Received: 2 January 2024 | Revised: 27 June 2024 | Accepted: 22 July 2024

IET Image Processing

DOI: 10.1049/ipr2.13195

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Face super resolution with a high frequency highway

Dan Zeng[1] | Wen Jiang[1] | Xiao Yan[1] | Weibao Fu[1] | Qiaomu Shen[1] | Raymond Veldhuis[2] | Bo Tang[1]

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

[2]Faculty of EEMCS, University of Twente, Enschede, The Netherlands

Correspondence
Dan Zeng, Department of Computer Science and Engineering and Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, Shenzhen, China.
Email: zengd@sustech.edu.cn

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

Abstract

Face shape priors such as landmarks, heatmaps, and parsing maps are widely used to improve face super resolution (SR). It is observed that face priors provide locations of high-frequency details in key facial areas such as the eyes and mouth. However, existing methods fail to effectively exploit the high-frequency information by using the priors as either constraints or inputs. This paper proposes a novel high frequency highway (H_2F) framework to better utilize prior information for face SR, which dynamically decomposes the final SR face into a coarse SR face and a high frequency (HF) face. The coarse SR face is reconstructed from a low-resolution face via a texture branch, using only pixel-wise reconstruction loss. Meanwhile, the HF face is directly generated from face priors via an HF branch that employs the proposed inception–hourglass model. As a result, H_2F allows the face priors to have a direct impact on the SR face by adding the outputs of both branches as the final result and provides an extra face editing function. Extensive experiments show that H_2F significantly outperforms state-of-the-art face SR methods, is general for different texture branch models and face priors, and is robust to dataset mismatch and pose variations.

1 | INTRODUCTION

Face images are ubiquitous in many applications [1–3] such as face recognition, person re-identification, and face image editing. However, in scenarios such as portal control, video surveillance, and traffic monitoring, face images often come with low resolution, which degrades the performance of face-related applications and harms the experience of human inspectors. Face super-resolution (SR), also known as face hallucination, recovers high-resolution (HR) face images from low-resolution (LR) ones and is useful in these applications.

Face 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 estimating 64 pixels of SR image from 1 pixel of LR input, which is challenging. In this paper, we aim to super-resolve the input LR face images from 16×16 to 128×128 with a magnification factor of 8.

Many deep learning methods have been proposed for face SR and we provide a detailed discussion on them in Section 2. Early methods, such as SRCNN [4] and VDSR [5], utilize only texture information (i.e. pixel-wise reconstruction loss). As illustrated in Figure 1, their SR results suffer from severe blur and over-smoothing because face SR is an ill-posed problem, and these methods can only fit an average face. To enhance face SR with more information, later methods incorporate shape priors or shape cues (e.g. landmarks, heatmaps, and parsing maps), which describe key facial components such as eyes, nose, mouth, and contour. Some methods [6, 7] use face priors as constraints, e.g. SuperFAN [6] trains the SR model by enforcing the SR face to produce face priors similar to the ground-truth HR face. Other methods [8, 9] use face priors as model inputs, e.g. DIC [9] fuses heatmaps into the indeterminate features by an attentive fusion module, which is then processed by convolution operations. Using face priors, both SuperFAN and DIC provide more visually plausible SR images than the texture-only methods as shown in Figure 1.

Dan Zeng and Wen Jiang contributed equally to this study.

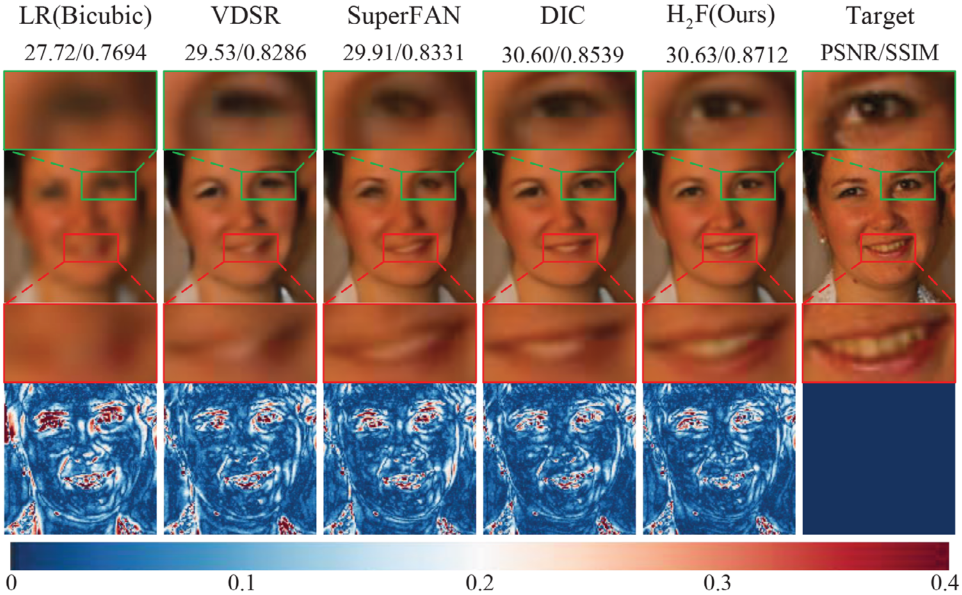


FIGURE 1 SR faces produced by some representative methods. Larger PSNR/SSIM indicates better image quality. The bottom row highlights the differences between the SR faces and the target HR in luminance, which shows that H_2F has the smallest difference in the eyes and mouth (i.e. fast-changing areas). The zoom-in figures show that H_2F better preserves details in the eyes and mouth in its SR face.

We take a signal processing perspective to explain the contribution of face priors in face SR. The differences between the SR faces and the target HR faces are plotted in the bottom row of Figure 1, where a warmer colour indicates a bigger difference. For the texture-only methods, 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 high frequency (HF) signals[1] in an image. Compared with the texture-only methods, SuperFAN and DIC reduce differences in the fast-changing areas and thus face priors improve SR by providing HF signals. However, existing methods use the face priors in an indirect manner, which may not be very effective. The face priors can only guide gradients in model training (when used as constraints) or serve as some feature maps in convolution operations that generate the final SR image (when used as inputs). In these methods, the HF information related to shape priors is prone to phase out after a series of convolution operations, resulting in inferior SR results. Additionally, the tight coupling of shape priors with the LR input for learning the final SR face makes it challenging to understand the root cause of the gain in HF information.

As shown in Figure 1, HF signal loss is inevitable in face SR, regardless of the strength of the SR model, due to the ill-ness nature of the task. However, incorporating shape priors into the SR model can mitigate such loss, as the solution space is constrained by facial knowledge. To better leverage the prior information, we propose to dynamically decompose a final SR face into a coarse SR face and an HF face. We expect that an HF face can be directly generated from face priors (i.e. in a direct manner) to facilitate the understanding of HF information gain and prevent possible smoothing of HF details during learning. To achieve this goal, we seek to answer two research questions. ❶ Is it possible for the model to learn HF information from shape priors? ❷ In the absence of the ground-truth HF face data, how can the model be effectively supervised during training?

The answer to the first question ❶ is affirmative. Face shape priors contain information (e.g. locations) about high-frequency details such as edges and boundaries of facial components. These details reflect the colours and textures that can be easily found in LR input; for example, the colour of the teeth is typically white, and the colour of the mouth is red. In other words, shape priors are closely related to missing HF signals, enabling the model to learn HF information from them. In this paper, we use face priors in a more direct manner by generating HF faces from these priors. It is worth noting that we focus on HF information in the image domain rather than the frequency domain to enhance the image sharpness. One rationale behind this choice is that we prefer a more intuitive and easy way to help understand the HF information gain of face SR.

The quick answer to the second question ❷ involves employing a dual reconstruction learning mechanism and leveraging the correlation between shape priors and missing HF signals. Specifically, to address the challenge of missing ground-truth HF face, we formulate a novel high frequency highway (H_2F) framework for face SR. This framework consists of two parallel branches: a texture branch and a high frequency branch. The texture branch recovers a coarse SR face from the LR image using pixel-wise reconstruction loss, while the high frequency branch extracts face priors and then reconstructs an HF face from the priors using the proposed inception–hourglasses model. The outputs of both branches are simply added to generate the final SR face. Instead of explicitly supervising the high frequency branch with ground-truth HF face, we employ

[1] In signal processing, high-frequency signals change rapidly over space or time.

additional pixel-wise reconstruction loss to enforce the final SR face to be similar to the ground-truth HR face. This design allows the high frequency branch to implicitly target the missing parts in the texture branch during learning, thereby eliminating the need for ground-truth HF face. As demonstrated in our experimental results in Section 4, H_2F employs dual reconstruction supervision for both the texture branch and the final output, which not only facilitates the learning of high-frequency signals from shape priors but also significantly enhances the SR learning process.

The H_2F framework provides a better way to utilize prior information for face SR. Figure 1 shows that H_2F preserves more details in the SR face and introduces fewer variations in the fast-changing areas compared to other methods. Interestingly, H_2F also supports explainable and intuitive SR faces by manipulating the face priors to edit the final SR face. One of the key advantages of H_2F is its versatility and flexibility. The framework allows existing methods such as VDSR [5] and SRResNet [10] as the texture branch model and any hourglasses-based structure as the high frequency branch. Its straightforward structure also makes H_2F easy to train. More importantly, H_2F can generate higher-quality SR faces than existing methods and is robust to cross-dataset and pose variations.

In summary, we made the following contributions to this paper:

- We explain the role of face priors in face SR as providing high-frequency signals, which motivates H_2F, the first to construct an HF face directly from face priors via a high frequency highway. H_2F can produce explainable results that edit the final SR face by manipulating the face priors, thereby simplifying the understanding of HF information gain.
- Following an in-depth exploration of two questions, we propose an innovative framework, H_2F, indeed an attempt to better utilize prior information for face SR. With its dual-reconstruction learning mechanism, H_2F dynamically decomposes a final SR face into a coarse SR face and an HF face, which effectively prevents the smoothing of HF details during learning.
- We conduct extensive experiments to evaluate the performance of H_2F and show that H_2F outperforms state-of-the-art methods in the quality of SR images and is robust to dataset mismatch and pose variations. Furthermore, H_2F is general, supporting many existing SR models as the texture branch model, and is easy to train.

