IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 10, 2020, accepted September 19, 2020, date of publication September 22, 2020,
date of current version October 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3025972

Guided Cascaded Super-Resolution Network for Face Image

LIN CAO, JIAPE LIU, KANGNING DU, YANAN GUO, AND TAO WANG

Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, Beijing Information Science and Technology University, Beijing 100101, China
School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing 100101, China

Corresponding author: Kangning Du (kangningdu@outlook.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671069, in part by the Qin Xin Talents Cultivation Program of Beijing Information Science and Technology University under Grant QXTCP A201902, and in part by the General Foundation of Beijing Municipal Commission of Education under Grant KM202011232021.

ABSTRACT The image super-resolution algorithm can overcome the imaging system's hardware limitation and obtain higher resolution and clearer images. Existing super-resolution methods based on convolutional neural networks(CNN) can learn the mapping relationship between high-resolution(HR) and low-resolution(LR) images. However, when the reconstruction target is a face image, the reconstruction results often have problems that the face area is too smooth and lacks details. We propose a guided cascaded face super-resolution network, called guided cascaded super-resolution network (GCFSRnet). GCFSRnet takes the LR image and a high-quality guided image as inputs, and it consists of a pose deformation module and a super-resolution network. Firstly, the pose deformation module converts the guide image's posture into the same as the low-resolution face image based on 3D fitting and 3D morphable model (3DMM). Then, the LR image and the deformed guide image are used as input of the super-resolution network. The super-resolution networks are formed by a cascade of two layers of networks, which extract different features. During the reconstruction process, the guide image can provide real facial details and help generate subtle facial textures. The cascade structure of a super-resolution network can gradually extract features and restore different levels of image details. The experimental results on the CASIA Web Face and CelebA datasets show that the proposed method can generate facial images with clear outlines and rich details, which are superior to other state-of-the-art methods such as SRResNet, SRGAN, VDSR, DBPN, etc.

INDEX TERMS Face super-resolution, 3D morphable model, guide image, pose deformation, cascade structure.

I. INTRODUCTION

Super-resolution reconstruction is one of the classic computer vision problems, which aims to recover high-frequency details from low-resolution videos or images. The human face is a common object in digital media. Using a super-resolution algorithm to improve the human face's clarity has high application value in areas such as face recognition, smart security, etc.

In general, super-resolution methods can be divided into three categories: interpolation-based methods [1]–[3], restoration-based methods [4], [5], and learning-based methods [6]–[11]. The interpolation-based methods assume that the image structure is piece-wise smooth and uses a basis function or kernel function to perform image interpolate to obtain an enlarged image. The interpolation-based method is fast and can maintain the image's edge, but the reconstructed model is usually unstable, making it difficult to reconstruct the details of the image accurately. The restoration-based method can gradually restore the HR images by solving the inverse process of image degradation, but it requires much calculation, and the degradation models are not unique. The learning-based method uses machine learning theory to learn the mapping function between HR and LR image pairs from the training database. Since the learning-based method can autonomously learn the mapping function, it has become an important super-resolution reconstruction method [12].

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

With the rapid development of computing power, the super-resolution method based on Convolutional Neural Networks (CNN) has gained the most success in learning-based methods. Dong *et al.* first proposed Super-Resolution CNN(SRCNN) [7] in 2016. It stacks three convolutional layers to learn the non-linear mapping relationship between HR and LR image pairs to achieve image super-resolution reconstruction. However, due to the small number of network layers, only shallow image information can be extracted, and there is a blur phenomenon in the reconstructed image. Based on SRCNN, Wang *et al.* [8] extended the depth of the network to 20 layers to extract deep features and used the residuals of HR and LR image pairs as the learning target to reduce calculation.

For face images, Yu *et al.* proposed UR-DGN [9]. It is the first time that a Generative Adversarial Networks(GANs) is introduced into the face super-resolution method. UR-DGN uses the approximately aligned frontal face image to train the model and resolve the artifacts in the reconstruction results caused by the input image's very low resolution. However, since the pictures in the dataset only contain front faces, the reconstruction results have a large error for side face images and different pose face images. Zhu *et al.* [10] reconstructed face images with clear contours by alternately optimizing the face SR task and the face dense field estimation task, but the non-end-to-end network structure makes the learning process very complicated. Song *et al.* [11] divide the LR image into five regions according to the facial features, then train networks for different regions, respectively, and stitches the reconstructed images together. This method can generate an image with rich facial features, but due to the stitching edges between adjacent regions, there will be noticeable discontinuous regions in the reconstructed image.

Unlike Single Image Super-Resolution(SISR), the video can be regarded as a collection of frame images, and there are only slight differences between adjacent frames. Therefore, Video Super-Resolution(VSR) can effectively use the adjacent frames' information to assist the reconstruction of the current frame image. In the VSR method proposed by Sajjadi in 2018 [13], the reconstructed frame image at the previous moment is added to the reconstruction process of the current frame image. It can obtain better auxiliary effects and improve the frame image clarity. To eliminate the deviation between adjacent frames, motion estimation and compensation is used to align two adjacent frame images.

In some cases, it is necessary to improve facial clarity for specific characters. For example, restore old photos or compressed images with low quality. Inspired by VSR [13], considering that the face super-resolution pays more attention to facial areas' performance, it should be helpful to use a high-definition face image with the same identity with LR image, we called guide image, to assist the reconstruction. Therefore, we introduce a Guided Cascaded Face Super-Resolution network called GCFSRnet. The proposed GCFSRnet consists of a pose deformation module and a super-resolution network. The pose deformation module firstly performs 3D fitting on the input images using 3D Morphable Model (3DMM) and then deforms the guide image to the same pose as the LR face image.

Since gradually increasing the image's clarity is conducive to obtaining the essential nature of image from coarse-to-fine change [15], the SR networks are formed by a cascade of two layers of networks, which extract different features. And then, each layer of the SR network takes the distorted guide image together with the low-resolution image as a common input to generate subtle facial textures. Finally, an adversarial network is used to generate more realistic textures for reconstructed images. In summary, this paper's main innovations are that a high-resolution face image as the guide image is added into the process of face super-resolution reconstruction, and a complete scheme is designed to make full use of the guide image so that the SR network can generate more details.

The rest of this article is arranged as follows: section II introduces the related works in this article; section III introduces our proposed method; section IV introduces the dataset, experimental setting, and experimental results; section V summarizes the full paper.

II. RELATED WORKS

A. FACIAL LANDMARK DETECTION

Facial landmark detection is a fundamental task in face image processing and has broad application scenarios. In recent years, landmark detection technology has gradually matured, and many open-source models have been used. However, when the image's resolution is lower than 30 pixels, the detection accuracy of medium and large poses face at standard facial resolution (192×192) has decreased by 15% and 30%, respectively [16]. Therefore, the existing research on facial landmark detection mainly focuses on improving ultra-low resolution facial images' detection accuracy.

In the face alignment method (FAN) [17] proposed by Yang *et al.* in 2017, a stacked HourGlass (SHG) structure is used to estimate face landmarks. The SHG structure was first proposed by Newell *et al.* [18] in 2016 to estimate human pose. This structure can make full use of multi-scale feature mapping to capture information at different scales. Benefit from the excellent performance of SHG structure, various algorithms based on SHG continuously proposed, such as the face alignment algorithm FAN [17] proposed in 2017 and the LAB [19] algorithm was proposed in 2018.

