Received: 20 August 2022 | Revised: 2 June 2023 | Accepted: 25 June 2023

DOI: 10.1049/ipr2.12863

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

# Cascaded face super-resolution with shape and identity priors

Dan Zeng[1,2] | Zelin Li[1,2] | Xiao Yan[1,2] | Wen Jiang[1,2] | Xinshao Wang[3] | Jiang Liu[1,2] | Bo Tang[1,2]

[1]Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, Shenzhen, China

[2]Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China

[3]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China

**Correspondence**
Bo Tang, Research Institute of Trustworthy Autonomous Systems and Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China.
Email: tangb3@sustech.edu.cn

**Funding information**
National Natural Science Foundation of China, Grant/Award Number: 62206123

**Abstract**

Despite impressive progress in face super-resolution (SR), it is an open challenge to reconstruct a reliable SR face that preserves authentic facial characteristics. Here, the problem of super-resolving low-resolution (LR) faces to high-resolution (HR) ones is addressed. To tackle the ill-posed nature of face SR, the cascaded super-resolution network (CSRNet) is proposed to utilize shape and identity priors jointly and progressively, the first to explore multiple priors. Specifically, CSRNet adopts a cascaded structure to transform an LR face to HR face progressively via multiple stages. At each stage, CSRNet forces its output face image to match both the shape priors and identity priors extracted from the ground-truth HR face. The shape priors estimated in one stage are merged into the inputs of its subsequent stage to provide rich information for the face SR. To generate realistic yet discriminative faces, the cascaded super-resolution generative adversarial network (CSRGAN) is also proposed to incorporate the adversarial loss and identification loss into CSRNet. Extensive experiments on popular benchmarks show that the CSRNet and CSRGAN outperform existing face SR state-of-the-art methods, both quantitatively and qualitatively, and detailed ablation studies show the advantage of this method.

## 1 | INTRODUCTION

Face super-resolution (SR), also known as face hallucination, is to recover a high-resolution (HR) face image from its low-resolution (LR) counterpart. Face SR plays an important role in many applications such as face recognition [1, 2], person re-identification [3], and face image editing [4, 5], where LR face images are common. Image SR is an ill-posed problem by nature as there are an overwhelming number of plausible HR solutions that explain the observed LR images equally well. Many details in the HR image are not present in the input LR image and the model needs to fill in these details. Specifically, super-resolving an image with a large magnification factor (i.e. 8×) requires to estimate 64 pixels of SR image from 1 pixel of LR input, which is challenging. As a result, early methods (e.g. VDSR [6]) that directly map LR images to HR images often produce unrealistic and over-smoothed images (see column two in Figure 1).

Face SR is different from general image SR in that face images come with facial priors that can be exploited to help tackle the ill-posed problem. *Shape priors* such as landmarks, heatmaps, and parsing maps describe both the global structures (e.g. face contour) and local details (e.g. the location and shape of eyes, nose, and mouth) of the face. Deep models including Retinaface [7], MTCNN [8], and HourGlass [9] are designed to generate shape priors for input face image. *Identity priors* provide semantic information about the person in the image (i.e. who is this person?), which is essential for enhancing authentic facial characteristics in the SR face. This semantic information can be extracted by deep models such as FaceNet [10, 11] and Arcface [12].

Many works use shape priors to enhance face SR. For example, SuperFAN [13] trains the SR model by enforcing the SR face to produce landmark heatmaps similar to that of the ground-truth HR face. FSRNet [14] fuses the estimated parsing maps with the LR face as input to the SR model. DIC [15] fuses heatmaps into the indeterminate features by an attentive fusion module, which is then processed by convolution operations. Relatively fewer works use identity priors to improve face SR. SICNN [16] feeds the SR face generated by the SR model to the face recognition network and jointly trains the face SR model and face recognition network. CSRIP [17] uses pre-trained face recognition networks to encourage the SR face and the ground-truth HR face to have similar classification results.

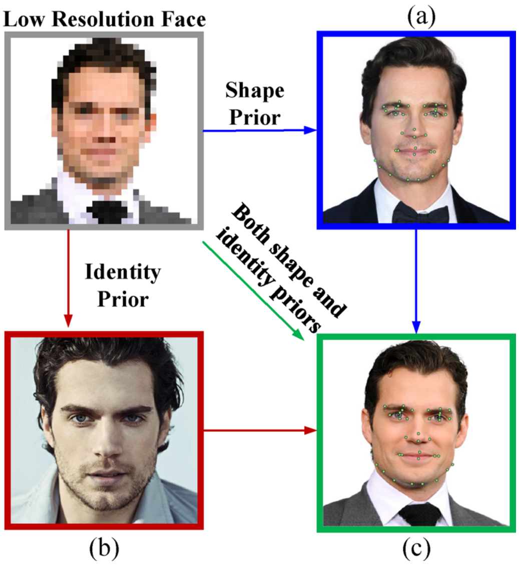


**FIGURE 1** SR faces generated by some representative methods with a magnification factor 8×. VDSR does not use facial priors, SuperFAN uses shape priors, CSRIP uses identity priors, and the proposed CSRNet uses both shape and identity priors. CSRNet provides more realistic details in key facial regions such as eyes, eyebrows, and teeth. Cascaded super-resolution generative adversarial network (CSRGAN) is CSRNet with the adversarial and identification losses. A larger PSNR/SSIM indicates better image quality.

As illustrated in Figure 1, models using either shape priors (e.g. SuperFAN) or identity priors (e.g. CSRIP) produce more visually plausible SR image than VDSR, which uses only pixel-wise reconstruction constraints.

Here, we propose the cascaded super-resolution network (CSRNet), the first to utilize shape and identity priors jointly and progressively for face SR to our best knowledge, and the idea is illustrated in Figure 2. With both shape and identity priors for face SR, CSRNet can preserve the visual and semantic quality of the SR face. Compared with existing methods including shape prior-based and identity prior-based face SR, CSRNet retains the most authentic facial characteristics.

The overall model structure of CSRNet is illustrated in Figure 3. CSRNet adopts a cascaded structure, which transforms an LR face to an SR face progressively via multiple stages. In each stage, CSRNet applies a face alignment network (FAN) to extract shape priors from the SR face. These shape priors are forced to match the shape priors extracted from the ground-truth HR face with the MSE loss and used as input for the next stage to facilitate SR learning. For the identity priors, CSRNet uses a COST face matcher [11] to extract discriminative embedding from the SR face in each stage and forces it to match the embedding extracted from the ground-truth HR face. Instead of constructing RGB face, CSRNet learns to predict the residue between the ground-truth HR face and simple bicubic interpolation of the LR face. To generate realistic yet discriminative faces, we incorporate the adversarial loss and identification loss into CSRNet and formulate CSRGAN. CSRGAN recovers realistic semantic meaning and generates more plausible details compared with other methods.

We conducted extensive experiments on the widely used datasets: Helen and CelebA. The results show that CSRNet provides better quantitative performance in terms of PSNR and SSIM than state-of-the-art face SR methods including SuperFAN, DIC, and CSRIP. The SR faces generated by CSRGAN is also more appealing by preserving more details. Detailed ablation studies verify that shape and identify priors are complementary and both contribute to better performance for face SR, which again show the advantage of our CSRNet method.

In summary, we made the following contributions here.

- To the best of our knowledge, CSRNet is the first deep face super-resolution model that utilizes shape priors and identity priors jointly and progressively. This is motivated by the fact that different facial priors are closely related and provide different perspectives of constraints for SR solution space.
- We design an effective pipeline for CSRNet, which uses a cascaded structure, extracts the shape priors on residual face image, and combines with semantic feature embeddings for identity loss. This enables both the intermediate SR images as well as final SR images to be equipped with facial structure information and identity knowledge. The use of complementary information in shape and identity priors via multiple stages is new and essential to enhance face SR.
- We conduct extensive experiments on benchmarks including CelebA and Helen which demonstrate the effectiveness of CSRNet using both shape and identity priors for face SR. Specifically, our CSRNet achieves state-of-the-art performance for the challenging task of super-resolving LR face images by an upscaling factor of 8. Additionally, our CSRGAN generates more realistic yet discriminative face images with adversarial and identity losses.

The rest of this paper is organized as follows. Related work is shown in Section 2. The proposed CSRNet is described in Section 3. Experimental evaluation is given in Section 4. Finally, we draw an overall conclusion in Section 5.

## 2 | RELATED WORK

Here, we review related works from three perspectives, namely model architecture, face SR with shape priors, and face SR with identity priors. A comparison between our CSRNet and some of the most relevant methods is provided in Table 1. It is worth noting that some recent methods, such as VQFR [18], improve face SR results by using reference image priors or import pre-trained generative adversarial network (GAN) priors [19–22], such as GFPGAN [20], GPEN [21] to enhance face SR. These methods focus on using additional information (i.e. reference image, or GAN priors which are trained with large-scale HR faces) to obtain high-definition (1024 × 1024) SR results. In contrast, we focus on exploring an effective way of combing multiple priors from LR image (i.e. without exploiting other information) for face SR.

