Neural Processing Letters (2024) 56:97
https://doi.org/10.1007/s11063-024-11562-8

CFGPFSR: A Generative Method Combining Facial and GAN Priors for Face Super-Resolution

Jinbo Liu[1] · Zhonghua Liu[1,2] · Weihua Ou[3] · Kaibing Zhang[4] · Yong Liu[1]

Accepted: 11 February 2024 / Published online: 9 March 2024

Abstract
In recent years, facial prior has been widely applied to enhance the quality of super-resolution (SR) facial images in face super-resolution (FSR) methods based on deep learning. However, most of the existing facial prior-based FSR methods have insufficient attention to local texture details, which can cause the generated SR facial images with overly smooth and unrealistic texture details, and show obvious artifacts under large magnification. With the help of GAN prior, recent advances can produce excellent results in terms of fidelity and realness. A generative framework for FSR is proposed in this work, which combines GAN and facial prior, termed CFGPFSR. Firstly, we pre-train a face StyleGAN2 and a face parsing network (FPN) that can generate decent parsing maps, in which the proposed CFGPFSR exploits rich and varied priors encapsulated in the face StyleGAN2 (GAN prior) and face parsing maps extracted from the FPN (facial prior) for FSR. Moreover, we introduce the Channel-Split Spatial Feature Transform (CS-SFT) method to further improve FSR performance. GAN and facial priors are introduced into the FSR process through the designed CS-SFT layers so that SR facial images obtain a promising balance between fidelity and realness. Unlike GAN inversion methods which necessitate costly image optimization at runtime, the proposed CFGPFSR can jointly recover facial details by only utilizing one forward pass. Experimental results on synthetic and real images indicate that the proposed CFGPFSR obtains remarkable performance in 16 × SR task, and some of its metrics such as peak signal to noise ratio (PSNR) and structural similarity (SSIM) are higher than that of the comparison methods. Meanwhile, it shows impressive results in reconstructing high-quality facial images.

Keywords Face super-resolution · Facial prior · GAN prior · CFGPFSR

✉ Zhonghua Liu
lzhua_217@163.com

1 Information Engineering College, Henan University of Science and Technology, Luoyang, China

2 School of Information Engineering, Zhejiang Ocean University, Zhoushan, China

3 School of Big Data and Computer Science, Guizhou Normal University, Guiyang, China

4 College of Electronics and Information, Xi'an Polytechnic University, Xi'an, China

1 Introduction

In recent years, a growing number of reconstruction methods have been developed to provide promising solutions for face super-resolution (FSR). Face super-resolution (FSR), *a.k.a.* face hallucination (FH), refers to producing high resolution (HR) face images from the corresponding low resolution (LR) inputs. FSR is a fundamental issue in face analysis, which has been extensively used to improve tasks involving faces, e.g., face alignment, face parsing, face recognition. This is due to the majority of existing techniques would suffer severely when given quite LR face images.

FSR is a special example of the task of single image super-resolution (SISR) [1, 2], which is an inherently ill-posed problem because there are always numerous possible HR counterparts for every LR image. In contrast to SISR, FSR normally needs to handle with very large upscaling factors (8 to 64) and only considers face images instead of arbitrary scenes. Face images have high structural similarity and finer texture details than other images. Therefore, FSR is not only required to produce massive amounts of finer local details and correct facial structure, but also to retain identity information. FSR has distinctive solution pipelines, which usually involve a number of extra facial priors, such as parsing map and attribution map. In order to enhance the performance of FSR, many researchers introduce specific prior information in face images into the FSR problem. Chen et al. [3] introduce the facial landmarks and parsing maps to improve recovery performance. For improving the SR quality, Yu et al. [4] estimate facial component heatmaps, which can supply localization of facial components. Li et al. [5] first utilize attribute information to guarantee correct attributes, and then design two parallel branches to produce facial prior and SR results. Ma et al. [6] optimizes face recovery and landmark estimation alternatively. Facial attributes, such as age, gender, and others, are also usually exploited in some face hallucination methods [7, 8]. However, these methods also have a number of limitations. On the one hand, though the facial priors is utilized to provide a good face structure for the SR face image, the lack of attention to local texture details results in overly smooth and unrealistic local texture details in the reconstructed image. In addition, there are obvious artifacts in the generated SR facial image under large magnification. On the other hand, a simple concatenation operation is used by most existing methods to incorporate the facial priors into the FSR process, which is not sufficiently clear and results in the facial priors may not be adequately utilized. Therefore, it is important to explore the more powerful and effective schemes that use facial priors.

Recently, some works have found that a well pre-trained face GAN model on a large-scale face dataset contains wealthy and diverse face information, in which the convolutional blocks in the face GAN model are able to capture fine facial texture and detail information. This information can be utilized as prior knowledge, also known as GAN prior, to effectively produce realistic face details to enhance the performance of FSR. For instance, some methods [9–11] try to use GAN inversion. They firstly convert the input image into a latent code for the pre-trained GAN by 'invert'. After that, they reconstruct the image by costly image specific optimization. Although the output images appear to be real, they usually generate images with low fidelity since it is insufficient that the low-dimension latent codes to accurately guide reconstruction. Unlike GAN inversion methods, GLEAN [12] obtains excellent performance on large-factor SR tasks, which employs the intermediate features of a pre-trained StyleGAN2 [13] as a latent bank. To enhance the performance of blind face restoration, GFPGAN [14] and GPEN [15] respectively introduce a pre-trained face GAN models as the GAN prior to reconstruct fine facial details. The utilization of GAN priors is the key for the success of such methods, which also inspires after works to explore more applications.