B. FACE CORRECTION BASED ON 3DMM

As a technical means to verify human identity, face recognition has completed complex and heavy tasks in many scenes because of its non-contact and convenience. Most face recognition methods require that the person being verified provides frontal photos. However, human faces that can be collected in natural environments usually have large posture and expression changes. It is necessary to use a face correction algorithm to correct non-frontal face images to obtain an ideal frontal image. Existing research on face pose correction can be

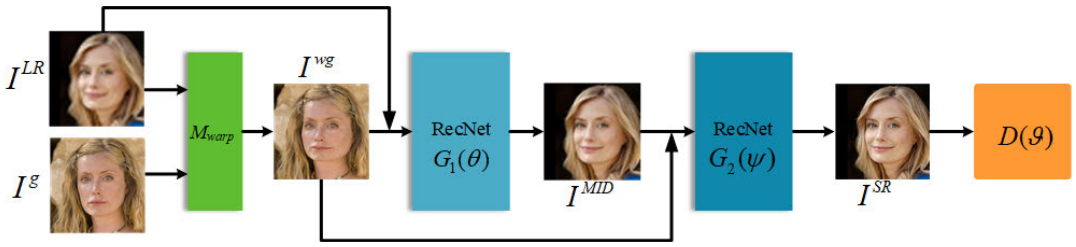


FIGURE 1. Network structure of the proposed GCFSRnet.

divided into 2D-based and 3D-based methods. Among them, the 3D-based methods can use 3D structural information to simulate the conversion between postures. 3D structural information can be obtained through a 3D morphable model (3DMM). 3DMM is a face 3D deformable model proposed by Blanz and Vetter *et al.* [14]. A common way to correct face posture based on 3DMM is to use a 3D face model to match a 2D face image, and then rotate the 3D model to obtain the frontal image.

In the face correction algorithms proposed in recent years, Hassner *et al.* [20] uses a unique and fixed 3D face model to approximate the fitted shape of input faces. This method is very effective for the human face's frontal area but will cause severe texture loss and artifacts on the contour and near-contour surfaces. Zhu *et al.* [21] normalized the 3D model by projecting the landmark of the 2D face onto the 3D face model and then filled the invisible areas by Poisson editing to obtain a smooth and natural normalized result. This method can retain more original information and has better performance than the 2D-based methods.

C. GAN

The proposal of GAN [22] solves the problem that the generated image is too smooth when using CNN for reconstruction. The GAN trains the generator and the discriminator simultaneously based on the game idea until the two reach the Nash equilibrium. This training process can cause the generator to produce images with realistic textures. Since Ledig *et al.* first proposed SRGAN, which combined GAN with SR, in 2016, GAN has been applied to most CNN-based SR tasks. However, the original GAN is prone to gradient disappearance during training; it is difficult to achieve the Nash equilibrium state. Therefore, the GAN based image SR method is also continuously improved with the improvement of GAN, such as DCGAN [23] proposed by Dosovitskiy *et al.*, WGAN [24] proposed by Arjovsky *et al.* and the improved WGAN-GP [25] based on WGAN.

III. METHOD

A. OVERVIEW OF GCFSRnet

In this work, we propose a novel guided cascaded face super-resolution network called GCFSRnet. The GCFSRnet consists of two parts: a *pose deformation module* noted as WarpNet (M_{warp}) and a super-resolution network noted as RecNet (G_1, G_2). The architecture of GCFSRnet is shown in Fig. 1. Note that the guide image should have the same identity and same size as the ground truth(GT) image, and the LR image is degraded from the GT image during training. Denote I^{LR} as the LR input image, I^{g} as the guide image, I^{SR} and I^{HR} as the recovered image and the GT image. As shown in Fig. 1, we first construct the pose deformation module M_{warp} on I^{LR} and I^{g} to get the warped guide image I^{wg},

$$I^{wg} = M_{warp}(I^{LR}, I^{g}) \tag{1}$$

The warped guide image I^{wg} supposed to have the same pose and face shape as I^{LR}. Then I^{wg} and I^{LR} are concatenated as the input of the SR network.

The SR network uses a two-layer cascade framework. The first layer is a *prior recover network* (G_1) and the second layer is a *structural constrain network* (G_2). The input image goes through the prior recover network G_1 to get the preliminary reconstructed image I^{MID},

$$I^{MID} = G_1(I^{LR}, I^{wg}) \tag{2}$$

Then, the structural constrain network(G_2) is utilized to recover the SR image by taking I^{MID} and I^{wg} as input,

$$I^{SR} = G_2(I^{MID}, I^{wg}) = G(I^{LR}, I^{wg}) \tag{3}$$

During training, the reconstructed image I^{SR} is used as the input of the discriminator together with the GT image I^{HR}. The discriminator randomly selects an image to determine whether it is a real high-quality image.

B. THE POSE DEFORMATION MODULE

In order to ensure that the SR network to obtain enough information from the input, we select another image of the same identity as the LR image to be the guide image, which can provide real high-frequency information for the reconstruction. To provide more additional high-definition information, the face area in the selected guide image is supposed to be clear, frontal, and open.

In the latest proposed face posture correction method [21], a non-frontal 2D face image is projected onto the 3DMM face model. Then the non-frontal face model is standardized by rotating the 3D model of the current image. Inspired by [21], the pose deformation module separately performs 3DMM projection on the guide image and the LR image and solve the rotation matrix between the projected 3DMM face model

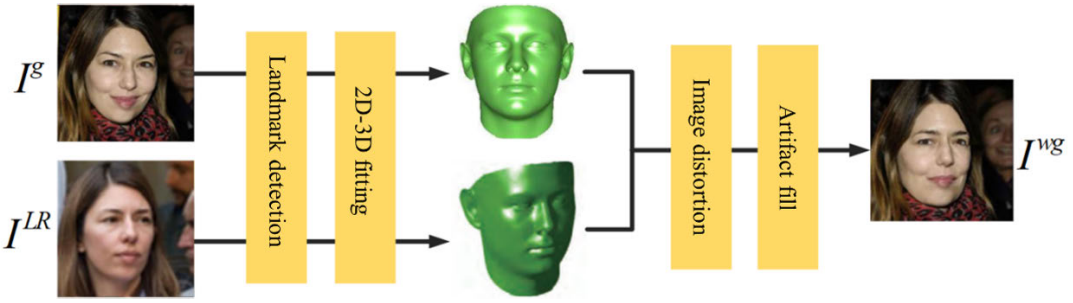


FIGURE 2. Flow chart of the pose deformation module.

and the standard 3DMM face model. To maintain posture consistency, the 3DMM model of the deformed guide image should have the same rotation matrix as the 3DMM model of the LR image. Therefore, through the inverse rotation matrix and parameter transformation, the guide image can be transformed into the same pose as the LR input. The overview of the pose deformation module is shown in Fig. 2.