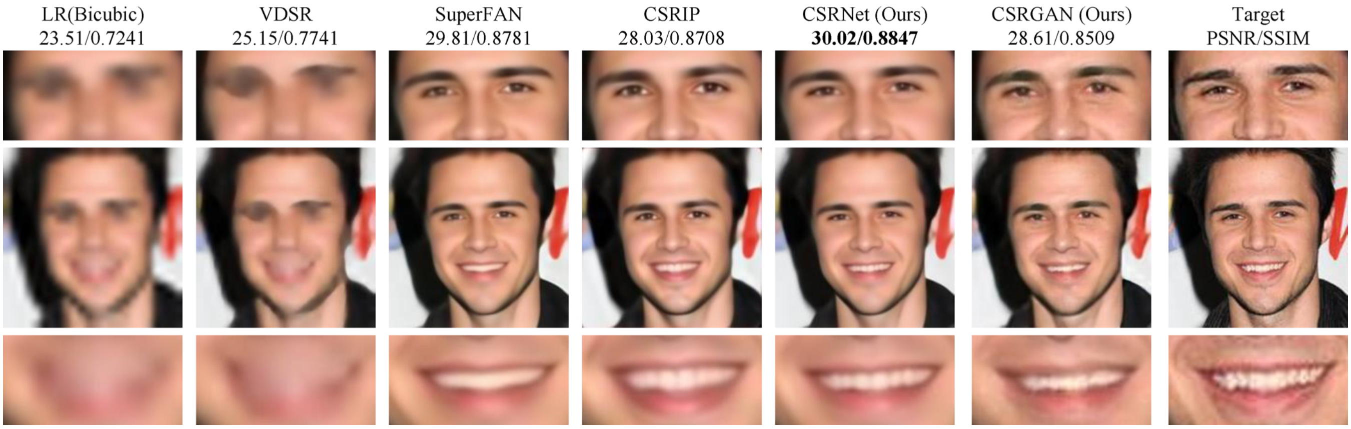


**FIGURE 2** A schematic illustration of the idea behind our CSRNet, which uses both shape and identity priors to preserve the visual and semantic quality of the SR face. (a), (b), and (c) are three plausible super-resolved faces for the input low-resolution face. Compared with (a) or (b), (c) retains the most authentic facial characteristics. (Best viewed in colour).

**TABLE 1** A comparison between CSRNet and some representative image SR methods, where VDSR, SRGAN, and LapSRN are general image SR methods, and SuperFAN, DIC, SICNN, and CSRIP are face SR methods, of which SuperFAN and DIC are face SR with shape priors, while SICNN and SCRIP are face SR with identity priors. Our CSRNet is a cascaded model that makes the most of identity as well as shape priors to enhance SR face with scaling factors of 2, 4, and 8.

| Method | VDSR | SRGAN | LapSRN | SuperFAN | DIC | SICNN | CSRIP | CSRNet (Ours) |
|---|---|---|---|---|---|---|---|---|
| Cascaded model | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | ✓ |
| Identity prior | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ |
| Shape prior | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ |
| Scaling factor | 4 | 4 | 2, 4, 8 | 4 | 8 | 8 | 2, 4, 8 | 2, 4, 8 |

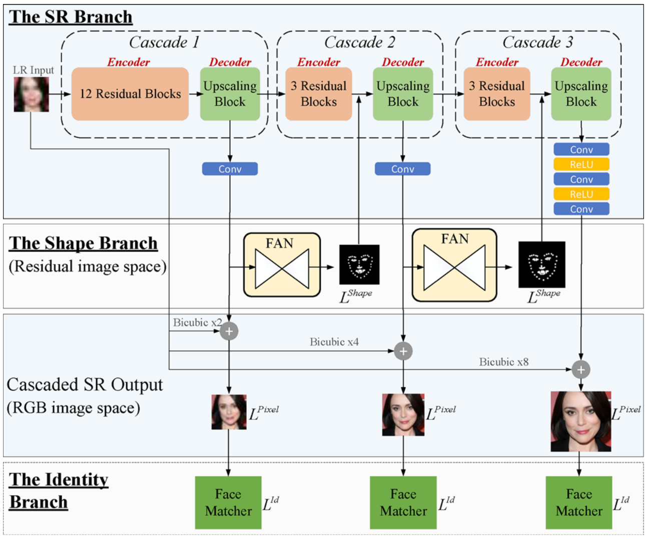


**FIGURE 3** The overall model structure of CSRNet. Our CSRNet consists of three branches: the SR branch, the shape branch, and the identity branch. In each cascade, the SR branch super-resolves the face image by a factor of 2×. The shape branch extracts the shape priors from the residual image and uses the shape priors obtained in one stage as the input features of the SR branch in the next stage. The identity branch uses the face matcher to extract semantic embedding from the model generated SR face. Given an LR input, CSRNet produces SR faces with different magnification factors (e.g. 2×, 4×, and 8×) in cascaded SR output.

## 2.1 | Model architectures

In terms of model structure, existing face SR methods can be roughly classified as *direct* and *cascaded*. The direct methods super-resolve an LR image to the desired spatial resolution in one shot. Some direct methods [6, 23, 24] first interpolate the LR input (e.g. via bicubic interpolation) to high resolution and then apply the model to adjust the coarse SR image. For example, VDSR [24] increases the depth of the network to 20 layers by using a global residual connection. However, the predefined interpolations may result in sub-optimal SR results. To address this issue, learning-based up-sampling is introduced as an alternative to predefined interpolation. The others, for example, CARN [25], SRGAN [26], URDGN [27], and EDSR [28], learn a mapping that works directly on LR input using models such as transposed convolution layer [29] and sub-pixel layer [30] at the end of deep CNNs for face super-resolution. In contrast, the progressive methods, for example, LapSRN [31], MS-LapSRN [32], and ProSR [33], learn to super-resolve the input LR image (i.e. learn to predict the residuals between the ground-truth HR image and LR input) via multiple stages, in which each stage uses a small upscaling factor (e.g. 2×) and the output of one stage is used as input for the next stage. The benefit of the progressive structure (also called cascaded structure) is that it allows supervision signal in every stage (especially the early stages), which makes the model easy to train. In addition, this progressive SR trains one model to meet the needs of multi-scale SR reconstruction while the direct architecture frameworks require training different models for varied scaling factors. Generally, without considering the face priors, image SR methods usually adopt the pixel-wise reconstruction loss (i.e. mean square error) to encourage the texture of the model generated SR image to be similar to the ground-truth HR image.

### 2.2 | Face SR with shape priors

Many methods utilize facial shape priors to enhance face SR. SuperFAN [13] introduces a FAN to extract heatmaps from face images and uses the MSE loss to guarantee that the target HR face and the generated SR face have consistent shape priors. PFSR [34] extends SuperFAN by gradually super-resolving LR face images to high resolutions. In JASRNet [35], FAN and the SR network share a common encoder, which extracts shallow features from the LR face images. Instead of using shape priors to provide supervision signals, some methods use shape priors as input for face SR. MTUN [36] concatenates component-based heatmaps with other input feature maps as input for the SR model. FSRNet [14] consists of a coarse SR network and a refined network. The refined network estimates parsing maps from the coarse SR image and feeds the parsing maps as well as other feature maps to an encoder–decoder network, which generates the final SR image. CBN [37] super-resolves the input LR face step by step and uses a gate network to fuse the coarse SR face and the dense correspondence filed of each stage. To improve SR performance, CBN extracts face prior from the intermediate SR faces instead of the input LR face. DIC [15] learns an attentive fusion module that uses the shape priors as attention weights for aggregating the feature representations in the SR model. Hu et al. [38] concatenate 3D facial priors and feature representations of the face SR model to produce a sharp SR face. PCRCN [39] adopts a cascaded recurrent network for face SR, and facial parsing priors are extracted and refined in each cascaded unit to facilitate facial details recovery. RCNet [40] estimates the facial landmarks on the coarse SR face (i.e. rather than the LR face) to achieve better accuracy. It then integrates the facial landmarks with face features to enhance face SR.

Compared with existing work, our CSRNet uses shape priors as both supervision and model inputs and extracts shape priors at multiple resolutions in the cascaded structure. The shape priors estimated in one stage are merged into the inputs of its subsequent stage to provide rich information for face SR. In addition, we collect shape priors from the residue face instead of SR faces generated by the model, which ensures that the filled residues are meaningful. We also show that using residue face yields lower landmark error.

### 2.3 | Face SR with identity priors

There are few works proposed to enhance the identity information in the reconstructed SR faces. SICNN [16] minimizes the identity difference between model generated SR face and ground-truth HR face by co-training the SR model and face recognition model. Specifically, it uses the pixel-wise reconstruction loss for the SR model, and adopts super-identity and recognition losses for the face recognition model. CSRIP [17] adopts the cascaded structure and extracts identity priors on the residue face instead of the model generated SR face. In each stage, it encourages the SR face and ground-truth HR face to have similar classification results for a pre-trained face recognition model. Since the face images from the same person have the same identity information regardless of the image resolution, DIDnet [41] consists of a face SR network and a degradation network through a dual loop. The SR network ensures the super-resolved face and the ground-truth HR face have the same identity features in the HR space. The degradation network ensures the generated LR face and the input LR have the same identity features in the LR space. Identity-aware FSR [42] consists of a face SR network and an identity-aware feature extractor. The SR network reconstructs an HR face from the input LR face and the feature extractor extracts the identity features of the reconstructed HR face. The identity features are decoupled into magnitude-related and angle-related features for explicit supervision to preserve identity information. EIPNet [43] utilizes a lightweight edge block and identity prior information to tackle the distortion in facial components so as to enhance SR. EIPNet is comprised of three residual blocks with edge blocks embedded in multiple scales to provide structural information in each ×2 upscaling process, and uses luminance–chrominance error to align global shape and colours. In addition, EIPNet encourages the final SR face and ground-truth HR face to have the same class-encoded vector by utilizing the identity loss.