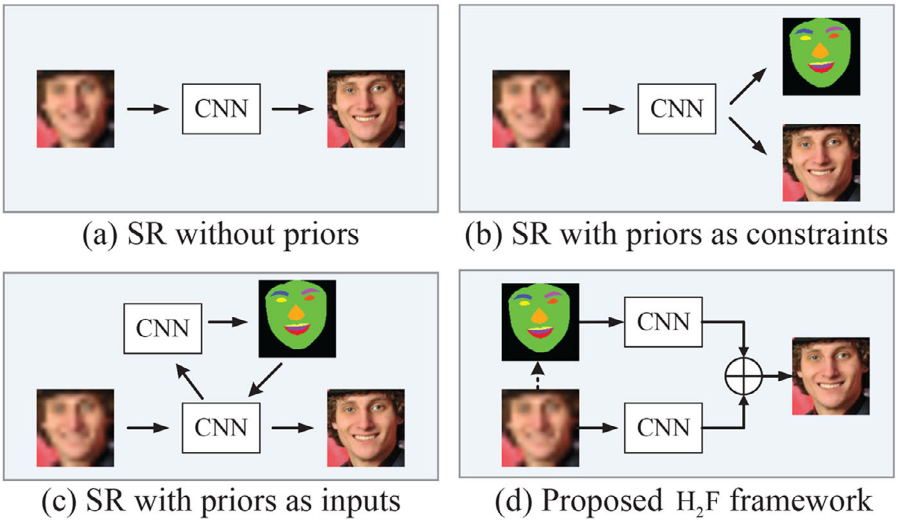
2 | RELATED WORK

In this section, we review related work in the deep face SR framework and emphasize our key idea of generating high-frequency faces directly from the shape priors, which sets it apart from existing methods. It is worth noting that some recent methods enhance face SR by using reference image priors, such as VQFR [11], or by importing pre-trained generative adversarial network (GAN) priors such as DebiasFR [12]. Additionally, some methods consider multiple priors for face SR [13, 14]. Specifically, GAN priors [15] and denosing diffusion probabilistic models (DDPM) [16], which gradually became the mainstream priors used in the SR models. PULSE [17] spearheaded the use of pre-trained GAN priors for face SR, followed by methods like GFPGAN [18] and GPEN [19], which introduce encoder–generator architectures for end-to-end restoration. Recent advancements include VQFR [11] and Restoreformer [20], which leveraged the vector quantization codebook from VQGAN [21] to harness generative priors better. Codeformer [22] focuses on enhancing codebook query accuracy. There are also works exploring diffusion generative priors for face SR. DifFace [23] capitalizes on the diffusion process of DDPM for degradation removal. PGDiff [24] employs facial semantics guidance at each step of reverse diffusion, facilitating effective degradation removal while preserving semantic content.

However, our paper primarily focuses on exploring the more effective use of shape priors for face SR, rather than comparing various prior knowledge applications.

2.1 | Deep SR without priors

Many methods utilize only the texture information (i.e. pixel-wise reconstruction) for natural image super-resolution and can be directly applied to face image super-resolution, as illustrated in Figure 2a. SRCNN [4] is the first work that learns a convolution neural network (CNN) model to map interpolated LR images to HR images. VDSR [5] increases network depth to 20 layers using a global residual connection. RCAN [25] adopts a 'residual in residual' structure with long skip connections to form a very deep network. In a similar vein, RDN [26] makes heavy use of residual connections and dense connections to increase network depth to retain high-frequency details in super-resolved images. HAN [27] improves face SR by modelling both intra-layer and inter-layer feature dependencies through an attention mechanism. ClassSE [28] accelerates SR networks by applying different class-module and SR-module tailored to the restoration difficulty of image patches. AdderSR [29] introduces an energy-efficient SR adder network, utilizing additions instead of multiplications for output features calculation. LESRCNN [30], a lightweight enhanced SR method, employs a heterogeneous architecture in its feature extraction module to reduce both the parameter count and training complexity. LKASR [31] uses large kernel attention in lightweight image SR, designing a module that combines the advantages of CNN and transformer for better feature extraction while maintaining the total parameters. Methods such as URDGN [32] and SRGAN [10] adopt a generator-discriminator framework to generate photo-realistic and visually pleasing SR images. SPARNet [33] improves the vanilla residual blocks by injecting the proposed face attention units (FAUs) to emphasize key face components. Repeating FAUs in the network architecture, SPARNet can further improve performance due to spatial attention learning. CTCNet [34] is a hybrid CNN-transformer architecture for face SR that reduces network feature



(a) SR without priors

(b) SR with priors as constraints

(c) SR with priors as inputs

(d) Proposed H_2F framework

FIGURE 2 A comparison of different deep face SR frameworks.

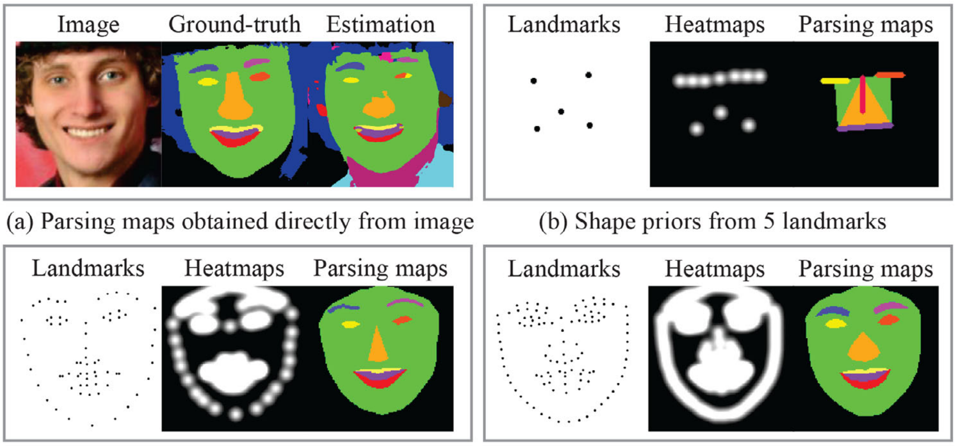
redundancy with an efficient feature enhancement unit. It also uses a feature refinement module to strengthen different face structure information and enhance the extracted features. SFMNet [35] introduces the Fourier transformer to build a spatial-frequency mutual network for face SR. The frequency branch leverages the Fourier transformer to achieve an image-size receptive field and captures the global facial structure. The spatial branch focuses on extracting local facial features. In a similar vein, SFNet [36] conducted frequency decomposition and recalibration with different receptive fields by the proposed MDSF and MCSF modules to enhance image restoration. The MDSF module constructs dynamic filters to decompose feature maps into various frequency components and emphasize them using channel attention. The MCSF aims to enlarge the receptive field and conduct frequency selection. In short, deep SR methods without priors focus on proposing effective network architecture to enhance the image SR results. However, without special consideration of face priors, these methods cannot achieve the optimal results for face SR.

2.2 | Face shape priors

Commonly utilized face-related structure and shape priors, in ascending order of granularity, facial landmarks, heatmaps, and parsing maps. We provide an illustration of them in Figure 3. These priors provide rich information about facial geometry including global structure (e.g. facial contour) and local details (e.g. eyes, nose, and mouth). Some face alignment networks (FANs) are designed for facial landmark detection, such as OpenFace [37] and Retinaface [38]. Given facial landmarks, heatmaps and parsing maps can be generated using predefined templates or pre-trained models such as GFC [39]. Additionally, parsing maps can be generated directly from face images directly using segmentation networks such as face-parsing [40, 41] and MaskGAN [42]. As face SR is inherently an ill-posed problem, many methods use face priors to complement pixel-level RGB cues. We categorize these methods based on the utilization of face priors, i.e. either as constraints or inputs, and provide a schematic illustration of them in Figures 2b and 2c, respectively.

2.3 | Face SR with priors as constraints

These works typically conduct face SR and prior estimation in a multi-task learning framework, allowing the priors to provide supervision for the SR model. SuperFAN [6] introduces a face alignment network to guarantee the consistency of face-related structures (i.e. facial heatmaps) between the target HR face and model-generated SR face. PFSR [43] extends SuperFAN by progressively super-resolving LR face images to higher resolutions. In JASRNet [7], the facial heatmaps estimation network (i.e. FAN) and the SR network share a common encoder, which extracts shallow features from the LR face images. VSR-DUF [44] shares a similar design philosophy for video SR. In this model, the filter generation (FG) branch and the residual generation (RG) branch share a common network for feature extraction from multiple neighbouring frames. The output of the FG branch is first applied to the input LR frame to obtain a filtered face, and then the output of the RG branch is added to produce the SR frame. CompositeNet [45] seamlessly integrates the advantages of CNNs and transformers and leverages multi-order head attention face priors for face SR. It consists of two parallel networks (i.e. the FHT-based network and the MOHA-based CNN) and the PC module. In particular, the proposed MOHA-based CNN captures both spatial and channel dependencies of facial priors and 2D information of face images to enhance recovery performance. Face SR methods that use priors as constraints can significantly improve the accuracy of key components of SR face. However, these methods implicitly introduce the priors as gradient constraints, which makes it difficult to explain the SR face and determine whether the priors are used appropriately.



(a) Parsing maps obtained directly from image

(b) Shape priors from 5 landmarks

(c) Shape priors from 68 landmarks

(d) Shape priors from 106 landmarks

FIGURE 3 An illustration of the face priors generated by different methods. For (a), ground-truth is generated by manual labelling and estimation is produced by MaskGAN. For (b)–(d), the priors are generated using different numbers of landmarks.

2.4 | Face SR with priors as inputs

These works integrate the face priors as input features of the face SR model. MTUN [46] concatenates component-based heatmaps with other input feature maps and feeds them into the SR model. FSRNet [8] 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 (along with other feature maps) into an encoder-decoder network to generate the final SR image. CBN [47] progressively super-resolves the LR face through multiple stages, using a gate network to fuse the coarse SR face with the dense correspondence filed (a type of face prior) at each stage. CBN extracts face priors from intermediate SR faces instead of the input LR face to improve accuracy. DIC [9] learns an attentive fusion module that uses the prior as attention weights to aggregate feature representations in the SR model. Hu et al. [48] concatenates 3D facial priors to feature representations of the SR model to produce a sharp face and address pose variations in LR faces. DFDNet [49] explicitly generates deep face components dictionaries from HR images using K-means clustering and then incorporates these dictionaries into the SR model to guide degraded face restoration. DCLNet [50] adopts a cascaded structure to transform an LR face to an HR face and estimate landmarks progressively across multiple stages. At each stage, it jointly learns the face degradation and degraded landmarks estimation tasks, providing additional priors as inputs to enhance the LR-HR feature mapping. FSRCH [51] embeds face priors into LR images for face SR, extracting these face priors from the original HR images, rather than LR ones, to ensure accuracy. However, it requires HR images during testing, which is not available in our application. Although the comparison is unfair to us, we include their results for comparison to demonstrate the effectiveness of our method. These face SR methods that utilize priors as inputs can have a more direct guide in feature representation and generally produce better SR faces with finer texture details corresponding to the priors.