3DMM is used as an average model to describe the shape of a human face. As in (4), a face image of any shape can be approximated as S by adjusting the correlation coefficient:

$$S = \bar{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp} \tag{4}$$

where $\bar{S}$ is the average shape of the three-dimensional face model, A_{id} and A_{exp} are the identity base and expression the base of the face model, α_{id} and α_{exp} are shape parameters and expression parameters respectively. To fit the 3DMM model, the most common method is to use a weak perspective projection [26], any 2D face image can be projected onto the 3D face model as:

$$s_{2d} = fPR(\alpha, \beta, \gamma)(S + t_{3d}) \tag{5}$$

In (5), s_{2d} is the image plane position of the 2D landmark which can use landmark detection algorithm to detect, f is the scale factor, P is the orthographic projection matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, $R(\alpha, \beta, \gamma)$ is a 3×3 rotation matrix controlled by three rotation angles, and t_{3d} is the translation parameter. When there is a deviation between the 2D face pose and the front face, the 2D and 3D landmarks do not exactly match. In this case, the matching process of the 3DMM needs to be corrected by landmark marching [21]:

$$s_{2d\_land} = fPR[\bar{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp} + t_{3d}]_{land} \tag{6}$$

In (6), the subscript *land* indicates the landmarks of the corrected 3D face. With fixed parameters, the 3DMM fitting parameters $\alpha_{id}, \alpha_{exp}, f, R, t_{3d}$ in (6) can be solved in an iterative manner. Then, the landmarks of the face and surrounding area are used to obtain the 3D grid bins. Finally, multiply the non-frontal 3D face model with the inverse rotation matrix R^{-1} to obtain the frontal image:

$$S_{img\_rn} = R^{-1}S_{img} \tag{7}$$

where S_{img} is the three-dimensional grid map including a non-frontal three-dimensional face model and landmarks, R is a rotation matrix estimated during the fitting process, and S_{img_rn} is a standard normalized three-dimensional grid map. Since (7) is only for the standard frontalization process of non-frontal face images, it cannot be directly applied to the discussed scenario. The deformation between the guide image I^g and the LR image I^{LR} can be obtained by solving (6) to get the respective fitted 3D face model, and the two can be converted by (8):

$$\begin{aligned} S_{img\_rn\_lr} &= R_{lr}^{-1}S_{img\_lr} \\ S_{img\_rn\_g} &= R_g^{-1}S_{img\_g} = R_{wg}^{-1}S_{img\_wg} \end{aligned} \tag{8}$$

where $S_{img\_rn\_lr}$ and $S_{img\_rn\_g}$ are standard normalized 3D grid maps with I^{LR} and I^g as the identity base respectively, S_{img_lr}, S_{img_g}, S_{img_wg} and R_{lr}, R_g, R_{wg} are LR images, guide images and the 3D grid map and rotation matrix of the deformed guide image to be solved. Because the distorted image should have the same rotation matrix as I^{LR}, which means $R_{wg} = R_{lr}$. According to the above conditions, S_{img_wg} can be expressed as (9):

$$S_{img\_wg} = R_{lr}R_g^{-1}S_{img\_g} \tag{9}$$

After the posture and expression are normalized, it is necessary to adjust further the landmarks of the boundary of the distorted image to eliminate the change in the position of the landmarks and preserve the original structure of the image. Finally, through the light adaptation [21] and edge filling [6] the image artifacts caused by the large yaw angle compensated and final distorted image I^{wg} is obtained.

C. THE SUPER-RESOLUTION NETWORK

In order to enhance the reconstruction accuracy of the SR network in the face region, GCFSRnet brings the structural information of the face image into the SR process. However, when the input image's resolution is too low, it is difficult to extract the accurate structural information of the face directly. Therefore, the SR network is designed as a two-layer cascade structure, consisting of a prior recover network G_1 and a structural constrain network G_2, as shown in Fig. 1.

The prior recover network restores high-frequency details based on the inherent before the LR image, thereby improving

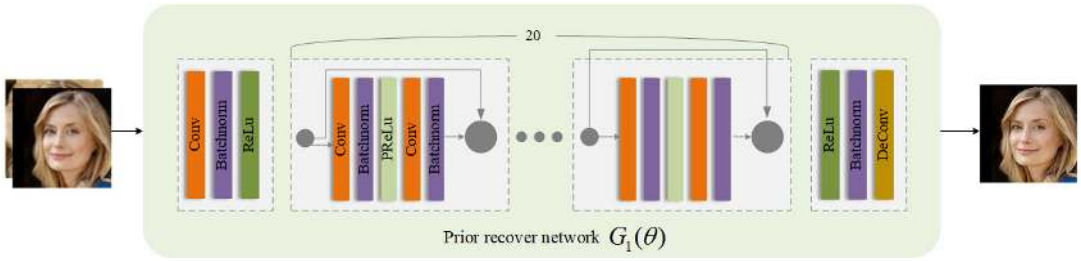


FIGURE 3. The diagram of prior recover network.

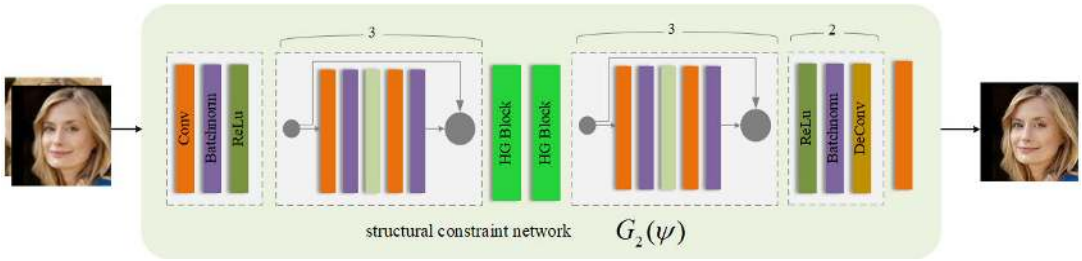


FIGURE 4. The diagram of structural constraint network.

face structure information extraction accuracy. Then, the structural constraints network enhances the facial region's reconstruction accuracy by additionally estimates the structural information of the face.

1) THE PRIOR RECOVER NETWORK

The detailed structure of the prior recover network is shown in Fig. 3. In order to extract deep features from LR input, recover face image with rich high-frequency information, and avoid gradients exploding/vanishing phenomenon, the residual block [27] is utilized as the basic architecture for feature extraction in the prior recover network. The residual block is composed of two convolutional layers with a kernel size of 3×3. Each layer is connected to the batch normalization layer and uses PReLU as an activation function. The LR input first goes through a convolutional layer with a kernel size of 3×3. To reduce calculation, the stride is set to 2, and the feature map size is half of the input size. Then 20 residual blocks are used to extract the feature, and the deconvolution layer is utilized to up-sample the feature map to the initial size of the input. Finally, a convolution layer is used to reconstruct the HR image.

2) THE STRUCTURAL CONSTRAINT NETWORK

The structural constraint network is shown in Fig. 4. The HourGlass(HG) structure [18] is utilized to estimate facial landmark heatmaps. Fig. 5 is the basic structure of a four-layer HG block. Using skip connection between symmetric layers, the HG block can effectively integrate multi-scale features and retain spatial information at different scales. Stacked hourglass(SHG) can capture the spatial information of each landmark of the face. Besides, the stacked hourglass

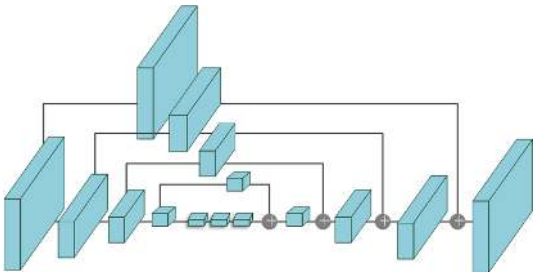


FIGURE 5. The structure of a Hourglass block.

can significantly improve the estimation accuracy by multiplexing the previous level's information.

In the structural constraint network, the upper layer's reconstruction results are extracted by the shallow features, and then a two-layer SHG block is used to estimate the facial landmark. After decoding the features through three residual blocks, two deconvolution layer recover the feature map to the original size. Finally, a single convolutional layer recovers the feature map to the high-resolution image. Introducing the HG block into the structural constraint network can not only strengthen the facial structure consistency of the reconstructed image and the GT image but also extract more high-frequency information and improve the visual effect of the reconstructed image.