Our CSRNet is different from these methods in that we use an additional semantic embedding loss to constrain identity priors and extract identity priors from the model generated SR face. In addition, we incorporate multi-scale identity information into the training CSRNet and enforce SR face of multiple resolutions and corresponding HR face to have the same identity.

## 3 | CASCADED SUPER-RESOLUTION NETWORK

Here, we first provide an overview of CSRNet and then describe the three branches that constitute CSRNet in detail, that is, the SR branch, the shape branch, and the identity branch.

### 3.1 | Overview of CSRNet

CSRNet consists of three branches as illustrated in Figure 3: (i) the SR branch (top row) that progressively transforms the input LR face to higher resolutions using multiple cascades (i.e. stages); (ii) the shape branch (second row) that uses the FAN to extract shape priors (heatmaps in our case) from the output of the SR branch; (iii) the identity branch (bottom row) that extracts semantic embedding (using the face matcher) from the SR face produced by each stage.

In each stage, the SR branch super-resolves the face image by a factor of 2×, and thus the number of pixels is multiplied by 4×. Denote a training sample as $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_S)$, in which $\mathbf{x}$ is the input LR face, $\mathbf{y}_s$ is the ground-truth HR face for stage $s$ and there are a total of $S$ stages. At stage $s$, the SR branch predicts $\mathbf{y}_s - \mathbf{y}_s^B$, in which $\mathbf{y}_s^B$ is the bicubic interpolation of the LR face $\mathbf{x}$ for stage $s$. We adopt this residue design as it is reported to provide better performance and easier convergence property than directly predicting $\mathbf{y}_s$ [33]. Denote the output of the SR branch at stage $s$ as $\mathbf{y}_s^{SR}$, the shape branch extracts the shape priors from $\mathbf{y}_s^{SR}$ (i.e. the residual image) rather than $\hat{\mathbf{y}}_s = \mathbf{y}_s^B + \mathbf{y}_s^{SR}$ (i.e. the model generated SR face), and the shape priors obtained in one stage are fused used as input features for the SR branch in the succeeding stage. The cascade stage $s$ (when $s \geq 2$) comprises a low-resolution encoder–decoder feature extractor $f_s^l(\cdot)$ and a shape prior predictor $f_s^p(\cdot)$, which is a FAN. The output of the cascade block is indicated by $F_s$, which is convoluted to produce $\mathbf{y}_s^{SR}$. The convolution operation is denoted by $Conv(\cdot)$. The final non-linear operation, achieved by stacking several convolutional and ReLU layers, is denoted by $O_S(\cdot)$. Therefore, the face SR process can be formulated by:

$$\mathbf{y}_1^{SR} = Conv\left(f_1^l(\mathbf{x})\right), \tag{1}$$

$$\mathbf{y}_s^{SR} = Conv\left(f_s^l(F_{s-1}), f_s^p\left(\mathbf{y}_{s-1}^{SR}\right)\right), \tag{2}$$

$$\mathbf{y}_S^{SR} = O_S\left(f_S^l(F_{S-1}), f_S^p\left(\mathbf{y}_{S-1}^{SR}\right)\right), \tag{3}$$

where $\hat{\mathbf{y}}_s = \mathbf{y}_s^B + \mathbf{y}_s^{SR}$ is our final generated SR face. CSRNet employs the shape priors in each upscaling process to preserve the high-frequency component. In contrast, the identity branch operates on the reconstructed SR face $\hat{\mathbf{y}}_s$ as texture details are essential for discriminative feature extraction.

In the training phase, the identity branch uses pre-trained models and is not updated during training while the FANs (in the shape branch) and the SR branch model are trained from scratch. In the inference phase, the identity branch is removed as it is not involved in computing the output SR face. However, the FANs are kept because they provide shape priors for the SR branch as input. The input LR face and the ground-truth HR faces for different stages are generated by down-sampling an HR face with different factors. The loss function for training CSRNet is defined as

$$\mathcal{L} = \mathcal{L}^{\text{Pixel}} + \mathcal{L}^{\text{Shape}} + \mathcal{L}^{\text{Id}}, \tag{4}$$

where $\mathcal{L}^{\text{Pixel}}$ is the pixel-wise reconstruction loss (i.e. mean square error) between the generated SR faces in cascaded SR output (Figure 3, third row) and the ground-truth HR faces, $\mathcal{L}^{\text{Shape}}$ measures the differences between the shape priors extracted from the SR faces and the ground-truth HR faces, and $\mathcal{L}^{\text{Id}}$ quantifies the differences in identity information (i.e. semantic embedding) between the SR faces and the ground-truth HR faces. Although it is possible to assign a weight factor for each of the three loss terms, we found that training with Equation (4) already results in good performance, and thus did not include weight factors due to the additional complexity of parameter tuning.

Before introducing the details of each branch and loss term, we would like to discuss the rationale behind CSRNet's overall structure. The pixel-wise reconstruction loss $\mathcal{L}^{\text{Pixel}}$ has been widely used in the literature to ensure that the texture of the SR image is similar to the ground-truth HR image. However, texture similarity does not necessarily lead to visual similarity. The shape priors contain both the global structure of the face (e.g. contour) and the shape of important facial components such as eyes, nose, and mouth (see an example in the shape branch of Figure 3). By encouraging the SR face and the HR face to have similar shape priors with $\mathcal{L}^{\text{Shape}}$, the SR face can preserve the facial structure information in the HR face, and thus looks more visually plausible. In a similar vein, the identity loss $\mathcal{L}^{\text{Id}}$ forces the SR face and the HR to be consistent in identity information such that they look similar to human inspectors. The cascaded structure also allows CSRNet to introduce supervision signals at each stage, which makes the model easy to train. To our best knowledge, CSRNet is the first to jointly utilize the shape and identity priors, and thus provides state-of-the-art performance for face SR. In the experiments, we found that the shape priors and identity priors are complementary and both of them lead to performance improvement.

In addition to jointly unitizing shape and identity priors, there are several other key designs in CSRNet that are different from existing works. First, compared with methods that utilize the cascaded structure, CSRNet fuses the shape priors as input features for face SR to strengthen the guidance of landmark maps. Second, different from existing works that extract shape priors from the SR face, we extract shape priors from the residual face as it provides better accuracy for shape prior detection and ensures the filled residues are meaningful. Third, instead of applying face recognition models on the residual face as in CSRIP, we extract identity priors from the SR face because the human vision system recognizes the original face rather than the residue.

## 3.2 | The SR branch

As illustrated in the top row of Figure 3, the SR network progressively super-resolves an LR face to higher resolutions using multiple cascades of CNNs (e.g. C1, C2, and C3). Each cascade (also called a stage) upscales the face with a factor of 2× and the overall scaling factor of CSRNet is the multiply of all cascades (e.g. 8× for Figure 3). By decomposing a difficult task into successive simple tasks, the cascaded structure of the SR branch allows for intermediate supervision at each cascade, enabling us to better constrain the SR solution than directly super-resolving the LR face with a large scaling factor. In addition, this cascaded design can lower the learning difficulty as well.

Each cascade consists of an encoder and a decoder. The encoder stacks multiple residual blocks and each block has five successive layers (i.e. Conv-BN-ReLU-Conv-BN). A skip

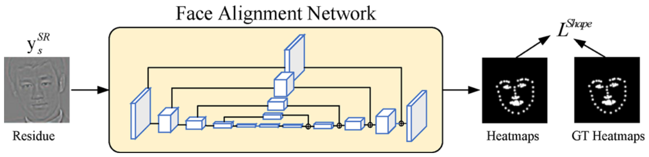


**FIGURE 4** The shape branch for a cascade. The HourGlass network is used as the face alignment network to detect 40 landmarks from the input residue. The shape loss encourages the model generated residual face to have shape prior (i.e. facial heatmaps) consistent with the ground-truth residue to preserve structure information in SR face.

connection is added between the first convolution (Conv) layer and the last batch normalization (BN) layer of each residual block. Each decoder consists of three successive layers (i.e. BN-ReLU-DeConv). The encoder does not change the resolution of the face while each decoder scales up the face by a factor of 2× with its deconvolution (DeConv) layer. We adopt the asymmetric pyramid architecture [33], in which a lower cascade is more complex than a higher cascade. Specifically, we use 12, 3, and 3 residual blocks for C1, C2, and C3, respectively. The advantage of the asymmetric pyramid architecture is that it enables a large upscaling factor while remaining efficient (by avoiding using complex models for the higher cascades).