Compared with existing methods, our H_2F framework utilizes the face priors in a more direct manner, explicitly learning a detailed image (also termed as HF face) from these face priors, aligning well with Occam's razor principle. As illustrated in Figures 2d and 4, the HF face is generated from the face priors and then combined with a coarse SR image to produce the final SR image. The HF branch and the texture branch simultaneously process the same LR input, ensuring proper alignment for the key facial components. In this way, the H_2F framework is more effective in preserving the high-frequency details present in the face priors compared to existing methods.

3 | A HIGH FREQUENCY HIGHWAY FRAMEWORK

In this part, we first provide an overview of the H_2F framework and then introduce our hourglasses-based high frequency branch, which reconstructs high-frequency face images directly from face priors. Additionally, for a comprehensive approach, we incorporate an adversarial loss to supervise the training of our model. This results in a GAN variant of H_2F, termed as H_2FGAN.

3.1 | H_2F framework overview

As illustrated in Figure 4, the H_2F framework consists of a texture branch network ($\mathcal{T}$) and a high frequency branch network ($\mathcal{H}$). The outputs of both branches have the same dimensionality (i.e. spatial size and channel number) as the required SR image and are fused via simple addition to produce the final result. The texture branch generates a coarse SR face using only the texture information, while the high frequency branch provides a high frequency face that captures details in rapidly changing facial areas such as the eyes and mouth. By

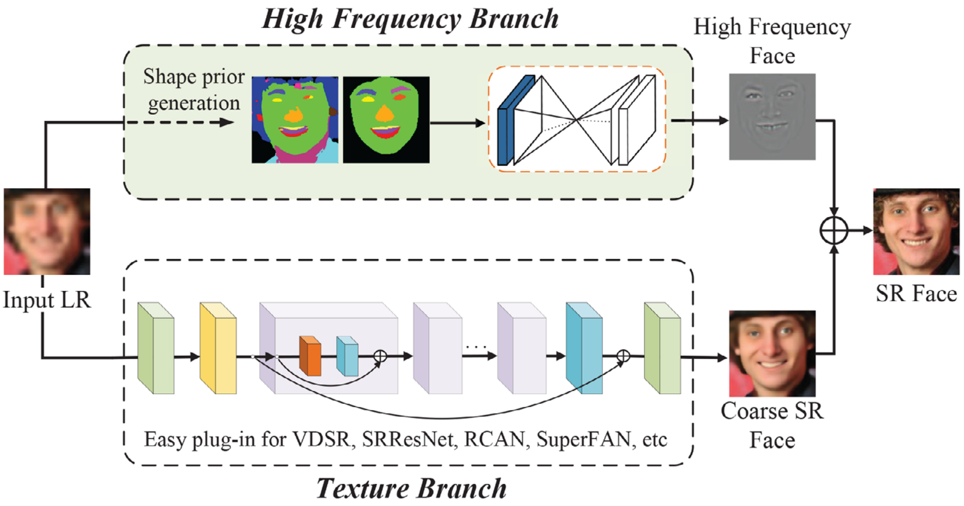


FIGURE 4 An illustration of the proposed H$_2$F framework, which dynamically decomposes the final SR face into a coarse SR face and a high frequency (HF) face. The framework is general in supporting many existing SR models as an easy plug-in in the texture branch. Utilizing a dual reconstruction learning mechanism and the inherent correlation between shape priors and HF information, HF faces can be effectively generated from the shape priors.

design, the two branches complement each other in the face SR task.

For the texture branch, the H$_2$F framework can use existing texture-only SR models such as VDSR [5], SRResNet [10], and RCAN [25]. In the experiments, we observed that applying a stronger SR model in the texture branch leads to better performance for the entire H$_2$F framework, as detailed in Section 4.4. The modular structure of H$_2$F can not only easily plug into existing SR models, but also enjoy the performance benefits of more advanced SR models in the future. Denote an input LR face image as $\mathbf{x}$, we express $\mathbf{y}_\mathrm{t}$ as the output of the texture branch $\mathcal{T}$,

$$\mathbf{y}_\mathrm{t} = \mathcal{T}(\mathbf{x}). \tag{1}$$

The high frequency branch first extracts 2D face shape priors (i.e. landmarks, parsing maps, and heatmaps) from the input LR image and then reconstructs the high frequency face image using these priors. Hence, the output of the high frequency branch can be expressed as

$$\mathbf{y}_\mathrm{h} = \mathcal{H}(\mathcal{P}(\mathbf{x})), \tag{2}$$

where $\mathcal{P}$ and $\mathcal{H}$ are the prior extraction network and high frequency face image generation network, respectively. $\mathbf{y}_\mathrm{h}$ is the output high frequency face. The final output of the H$_2$F framework $\tilde{\mathbf{y}}$ is the pixel-wise summation of the two branches

$$\tilde{\mathbf{y}} = \mathbf{y}_\mathrm{t} + \mathbf{y}_\mathrm{h}. \tag{3}$$

Given a training dataset $\mathcal{S} = \left\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right\}_{i=1}^{N}$ containing N samples, where $\mathbf{y}^{(i)}$ is the ground-truth HR image of LR image $\mathbf{x}^{(i)}$. With the dual-reconstruction learning mechanism, H$_2$F dynamically decomposes a final SR face into a coarse SR face and an HF face, which effectively prevents the smoothing of HF details during learning. The dual reconstruction loss function of the H$_2$F framework is defined as:

$$\mathcal{L}^{\mathrm{H_2F}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left\{\left\|\mathbf{y}^{(i)} - \tilde{\mathbf{y}}^{(i)}\right\|^2 + \alpha\left\|\mathbf{y}^{(i)} - \mathbf{y}_t^{(i)}\right\|^2\right\}, \tag{4}$$

where θ are the parameters for the H$_2$F framework, which contains both the texture and the high frequency branch networks. $\mathcal{L}^{\mathrm{H_2F}}(\theta)$ employs dual reconstruction supervision to enhance the SR learning which combines the pixel-wise reconstruction loss of the texture branch with that of the entire framework, and α is a positive weight term for the loss of the texture branch. $\tilde{\mathbf{y}}^{(i)}$ represents the final output and $\mathbf{y}_t^{(i)}$ represents the coarse SR face of the texture branch. We empirically set the value of α to 1, given that the final SR is obtained by simply adding the outputs of two branches. Slightly tuning with this hyperparameter may further enhance SR results. To address the absence of ground-truth HF face, we employ the term $\left\|\mathbf{y}^{(i)} - \tilde{\mathbf{y}}^{(i)}\right\|^2$ to ensure the outputs of the high frequency branch complement the texture branch outputs. This facilitates the learning of missing details in the texture branch by the high frequency branch. The entire H$_2$F framework can be trained end-to-end.

3.2 | High frequency branch

As illustrated in Figure 5, the high frequency branch consists of a shape prior generation module ($\mathcal{P}$) and a high frequency face image generation module ($\mathcal{H}$), which are detailed as follows.

3.2.1 | Shape prior generation

We first resize the LR face to the desired resolution via Bicubic interpolation. Then, we use two different methods to

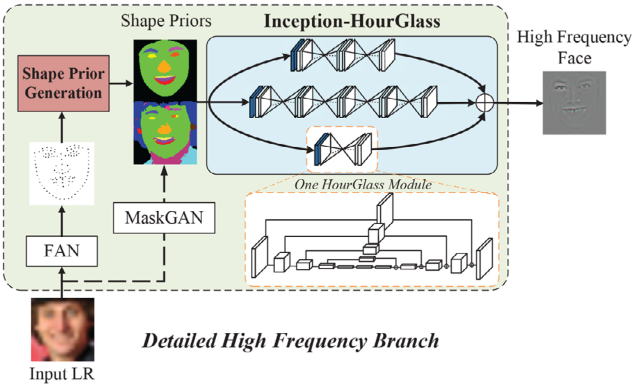


FIGURE 5 An illustration of high frequency branch, which consists of shape prior generation and high frequency face generation. It fills in texture according to the priors preserved in the fine-grained characteristic of input LR face (e.g. the shape of the eyes, edges of the mouth, and the face contour) that are missing in the texture branch. The proposed inception–hourglass is used for high frequency face generation.

generate the face priors. The first is direct generation by applying segmentation networks such as face-parsing [40, 52] and MaskGAN [42] to the interpolated LR image. The second is indirect generation via landmarks, which we elaborate on as follows.

We apply a pre-trained face alignment network (FAN) such as RetinaFace [38] and OpenFace [37] to detect the facial landmarks on the interpolated LR face image. Specifically, we utilize 3 sets of landmarks containing 5, 68, and 106 points, respectively. RetinaFace is used to detect 5 landmarks and 106 landmarks, while OpenFace is used to detect 68 landmarks. In the experiments, we observed that using more landmarks (i.e. providing better structural information in the priors) leads to better performance. We use two different schemes to generate shape priors based on sparse landmarks (i.e. 5) and dense landmarks (i.e. 68, 106). As for 68 and 106 landmarks, we group them into 9 categories according to semantic meanings: left eyebrow, right eyebrow, left eye, right eye, nose, upper lip, lower lip, inner mouth, and facial contour. Each group's landmarks are connected to form an enclosed area. Adjustments such as landmarks interpolation and incorporation of average face knowledge are made to obtain the complete yet coarse overall facial structure (i.e. coarse shape cue). With 5 landmarks, we generate 6 parsing maps corresponding to the left eye, right eye, nasal bridge, mouth, nose region, and tight face region, respectively, as shown in Figure 6. Specifically, for each of the 5 landmarks that corresponds to a facial component (e.g. marked in red), we interpolated an additional landmark (e.g. marked in blue) to fit an average face. The same landmark grouping rule is also used when generating the heatmaps.