D. ADVERSARIAL NETWORK ARCHITECTURE

In order to generate more realistic image details, the GCFSRnet is regarded as a GAN, which uses the discriminator network to promote the generator to generate a more realistic result. The discriminator architecture is as shown in Fig. 6. It contains 8 convolutional layers. Compared with

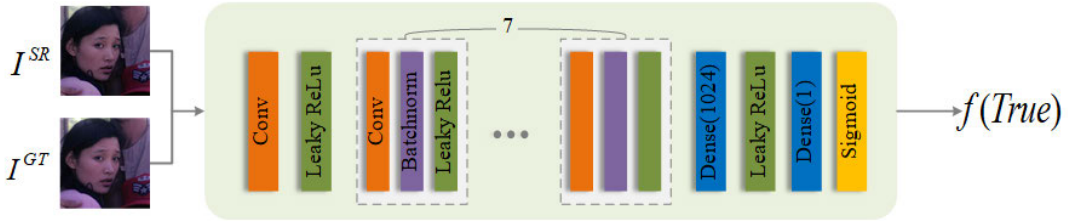


FIGURE 6. The structure of the adversarial network.

the VGG network [29], the convolution kernel has increased from 64 to 512. We follow the architectural guidelines summarized by Radford *et al.* [23] and use LeakyReLU activation ($\alpha = 0.2$) and avoid max-pooling setting in the network. After 8 convolutional layers, the resulting feature maps are followed by two dense layers. Then a final sigmoid activation function is used to obtain a probability for classification.

E. LOSS FUNCTIONS

1) RECONSTRUCTION LOSS

Define the training set A and B as $\{I^{LR}, I^{HR}\}_{i=1}^{N}$, N is the number of training sample pairs. Then the reconstruction loss l_{pixel} is used to minimize the pixel distance between the reconstructed image and the real HR image, it is defined as:

$$l_{pixel} = \frac{1}{2N}\left\{\left\|I_i^{HR} - I_i^{MID}\right\|^2 + \left\|I_i^{HR} - I_i^{SR}\right\|^2\right\} \quad (10)$$

where I_i^{MID}, I_i^{SR} and I_i^{HR} are the output of G_1, the output of G_2 and the corresponding GT image of the i-th image respectively.

2) PERCEPTUAL LOSS

The reconstruction loss alone can constrain the two images' consistency at the pixel level and improve the PSNR value, but the reconstructed image should also maintain the right consistency with the real image in high-level features. SRGAN and the perceptual loss proposed by Johnson *et al.* make the reconstructed image closer to the real image in detail by constraining the two pictures' similarity in the feature space. SRGAN uses the advanced features after the fourth convolutional layer before the fifth pooling layer of VGG-19 [29] structure and formulated the perceptual loss as (11):

$$l_{feature/i,j} = \frac{1}{W_{i,j}H_{i,j}}\sum_{x=1}^{W_{i,j}}\sum_{y=1}^{H_{i,j}}(\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G(I^{LR}))_{x,y})^2 \quad (11)$$

where G is a generator of the SR network, $\phi_{i,j}$ represents the feature map of the i-th pooling layer (after the activation layer) before the j-th convolution layer in the VGG-19 network, $W_{i,j}$ and $H_{i,j}$ are the dimension of feature map in width and height respectively. In our work, i and j are the same as SRGAN, respectively 5 and 4.

3) LANDMARK LOSS

To enhance the recovered image's spatial consistency and the GT image, the SHG in a structural constraint network is utilized to generate the landmark heatmap. The landmark loss can be defined as (12):

$$l_{landmark} = \frac{1}{N}\sum_{n=1}^{N}\sum_{ij}(M_{i,j}^{\tilde{n}} - M_{i,j}^{n})^2 \quad (12)$$

where $M_{i,j}^{n}$ and $M_{i,j}^{\tilde{n}}$ represent the n-th landmark heatmap of the recovered image and the GT image at position (i, j).

4) ADVERSARIAL LOSS

To generate realistic image details, this paper uses a discriminator network to constrain the adversarial loss. During the training process, the discriminator randomly selects one of the reconstructed image and the real image to distinguish the actual data distribution from the generated data distribution. The discriminator interacts with the generator to push the generator to produce realistic and detailed images. In order to reduce the difficulty of network training, this paper uses WGAN-GP [25] to define the adversarial loss as:

$$l_{WGAN} = \mathbb{E}_{I'\sim\mathbb{P}_g}[D(I^{SR})] - \mathbb{E}_{I\sim\mathbb{P}_r}[D(I^{HR})] + \lambda\mathbb{E}_{I'\sim\mathbb{P}_{I'}}[(\left\|\nabla_{I'}D(I')\right\|_2 - 1)^2] \quad (13)$$

where G and D are the SR network and adversarial network, $\mathbb{P}_g$ and $\mathbb{P}_r$ are the generated sample $I' = G(I^{LR})$ and the GT sample respectively. Randomly sample a pair of true and false samples, and randomly interpolate between the two samples to obtain the sample distribution $\mathbb{P}_{I'}$.

5) OVERALL OBJECTIVE

The overall objective function of the GCFSRnet can be written as in (14):

$$l_{SR} = \alpha l_{pixel} + \beta l_{feature} + \mu l_{landmark} + \delta l_{WGAN} \quad (14)$$

where α, β, μ, δ are the weighting coefficients for l_{pixel}, $l_{landmark}$, l_{feture} and l_{WGAN} respectively.

IV. EXPERIMENTS

A. DATASET

Since the GCFSRnet needs to obtain a high-quality frontal face image, which has the same identity as the LR image, the CASIA WebFace dataset [30], which is commonly used

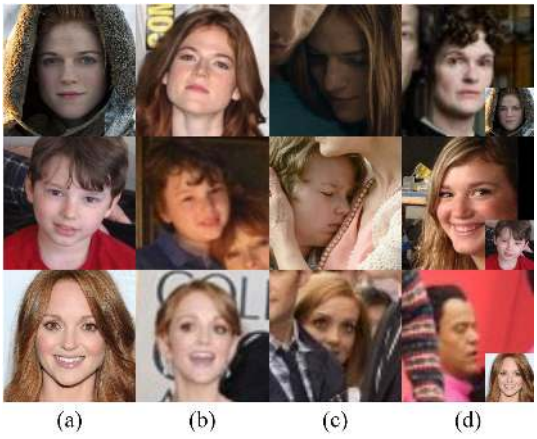


FIGURE 7. Example images of CASIA WebFace.

in face recognition tasks, is used as the training set, then the CASIA WebFace and CelebA are used as the test set. The CASIA WebFace contains 494,414 unaligned face images of 10575 people at various ages, expressions, and lighting environments with the image size of 256 × 256. CelebA dataset contains 202,599 face images of 10,177 celebrities. However, as shown in Fig. 7, the quality of the images in CASIA WebFace is quite different, and there are cases where the identity label is wrong. So we exclude images with poor quality (Fig. 7-b), large occlusion in face area (Fig. 7-c), and incorrect identity label (Fig. 7-d) in CASIA WebFace dataset. Then the selected 5156 images corresponding to 300 people in CASIA WebFace are used as the training set, 497 images for 30 people in CASIA WebFace, and 800 images for 50 people in CelebA are used as the test set, and there is no intersection between the training set and test set. In addition, each person has a corresponding guide image. The guide image is supposed to be frontal, open eyes, clear, and unobstructed, and the guide image and its degraded image are not used as the HR-LR pair in training.

