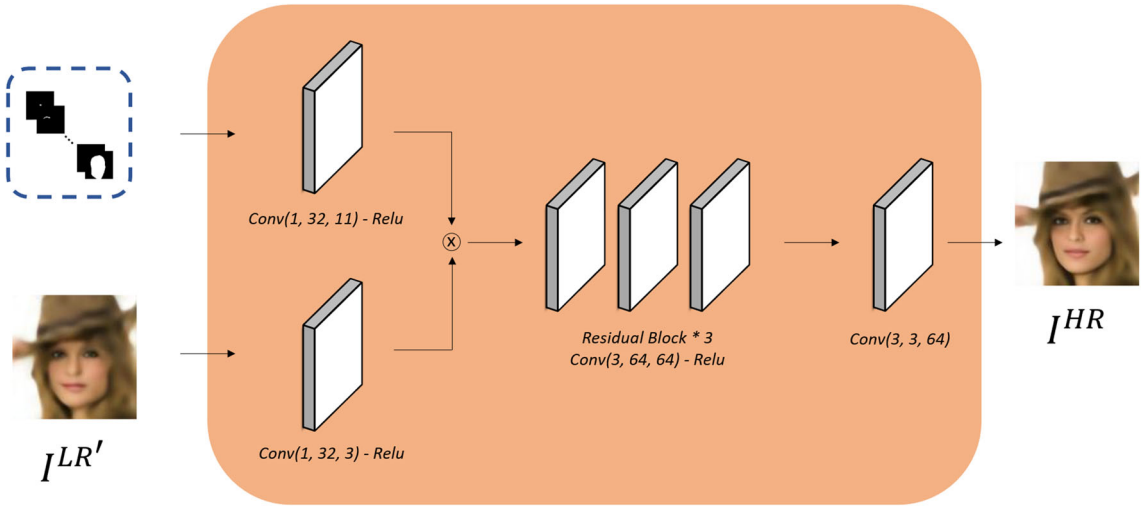

Fig. 6 Network architecture of the reconstruction network. The parsing map reconstructed by the estimation network and the coarse SR image reconstructed by the base network are processed by the preprocessing layer, and the output is sent into the mapping layer and finally reconstructed through the convolutional layer

Table 1 Comparison of different methods. Our method uses less prior information than other FSR methods

Methods	Prior
Bicubic	None
SRCNN	None
VDSR	None
FSRNet	Parsing map, landmark, heatmap
GAMFSR(ours)	Parsing map

4.1 Dataset and study environments

For the training work, we select 20000 pictures to train our model on CelebA-Mask-HD dataset. We select 100 and 50 pictures for testing in CelebA-Mask-HD and Helen datasets, respectively. For CelebA, we downsample all selected pictures to 128×128 as HR images and then resize them to 16×16 as LR images using the bicubic interpolation method. For Helen, we crop out the background, keeping the part of the face image.

To make full use of image information, we train our model in RGB color space. In YCbCr color space, the Y (luminance) channel contains the most sensitive information to the human eyes. Therefore, we use Y channel to evaluate the results of the methods.

For implementation, our experiments are based on the environment of Windows 10, Python 3.6 and Pytorch 1.8 [40] with NVIDIA RTX 3060. Due to the differences in environment, we re-experimented the previous methods on our platform as much as possible.

All comparison models are trained by 100 epochs, we use main square error as loss function, and our model is optimized by Adam [41] with the parameters of beta1 = 0.9 and beta2 = 0.99. Learning rate for training work is 2.5e−4, and for optimization, it is 1e-4.

The peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index are widely used as metrics for image evaluation. Therefore, these two metrics are introduced as metrics for the evaluation of performance.

4.2 Implementation details

Table 2 summarizes the primary architecture details of the proposed framework. For the training details, we adopt the method of training progressively and final optimization, which can accelerate the convergence of the network and find the optimal solution. In order to achieve super-resolution with a scale factor of 8, that is, a super-resolution of a 16×16 input image to get a result of 128×128, we need to set three layers in the base network. The first step is to train the first layer in the base network to achieve a resolution from 16×16 to 32×32. Since each layer in the base network has the same network structure, in the second step, the output results and the parameters of the first layer are loaded into the second layer and then optimize the parameters to achieve a resolution improvement from 32×32 to 64×64. The last step is to load the parameters of the first two steps into the first two layers of the base network, and a small learning rate is set for fine-tuning. Then, the estimation network, the reconstruction network and the last layer in the base network are trained as a whole. Each step is trained by 100 epochs.