Figure 3 illustrates the face priors generated by different methods. In the experiments, we found that generating priors directly through segmentation networks yields superior performance compared to using landmarks. However, employing a varying number of landmarks to generate priors with various granularity offered valuable insights into how the quality of priors affects the performance of H_2F.

3.2.2 | High frequency face generation

HF faces can be effectively generated from the shape priors, leveraging a dual reconstruction learning mechanism and the inherent correlation between shape priors and HF information. The high frequency image generation module ($\mathcal{H}$), taking face priors as inputs, produces face images with dimensions identical to the target HR image. The hourglass network [53] was first proposed for human pose estimation to capture various spatial relationships associated with the body. Subsequently, hourglass has been adapted for facial landmark estimation in several studies [8, 9, 46], essentially predicting face shape priors for a face image. We employ the hourglass network in a novel manner: using the face shape priors as inputs and tasking the network with generating images. This is motivated by the observation that the face shape priors inherently contain high-frequency information (e.g. edges, contours) in critical facial components like eyes and mouth. Generating an HF image directly from the face shape priors is more effective in preserving high-frequency information than existing methods, which typically use the face priors as constraints or intermediate inputs. The hourglass network is an encoder-decoder CNN structure with the same spatial size for both input and output and thus can be trained to approximate the difference between the ground-truth HF face and texture branch output with proper supervision.

We adopt an inception–hourglass structure for high frequency face generation. The inception architecture was first proposed in GoogLeNet [54], which was successful in improving the performance of CNNs. As shown in Figure 5, the inception–hourglass structure contains three branches and from bottom to top, the branches stack 1, 4, and 2 hourglass networks, respectively. Our experiments demonstrate that both the number of branches and the number of hourglass networks in each branch can be adjusted. As illustrated in the lower right corner of Figure 5, the hourglass network uses a skip connection between symmetrical layers so that it can preserve spatial information at different scales. It's important to note that the Inception-hourglass network, although effective, is not tied to the core concept of our HF branch and could be substituted with alternative architectures such as U-Net [55].

3.3 | GAN version of H_2F

It is common to apply adversarial loss to further enhance the perceptual quality of SR image as GAN [8–10] has been widely used to generate photo-realistic and visually pleasing images. To complement our method, we introduce H_2FGAN, a GAN variant of our H_2F model for face SR. Specifically, we build a discriminator D to distinguish the super-resolved face image and ground-truth HR image [9] by minimizing

$$\mathcal{L}^{\mathrm{dis}} = -\mathbb{E}[\log D(\mathbf{y})] - \mathbb{E}[\log(1 - D(G(\mathbf{x})))], \tag{5}$$

where $\mathbb{E}$ is the expectation over a probability distribution.

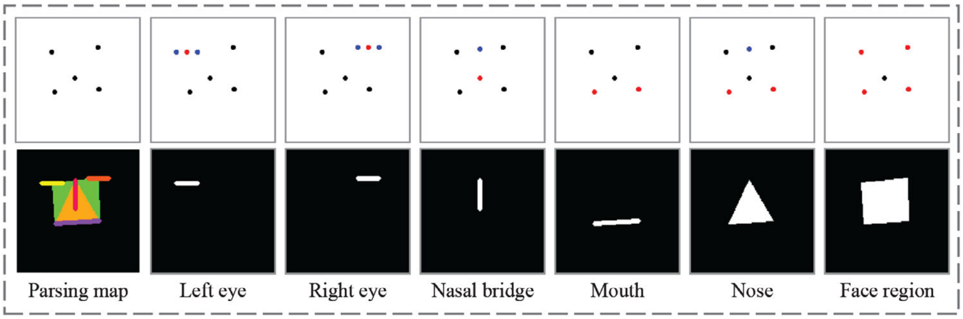


FIGURE 6 An illustration of generating 6 parsing map from 5 landmarks.

Meanwhile, our generator network (i.e. H_2F) is encouraged to deceive the discriminator D by using adversarial loss [9]:

$$\mathcal{L}^{\text{adv}} = -\mathbb{E}[\log D(G(\mathbf{x}))] \tag{6}$$

Additionally, we incorporate perceptual loss $\mathcal{L}^{\text{perc}}$ to enhance the perceptual quality of SR face images. Specifically, we use a pre-trained VGG16 [56] as a feature extractor $f(\cdot)$ for images, and we aim to minimize the Euclidean distance between features of ground-truth face image $f(\mathbf{y})$ and those of super-resolved image $f(\tilde{\mathbf{y}})$:

$$\mathcal{L}^{\text{perc}} = \left\| f(\mathbf{y}) - f(\tilde{\mathbf{y}}) \right\|^2 \tag{7}$$

Finally, the discriminator is optimized by minimizing $\mathcal{L}^{\text{dis}}$ in Equation (5), and the generator is optimized by minimizing the overall objective function:

$$\mathcal{L} = \mathcal{L}^{\text{H}_2\text{F}} + \lambda_{\text{adv}} \cdot \mathcal{L}^{\text{adv}} + \lambda_{\text{perc}} \cdot \mathcal{L}^{\text{perc}} \tag{8}$$

where λ_{adv} and λ_{perc} are trade-off weight terms for the adversarial loss and perceptual loss, respectively. $\mathcal{L}^{\text{H}_2\text{F}}$ is our proposed H_2F loss detailed in Equation (4). The overall objective functions are used to train photo-realistic H_2FGAN model while the PSNR-oriented model H_2F is trained solely with $\mathcal{L}^{\text{H}_2\text{F}}$. It is worth noting that enhancing the perceptual quality of H_2FGAN is not the main focus of our paper. Our primary focus is on exploring the efficacy of the proposed H_2F framework.

ALGORITHM 1 Training procedure of H_2F.

```
Input: Training dataset $\mathcal{S}$, the face prior extraction model $\mathcal{P}$;
Positive weight term for the loss $\alpha$;
Outputs: $H_2F$ network weights $\theta$
Initialize: Texture branch network $\mathcal{T}$ and high frequency branch network $\mathcal{H}$
for $t = 0 \ldots T-1$ do
    $\mathbf{x}, \mathbf{y} \leftarrow$ SampleMiniBatch($\mathcal{S}$)
    Extract high-quality face priors by using off-the-shelf $\mathcal{P}$:
    $\mathcal{P}(\mathbf{x}) \leftarrow$ Forward($\mathbf{x}, \mathcal{P}$)
    Obtain the output of high frequency branch:
    $\mathbf{y}_\text{h} \leftarrow$ Forward($\mathcal{P}(\mathbf{x}), \mathcal{H}$)
    Obtain the output of texture branch:
    $\mathbf{y}_\text{t} \leftarrow$ Forward($\mathbf{x}, \mathcal{T}$)
    Obtain the final SR results of $H_2F$ network:
    $\tilde{\mathbf{y}} \leftarrow \mathbf{y}_\text{t} + \mathbf{y}_\text{h}$
    Compute the loss $\mathcal{L}^{\text{H}_2\text{F}}$ according to Equation (4),
    $\mathcal{L}^{\text{H}_2\text{F}} \leftarrow \|\mathbf{y} - \tilde{\mathbf{y}}\|^2 + \alpha \|\mathbf{y} - \mathbf{y}_\text{t}\|^2$
    Update the $H_2F$ network.
end for
return $\theta$
```

3.4 | Algorithm for H_2F

For the GAN version of H_2F (aka H_2FGAN), we utilize loss $\mathcal{L}$ in Equation (8) to guide the generator learning and use $\mathcal{L}^{\text{dis}}$ in Equation (5) to guide the discriminator learning to improve the perceptual quality. For the proposed H_2F framework, we use $\mathcal{L}^{\text{H}_2\text{F}}$ introduced in Equation (4) to train the SR network.

We list detailed step-by-step pseudo-code for H_2F in Algorithm 1. Given training dataset and off-the-shelf face prior extraction model, the high-quality face priors are first extracted and then used as the input of the high frequency branch to obtain the high frequency face. The texture branch generates a coarse SR face using only the texture information. The final output of the framework is the pixel-wise summation of the two branches. For loss calculation, both coarse face and final SR face are used as the dual construction supervision to enhance the SR learning.

4 | EXPERIMENTAL EVALUATION

In this section, we first introduce the experiment settings and then present the main results, comparing the SR performance of H_2F with state-of-the-art baselines. Furthermore, we demonstrate that H_2F provides an extra face editing function, allowing the face priors to have a direct impact on the SR face. Finally, we provide experimental results that highlight the merits and insights (such as flexibility and effectiveness) of the H_2F

framework. Code required to reproduce our experiment results is available at https://github.com/danzeng1990/H2F-FaceSR.

4.1 | Experiment settings

4.1.1 | Datasets

We mainly conduct our experiments on two widely used benchmark datasets for face SR, i.e. CelebA [57] and Helen [40, 58]. For CelebA, we use 168,854 images for training and 1,000 images for testing. For Helen, we use 2,005 images for training and 50 images for testing. We adopt the same train/test split as used in DIC [9], ensuring consistency across all methods.

4.1.2 | Performance metrics

We are mainly concerned with the quality of the model-generated SR images. For quantitative evaluation, we adopt PSNR and SSIM [59], two widely used performance measures in SR literature. Following convention, the SR images are converted from the RGB space to the YCrCb space, and only the illuminance channel is used to calculate the two measures. For PSNR and SSIM, a larger value means better performance. Moreover, we include the result of face alignment as an additional performance metric. Retinaface [38] is used to detect 106 landmarks on the SR faces. We report the detection rate, which is the percentage of SR faces that pass landmark detection. For the detected SR faces, the normalized root mean squared error (NRMSE) is calculated w.r.t. landmarks on the ground-truth HR faces. We normalize NRMSE by the height of the image and note that higher detection rate and smaller NRMSE mean better performance. For qualitative evaluation, we illustrate the SR images generated by different methods and focus on key facial components such as mouth and eyes. Specifically, we use a variation of our H$_2$F method, known as H$_2$FGAN (a GAN version of H$_2$F), to demonstrate the improved visual quality of the SR images.