Without any pre-alignment operation, the LR image is generated according to Eq.15.

$$I^{d,s} = \left(\left(I \bigotimes \mathbf{k}_\rho\right) \downarrow_s + \mathbf{n}_\sigma\right)_{JEPG_q} \tag{15}$$

In Eq.15, the image I is first convolved with a Gaussian blur kernel $\mathbf{k}_\rho$ with a standard deviation of ρ, then down-sampled by factor s and added additive white Gaussian noise $\mathbf{n}_\sigma$ with a noise coefficient of σ. Finally, we use JPEG compress quality coefficient q on the image is able to obtain the degraded image $I^{d,s}$. In order to keep the size of the output image as same as the input, we use Bicubic interpolation to upsample $I^{d,s}$ to the initial size as the LR input. Because this experiment use 4x scale face super-resolution reconstruction as an example, so we set the down-sampling parameter $s = 4$ in Eq.15. In the general image compression quality factor q is about 80. In order to show the SR algorithms can reconstruct image with severe degradation we set $q = 50$ in this paper. The multiplicative noise factor and the additive noise factor

FIGURE 8. Distortion results of the pose deformation module.

are used using a common settings $\rho = 1$ and $\sigma = 3$. So the parameters of Eq.15 are set as $s = 4$, $\rho = 1$, $\sigma = 3$, $q = 50$.

$$I^{LR} = \left(I^{d,s}\right) \uparrow_s \tag{16}$$

B. TRAINING DETAILS

In order to determine each hyper parameter of the loss function, we use a grid search method: set the value of each parameter to: 0.005, 0.01, 0.05, 0.1, 0.5, 1, and calculate the average PSNR of the training set under different parameters, the parameters with the highest PSNR are selected as the final parameters, which are $\alpha = 1$, $\beta = 0.1$, $\mu = 0.005$, and $\delta = 0.01$.

TABLE 1. The PSNR of the different hyper parameters for overall loss.

	0.005	0.01	0.05	0.1	0.5	1
α	24.86	26.32	27.78	28.93	29.54	29.65
β	26.39	27.45	28.77	29.65	28.98	28.32
μ	29.65	29.54	29.30	29.14	28.75	28.43
δ	29.43	29.65	29.55	29.20	28.77	27.63

In addition, we also analyzed the different effects of different parameters on the overall loss. For this reason, we added a set of ablation study to demonstrate the hyper parameters (α, β, μ, and δ). The experimental results are shown in Table 1. α, β, μ, and δ represent the four hyper parameters. We fix the three parameters of $\alpha = 1$, $\beta = 0.1$, $\mu = 0.005$, and $\delta = 0.01$ in turn, and then set the remaining one parameter to 0.005, 0.01, 0.05, 0.1, 0.5, 1 and calculate the average PSNR of the training set. The results in each row in Table 1 represent the average PSNR of the training set for different parameter values. It can be seen from the table that as the value of α increases, the PSNR result gradually increases, which shows that α has the greatest impact on the overall loss. Compared with α, the influence of parameter β is second. At the same time, the influence of the parameters μ, and δ on the overall

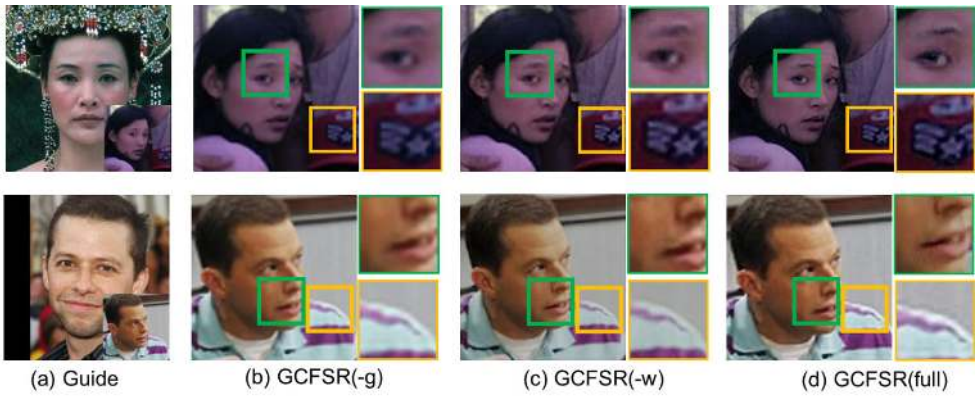


FIGURE 9. Reconstruction results of GCFSR(-g), GCFSR(-w) and GCFSRnet with upscale factor of 4.

loss is significantly smaller than that of α and β, and the influence of μ is the smallest.

The GCFSRnet is trained using the RMSprop algorithm with an initial learning rate of 0.001. The discriminator is trained using the Adam algorithm with learning rate 0.0001, $\beta_1 = 0.1$, $\beta_2 = 0.999$, e $= 10^{-8}$. The tradeoff parameters are set as $\alpha = 1$, $\beta = 0.1$, $\mu = 0.005$, and $\delta = 0.01$.

The training process of GCFSRnet can be divided into two steps. At first, the GCFSRnet is trained with the mini-batch size of 16. The discriminator does not participate in network training. Therefore the adversarial loss weight is set to 0. After 100 epochs, the adversarial loss weight is set to 0.1, the learning rate of RMSprop decays by a factor of 0.9 after each epoch. The batch size is 4 at this stage, and the training is stopped after 30 epochs.

C. ABLATION STUDY

Fig. 8 shows the distortion results of the pose deformation module in GCFSRnet. After the guide image and LR image pass through the pose deformation module, it can generate a natural and smooth facial deformation result with a noticeable deformation effect. For the image's face area, the distorted image is clear and undistorted; it has the same posture as the LR image and can provide sufficient high-frequency information during the SR process.

In order to verify the effectiveness of the guide image and the pose deformation module, we further build three networks i.e., GCFSR(-g), GCFSR(-c), and GCFSR(-w) for comparison. Among them, GCFSR(-g) neither use the guide image nor contain the pose deformation module. Only the LR image is used as the input. In GCFSR(-c), the guide image is only used in the pose deformation module and the first-layer prior recovery network as a common input. GCFSR(-w) does not contain a pose deformation module, and the original guide image without deformation is directly input into the SR network with the LR image. GCFSR(-g), GCFSR(-c) and GCFSR(-w) are all trained with the same dataset and parameters as the GCFSRnet.

To compare the experimental results more effectively, we divide the above networks into three groups for comparison, see Fig. 9, Fig. 10, and Fig. 11. Fig. 9 illustrates the influence of the pose deformation module, and Fig. 10 illustrates the influence of the addition of guide images in the SR network, and Fig. 11 illustrates the beneficial information of the guided image and the deformation module to the network.

In Fig. 11, where (b)(c)(e)(f) are obtained from the residual between the GCFSR method and the GCFSR(-g), GCFSR(-w) and GCFSR(-c). After the residual calculation, the residual value is stretched. Among them, (a) and (d) are the super-resolution results of the face obtained by the GCFSR method, (b) and (e) are the extra details obtained by the guided image in the GCFSR method, (c) is the extra details obtained by the pose deformation module, (e) is the extra details obtained by the cascade of two layers of networks. It can be clearly seen from (b) and (e) that adding a guided image can effectively increase the edge and high-frequency details of the image and increase the facial details; the deformation module obtained in (c) can also increase the high-frequency details and different facial parts of the image. It can be seen in (f) that use two cascaded networks can further capture the image information and make the image clearer and sharper.

As shown in Fig. 9, GCFSR(-g) without using the guide image cannot generate sharp contours in the facial features area, and the high-frequency information of the overall image of the reconstructed image is insufficient. For GCFSR(-w) that deforms the guide image without the pose deformation module, although its reconstruction results in the face area are significantly improved compared to GCFSR(-g), there are still unclear problems edges. In contrast, the reconstruction results generated by the GCFSRnet have a clearer facial contour, more accurate and sharp performance in the facial features area, and a more fine-grained visual effect in other non-face areas (as marked by the yellow box in Fig. 9).