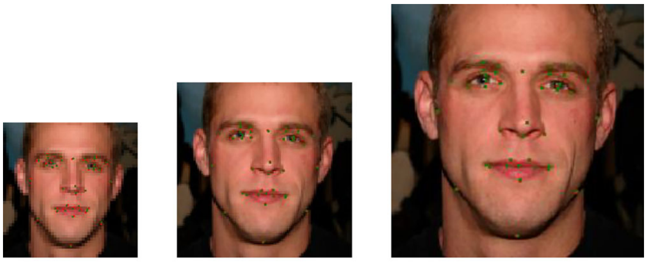
For C1 and C2, we use an FAN to extract shape priors from their outputs and these shape priors are used as input feature maps for the decoder in their succeeding cascade. To map the outputs of the cascades to residual face (i.e. $\mathbf{y}_s - \mathbf{y}_s^B$), we use one convolution layer for C1 and C2 but more convolution layers (i.e. three convolution layers) for C3 as it produces the final SR face. We apply the following pixel-wise reconstruction loss on the SR branch

$$\mathcal{L}^{\text{Pixel}} = \frac{1}{|\mathcal{T}|} \sum_{\mathbf{y}_s \in |\mathcal{T}|} \sum_{s=1}^{S} \left\| \mathbf{y}_s - \mathbf{y}_s^B - \mathbf{y}_s^{SR} \right\|^2, \tag{5}$$

where $\mathcal{T}$ is the training dataset which is composed of LR input and ground-truth HR faces with different magnification factors, $\mathbf{y}_s$ is the ground-truth HR face for stage $s$, and $\mathbf{y}_s^B$ and $\mathbf{y}_s^{SR}$ are the bicubic interpolation of the input LR face and the output of the SR branch at stage $s$, respectively. $\mathcal{L}^{\text{Pixel}}$ encourages the model generated SR face (i.e. $\mathbf{y}_s^B + \mathbf{y}_s^{SR}$) to approximate the ground-truth HR face and is applied at every stage, which provides successive supervision.

## 3.3 | The shape branch

As for the shape priors, we choose landmark heatmaps, which describe the location and shape of key facial components. As illustrated in Figure 4, heatmaps provide rich structure information about the face, including global structure (e.g. facial contour) and local details (e.g. eyes, nose, and mouth). Here, we use 8 heatmaps generated from 40 landmarks on a face image and these heatmaps correspond to different semantic components, that is, left eyebrow, right eyebrow, left eye, right eye, nose, facial contour, inner mouth, and mouth contour, respectively.



(a) 48 x 48 (b) 96 x 96 (c) 192 x 192

**FIGURE 5** 'An example face to illustrate the face alignment results of (a) 48 × 48, (b) 96 × 96', and (c) 192 × 192 by using different HourGlass networks. Green/red indicate training the network with residue face and RGB face, resppectively.

We use the HourGlass network [9] as the FAN to detect landmarks from face images. As illustrated in Figure 4, our HourGlass network consists of 4 residual modules and uses 64 channels for the feature maps. Instead of detecting landmark on the SR face as in existing works, we apply the HourGlass network on the residue face predicted by the SR branch in the residual image space. This is because the key facial components are prominent (i.e. having larger pixel values than their neighbours) in the residue image and thus landmark detection provides high accuracy. For a fair comparison, we train two FANs on the residue and SR faces, respectively, and evaluated the accuracy of landmark detection using normalized root mean square error (NRMSE) which is smaller the better. The results show that the NRMSE on images with a resolution of 48 × 48, 96 × 96, and 192 × 192 are 0.7482, 0.6206, and 0.6397 for residue face, and 0.7595, 0.7002, and 0.6433 for SR face. Such results support our design of extracting shape prior from the generated residual face. Figure 5 illustrates the face alignment results of training the HourGlass network on the RGB face and residue face, respectively.

The loss of the shape branch as illustrated in Figure 4 is defined as

$$\mathcal{L}^{\text{Shape}} = \frac{1}{|\mathcal{T}|} \sum_{\mathbf{y}_s \in |\mathcal{T}|} \sum_{s=1}^{S} \left\| \mathcal{P}_s(\mathbf{y}_s - \mathbf{y}_s^B) - \mathcal{P}_s(\mathbf{y}_s^{SR}) \right\|^2, \tag{6}$$

where $\mathcal{P}_s$ denotes the FAN model for cascade $s$, and $\mathcal{P}_s(\mathbf{y}_s - \mathbf{y}_s^B)$ is the shape prior extracted from the ground-truth residual face, and $\mathcal{P}_s(\mathbf{y}_s^{SR})$ is the shape prior from the model generated residual face. The MSE loss in $\mathcal{L}^{\text{Shape}}$ encourages the model generated residue face to have shape priors consistent with the ground-truth residue.

## 3.4 | The identity branch

Recall that the identity branch constrains the SR face to produce semantic embeddings similar to the ground-truth HR face. Identity prior constraints are critical because they make the SR face and HR face look similar for human inspectors. For semantic embedding, we use a face matcher (i.e. FaceNet) model to extract a 512 -dimensional feature embedding from

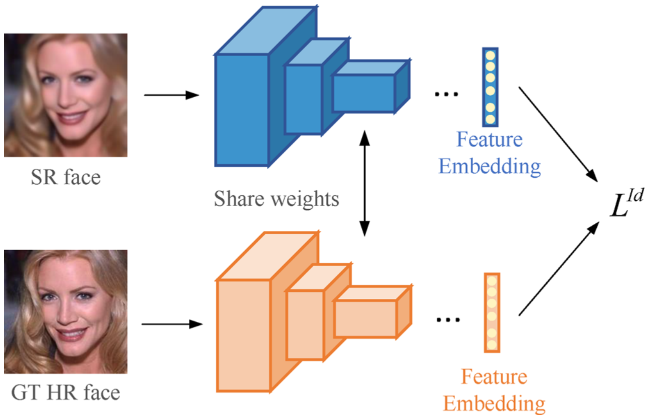


**FIGURE 6** The identity branch for a cascade. The FaceNet model is used to extract semantic feature embedding from the generated SR face. The identity loss encourages the model generated SR face to have semantic features consistent with the ground-truth HR face to preserve identity information in the SR face.

a face image as illustrated in Figure 6. The SR model is trained by minimizing the Euclidean distance between the embeddings of faces corresponding to the same person. Thus, the semantic embedding loss is defined as

$$\mathcal{L}^E = \frac{1}{|\mathcal{T}|}\sum_{\mathbf{y}_s\in|\mathcal{T}|}\sum_{s=1}^{S}\left\|\phi(\mathbf{y}_s)-\phi(\mathbf{y}_s^B+\mathbf{y}_s^{SR})\right\|^2, \qquad (7)$$

where $\phi$ is the face matcher model, $\phi(\mathbf{y}_s)$ and $\phi(\mathbf{y}_s^B+\mathbf{y}_s^{SR})$ are the embeddings of the ground-truth HR face and the model generated SR face, respectively. In practice, famous face recognition models require input images to have an image size greater than $100 \times 100$, thus CSRNet only applies faceNet to cascade C3.

## 3.5 | CSRGAN

Training SR models with the adversarial loss helps to generate more realistic images by using a discriminator network to distinguish the super-resolved images from the ground-truth HR images, and encouraging the SR network to deceive the discriminator [44]. Following this idea, we incorporate the adversarial loss into CSRGAN by using the CSRNet as a generator to ensure that CSRNet synthesizes realistic images. Moreover, we also employ an auxiliary classifier (i.e. discriminator classifier) to enhance the discriminative ability of SR images.

As illustrated in Figure 7, CSRGAN consists of CSRNet (i.e. a generator, **G**) and a discriminator classifier (i.e. **D**). CSRNet generates a super-resolved face from the LR input and the discriminator classifier outputs the probability that the input is real and its classification distribution over identities, which follows the same network structure as ACGAN [45]. The objective function has two parts: the log-likelihood of the correct source (i.e. HR or SR), $\mathcal{L}^S$, and the log-likelihood of the correct identities, $\mathcal{L}^C$.

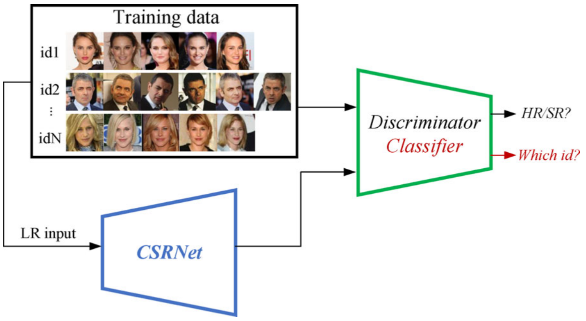


**FIGURE 7** The framework of the proposed CSRGAN. Our CSRGAN uses CSRNet as a generator to generate SR faces and a discriminator classifier to distinguish between SR face and the ground-truth HR face and predict the correct identity. As a result, CSRNet attempts to deceive the discriminator by generating realistic yet discriminative SR face.

$$\mathcal{L}^S = \mathbb{E}[\log P(S=\mathrm{HR}|\mathbf{x})] + \mathbb{E}[\log P(S=\mathrm{SR}|\mathbf{G}(\mathbf{x}))], \qquad (8)$$

$$\mathcal{L}^C = \mathbb{E}[\log P(C=id|\mathbf{x})] + \mathbb{E}[\log P(C=id|\mathbf{G}(\mathbf{x}))], \qquad (9)$$

where $\mathbb{E}$ is the expectation over a probability distribution. $S =$ HR means the source of face image is a ground-truth HR face and $S =$ SR indicates a face image from the generated SR face. $C = id$ represents the correct identity of the input face.