4.1.3 | Comparison methods

We compare H$_2$F with several state-of-the-art baselines. These include five models that utilize only texture information (SRCNN [4], VDSR [5], SRResNet [10], RCAN [25], and SFMNet [35]); two models that use face priors as constraints (SuperFAN [6], CompositeNet [45]); and four models that incorporate face priors as inputs (FSRNet [8], DIC [9], FSRCH [51], and DCLNet [50]). Bicubic interpolation serves as a naive baseline. The results of DIC, FSRNet, and DCLNet are reproduced from their papers as they use the same experiment settings as ours. For the other models, we train them under our experiment settings and conduct extensive hyper-parameter tuning for optimized performance. It is important to note that the open-source code for FSRCH and DCLNet is unavailable, so we include only their quantitative results from their original papers for comparison.

For H$_2$F, the shape priors are generated from the ground-truth HR faces during training (to accurately learn the correlation between shape priors and HF faces) and from the LR faces during the test phase. While using priors of LR faces for both training and testing might offer more consistency, we opt not to use the shape priors of LR faces for training due to the convergence issues. In the comparison with other methods in Section 4.2, we use MaskGAN to generate face priors for H$_2$F. For the inception–hourglass structure, we use 4 residual modules and 128 channels for the feature maps in one hourglass component.

4.1.4 | Implementation detail

We use bicubic interpolation to generate LR faces from HR faces. All experiments are performed with a magnification factor of 8× between the LR (16×16) and HR (128×128) faces, which corresponds to a 64× increase in image pixels and is a popular setting for face SR methods. For the Helen dataset, we rotate the original images by 90°, 180°, and 270°, and then flip the images horizontally to augment the training dataset (also used in DIC [9] and FSRNet [8]). H$_2$F is trained using the ADAM optimizer [60] under the default setting (i.e. $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e - 07$) with a batch size of 32. The learning rate is initialized as $5e^{-4}$, halved at the 40th epoch, and decreased to $1e^{-4}$ at the 50th epoch to train for another 3 epochs. We report the performance when the entire H$_2$F framework is trained end-to-end but observe that H$_2$F also works well when using a pre-trained face SR model in the texture branch. In this case, the high frequency branch only needs to be trained for a small number of epochs (e.g. 5). Additionally, H$_2$FGAN, a variant of H$_2$F to further enhance the qualitative results of SR images, is trained by setting $\lambda_{\text{adv}} = \lambda_{\text{perc}} = 0.01$. Both H$_2$F and H$_2$FGAN will be evaluated. All experiments are conducted using Tensorflow2 [61] on one NVIDIA Tesla V100 GPU.

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

In this part, H$_2$F is used for both quantitative and qualitative comparison with state-of-the-art face SR methods. Face detection as well as landmark detection on SR faces are used as downstream tasks to demonstrate the effectiveness of our method. For a fair comparison, H$_2$FGAN is used for qualitative comparison with other GAN-based SR methods (i.e. DICGAN).

4.2.1 | Quantitative results of H$_2$F

We report the PSNR and SSIM of H$_2$F and the baselines in Table 1. The results show that H$_2$F consistently outperforms all baselines in terms of PSNR and SSIM on the CelebA dataset

TABLE 1 A comparison of PSNR/SSIM with some state-of-the-art face SR methods.

	CelebA		Helen	
Method	PSNR	SSIM	PSNR	SSIM
Bicubic	23.58	0.6285	23.89	0.6751
SRCNN [4]	24.83	0.6500	24.04	0.6440
VDSR [5]	25.97	0.7010	25.84	0.7140
SRResNet [10]	25.82	0.7369	25.30	0.7297
RCAN [25]	26.79	0.7769	25.32	0.7267
SuperFAN [6]	26.12	0.7593	25.48	0.7550
FSRNet [8]	26.48	0.7718	25.90	0.7759
SFMNet [35]	26.48	0.8074	NA	NA
CompositeNet [45]	27.11	0.7861	26.82	0.7952
DIC [9]	27.37	0.7963	26.29	0.7933
DCLNet [50]	27.84	0.8124	**27.15**	0.7958
FSRCH [51]†	27.65	0.7946	NA	NA
H_2F(Ours)	**28.02**	**0.8142**	26.76	**0.7964**

Red/blue indicate the best/second performance. † represents shape priors extracted from HR images, which is not available in the common face SR application.

(with 1000 test images) and SSIM on the Helen dataset (with 50 test images). Although DCLNet [50] shows superior PSNR on the Helen dataset (with 50 test images), it is a cascaded network structure with complex CNN architectures and relies on multiple shape priors supervision. FSRCH [51], which utilizes the shape priors from HR images for face SR, is not practical for typical face SR applications. We also found that using better face priors leads to better performance for H_2F. The Helen dataset provides 11 ground-truth parsing maps for each image as illustrated in Figure 3a. We train and test H_2F using the ground-truth parsing maps and found that H_2F achieves **26.84** and **0.8043** in PSNR and SSIM, respectively, which are better than the results in Table 1. We will further illustrate that gains in both the texture branch and the high frequency branch translate into better performance for the H_2F framework in Sections 4.4 and 4.5.

Table 2 reports landmark detection performance of the SR faces generated by different methods. We do not include performance of DCLNet and FSRCH as there is no open-source code available. For both detection rate and NRMSE, H_2F consistently outperforms all baselines on the two datasets, which suggests that H_2F generates SR faces that match the landmarks in the ground-truth HR faces more accurately.

4.2.2 | Cross-dataset performance

We report the PSNR/SSIM for RCAN, DIC, and H_2F when the training dataset and test dataset are different in Table 3. The source domain refers to the dataset used for training, while the target domain is the one used for testing. The results show that all models suffer from performance degradation under dataset mismatch but H_2F still outperforms RCAN and DIC. In addition, the performance degradation caused by dataset mismatch is less significant for H_2F than for RCAN and DIC (comparing with the results in Table 1).

4.2.3 | Cross-pose performance

To evaluate the robustness of H_2F to pose variations, we use Pointing04 [62], a head pose estimation dataset that contains the face images of 15 persons. For each person, we choose 2 images with an angle of ±90° in the yaw direction (i.e. side faces), which results in 30 test images. We report the PSNR/SSIM for RCAN, DIC, and H_2F under pose variations in Table 4. The models are trained on the CelebA dataset in which most images are frontal faces but tested on 30 side face images from the Pointing04 dataset. The results show that H_2F outperforms both RCAN and DIC under pose variations.

4.2.4 | Qualitative results of H_2F

We illustrate some example SR images generated by different methods in Figure 7. The results show that both DIC and H_2F produce significantly more realistic face images than the other methods, which is in line with the PSNR/SSIM results in Table 1. Compared with DIC, H_2F is more successful in preserving the shape and details in important facial regions such as mouth, nose, and eyes. Specifically, for the face in the first row, H_2F keeps the lip shape of the target HR while DIC does not. For the face in the third row, DIC has severe blur in the nose region while H_2F does not. Preserving more details in key facial areas can help human inspectors better identify key facial features and distinguish different persons, and improve the performance of face related tasks (e.g. landmark detection as shown in Table 2).

4.2.5 | Comparison with DICGAN

For a fair comparison with DIC [9], we also train H_2FGAN (a GAN version of H_2F) and compare it with DICGAN (a GAN version of DIC). Figure 8 shows some example SR images generated by both models. The results indicate that both methods successfully generate photo-realistic face images from the LR input. H_2FGAN outperforms DICGAN in preserving the shape of the face and realistic details in crucial facial regions like the mouth, nose, and eyes. It is worth noting that our primary focus is on exploring the efficacy of the proposed H_2F framework, which we elaborate on in the following sections.

4.3 | Face editing function with H_2F

In H_2F, high frequency faces contain details in key facial areas and are explicitly generated from face priors, making it easier to understand the impact of the face priors on the final SR face. We demonstrate that H_2F provides an extra face editing function that allows the face priors to have a direct impact on the SR

TABLE 2 NRMSE (first row) and detection rate (second row) for landmark detection on the SR faces produced by different methods. Retinaface [38] is used to detect 106 landmarks on the SR faces.

Dataset	Bicubic	SRCNN	VDSR	SRResNet	RCAN	SuperFAN	FSRNet	DIC	H_2F (ours)
CelebA	0.1035	0.0807	0.0238	0.0449	0.0122	0.0125	0.0232	0.0107	**0.0083**
	83.5%	88.8%	93.1%	98.8%	100%	100%	100%	100%	**100%**
Helen	0.1064	0.1172	0.0812	0.0570	0.0426	0.0347	0.0461	0.0241	**0.0181**
	78.0%	74.0%	70.0%	90.0%	94.0%	94.0%	96%	98.0%	**100%**

Red/blue indicate the best/second performance.

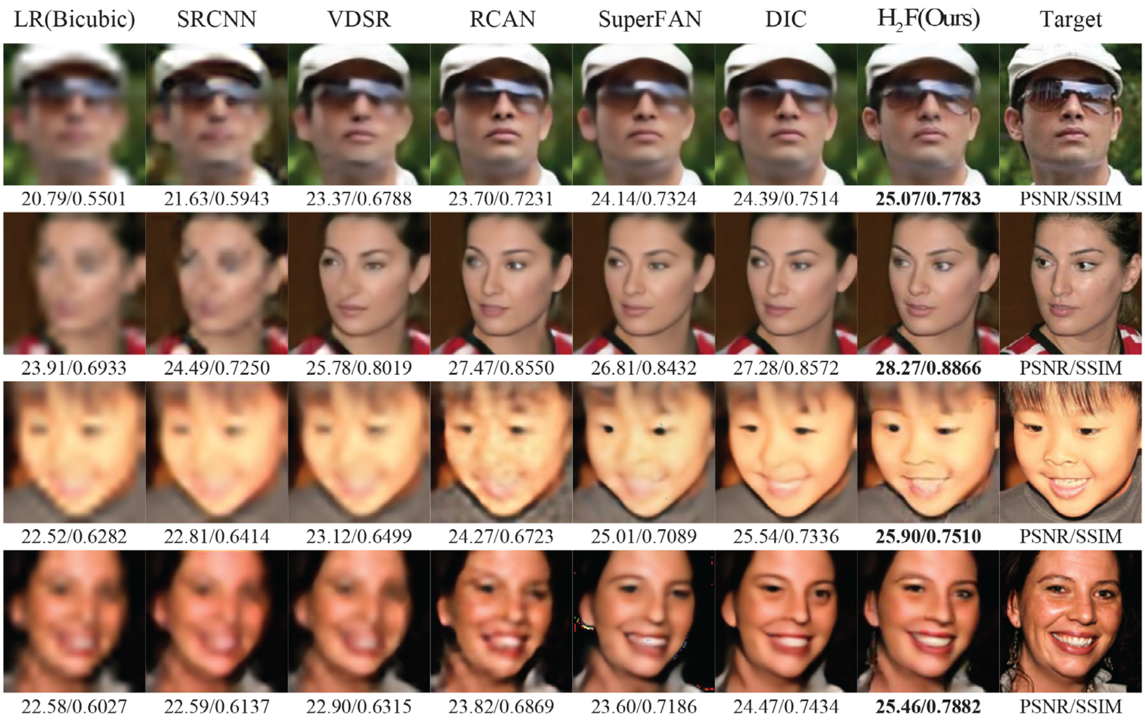


FIGURE 7 Example face images produced by different face SR methods. H_2F provides more realistic details in key facial regions such as eyes, nose and mouth, and is good at handling occlusion (1st row), pose rotation (2nd row) and expression variations (3rd and 4th row).