Notice that when the guide image is very close to the face pose in the LR image, GCFSR(-w) can also obtain satisfactory reconstruction results. At this time, the effect of the pose

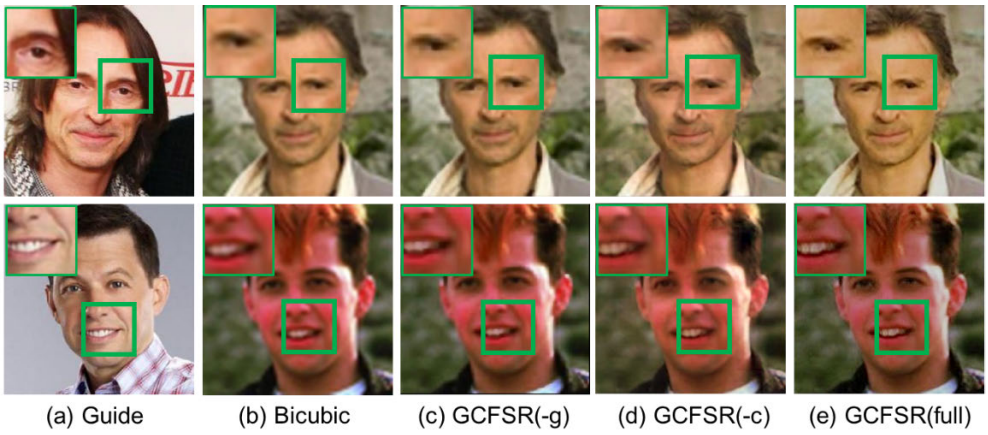


FIGURE 10. The reconstruction results of GCFSR(-g), GCFSR(-c) and GCFSRnet with upscale factor of 4.

FIGURE 11. The extra details obtained GCFSR(g), GCFSR(w) and GCFSR(c) with upscale factor of 4.

deformation module on the guide image is almost negligible. This is because the SR network improves the output image's consistency and the input image by minimizing pixel loss during the training process. When the guide image's face area can already basically coincide with the LR image, it is enough for the SR network to learn the high-frequency information of the current reconstruction area without the pose deformation network. However, under normal circumstances, the difference between the high-definition guide image and the LR image makes GCFSR(-w) actively add real high-frequency information to the SR process. This information always exists large deviations with the current reconstruction area at pixel position and cannot be used. Therefore, it is necessary to jointly utilize the guide image and the pose deformation module to reconstruct the realistic and rich-information face image.

In Fig. 10, the areas with noticeable facial features are highlighted. It can be seen from Fig. 10 that the GCFSR(-c), which added the guide image only once during the SR process, can produce clearer facial features compared to GCFSR(-g). The picture's overall clarity has improved, but it still exists apparent deficiency compared with the GCFSRnet. Due to the two cascaded guide images are used in GCFSRnet, the facial feature area generated by GCFSRnet is more affected by the guide image, and this area has a higher definition and sharper edge texture.

PSNR and SSIM, which are often used in image processing tasks, are used for quantitative comparison, and the results are shown in Table 2. Generally speaking, the higher value of PSNR and SSIM means better performance. The highest value is displayed in red font, and the second-highest value is marked in blue font. Consistent with visual quality, the GCFSRnet has achieved the highest value in both PSNR and SSIM indicators.

The above experiments prove that the guide image can enhance the image details of the face area. In order to verify whether the reconstructed image will change when the guide image is different, we set up two groups of comparison results based on age and gender, as shown in Fig. 12. Since GCFSRnet is not aiming at blind face reconstruction, under the upscale factor of 4, the input LR image can maintain the original contour and facial features. The overall reconstructed

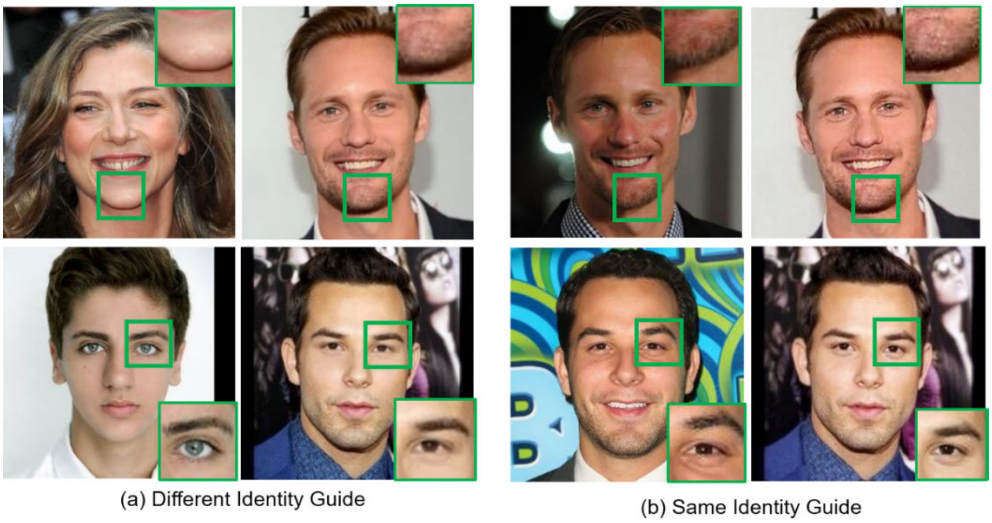


FIGURE 12. Reconstruction results when guide images have different identities.

FIGURE 13. Reconstruction results of GCFSRnet and other state-of-the-art methods at an upscale factor of 4 on CASIA WebFace dataset.

TABLE 2. Resulted mean PSNR and mean SSIM.

Algorithm	PSNR mean (dB)	SSIM mean
Bicubic	26.47	0.7656
GCFSR(-g)	28.51	0.8497
GCFSR(-c)	28.84	0.7841
GCFSR(-w)	28.97	0.8329
GCFSRnet	**29.65**	**0.8842**

image is not changed, but when zooming in on the five senses' details, subtle changes can be observed. Therefore, in the reconstruction task of the upscale factor of 4, the guide image's identity selection will have different degrees of influence on the expression of the reconstruction results in features and textures. However, when the current reconstruction object cannot find a suitable guide image of the same identity, the features similar to the other identities can also achieve good results.

D. COMPARISON WITH STATE-OF-THE-ART

To evaluate the reconstruction performance of the GCFSRnet at the upscale factor of 4 and 8, we compared with state-of-the-art SR methods, including SRGAN [22], SRResNet [22], VDSR [15], DBPN [31], FSRNet [28], SiGAN [32], ATMFN [33]. SRGAN, SRResNet, and VDSR only provide 4× upscale models, while DBPN, FSRNet, SiGAN, and ATMFN provide 4× and 8× upscale models. It is worth mentioning that FSRNet, SiGAN, and ATMFN are SR methods for face hallucination. As a comparison, we add the above methods to the comparison experiment. We use the above models' released pre-training model to conduct a comparative experiment at the corresponding upscale factor.