During training, **D** is trained to maximize $\mathcal{L}^S + \mathcal{L}^C$ and **G** is trained to maximize $\mathcal{L}^C - \mathcal{L}^S$. Specifically, $\mathcal{L}^S + \mathcal{L}^C$ encourages the discriminator to distinguish between super-resolved face images and the HR faces, and predicts the correct identity regardless of the source of input face. $\mathcal{L}^C - \mathcal{L}^S$ forces the model generated SR face $\mathbf{G}(\mathbf{x})$ to look realistic and have similar identity distribution as the input LR face $\mathbf{x}$.

# 4 | EXPERIMENTAL EVALUATION

Here, we first introduce the experiment settings and then present the main results that compare our CSRNet with state-of-the-art face SR methods. We also provide an ablation study to show the benefits of jointly utilizing shape and identity priors.

## 4.1 | Experiment settings

### 4.1.1 | Datasets and performance metrics

For CSRNet model training, we used the CASIA Webface dataset [46], which contains in total 291,915 images from 10,064 identities. Note that this CASIA dataset is a publicly available version with incorrectly labeled face images manually removed. We mainly conduct our experiments on two widely used benchmarks for face SR, that is, CelebA [47] and Helen [48, 49]. For performance test, we used 1000 images from

(a) CASIA Webface (b) Helen (c) CelebA

**FIGURE 8** Examples face images from training dataset. (a) CASIA Webface and test datasets, (b) Helen, and (c) CelebA.

**TABLE 2** Quantitative comparison between CSRNet and existing SR methods where 24 × 24 pixel images are super-resolved to the final resolution of 192 × 192 pixels using an upscaling factor of 8×. **Red**/blue indicates the best/second performance. Our CSRNet model is setting a new record of SR performance on both datasets.

| | | **Helen** | | **CelebA** | |
|---|---|---|---|---|---|
| **SR method** | **Scale** | **PSNR** | **SSIM** | **PSNR** | **SSIM** |
| Bicubic | ×8 | 25.34 | 0.7163 | 24.59 | 0.6819 |
| VDSR [6] | ×8 | 26.30 | 0.7455 | 25.57 | 0.7143 |
| SRGAN [26] | ×8 | 27.66 | 0.7987 | 26.80 | 0.7667 |
| LapSRN [31] | ×8 | 27.07 | 0.7722 | 26.21 | 0.7389 |
| NLSA [51] | ×8 | 27.77 | 0.7920 | 26.77 | 0.7583 |
| SICNN [16] | ×8 | 27.29 | 0.7793 | 26.43 | 0.7464 |
| CSRIP [17] | ×8 | 27.81 | 0.8109 | 26.99 | 0.7795 |
| SuperFAN [13] | ×8 | 28.51 | 0.8101 | 27.71 | 0.7825 |
| RCNet [40] | ×8 | 25.99 | 0.7411 | 25.18 | 0.7077 |
| PCRCN [39] | ×8 | 26.30 | 0.7286 | 26.30 | 0.7286 |
| DIC [15] | ×8 | 28.29 | 0.8016 | 27.13 | 0.7635 |
| DICGAN [15] | ×8 | 27.68 | 0.7737 | 26.99 | 0.7487 |
| CSRGAN (Ours) | ×8 | 27.64 | 0.7852 | 26.90 | 0.7581 |
| CSRNet (Ours) | ×8 | 28.71 | 0.8143 | 27.86 | 0.7867 |

the CelebA dataset and 330 images from the Helen dataset. As the face images in CASIA Webface are loosely cropped, we first take the central image patch with 200 × 200 pixels, and then resize the image to 192 × 192 pixel. For both CelebA and Helen, we crop the face region in each image based on the landmarks provided by the dataset. Our dataset pre-processing procedure follows CSRIP [17] and some examples of the pre-processed face images are illustrated in Figure 8.

To quantitatively measure the quality of the model generated SR faces, we use PSNR and SSIM [50], which are widely adopted in the image SR literature. Following the convention, the SR faces are converted from RGB to YCrCb space, and the illuminance channel is compared with (the illuminance channel of) the ground-truth HR faces to calculate the two measures. For both PSNR and SSIM, larger value means better performance.

### 4.1.2 | Implementation details

Following CSRIP [17], we used bicubic degradation to generate LR face images from HR faces. Specifically, the pre-processed faces (with a resolution of 192 × 192) are used as the ground-truth HR faces for cascade C3, and these faces are down-sampled to a resolution of 96 × 96 and 48 × 48 to serve as the ground-truth HR faces for C2 and C1, respectively. For the input LR faces, the pre-processed faces are down-sampled to 24 × 24. We train CSRNet using the ADAM optimizer with the default parameters and the batch size is 64. We first train CSRNet without identity branch using 1$e^{-3}$ as the initial learning rate for 47 epoches. Then we continue to include identity branch to fine-tune for another 4 epoch using 1$e^{-4}$. As for CSRGAN training, we use the pre-trained CSRNet model as the initial generator and fine-tune generator and discriminator with the ADAM optimizer for 4 epochs. All experiments are conducted using Tensorflow2 on two NVIDIA Titan RTX GPUs.

## 4.2 | Comparison with state-of-the-art methods

We compare the CSRNet model with 11 state-of-the-art SR models, that is, VDSR, SRGAN, LapSRN, NLSA, SICNN, CSRIP, SuperFAN, RCNet, PCRCN, DIC, and DICGAN. Specifically, we include four models that utilize only texture information, two models that incorporate identity priors (i.e. SICNN [16] and CSRIP [17]), and five models that exploits shape priors (i.e. SuperFAN [13], DIC [15], DICGAN [15], RCNet [40], and PCRCN [39]). Results for the texture-only models (except NLSA) and SICNN are produced from CSRIP, which retrains these models using the same training dataset as ours. For NLSA, SuperFAN, DIC, and DICGAN, we trained SR models on the CASIA dataset with extensive parameter tuning for performance. For RCNet and PCRCN, we implemented SR models as open-source code is not available and trained them under our experiment settings. We also included bicubic interpolation as a naive baseline. In addition, we compare our CSRNet with other cascaded SR models (i.e. LapSRN [31] and CSRIP [17]) with upscaling factors 2× and 4×. Code required to reproduce our experiment results is available at https://github.com/AnonymousExplorer/CSRNet

### 4.2.1 | Quantitative comparison

We report the PSNR and SSIM of CSRNet and the comparison methods for a scaling factor of 8× in Table 2. The results show that CSRNet consistently outperforms all baselines for both test datasets. We also observe that methods use either shape or identify priors (e.g. SuperFAN and CSRIP) perform better than the texture-only methods, which verifies the importance of face priors. By exploiting both shape and identity priors, CSRNet further outperforms methods that use a single type of prior.

**TABLE 3** Quantitative comparison between CSRNet and other progressive methods. **Red**/blue indicates the best/second performance. Our CSRNet model attains highly competitive performance on both datasets.

| | | Helen | | CelebA | |
|---|---|---|---|---|---|
| SR method | Scale | PSNR | SSIM | PSNR | SSIM |
| Bicubic | ×2 | 28.46 | 0.8983 | 27.92 | 0.8891 |
| LapSRN | ×2 | 30.23 | 0.9326 | 29.74 | 0.9262 |
| PCRCN | ×2 | 31.01 | 0.9481 | 30.59 | 0.9438 |
| CSRIP | ×2 | 32.41 | 0.9663 | 31.44 | 0.9609 |
| CSRGAN(Ours) | ×2 | 33.31 | 0.9654 | 32.73 | 0.9612 |
| CSRNet(Ours) | ×2 | 33.88 | 0.9693 | 33.31 | 0.9654 |
| Bicubic | ×4 | 26.32 | 0.7835 | 25.63 | 0.7539 |
| LapSRN | ×4 | 28.30 | 0.8523 | 27.52 | 0.8258 |
| PCRCN | ×4 | 25.01 | 0.8630 | 24.80 | 0.8454 |
| CSRIP | ×4 | 29.56 | 0.8952 | 28.66 | 0.8720 |
| CSRGAN(Ours) | ×4 | 29.62 | 0.8816 | 28.84 | 0.8599 |
| CSRNet (Ours) | ×4 | 30.47 | 0.8983 | 29.71 | 0.8792 |

Among the comparison methods, LapSRN and CSRIP adopt the cascaded structure and thus can generate SR faces with an upscaling factor of 2× and 4×. We compare the quality of the intermediate SR faces of CSRNet with those in Table 3. The results show that CSRNet also outperforms LapSRN and CSRIP in terms of intermediate results. Interestingly, we observe that the performance improvement of CSRNet over CSRIP is more significant for 2× than for 4×. Since CSRIP does not utilize shape priors, this phenomenon suggests that shape priors are important for the first cascade (from 24 × 24 to 48 × 48).

### 4.2.2 | Qualitative comparison

We illustrate some example SR images generated by different methods that are used for quantitative comparison with an upscaling factor of 8× in Figures 9 and 10. The results show that CSRNet produces face images more similar to the ground-truth, especially for important facial regions such as the eyes and mouth. In addition, although the qualitative performance of CSRGAN does not match CSRNet in terms of PSNR and SSIM, we observed that CSRGAN renders more realistic faces than CSRNet by giving more details (e.g. wrinkles and whiskers). This is because CSRGAN is trained to fool the discriminator instead of providing higher PSNR and SSIM. The face that CSRGAN works well suggests that using both shape and identify priors provides sufficient yet complementary information of the generator to produces realistic faces.