4.3 Ablation study

Effects of Base Network We conduct an experiment to investigate the effect of the base network. First, we compared the improved FSRCNN with the original FSRCNN in the task of single image super-resolution with scale factor 2. Figure 7 illustrates the relationship between the training epochs and PSNR of FSRCNN and its improved version. Figure 7 shows that the proposed improvement method in this paper has a faster convergence speed, resulting in better training results with the same training time. Improved FSRCNN can arrive at PSNR=36.82dB; however, the original FSRCNN is 36.46dB.

Effects of Global Attention Mechanism For the ablation study, we conduct a set of experiments to investigate the effect of our proposed GAM attention mechanism. We remove the attention module we proposed in the network structure, and this model is called GAMFSR-NG. The PSNR and SSIM performances on the CelebA-Mask-HQ and Helen datasets are presented in the table. As given in Table 3, when our network loses the guidance of our attention module, the network performance degraded severely since its inference ability of

Table 2 Detail and parameters of the proposed framework

Base network	Estimation network	Reconstruction network
Refined FSRCNN x 3	Conv (9,64,3)	
	ResBlock (128,64)*3	Cat (Conv (1,32,11), Conv (1,32,3))
	ResBlockWithAtt (128)	ResBlock (64)*3
	U-Net with GAM depth 3	Conv (3,3,64)
	Conv (1,11,128)	

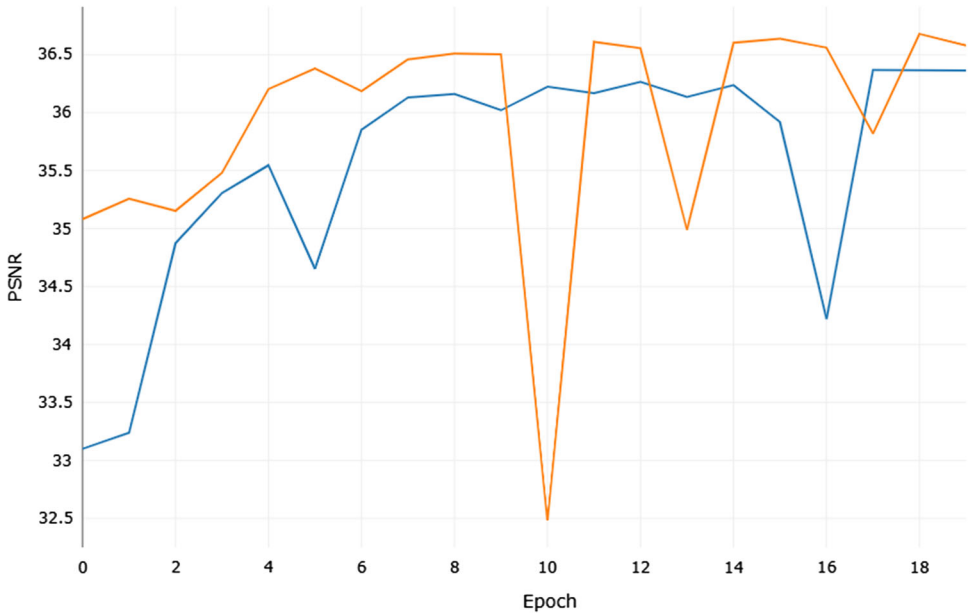

Fig. 7 Quantitative analysis of epoch. The orange line in the figure represents the training curve of the improved FSRCNN, while the blue line represents the original FSRCNN. The experiment is based on a 2x single image super-resolution task

Table 3 Results comparison of the methods we proposed

	PSNR	SSIM
GAMFSR-NA	25.94	0.7559
GAMFSR-RCAN [27]	26.12	0.7675
GAMFSR-CBAM [42]	26.17	0.7686
GAMFSR-CSAM [43]	26.01	0.7576
GAMFSR	**26.50**	**0.7793**

The best results are given in bold. NA, no global attention module

the parsing map prior information is weakened. And we also compared the effectiveness of our global attention module with other attention modules which achieve great achievements, such as RCAB, CBAM and CSAM. The results is also shown in Table 3. From the table, we can see our attention module GAM gets the best result.