Fig. 13 and Fig. 14 is the reconstruction results of GCFSRnet and other state-of-the-art methods at the upscale factor of 4 on the CASIA Web Face and CelebA dataset,

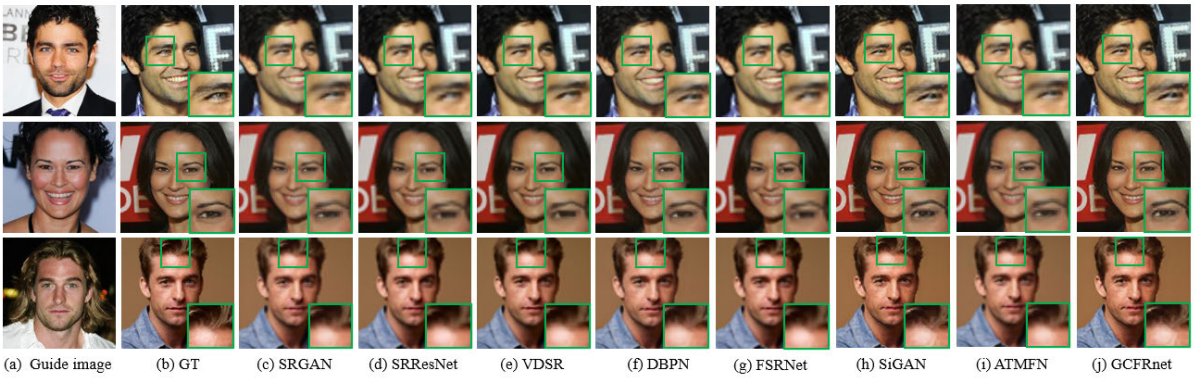


FIGURE 14. Reconstruction results of GCFSRnet and other state-of-the-art methods at upscale factor of 4 on CelebA dataset.

TABLE 3. Qualitative comparison with state-of-the-art methods.

		CASIA Web Face			CelebA		
		PSNR	SSIM	NIQE	PSNR	SSIM	NIQE
Bicubic	4	24.57	0.72	5.34	25.08	0.70	5.03
SRGAN	4	27.72	0.83	4.68	26.33	0.84	4.77
SRResNet	4	26.26	0.74	4.12	26.57	0.73	3.93
VDSR	4	25.97	0.76	2.31	26.04	0.78	3.21
DBPN	4	27.86	0.86	3.37	28.59	0.83	2.85
FSRNet	4	28.43	0.78	3.25	27.84	0.81	2.69
SiGAN	4	27.14	0.86	2.44	29.32	0.86	2.97
ATMFN	4	29.74	0.84	2.14	28.44	0.80	2.19
GCFSRnet	4	30.41	0.89	2.01	29.88	0.87	2.26
Bicubic	8	22.19	0.53	7.44	22.31	0.53	7.98
SRGAN	8	22.53	0.55	6.31	22.61	0.56	5.92
VDSR	8	23.39	0.56	5.65	23.5	0.55	4.63
DBPN	8	24.25	0.60	3.17	24.38	0.61	4.25
FSRNet	8	24.32	0.58	3.54	24.12	0.58	3.86
SiGAN	8	23.96	0.58	3.28	23.67	0.59	3.45
ATMFN	8	24.48	0.59	4.12	24.09	0.56	2.86
GCFSRnet	8	24.67	0.60	2.89	24.29	0.60	3.12

respectively. Table 3 summarizes quantitative results on the two datasets.

For scale factor 4, it can be seen from Table 3 that similar to the visible results of Fig. 13 and Fig. 14. The proposed GCFSRnet has obtained the highest values on the PSNR and SSIM. Besides, we used NIQE [34] as a reference-free image quality evaluation to compare the different methods' effectiveness. The smaller value of NIQE means better image quality. The results in Table 3 show that the proposed GCFSRnet has good results in these three quantitative methods. It can be seen that the reconstructed images of SRGAN and SRResNet are too smooth since the network of these methods is too shallow to obtain enough high-frequency information. Therefore the blur effect caused by the low resolution still exists. By increasing the network's depth and fitting images of different sizes, VDSR can effectively predict missing pixels and generate sharper reconstruction results. However, when zoomed in, the image's unnatural texture transitions can be clearly found, and the reconstructed image is not realistic. Compared with SRGAN, SRResNet, FSRNet, and ATMFN, the result of DBPN has slightly improved in sharpness and edge texture but still blurring. The result of SiGAN is similar to VDSR, with good sharpness and texture details, but the overall effect of the image is not realistic. While our method actively provides high-frequency information with the same identity for reconstruction and can reconstruct images with rich details and clear facial areas such as facial features and hair, and closer to the GT image. Besides, the LR image is reconstructed by two cascades of the SR network, which can gradually improve the image clarity, makes the GCFSRnet reconstruct the background with a more delicate texture.

For scale factor 8, Fig. 15 shows the comparison results of the GCFSRnet and other state-of-the-art methods. It can be seen that GCFSRnet can obtain the best visual effects, but the quantitative value is not the highest.

It can be seen that our method can obtain better visual effects in the face area, but the sharpness of the facial features and facial contour areas of the 8× reconstructed image is significantly reduced compared to the 4× reconstructed image, and the overall image is smoother. This is because the upper

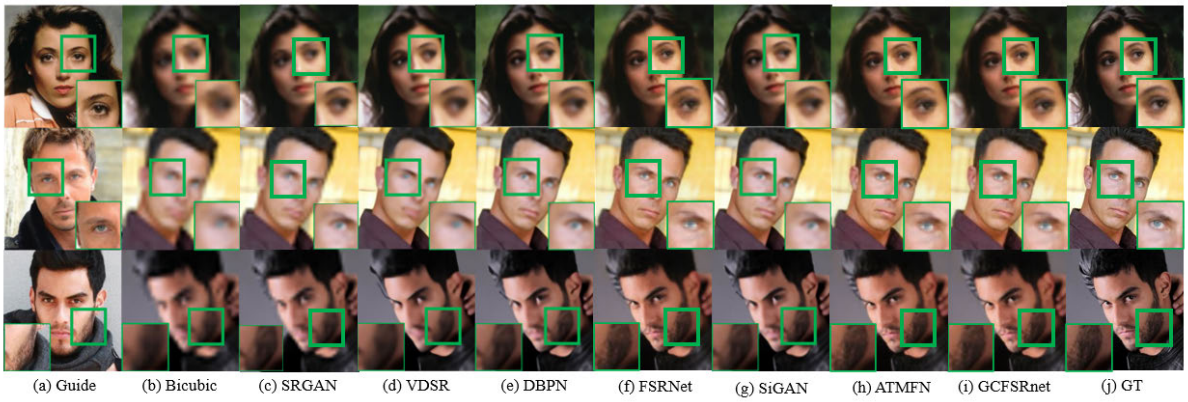


FIGURE 15. Reconstruction results of GCFSRnet and other state-of-the-art methods at upscale factor of 8.

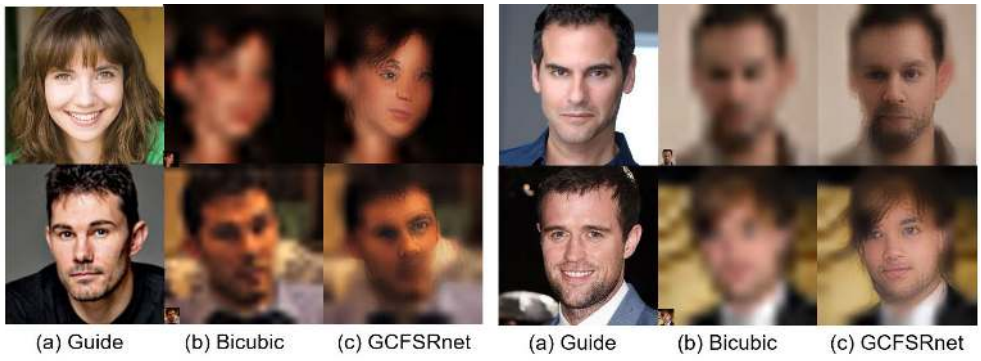


FIGURE 16. Reconstruction results of GCFSRnet at upscale factor of 16, the actual input image is shown in the lower left corner of column b.

limit of the reconstructed image's visual effect is affected by the quality of the training data. During the experiment, the image quality of our training set is not high and mixed. Therefore the training set requires manual screening of image data and labeling a guided image for each identity, making the training data scale smaller and the reconstruction effect limited. If we choose a face dataset with higher quality and larger scale for training, we may be able to obtain better visual effects.