For SR comparison of intermediate SR results, we compare the intermediate results of the upscaling factor of 2× and 4× produced by methods that use the cascaded structure in Figures 11 and 12, respectively. These example SR faces show that CSRNets produce more stable intermediate results which are more similar to the true HR face.

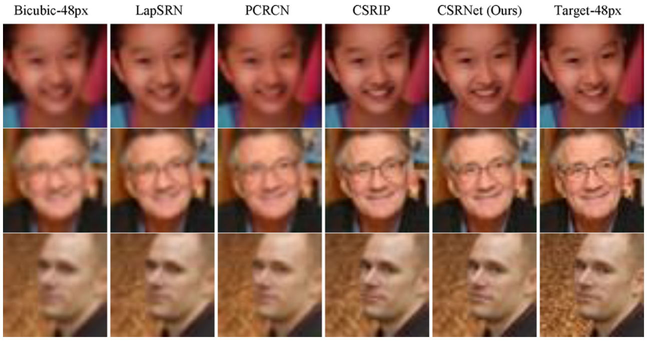


**FIGURE 9** Example face images generated by some representative face SR methods with an upscaling factor of 8× (i.e. from 24 × 24 to 192 × 192). For each sample, four methods including VDSR, SRGAN, LapSRN, and NLSA are texture-only methods that do not use face priors, while SICNN and C-SRIP use identity priors, and SuperFAN, RCNet, PCRCN, DIC, and DICGAN use shape priors. In contrast, our methods utilize both shape and identity priors. CSRNet provides more fine-grained details in key facial areas (e.g. eyes and mouth), resulting in the highest PSNR and SSIM values. CSRGAN further renders the photo-realistic SR faces. The target faces are from the CelebA dataset.

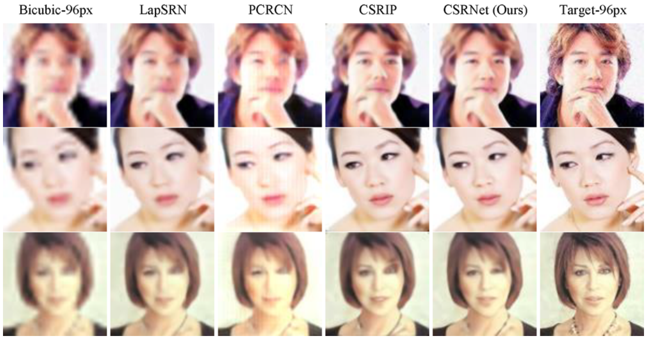


**FIGURE 10** Example face images generated by some representative face SR methods with an upscaling factor of 8× (i.e. from 24 × 24 to 192 × 192). For each sample, four methods including VDSR, SRGAN, LapSRN, and NLSA are texture-only methods that do not use face priors, while SICNN and C-SRIP use identity priors, and SuperFAN, RCNet, PCRCN, DIC, and DICGAN use shape priors. In contrast, our methods utilize both shape and identity priors. CSRNet provides more fine-grained details in key facial areas (e.g. eyes and mouth), resulting in the highest PSNR and SSIM values. CSRGAN further renders the photo-realistic SR faces. The target faces are from the Helen dataset.

### 4.2.3 | More qualitative comparison

We also train CSRGAN (a GAN version of CSRNet) and compare it with other representative GAN-based methods, including SRGAN and DICGAN. To evaluate their effectiveness, we compute LPIPS values on the Helen and CelebA datasets and show some example faces generated by different GAN-based SR methods in Figure 13. We also report the LPIPS values for each method on both datasets. Specifically, on the Helen dataset, we obtained LPIPS values of 0.2639, 0.1581, and

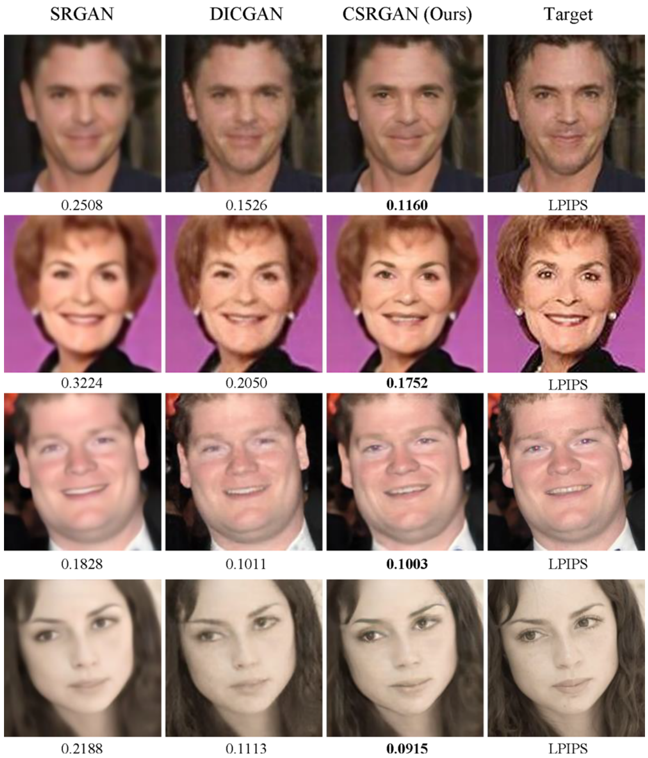


**FIGURE 11** Faces generated by representative cascaded face SR method with an upscaling factor of ×2, from 24 × 24 to 48 × 48.

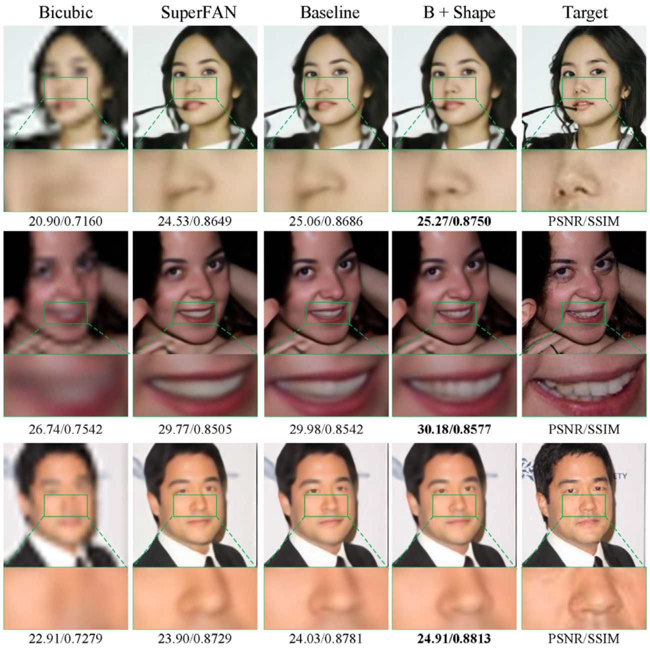


**FIGURE 12** Faces generated by representative cascaded face SR method with an upscaling factor of ×4, from 48 × 48 to 96 × 96.

0.1896 for SRGAN, DICGAN, and CSRGAN, respectively. On the CelebA dataset, we obtained LPIPS values of 0.2958, 0.1673, and 0.2046 for SRGAN, DICGAN, and CSRGAN, respectively. Our results show that both DICGAN and CSRGAN outperform SRGAN in terms of LPIPS. Furthermore, our proposed CSRGAN achieves comparable performance to DICGAN. It is worth noting that improving the perceptual quality of CSRGAN is not the main focus of our paper. Instead, we aim to explore the effectiveness of our proposed CSRNet in combining shape and identity priors.

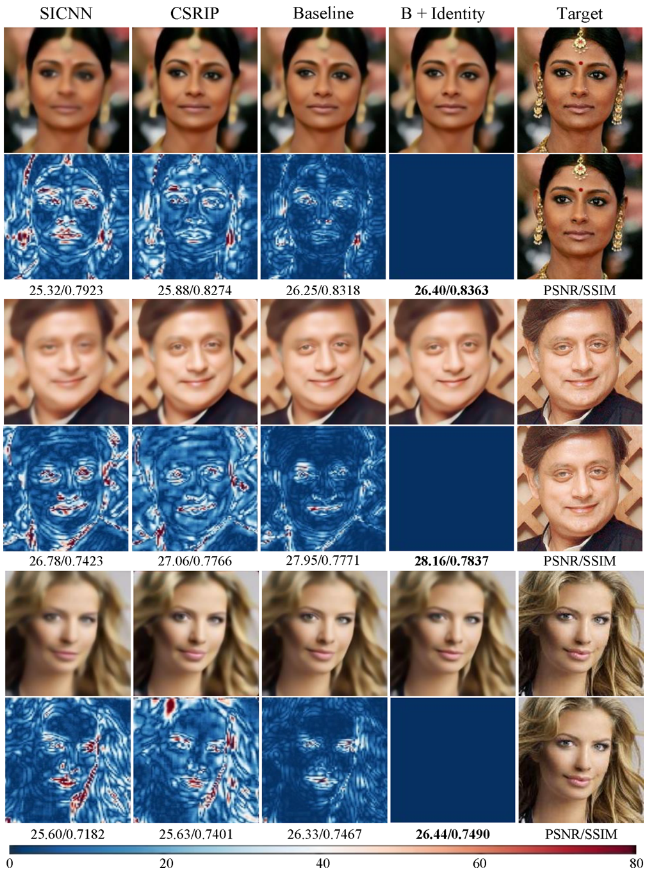


**FIGURE 13** Faces generated by representative GAN-based SR methods with an upscaling factor of 8×, from 24 × 24 to 192 × 192.