Effects of the prior information We set another experiment to investigate weather prior information has impact on the performance of the network. On the one hand, we set a group with no prior information is denoted as GAMFSR-NP, and on the other hand, we set another group uses landmarks instead of parsing map prior information is called GAMFSR-LM. Quantitative comparison results are shown in Table 4. From the table, we can learn that with the guidance of the parsing map prior information, the performance of the network can be improved significantly. However, GAMFSR-LM does not work better than GAMFSR-NP. The reason is that the prior estimation network and global attention module is designed for parsing map and cannot take advantage of landmark well.

Table 4 Quantitative results of different models

	GAMFSR-NP	GAMFSR-LM	GAMFSR
PSNR	26.21	26.14	**26.50**
SSIM	0.7668	0.7674	**0.7793**

The best results are given in bold. NP, no prior information, LM: landmarks

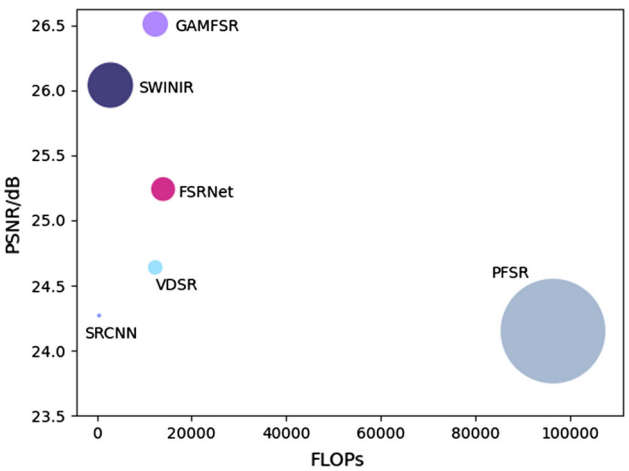

Fig. 8 An illustration of the average PSNR values versus the number of parameters and the computation cost on 128 × 128 images for each model with a scale factor of ×8 SR

Effects of the proposed network We compare the proposed GAMFSR method with other SR methods, all of which are state of the art at the time. It can be observed that our proposed method achieves the best PSNR and SSIM performance on CelebA-Mask-HQ dataset, and it is noteworthy that our method outperforms FSRNet on SSIM by a large

Table 5 Results comparison of different methods work on CelebA-Mask-HQ

	Bicubic	SRCNN	FSRNet	PFSR [44]	SwinIR [1]	GAMFSR
PSNR	23.20	24.27	25.24	24.15	26.04	**26.50**
SSIM	0.6749	0.7059	0.7324	0.6958	0.7649	**0.7793**

The best results are are given in bold

Fig. 9 Qualitative results of our GAMFSR with other methods, tested with ×8 SR, please zoom in for differences

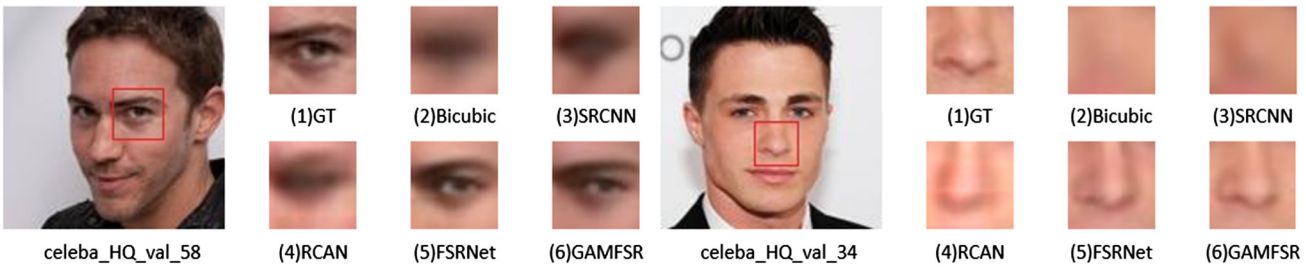

Fig. 10 Qualitative comparison of our GAMFSR with other methods, tested with ×8 SR, please zoom in for differences

margin. Therefore, our method has a strong inference ability. While preserving the pixel-wise accuracy of the SR image, the network also makes full use of the prior information provided by the parsing map to make the reconstructed image have higher structural similarity. We visualize the SR results generated by GAMFSR and other SR methods in the figure. It can be observed that the method we proposed has more detailed information than other SR methods. This is due to the precision of our base network for coarse SR results inference, the accurate prediction of the prior information of the parsing map by the GAM attention mechanism and the reconstruction ability of the reconstruction network.