E. SHORTCOMINGS AND PROSPECTS

Generally, face SR and face hallucination methods mainly targeted the upscale factor of 4 and 8, while the upscale factor of 16 is still a big challenge. In this paper, the result at the upscale factor of 16 is presented in Fig. 16. Although the GCFSRnet takes high-quality guide images as an extra input, the reconstructed images are still blurry and have apparent errors. This result is mainly because under high upscale factors, and it is challenging to detect accurate face landmarks from LR images, sometimes the landmarks are even undetectable, resulting in the deformation module not working. When this happens, the SR network cannot learn useful information from the guide image. Besides, when the upscale factor is too large, the reconstruction network can hardly obtain enough information from the LR image. It can only depend on generative adversarial training to simulate the real face image based on the small amount of input information, making the difference between the reconstructed image and the GT image is too large.

V. CONCLUSION

In this paper, we proposed a guided cascade face SR method, called GCFSRnet. GCFSRnet takes the LR image and a high-quality guide image of the same identity as inputs. It can be divided into a pose deformation module and an SR network. The pose deformation module converts the guide image's posture into the same as the low-resolution face image based on 3D fitting and 3DMM. The SR network extracts image features from the LR image and the deformed guide image through a cascade structure. During the SR process, the deformed guide image can provide high-frequency face information, which generates fine facial texture. The cascade structure of the SR network can enhance image clarity step by step and improve the guide image's utilization. Experiments on the CASIA Web Face and CelebA dataset show that our GCFSRnet can generate facial images with clear outlines and rich details.

REFERENCES

[1] H. S. Prashanth and H. L. Shashidhara, "Image scaling comparison using universal image quality index," in *Proc. Int. Conf. Adv. Comput., Control, Telecommun. Technol.*, Trivandrum, India, Dec. 2009, pp. 859–863.

[2] K. T. Gribbon and D. G. Bailey, "A novel approach to real-time bilinear interpolation," in *Proc. 2nd IEEE Int. Workshop Electron. Des. Test Appl.*, Perth, WA, Australia, 2004, p. 126.

[3] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 6, pp. 1153–1160, Dec. 1981, doi: 10.1109/TASSP.1981.1163711.

[4] M. Irani and S. Peleg, "Motion analysis for image enhancement: Resolution, occlusion, and transparency," *J. Vis. Commun. Image Represent.*, vol. 4, no. 4, pp. 324–335, 1993.

[5] Stark, Henry & Oskoui, Peyma, "High-resolution image recovery from image-plane arrays, using convex projections," *J. Opt. Soc. America. A, Opt. image Sci.*, vol. 6, pp. 1715–1726, Oct. 1989, doi: 10.1364/JOSAA.6.001715.

[6] W. T. Freeman and E. C. Pasztor, "Learning low-level vision," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1182–1189, doi: 10.1109/ICCV.1999.790414.

[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision*, vol. 8692, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014.

[8] 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)*, Santiago, China, Dec. 2015, pp. 370–378, doi: 10.1109/ICCV.2015.50.

[9] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Computer Vision*, vol. 9909, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016.

[10] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep Cascaded Bi-Network for Face Hallucination," *Computer Vision*, vol. 9909. Cham, Switzerland: Springer, 2014.

[11] Y. Song, J. Zhang, S. He, L. Bao, and Q. Yang, "Learning to hallucinate face images via component generation and enhancement," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 4537–4543, doi: 10.24963/ijcai.2017/633.

[12] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal Process.*, vol. 128, pp. 389–408, Nov. 2016, doi: 10.1016/j.sigpro.2016.05.002.

[13] M. S. M. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6626–6634.

[14] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Techn.*, 1999, pp. 1–8, doi: 10.1145/311535.311556.

[15] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1646–1654.

[16] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *Proc. ACM SIGGRAPH Papers*, 2003, pp. 300–325, doi: 10.1145/1201775.882269.

[17] Bulat, Adrian & Tzimiropoulos, Georgios, "How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks)," *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1021–1030, doi: 10.1109/ICCV.2017.116.

[18] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2017, pp. 483–499, doi: 10.1007/978-3-319-46484-8_29.

[19] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2129–2138.

[20] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4295–4304.

[21] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 787–796.

[22] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 105–114.

[23] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," Tech. Rep., 2016.

[24] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," Tech. Rep., 2017.

[25] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein GANs," Tech. Rep., 2016.

[26] A. M. Bruckstein, R. J. Holt, T. S. Huang, and A. N. Netravali, "Optimum fiducials under weak perspective projection," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Corfu, Greece, 1999, p. 67.

[27] 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.

[28] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-End learning face super-resolution with facial priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2492–2501.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[30] D. Yi, Z. Lei, S. Liao, and S. Li, "Learning face representation from scratch," Tech. Rep., 2016.

[31] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2472–2481.

[32] C. Hsu, L. Chung, S. Chia-Wen, and G. Wengtai, "SiGAN: Siamese generative adversarial network for identity-preserving face hallucination," Tech. Rep., 2018.

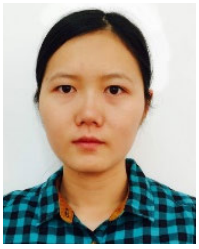
[33] K. Jiang, Z. Wang, P. Yi, G. Wang, K. Gu, and J. Jiang, "ATMFN: Adaptive-threshold-based multi-model fusion network for compressed face hallucination," *IEEE Trans. Multimedia*, early access, Dec. 18, 2020, doi: 10.1109/TMM.2019.2960586.

[34] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013, doi: 10.1109/LSP.2012.2227726.

LIN CAO received the B.Eng. degree in telecommunication engineering from Northeastern University, China, in 1999, and the Ph.D. degree in signal and information processing from the Institute of Electronics, Chinese Academy of Sciences, in 2005. He is currently a Professor with the Department of Electronic Engineering, Beijing Information Science and Technology University (BISTU). He teaches courses on digital signal processing, digital image processing, and soft design fundamentals. He is also the Dean of the School of Information and Communication Engineering and the Deputy Director of the Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, BISTU. He has published over 40 articles on image processing and pattern recognition. His research interests include radar signal processing and image understanding and recognition. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the China Education Society of Electronics.

JIAPE LIU received the B.S. degree in information and communication engineering from Beijing Information Science and Technology University, in 2017, where she is currently pursuing the master's degree of science in electronic and communication engineering. Her research interest includes face image super-resolution.

KANGNING DU received the B.Sc. degree in telecommunication engineering from Beijing Information Science and Technology University, in 2011, and the Ph.D. degree in communication and information system from the Institute of Electronics, Chinese Academy of Sciences, in 2016. He is currently a Teacher of electronic engineering with Beijing Information Science and Technology University. His research interests include radar signal processing and image understanding and recognition.



YANAN GUO received the B.Sc. degree from Hubei Polytechnic University, in 2014, and the Ph.D. degree from Yunnan University, in 2019. She is currently a Teacher of electronic engineering with Beijing Information Science and Technology University. Her research interests include machine learning and computer vision.

TAO WANG received the Ph.D. degree from the School of Electronic and Information Engineering, Beihang University, in 2019. He is currently a Teacher of electronic engineering with Beijing Information Science and Technology University. His research interests include radar signal processing, data fusion, and target localization and tracking in intelligent transportation system or vehicle intelligent assistance systems.