**TABLE 4** Ablation study for CSRNet under a scaling factor of 8×, from 24 × 24 to 192 × 192. *Baseline* means using only the SR branch. **Red** indicates the best performance.

| | Helen | | CelebA | |
|---|---|---|---|---|
| **SR method** | **PSNR** | **SSIM** | **PSNR** | **SSIM** |
| Baseline | 28.66 | 0.8141 | 27.82 | 0.7862 |
| B+Shape | 28.69 ↑ | 0.8142 ↑ | 27.84 ↑ | 0.7866 ↑ |
| B+Identity | 28.66 - | 0.8143 ↑ | 27.82 - | 0.7865 ↑ |
| CSRNet (Ours) | 28.71 ↑ | 0.8143 - | 27.86 ↑ | 0.7867 ↑ |

**TABLE 5** Comparison of model size between CSRNet and different baselines. *Baseline* means using only the SR branch.

| **Methods** | **Baseline** | **B+Shape** | **B+Identity** | **CSRNet (Ours)** |
|---|---|---|---|---|
| Parameters | 1.62M | 1.62M | 2.07M | 2.07M |

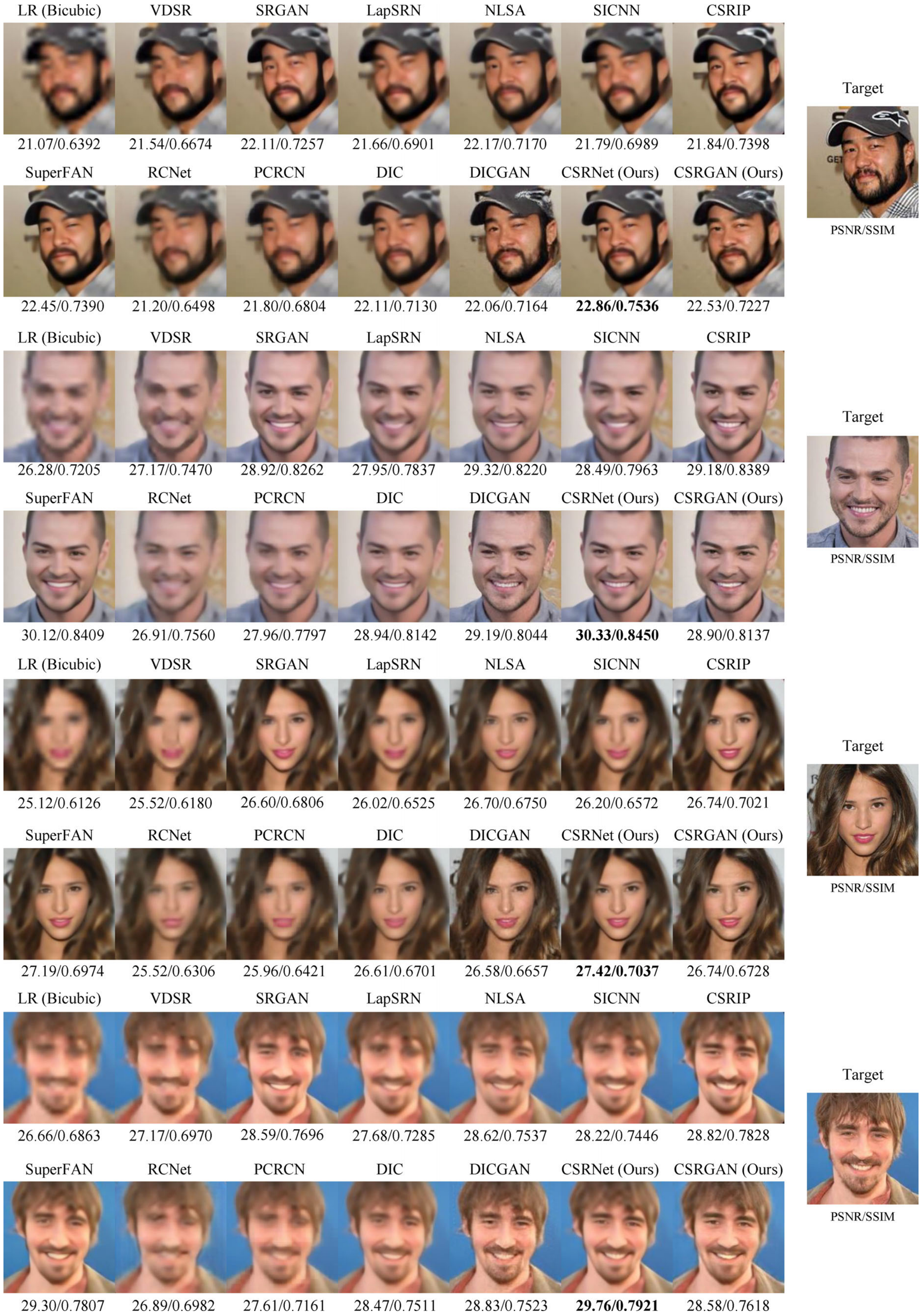


**FIGURE 14** The effect of shape priors on SR faces. "Baseline" uses only the SR branch and "B+Shape" incorporates shape priors. Compare with "Baseline", "B+Shape" produces more fine-grained facial components including the contour of nose (first and third example faces), the shape of teeth (second example face), etc. Compared with other existing shape prior-based SR methods, "B+Shape" provides a better way of utilizing shape priors.

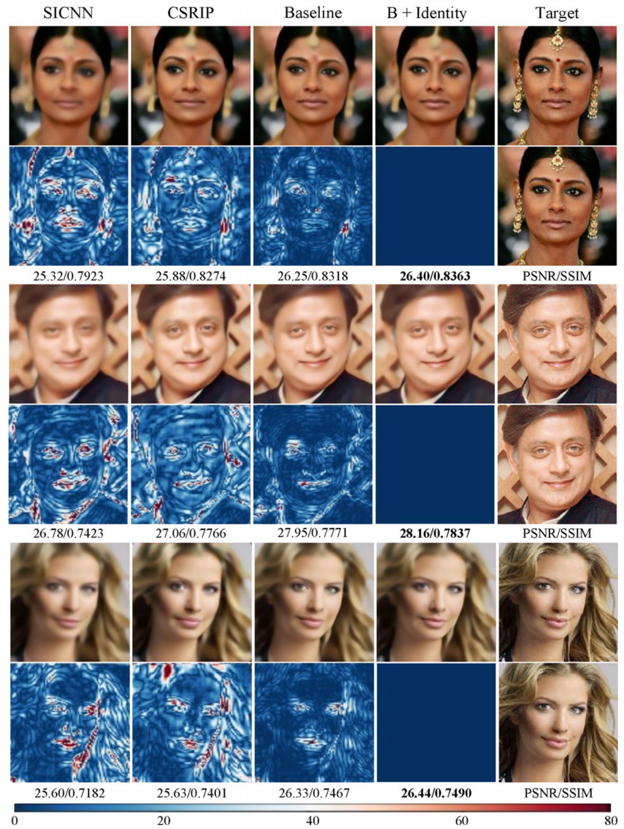


**FIGURE 15** The effect of identity priors on SR faces. Large PSNR/SSIM indicates better quality. "Baseline" uses only the SR branch and "B+Identity" incorporates identity priors. Compare with "Baseline", "B+Identity" alters the details in key facial regions (e.g. eyes, mouth, and face contour). Compared with SICNN and CSRIP, "B+Identity" is more effective in utilizing the identity priors.

## 4.3 | Ablation study

We conduct an ablation study of CSRNet and report the results in Table 4. *Baseline* means using only the SR branch and "+" means enabling different face priors for SR, including *B+Shape* and *B+Identity*. We make the following observations based on comparison results in Table 4. First, using either the shape priors or identity priors yields better performance than the baseline method (only SR branch). Second, for shape priors, heatmap constraints are essential to enhance the quality of the SR face and produce higher PSNR and SSIM. Third, for identity priors, semantic embedding loss (with FaceNet) is necessary, which mainly enhances the semantic information to produce higher SSIM. Most importantly, shape priors and identity priors are complementary, and CSRNet consistently provides the best performance by jointly utilizing them.

### 4.3.1 | Effects of model size

To ensure that the performance improvement of our proposed CSRNet is not simply due to additional parameters, we compare the model size of different baselines in Table 5. *Baseline* and *B+Shape* have the same network structure with 1.62M parameters, while *B+Identity* and *CSRNet* share the same network structure with 2.07M parameters. With the same number of parameters, *B+Shape* outperforms *Baseline*, indicating the effectiveness of utilizing shape priors. More importantly, *CSRNet* outperforms *B+Identity* despite having the same number of parameters, confirming that combining identity and shape priors is crucial to achieving improved face SR results.

### 4.3.2 | Effects of utilizing shape prior

We illustrate the effectiveness of CSRNet in utilizing the shape priors by comparing it with existing methods. Specifically, we compare with SuperFAN as well as our *Baseline* method in Figure 14. The qualitative results demonstrate that our *B+Shape* method (which does not use identity priors) better preserves the shape of the eye and the contour of the nose than *Baseline*. Compared with SuperFAN, *B+Shape* adopts the cascaded structure and extracts shape priors from the residue face. The superior result quality of *B+Shape* suggests that the two designs contribute to performance and provide a better way of utilizing shape priors.