TABLE 3 Cross-dataset performance (PSNR/SSIM) comparison.

Source domain	Target domain	Method	PSNR	SSIM
CelebA	Helen	RCAN	26.42	0.7809
		DIC	26.94	0.7944
		H_2F (ours)	**27.80**	**0.8252**
Helen	CelebA	RCAN	24.57	0.6677
		DIC	25.48	0.7217
		H_2F (ours)	**26.32**	**0.7964**

Red/blue indicate the best/second performance.

TABLE 4 Cross-pose performance (PSNR/SSIM) comparison.

Dataset	RCAN	DIC	H_2F (ours)
Pointing04 (±90°)	28.18/0.8214	28.30/0.8250	**29.03/0.8401**

Red/blue indicate the best/second performance.

face. Therefore, it is interesting to explore how changes in the face priors affect the SR face. The results are shown in Figure 9, in which the face priors are modified by adjusting the positions of the landmarks. The results show that editing the landmarks often leads to the intended changes in the SR faces, such as enlarging the eyes, shrinking the nose and changing the shape of the eyebrow. These results indicate that H_2F produces SR faces that can be easily explained and controlled by manipulating the face priors. It verifies that H_2F effectively exploits the high frequency signals learned from the face priors. Furthermore, we also believe H_2F provides a fresh perspective for face editing.

4.4 | Flexibility of H_2F

We evaluate the flexibility of H_2F by using different SR models in the texture branch. By default, the face priors are 9 parsing maps generated from 106 landmarks, and the

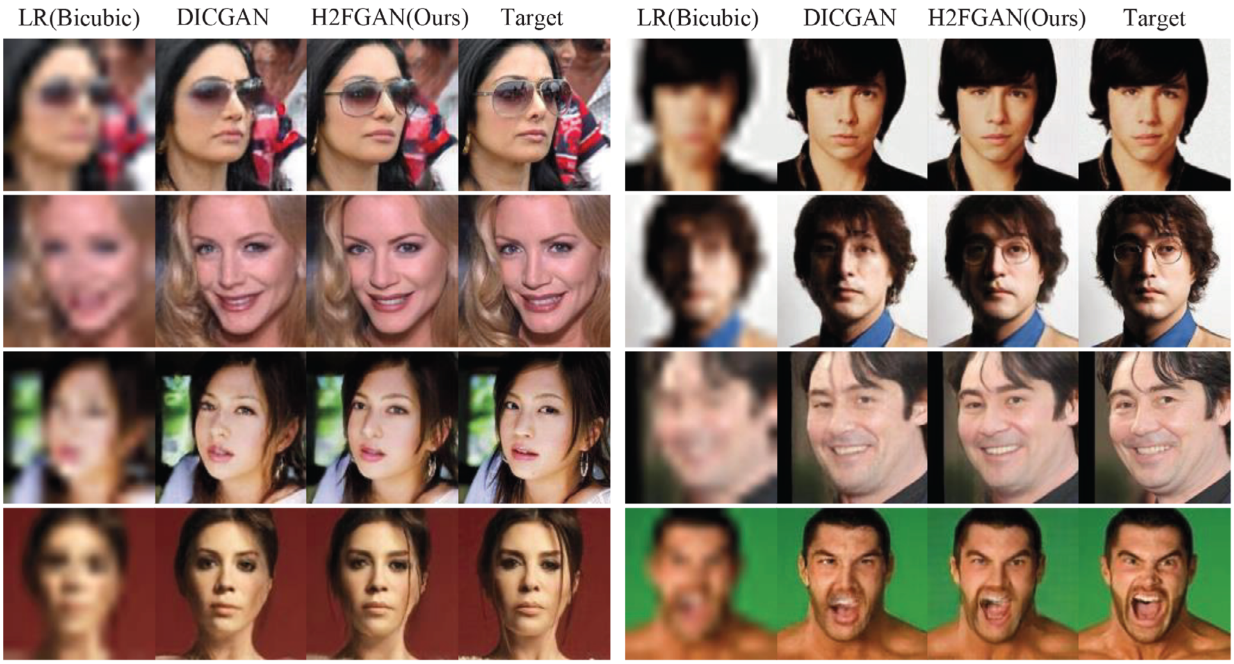


FIGURE 8 Example face images produced by GAN-based variants of different face SR methods. H_2FGAN is more succeed in handling occlusion (e.g. eyeglasses, sunglasses, and a pinch of hair), expression variations (e.g. happy, anger) and other fine-grained details preserving (e.g. teeth, shape of eyeball).

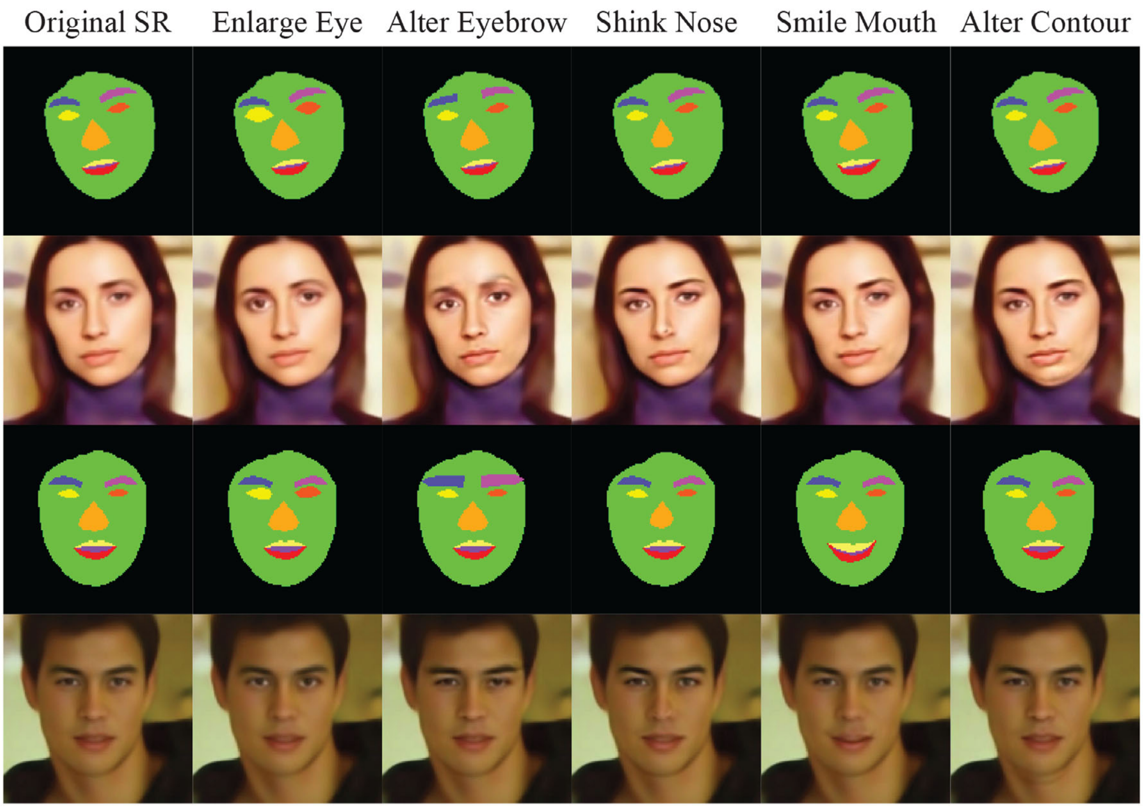


FIGURE 9 Face editing by changing the face priors including enlarge eye, alter eyebrow, shrink nose, alter mouth and face contour, respectively, in the high frequency branch of H_2F which produces explainable results.

TABLE 5 Performance of H_2F with different texture branch models.

Texture branch model	Coarse PSNR/SSIM	PSNR/SSIM
H_2F-VDSR	26.08/0.7435	**26.32/0.7574**
H_2F-SRResNet	26.43/0.7586	**26.59/0.7678**
H_2F-RCAN	26.84/0.7723	**26.88/0.7741**
H_2F-SuperFAN	26.88/0.7740	**26.95/0.7753**

inception–hourglass module has only 1 branch and 1 hourglass component. The experiment is conducted on the CelebA dataset. Note that the SuperFAN model here is different from the one in [6] as we remove the supervision of face alignment to make it a texture-only model. The results are presented in Table 5. The second column reports the performance of using only the texture branch (trained alongside the high frequency branch), and the third column reports the performance of the entire H_2F framework. We can make the following observations:

(1) Better coarse SR results: when trained alongside the high frequency branch, the texture branch model provides better performance than trained alone (see the standalone results in Table 1). For example, SRResNet achieves a PNSR of 25.82 when trained standalone, and this increases to 26.43 when trained in the H_2F framework. This phenomenon indicates that incorporating the high frequency branch enhances learning in the texture branch.

(2) Better final SR results: the combination of the texture branch and the high frequency branch leads to improved performance in all texture branch models, and a better texture branch model results in better performance for the H_2F framework. This suggests that using dual reconstruction supervision (i.e. for texture branch and final output) indeed enhances the SR learning process.

The two observations suggest that H_2F is a general framework that supports easy plug-in of texture branch models and may benefit from better SR models in the future.