Figure 8 compares the parameter and computational complexity with other methods. Moreover, compared to other in-prior SR methods, inferring less prior information in the process can usually make our network more robust and computationally faster.

To illustrate our method's superiority in terms of quantitative metrics, we compare our best method (GAMFSR) with several methods, including FSRNet, end-to-end generative super-resolution method SRCNN, VDSR and SwinIR [1]. For the fairness of comparison, we train all models on our platform.

Table 5 shows the quantitative results of comparison methods test on CelebA-HQ dataset. Because the prior information is not considered, the results of traditional methods such as bicubic, SRCNN and VDSR do not work well. In contrast, the face prior knowledge guided methods and trans-

Table 6 Results comparison of different methods work on Helen

	Bicubic	SRCNN	VDSR	FSRNet	PFSR	SwinIR	GAMFSR
PSNR	23.75	24.26	24.83	25.30	24.75	24.94	**25.44**
SSIM	0.6423	0.6634	0.6878	**0.7422**	0.7118	0.7130	0.7375

The best results are are given in bold

former methods such as SwinIR has better performance. For CelebA-Mask-HQ dataset, our best method (GAMFSR) gets 0.46dB higher than the second place method SwinIR on the metric of PSNR, and for SSIM, it is 0.0144 higher than it. In general, considering that our method uses far less prior information than FSRNet and has better performance. Therefore, it is obvious that GAMFSR works better than other methods on both quantitative and qualitative results. Qualitative comparisons of our method is illustrated in Figs. 9 and 10.

Table 6 shows the quantitative results of comparison methods test on Helen dataset. We carefully analyzed why our method does not perform well on the Helen dataset. Our training datasets are all from CelebA-Mask-HQ; however, the light intensity, color and structure in the Helen dataset are quite different from that in CelebA-Mask-HQ; therefore, we assume that our training results are on the Helen dataset due to the lack of training data with the above characteristics cause the poor performance.

5 Conclusion

In this paper, we propose a novel network structure (GAMFSR) for face image super-resolution. We reconstruct coarse SR results by a designed progressive network and based on the U-Net module armed with our global attention mechanism (GAM) which can make the network using less prior information achieve better performance by enhancing the expressive power of key components in parsing maps. Furthermore, we also study the effects of our global attention mechanism (GAM) on different prior information and the effects of different attention mechanisms with parsing map prior information. Experiments show that our method achieves the effect of improving the results of PSNR and SSIM. Quantitative and qualitative results of face super-resolution demonstrate the effectiveness of the method we proposed.

Data availability The raw/processed data required to reproduce these findings cannot be shared at this time as the data also form part of an ongoing study.

Declarations

Conflict of interest statement We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

1. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833–1844 (2021)
2. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Patt. Anal. Mach. Intell. **38**(2), 295–307 (2015)
3. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654 (2016)
4. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European Conference on Computer Vision. Springer: London (2014) pp 184–199
5. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Learning face hallucination in the wild. In: Twenty-ninth AAAI Conference on Artificial Intelligence (2015)
6. Huang, W., Chen, Y., Mei, L., You, H.: Super-resolution reconstruction of face image based on convolution network. In: International Conference on Intelligent and Interactive Systems and Applications, pp. 288–294. Springer (2017)
7. Rajput, S.S., Arya, K., Singh, V.: Robust face super-resolution via iterative sparsity and locality-constrained representation. Inform. Sci. **463**, 227–244 (2018)
8. Song, Y., Zhang, J., He, S., Bao, L., Yang, Q.: Learning to hallucinate face images via component generation and enhancement. arXiv preprint arXiv:1708.00223 (2017)
9. Jiang, J., Yu, Y., Hu, J., Tang, S., Ma, J.: Deep cnn denoiser and multi-layer neighbor component embedding for face hallucination. arXiv preprint arXiv:1806.10726 (2018)
10. Chen, C., Li, X., Yang, L., Lin, X., Zhang, L., Wong, K.-Y.K.: Progressive semantic-aware style transformation for blind face restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11896–11905 (2021)
11. Lu, Y., Tai, Y.-W., Tang, C.-K.: Attribute-guided face generation using conditional cyclegan. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 282–297 (2018)
12. Yu, X., Fernando, B., Hartley, R., Porikli, F.: Super-resolving very low-resolution face images with supplementary attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 908–917 (2018)
13. Li, M., Sun, Y., Zhang, Z., Xie, H., Yu, J.: Deep learning face hallucination via attributes transfer and enhancement. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 604–609. IEEE (2019)
14. Bayramli, B., Ali, U., Qi, T., Lu, H.: Fh-gan: Face hallucination and recognition using generative adversarial network. In: International Conference on Neural Information Processing, pp. 3–15. Springer (2019)
15. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet domain generative adversarial network for multi-scale face hallucination. Int. J. Comput. Vis. **127**(6), 763–784 (2019)
16. Lai, S.-C., He, C.-H., Lam, K.-M.: Low-resolution face recognition based on identity-preserved face hallucination. In: 2019 IEEE