### 4.3.3 | Effects of utilizing identity prior

We illustrate the effectiveness of CSRNet in utilizing the identity priors by comparing with existing methods that use identity priors in Figure 15. We use three example faces to demonstrate the difference between these methods. Specifically, the differences between the SR faces of *B+Identity* and other SR faces are plotted in every second row of example faces, and warmer colour indicates a bigger difference. Compared with SICNN and CSRIP, the differences are large for facial areas that have a big variation in pixel values (e.g. eyes and mouth), and these fast-changing areas correspond to the import facial regions in an image. These differences results show that *B+Identity* (which does not use shape priors) better preserves fine-grained details for facial regions especially the shape of the mouth and the location of the eyeball. Compared with *Baseline*, the variations in pixel value are smaller than other existing methods. However, we can still observe changes in key facial regions like eyes, mouth, and face contour that mainly present the identity and semantic information. In short, the results in Figure 15 indicate that our designs are effective in utilizing the identity priors.

## 5 | CONCLUSIONS

Here, we propose CSRNet, the first deep face super-resolution (SR) model that jointly utilizes shape priors and identity priors. CSRNet adopts a cascaded structure that progressively transforms a low-resolution face image to high resolutions, and forces both the shape and identity priors of the model generated SR face to match their counterparts extracted from the ground-truth high-resolution face. In this way, the intermediate

SR images as well as final SR images are equipped with facial structure information and identity knowledge. The use of complementary information in shape and identity priors via multiple cascades is new and essential to enhance face SR. Extensive experiments on widely used benchmarks including CelebA and Helen demonstrate that our proposed CSRNet outperforms state-of-the-art face SR methods and CSRGAN generates more realistic yet discriminative face images with adversarial and identity losses. Finally, a detailed ablation study indicates that shape and identity priors are complementary in that they constrain the SR face from different aspects.

## AUTHOR CONTRIBUTIONS

Dan Zeng: Conceptualization, formal analysis, investigation, methodology, software, supervision, writing - original draft. Zelin Li: Investigation, software, validation, visualization. Xiao Yan: Conceptualization, writing - review and editing. Wen Jiang: Data curation, investigation, software. Xinshao Wang: Conceptualization, methodology. Jiang Liu: Funding acquisition, supervision. Bo Tang: Funding acquisition, project administration, supervision.

## ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 62206123).

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

CelebA: The data that support the findings of this study are openly available at https://doi.org/10.1109/ICCV.2015.425, reference [47].

Helen: The data that support the findings of this study are openly available at https://doi.org/10.1007/978-3-642-33712-3_49 and https://doi.org/10.1109/CVPR.2013.447, references [48, 49].

## ORCID

*Dan Zeng* https://orcid.org/0000-0002-9036-7791

## REFERENCES

1. Masi, I., Wu, Y., Hassner, T., Natarajan, P.: Deep face recognition: A survey. In: 2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI), pp. 471–478, Brazil, 29 October 2018
2. Adjabi, I., Ouahabi, A., Benzaoui, A., Taleb.Ahmed, A.: Past, present, and future of face recognition: A review. Electronics 9(8), 1188 (2020)
3. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. 44(6), 2872–2893 (2021)
4. He, Z., Zuo, W., Kan, M., Shan, S., Chen, X.: Attgan: Facial attribute editing by only changing what you want. IEEE Trans. Image Process. 28(11), 5464–5478 (2019)
5. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.F., Barriuso, A., et al.: Datasetgan: Efficient labeled data factory with minimal human effort. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10145–10155, Nashville, 19 June 2021
6. 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, Las Vegas, 26 June 2016
7. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. arXiv:190500641 (2019)
8. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. 23(10), 1499–1503 (2016)
9. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision, pp. 483–499, Amsterdam, The Netherlands, 8 October 2016
10. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823, Boston, 7 June 2015
11. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence, vol. 31, California, 4 February 2017
12. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699, California, 16 June 2019
13. Bulat, A., Tzimiropoulos, G.: Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 109–117, Salt Lake City, 18 June 2018
14. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: FSRNnet: 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, Salt Lake City, 18 June 2018
15. 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, 14 June 2020
16. Zhang, K., Zhang, Z., Cheng, C.W., Hsu, W.H., Qiao, Y., Liu, W., et al.: Super-identity convolutional neural network for face hallucination. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 183–198, Munich, Germany, 8 September 2018
17. Grm, K., Scheirer, W.J., Štruc, V.: Face hallucination using cascaded super-resolution and identity priors. IEEE Trans. Image Process. 29, 2150–2165 (2020)
18. Gu, Y., Wang, X., Xie, L., Dong, C., Li, G., Shan, Y., et al.: VQFR: Blind face restoration with vector-quantized dictionary and parallel decoder. arXiv:220506803 (2022)
19. Menon, S., Damian, A., Hu, S., Ravi, N., Rudin, C.: Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2437–2445 (2020)
20. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9168–9178 (2021)
21. Yang, T., Ren, P., Xie, X., Zhang, L.: GAN prior embedded network for blind face restoration in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 672–681 (2021)
22. Zhao, Y., Su, Y.C., Chu, C.T., Li, Y., Renn, M., Zhu, Y., et al.: Rethinking deep face restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7652–7661 (2022)
23. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. 38(2), 295–307 (2015)
24. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: European Conference on Computer Vision, pp. 391–407, Amsterdam, The Netherlands, 8 October 2016
25. Ahn, N., Kang, B., Sohn, K.A.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the

European Conference on Computer Vision (ECCV), pp. 252–268, Munich, Germany, 8 September 2018

26. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690, Hawaii 21 July 2017
27. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: European Conference on Computer Vision, pp. 318–333, Amsterdam, The Netherlands, 8 October 2016
28. Lim, B., Son, S., Kim, H., Nah, S., MuLee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops, pp. 136–144, Hawaii, 21 July 2017
29. Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R.: Deconvolutional networks. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2528–2535, California 13 June 2010
30. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874–1883, Las Vegas, 26 June 2016
31. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep Llaplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 624–632, Hawaii, 22 July 2017
32. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Fast and accurate image super-resolution with deep Laplacian pyramid networks. IEEE Trans. Pattern Anal. Mach. Intell. 41(11), 2599–2613 (2018)
33. Wang, Y., Perazzi, F., McWilliams, B., Sorkine-Hornung, A., Sorkine-Hornung, O., Schroers, C.: A fully progressive approach to single-image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops, pp. 864–873, Salt Lake City, 18 June 2018
34. Kim, D., Kim, M., Kwon, G., Kim, D.S.: Progressive face super-resolution via attention to facial landmark. arXiv:190808239 (2019)
35. Yin, Y., Robinson, J.P., Zhang, Y., Fu, Y.: Joint super-resolution and alignment of tiny faces. arXiv:191108566 (2019)
36. 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, Munich, Germany, 8 September 2018
37. Zhu, S., Liu, S., Loy, C.C., Tang, X.: Deep cascaded bi-network for face hallucination. In: European Conference on Computer Vision, pp. 614–630, Amsterdam, The Netherlands, 8 October 2016
38. Hu, X., Ren, W., LaMaster, J., Cao, X., Li, X., Li, Z., et al.: Face super-resolution guided by 3d facial priors. In: European Conference on Computer Vision, pp. 763–780, Glasgow, 23 August 2020
39. Liu, S., Xiong, C., Shi, X., Gao, Z.: Progressive face super-resolution with cascaded recurrent convolutional network. Neurocomputing 449, 357–367 (2021)
40. Leng, J., Wang, Y.: RCNet: Recurrent collaboration network guided by facial priors for face super-resolution. In: 2022 IEEE International Conference on Multimedia and Expo (ICME), pp. 01–06. IEEE, New York (2022)
41. Cheng, F., Lu, T., Wang, Y., Zhang, Y.: Face super-resolution through dual-identity constraint. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6, Shenzhen, China, 5 July 2021
42. Chen, J., Chen, J., Wang, Z., Liang, C., Lin, C.W.: Identity-aware face super-resolution for low-resolution face recognition. IEEE Signal Process Lett. 27, 645–649 (2020)
43. Kim, J., Li, G., Yun, I., Jung, C., Kim, J.: Edge and identity preserving network for face super-resolution. Neurocomputing 446, 11–22 (2021)
44. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv:14111784 (2014)
45. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: International Conference on Machine Learning, pp. 2642–2651, Sydney, Australia, 7 August 2017
46. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv:14117923 (2014)
47. 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, Santiago, Chile, 7 December 2015
48. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: European Conference on Computer Vision, pp. 679–692, Florence, Italy, 7 October 2012
49. Smith, B.M., Zhang, L., Brandt, J., Lin, Z., Yang, J.: Exemplar-based face parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3484–3491, Portland, OR, 23 June 2013
50. 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)
51. Mei, Y., Fan, Y., Zhou, Y.: Image super-resolution with non-local sparse attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3517–3526 (2021)

**How to cite this article:** Zeng, D., Li, Z., Yan, X., Jiang, W., Wang, X., Liu, J., Tang, B.: Cascaded face super-resolution with shape and identity priors. IET Image Process. 17, 3309–3322 (2023). https://doi.org/10.1049/ipr2.12863