4.5 | Effectiveness of a high frequency highway

In this part, we focus on three aspects of the H_2F framework. First, we demonstrate the performance gain of using a high frequency highway. Then, we explore the effect of using different face priors in H_2F. Finally, we conduct an ablation study on the inception–hourglass design. By default, the texture branch model is VDSR [5], the face priors are 9 parsing maps generated from 106 landmarks, and the inception–hourglass module has only 1 branch and 1 hourglass component. The dataset is CelebA.

4.5.1 | Gain of the high frequency branch

Table 5 clearly shows that the high frequency branch improves performance by complementing the texture branch. We illustrate some coarse SR faces and high frequency (HF) faces in Figure 10 to visualize the contribution of the high frequency branch. Two observations can be made:

TABLE 6 Performance of H_2F with face priors of varying richness.

Shape priors	# of landmarks	# of h or p	PSNR	SSIM
Heatmaps	5	$h = 6$	25.98	0.7429
	68	$h = 9$	25.98	0.7418
	106	$h = 9$	26.21	0.7542
Parsing maps	5	$p = 6$	26.05	0.7456
	68	$p = 9$	26.27	**0.7575**
	106	$p = 9$	**26.32**	0.7574

h and p denote the number of heatmaps and parsing maps, respectively.

(1) The HF face contains details in important facial areas such as eyes, nose and mouth, and coarse SR is significantly different from the final SR in these areas.

(2) The HF face is more prominent with a weaker texture branch model, for example, HF faces contain more details for VDSR than SuperFAN. This is because better models lose fewer high frequency details.

The two observations show that the high frequency branch effectively adds high frequency signals missing from the texture branch and can adapt to different texture branch models.

4.5.2 | Effect of different face priors

We report the performance of H_2F using face priors with varying levels of rich information in Table 6. We use h and p to denote the number of heatmaps and parsing maps, respectively, and k to denote the number of landmarks on which the face priors are generated. Note that parsing maps provide more information than heatmaps as they incorporate the semantic meaning of facial components. The results in Table 6 show that using face priors with rich information generally results in better performance for the H_2F framework. For example, using the same number of landmarks, parsing maps provide better performance than heatmaps. For both heatmaps and parsing maps, increasing the number of landmarks improves performance with only one exception. Furthermore, as a glimpse of the results in Section 4.2 indicates, using more precise shape priors often leads to better results. That is, on the Helen dataset, the PSNR/SSIM obtained using the predicted priors and ground truth priors are 26.84/0.8043 and 26.76/0.7964, respectively.

4.5.3 | Inception–hourglass design

We adopt an inception–hourglass model in H_2F to map the face priors to HF face in the high frequency branch, which uses multiple branches of hourglass components. We evaluate the performance gain of the inception–hourglass design in Table 7, in which “w/o” means using only one branch and the numbers

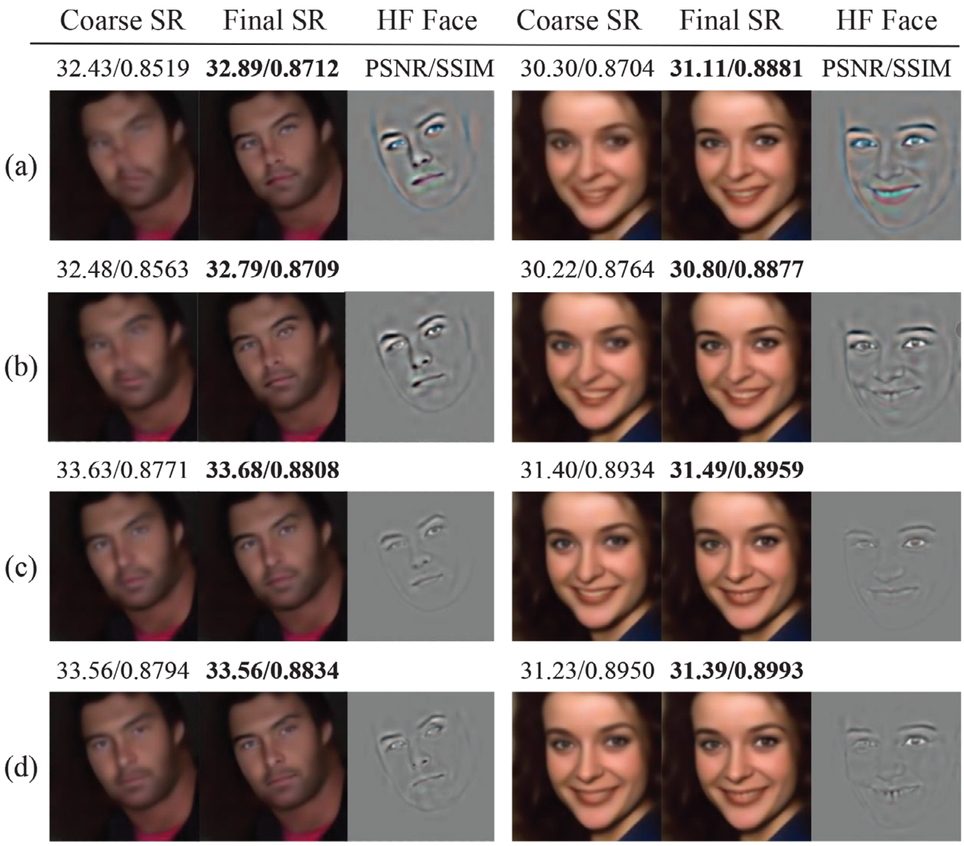


FIGURE 10 The coarse SR (1st and 4th columns), high frequency (HF) face (3rd and 6th columns), and final SR (2nd and 4th columns) when using different models in the texture branch. From top to bottom, the texture branch models are (a) VDSR, (b) SRResNet, (c) RCAN, and (d) SuperFAN, respectively.

TABLE 7 Performance of different inception–hourglass designs.

	# of hourglasses				
Inception design	**1**	**2**	**4**	**PSNR**	**SSIM**
w/o	✓			26.32	0.7574
		✓		26.32	0.7575
Two branches	✓	✓		26.37	0.7586
Three branches	✓	✓	✓	**26.40**	**0.7629**

"w/o" means using only one branch with one hourglass component.

(i.e. 1, 2, 4) indicate the number of hourglass components in each branch. The results show that using more hourglass components in one branch leads to little to no performance gain. However, using multiple branches provides better performance than one branch, which verifies the effectiveness of the inception–hourglass design.

4.6 | More ablation study

In this part, we verify the performance gain of using H_2F framework does not come from a more complex network structure, nor from the incorporation of shape priors. Instead, the H_2F framework is the main reason for achieving good performance.

TABLE 8 A comparison with H_2F and baseline++ methods. Baseline++ is extended by adding more feature maps to ensure methods have an equivalent number of parameters.

Model	**# of params**	**PSNR**	**SSIM**
VDSR++	1,496,448	26.11	0.7425
H_2F-VDSR	1,455,683	**26.32**	**0.7574**
SRResNet++	2,114,115	26.28	0.7533
H_2F-SRResNet	2,020,422	**26.59**	**0.7678**
SuperFAN++	2,373,719	26.67	0.7749
H_2F-SuperFAN	2,419,866	**26.95**	**0.7753**

We compare the performance of H_2F and "baseline++" SR methods in Table 8 to ensure the improvement of H_2F is not due to additional parameters in the SR model. To ensure fairness, the "baseline++" methods are designed to have an equivalent number of parameters as H_2F by adding more feature maps. However, this does not guarantee equal parameters. The methods are trained with a similar amount of parameters for a fair comparison. Two observations can be made:

(1) The baseline++ methods (e.g. VDSR++, SRResNet++, SuperFAN++) perform better than the corresponding baselines (e.g. VDSR, SRResNet, SuperFAN) as expected. For example, SRResNet achieves PSNR of 25.82 and SSIM

TABLE 9 Ablation study of using H_2F framework with the same shape priors. The method not applying H_2F utilizes the same input LR face and shape prior as input for SR results.

	VDSR		SRResNet		SuperFAN	
Model	w/o H_2F	w/ H_2F	w/o H_2F	w/ H_2F	w/o H_2F	w/ H_2F
PSNR	26.26	**26.32**	26.49	**26.59**	26.84	**26.95**
SSIM	0.7557	**0.7574**	0.7643	**0.7678**	**0.7768**	0.7753

of 0.7369, but SRResNet++ achieves PSNR of 26.28 and SSIM of 0.7533.

(2) Importantly, with the same number of parameters, H_2F consistently outperforms the baseline++ methods, which verifies the effectiveness of H_2F.

In addition, Table 9 compares the performance of different SR methods with the same shape prior and input LR face. The comparison is made both with (denoted as "w/") or without (denoted as "w/o") the use of the H_2F framework. The methods not applying H_2F use the same input LR face and shape prior as input for SR results. The results clearly show that H_2F produces better results even when the same shape priors are used, thus confirming that the H_2F framework is the main factor for achieving good performance.

5 | DISCUSSION

Our H_2F is an innovative framework that directly utilizes the face priors for face SR. To the best of our knowledge, we are the first to extract an HF face directly from the shape priors through our high frequency highway, which addresses the potential for smoothing HF details during learning. Furthermore, with the dual-reconstruction learning mechanism, H_2F can learn the HF face without the need for a ground-truth HF face. Our H_2F demonstrates a deep understanding of existing methods. While it delves deeply into these methods, it's not merely about exploring what is known. We explain the contribution of face priors in face SR from a signal processing perspective and it is non-trivial to understand the root cause behind the increased HF information.

In our experiments, we are consistent with the widely-used face SR setting, superresolving LR images from 16×16 to 128×128 of HR images with an 8× magnification factor. We have included several SISR methods as they can be integrated into the texture branch model of H_2F, demonstrating the general applicability and flexibility of our method. Our benchmarks are based on two widely used datasets, CelebA and Helen. We consciously avoid 1024×1024 high-definition images as our desired resolution, as such settings are more suitable for GAN prior-based FSR methods (i.e. not based on shape priors).

6 | CONCLUSION

In this paper, we propose a H_2F framework for face superresolution. The proposed framework is motivated by the observation that face priors provide information (e.g. locations) of high-frequency details in key facial areas. H_2F dynamically decomposes a final SR face into two components: a coarse SR face and an HF face. The HF face is directly learned from shape priors, as they are strongly related to missing high-frequency signals. To address the challenge of missing ground-truth HF faces, we propose dual reconstruction losses (i.e. for the texture branch and final output), where the high frequency branch takes the missing parts in the texture branch as learning targets.