International Conference on Image Processing (ICIP), pp. 1173–1177. IEEE (2019)
17. Li, X., Li, W., Ren, D., Zhang, H., Wang, M., Zuo, W.: Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2706–2715 (2020)
18. Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., Yang, R.: Learning warped guidance for blind face restoration. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 272–289 (2018)
19. Schaefer, S., Mcphail, T., Warren, J.: Image deformation using moving least squares. ACM Trans. Graph. **25**(3), 533–540 (2006)
20. Baker, S., Kanade, T.: Hallucinating faces. In: Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), pp. 83–88. IEEE (2000)
21. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738 (2015)
22. Lee, C.-H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5549–5558 (2020)
23. Liu, C., Shum, H.-Y., Zhang, C.-S.: A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1, p. IEEE (2001)
24. Wang, X., Tang, X.: Hallucinating face by eigentransformation. IEEE Trans. Syst. Man Cybernet. Part C (Appl. Rev.) **35**(3), 425–434 (2005)
25. Jian, M., Lam, K.-M.: Face super-resolution based on singular value decomposition. In: Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1–5. IEEE (2012)
26. Keys, R.: Cubic convolution interpolation for digital image processing. IEEE Trans. Acoust. Speech Signal Process. **29**(6), 1153–1160 (1981)
27. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks (2018)
28. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 105–114 (2017). https://doi.org/10.1109/CVPR.2017.19
29. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0 (2018)
30. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
31. Song, Y., Zhang, J., He, S., Bao, L., Yang, Q.: Learning to hallucinate face images via component generation and enhancement. arXiv preprint arXiv:1708.00223 (2017)
32. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2492–2501 (2018)
33. Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W.: Feedback network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3867–3876 (2019)
34. Ma, C., Jiang, Z., Rao, Y., Lu, J., Zhou, J.: Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5569–5578 (2020)
35. Yu, X., Fernando, B., Ghanem, B., Porikli, F., Hartley, R.: Face super-resolution guided by facial component heatmaps. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 217–233 (2018)
36. Li, M., Sun, Y., Zhang, Z., Yu, J.: A coarse-to-fine face hallucination method by exploiting facial prior knowledge. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 61–65 IEEE (2018)
37. Zhang, Y., Wu, Y., Chen, L.: Msfsr: A multi-stage face super-resolution with accurate facial representation via enhanced facial boundaries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 504–505 (2020)
38. Wang, H., Hu, Q., Wu, C., Chi, J., Wu, H.: Dclnet: dual closed-loop networks for face super-resolution. Knowl. Based Syst. **222**(33), 106987 (2021)
39. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision, pp. 483–499. Springer (2016)
40. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
41. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
42. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
43. Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., Zhang, K., Cao, X., Shen, H.: Single image super-resolution via a holistic attention network. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16, pp. 191–207. Springer (2020)
44. Kim, D., Kim, M., Kwon, G., Kim, D.-S.: Progressive face super-resolution via attention to facial landmark. arXiv preprint arXiv:1908.08239 (2019)

Jinlu Zhang was born in Heilongjiang, China, in 1998. He received the master's degree in control science and engineering from Heilongjiang University in 2023. His main research interests are intelligent detection.

Mingliang Liu was born in Heilongjiang province, China, in 1980. He received the Ph.D. degree in forestry engineering automation from Northeast Forestry University in 2017. He is now a full professor and a Master's tutor in the School of Electrical Engineering from Heilongjiang University. His research interests include intelligent detection, fault diagnosis, signal processing, and pattern recognition.

Xiaohang Wang was born in Xinjiang, China, in 1996. He received the master's degree in control science and engineering from Heilongjiang University in 2022. His main research interests are intelligent detection.