Extensive experiments demonstrate that H_2F significantly outperforms state-of-the-art face SR methods and is robust to dataset mismatch and pose variations. Additionally, the framework provides an extra face editing function and allows the use of many existing SR models in its texture branch, and effectively supports various face priors using the inception–hourglass in its high frequency branch.

By leveraging our H2F framework, we can obtain finer texture details and explainable SR results. However, the H2F framework also requires slightly larger computational resources due to the incorporation of the high-frequency branch network. Our potential limitation is the impact on inference speed due to the added high frequency highway. However, this tradeoff allows us to gain deeper insights into the high-frequency information and obtain superior SR results. In the future, we will consider integrating Transformer architecture or CNN-transformer based architectures into our texture branch model to push performance boundaries further. Specifically, it is worth integrating the renowned transformer architecture, SwinIR [63], which utilizes only texture information as our plug-in texture branch. We leave this for future research.

AUTHOR CONTRIBUTIONS

Dan Zeng: Conceptualization; formal analysis; investigation; methodology; writing—original draft. **Wen Jiang**: Investigation; software; validation. **Xiao Yan**: Conceptualization; writing—review and editing. **Weibao Fu**: Data curation; investigation; software. **Qiaomu Shen**: Conceptualization; visualization. **Raymond Veldhuis**: Supervision; writing—review and editing. **Bo Tang**: 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

The data that support the findings of this study are openly available in https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html and http://www.ifp.illinois.edu/~vuongle2/helen/ at https://doi.org/10.1109/ICCV.2015.425, https://doi.org/10.1109/TIP.2003.819861, https://doi.org/10.1109/CVPR.2013.447 reference number [57,40,58] in the revised version.

ORCID

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

REFERENCES

1. Zhang, W., Chen, Y., Yang, W., Wang, G., Xue, J.-H., Liao, Q.: Class-variant margin normalized softmax loss for deep face recognition. IEEE Trans. Neural Networks Learn. Syst. 32(10), 4742–4747 (2020)
2. Miao, J., Wu, Y., Yang, Y.: Identifying visible parts via pose estimation for occluded person re-identification. IEEE Trans. Neural Networks Learn. Syst. 33(9), 4624–4634 (2021)
3. 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)
4. 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)
5. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1646–1654. IEEE, Piscataway, NJ (2016)
6. 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 (CVPR), pp. 109–117. IEEE, Piscataway, NJ (2018)
7. Yin, Y., Robinson, J.P., Zhang, Y., Fu, Y.: Joint super-resolution and alignment of tiny faces. arXiv:191108566 (2019)
8. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: FSRNet: end-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2492–2501. IEEE, Piscataway, NJ (2018)
9. 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 Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5569–5578. IEEE, Piscataway, NJ (2020)
10. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., 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 (CVPR), pp. 4681–4690. IEEE, Piscataway, NJ (2017)
11. Gu, Y., Wang, X., Xie, L., Dong, C., Li, G., Shan, Y., Cheng, M.-M.: VQFR: blind face restoration with vector-quantized dictionary and parallel decoder. arXiv:220506803 (2022)
12. Li, Z., Zeng, D., Yan, X., Shen, Q., Tang, B.: Analyzing and combating attribute bias for face restoration. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 1151–1159. ACM, New York (2023)
13. Kim, J., Li, G., Yun, I., Jung, C., Kim, J.: Edge and identity preserving network for face super-resolution. Neurocomputing 446, 11–22 (2021)
14. 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 Proc. 17(11), 3309–3322 (2023)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410. IEEE, Piscataway, NJ (2019)
16. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv:201002502 (2020)
17. 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. IEEE, Piscataway, NJ (2020)
18. 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. IEEE, Piscataway, NJ (2021)
19. 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. IEEE, Piscataway, NJ (2021)
20. Wang, Z., Zhang, J., Chen, R., Wang, W., Luo, P.: RestoreFormer: high-quality blind face restoration from undegraded key-value pairs. arXiv:2201.06374 (2022)
21. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. arXiv:2012.09841 (2020)
22. Zhou, S., Chan, K.C.K., Li, C., Loy, C.C.: Towards robust blind face restoration with codebook lookup transformer. In: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems, pp. 30599–30611. ACM, New York (2022)
23. Yue, Z., Loy, C.C.: DifFace: blind face restoration with diffused error contraction. arXiv:2212.06512 (2022)
24. Yang, P., Zhou, S., Tao, Q., Loy, C.C.: PGDiff: guiding diffusion models for versatile face restoration via partial guidance. In: NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems, pp. 32194–32214. ACM, New York (2023)
25. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301. Springer, Cham (2018)
26. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2472–2481. IEEE, Piscataway, NJ (2018)
27. Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., Zhang, K., Cao, X., Shen, H.: Single image super-resolution via a holistic attention network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 191–207. Springer, Cham (2020)
28. Kong, X., Zhao, H., Qiao, Y., Dong, C.: ClassSR: a general framework to accelerate super-resolution networks by data characteristic. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12016–12025. IEEE, Piscataway, NJ (2021)
29. Song, D., Wang, Y., Chen, H., Xu, C., Xu, C., Tao, D.: AdderSR: towards energy efficient image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15648–15657. IEEE, Piscataway, NJ (2021)
30. Tian, C., Zhuge, R., Wu, Z., Xu, Y., Zuo, W., Chen, C., Lin, C.-W.: Lightweight image super-resolution with enhanced CNN. Knowl.-Based Syst. 205, 106235 (2020)
31. Feng, H., Wang, L., Li, Y., Du, A.: LKASR: large kernel attention for lightweight image super-resolution. Knowl.-Based Syst. 252, 109376 (2022)
32. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 318–333. Springer, Cham (2016)
33. Chen, C., Gong, D., Wang, H., Li, Z., Wong, K.-Y.K.: Learning spatial attention for face super-resolution. IEEE Trans. Image Process. 30, 1219–1231 (2020)
34. Gao, G., Xu, Z., Li, J., Yang, J., Zeng, T., Qi, G.-J.: CTCNet: a CNN-transformer cooperation network for face image super-resolution. IEEE Trans. Image Process. 32, 1978–1991 (2023)
35. Wang, C., Jiang, J., Zhong, Z., Liu, X.: Spatial-frequency mutual learning for face super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22356–22366. IEEE, Piscataway, NJ (2023)
36. Cui, Y., Tao, Y., Bing, Z., Ren, W., Gao, X., Cao, X., Huang, K., Knoll, A.: Selective frequency network for image restoration. Paper presented at the eleventh international conference on learning representations, Kigali, Rwanda, 1–5 May 2023
37. Zadeh, A., Chong Lim, Y., Baltrusaitis, T., Morency, L.P.: Convolutional experts constrained local model for 3D facial landmark detection. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 2519–2528. IEEE, Piscataway, NJ (2017)
38. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: RetinaFace: single-stage dense face localisation in the wild. arXiv:190500641 (2019)
39. Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3911–3919. IEEE, Piscataway, NJ (2017)

40. 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 (CVPR), pp. 3484–3491. IEEE, Piscataway, NJ (2013)
41. Song, L., Cao, J., Song, L., Hu, Y., He, R.: Geometry-aware face completion and editing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 2506–2513. ACM, New York (2019)
42. Lee, C.H., Liu, Z., Wu, L., Luo, P.: MaskGAN: towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5549–5558. IEEE, Piscataway, NJ (2020)
43. Kim, D., Kim, M., Kwon, G., Kim, D.-S.: Progressive face super-resolution via attention to facial landmark. arXiv:190808239 (2019)
44. Jo, Y., Oh, S.W., Kang, J., Kim, S.J.: Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3224–3232. IEEE, Piscataway, NJ (2018)
45. Wei, F., Wang, S., Yang, J., Sun, X., Wang, Y., Chen, Y.: A composite network model for face super-resolution with multi-order head attention facial priors. Pattern Recognit. 139, 109503 (2023)
46. 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. Springer, Cham (2018)
47. Zhu, S., Liu, S., Loy, C.C., Tang, X.: Deep cascaded bi-network for face hallucination. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 614–630. Springer, Cham (2016)
48. Hu, X., Ren, W., LaMaster, J., Cao, X., Li, X., Li, Z., Menze, B., Liu, W.: Face super-resolution guided by 3D facial priors. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 763–780. Springer, Cham (2020)
49. Li, X., Chen, C., Zhou, S., Lin, X., Zuo, W., Zhang, L.: Blind face restoration via deep multi-scale component dictionaries. In: European Conference on Computer Vision, pp. 399–415. Springer, Cham (2020)
50. Wang, H., Hu, Q., Wu, C., Chi, J., Yu, X., Wu, H.: DCLNet: dual closed-loop networks for face super-resolution. Knowl.-Based Syst. 222, 106987 (2021)
51. Lu, T., Wang, Y., Zhang, Y., Jiang, J., Wang, Z., Xiong, Z.: Rethinking prior-guided face super-resolution: a new paradigm with facial component prior. IEEE Trans. Neural Networks Learn. Syst. 35(3), 3938–3952 (2022)
52. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.-F., Barriuso, A., Torralba, A., Fidler, S.: 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. IEEE, Piscataway, NJ (2021)
53. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 483–499. Springer, Cham (2016)
54. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9. IEEE, Piscataway, NJ (2015)
55. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, pp. 234–241. Springer, Cham (2015)
56. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:14091556 (2014)
57. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV), pp. 3730–3738. IEEE, Piscataway, NJ (2015)
58. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 679–692. Springer, Berlin, Heidelberg (2012)
59. 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)
60. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:14126980 (2014)
61. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: TensorFlow: a system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pp. 265–283. ACM, New York (2016)
62. Gourier, N., Letessier, J.: The pointing 04 data sets. In: Proceedings of Pointing 2004, ICPR International Workshop on Visual Observation of Deictic Gestures, pp. 1–4. ICPR, Cambridge (2004)
63. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: Image restoration using Swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833–1844 (2021)

How to cite this article: Zeng, D., Jiang, W., Yan, X., Fu, W., Shen, Q., Veldhuis, R., Tang, B.: Face super resolution with a high frequency highway. IET Image Process. 18, 3570–3586 (2024). https://doi.org/10.1049/ipr2.13195