Therefore, we propose a generative framework for FSR combining GAN and facial priors, called CFGPFSR, which can make the SR results obtaining a promising balance of realness and fidelity just by one forward pass. Face parsing maps provide the segmentation of facial components, and accurate face parsing maps can effectively promote the recovery of facial structure and facial components. Therefore, we pre-trained a FPN to generate decent face parsing maps as our facial prior. Due to the excellent performance of StyleGAN2, a face StyleGAN2 is pre-trained as our GAN prior. More specifically, our CFGPFSR consists of an encoder, a GAN Prior Bank and a decoder, in which the GAN Prior Bank is our pre-trained face StyleGAN2. The encoder is utilized to extract latent vectors and multi-resolution convolutional features, which capture significant high-level cues as well as spatial structure of the LR image. After that, the latent vectors are directly used to condition the GAN Prior Bank to generate a set of intermediate features. In order to obtain higher quality SR results, the Channel-Split Spatial Feature Transform (CS-SFT) [14] method is introduced. The multi-resolution convolutional features extracted from the encoder and face parsing maps extracted from the pre-trained FPN are passed to the GAN Prior Bank. Such cues are further utilized to condition the intermediate features generated by the GAN Prior Bank through the designed CS-SFT layer, which enables the proposed CFGPFSR to effectively incorporate GAN and facial priors. Finally, the decoder produces the final result by integrating the features from both the encoder and the GAN Prior Bank. The main contributions of this paper can be summarized as follows:

(1) A generative framework is presented for FSR termed CFGPFSR, which jointly utilizes GAN and facial priors at high level to guide the SR process for generating highly realistic and faithful SR face images.
(2) We pre-train a face StyleGAN2 with rich and varied priors and a face parsing network (FPN) that can generate decent parsing maps. By introducing the CS-SFT instead of a simple concatenation operation, our CFGPFSR model incorporates prior information encapsulated in the face StyleGAN2 (GAN prior) and face parsing maps extracted from the FPN (facial prior) into the FSR process through the designed CS-SFT layers, which allows the SR results to obtain a promising balance of realness and fidelity.
(3) Qualitative and quantitative experiments on synthetic and real images demonstrate that in 16 × SR task, our proposed framework achieves superior results in terms of both image quality and facial details restoration.

2 Related Work

2.1 Face Super-Resolution

Face super-resolution (FSR), which is also named face hallucination (FH), is a special example of the task of single image super-resolution (SISR). FH is first proposed by Baker and Kanade [16], who show that it is possible to achieve high magnification SR on specific types of images (*e.g.*, face and text). Since then, a number of methods have been proposed to improve FH performance, which can be broadly categorized into subspace based methods [17–19] and component based methods [20, 21]. Subspace based methods always focus on principal component analysis (PCA), which requires precisely aligned faces. Component-based methods involve the detection of facial landmarks, which is very difficult to detect directly from LR face images. Neither of these efforts achieves satisfactory results for FSR.

Recently, due to the excellent learning ability, deep convolutional neural networks [22] have achieved remarkable progress in FSR. Cao et al. [23] propose Attention-FH, which uses reinforcement learning to discover attended patches and then enhances the facial part sequentially. Chen et al. [24] propose RBPNet, which recovers HR images from LR images by iterative residual learning. Jiang et al. [25] design two individual branches for learning global facial contours and local facial component details, and then fuse the results of the two branches producing final SR results. Chen et al. [26] develop SPARNet, which introduces spatial attention into the generator and utilizes a multi-scale discriminator to enhance image quality. Some works also introduce efficient learning [27, 28] into FSR. Gao et al. [29] propose an efficient joint learning network, which is composed of denoising module and SR module, to obtain high-quality HR facial images. Compared to subspace-based methods and component based methods, the methods based on deep convolutional neural networks achieve more superior FSR results. In order to further enhance the quality of SR images, some works introduce the facial prior into FSR tasks to guide the network in generating SR images.

2.2 Facial Prior

In contrast to general images, the face images have specific prior information such as face parsing maps and landmark heatmaps. Chen et al. [3] concatenate predicted facial landmark heatmaps and parsing maps from LR with the intermediate features of the network. Yu et al. [4] concatenate the intermediate features of the network with predicted facial component heatmaps for FSR. Ma et al. [6] propose DIC, which iteratively recovers SR results and predicts facial landmark heatmaps. Wang et al. [30] design a progressive facial prior estimation framework and a new prior-guided feature enhancement module to guide the face image super-resolution by generating multiple facial component maps. However, most existing methods focus more on global shape and structure information, but not enough on local texture information. Furthermore, most methods simply incorporate the facial priors into the FSR process through a simple concatenation operation, which results in making insufficient use of facial prior information.

In recent years, a number of researches have introduced the face GAN prior into FSR tasks to reconstruct face images by utilizing the prior information encapsulated in well pre-trained face GAN models, which achieves impressive results in recovering high-quality facial images.

2.3 GAN Prior

The pre-trained GAN priors [13, 31] have been extensively adopted in GAN inversion methods [9–11], whose principal intent is to search the latent codes which are closest to the input image. Gu et al. propose mGANprior [9], which optimizes several latent codes to enhance the visual quality of the output images. Menon et al. propose PULSE [10], which iteratively optimizes the latent code of StyleGAN by using L1 loss function to constrain the input and downsampled output. For improving the performance of the model, Pan et al. [11] not only optimize the latent code, but also finetune the pre-trained GAN prior. However, due to low-dimensional latent codes being inadequate to guide image recovery, such methods often produce unsatisfactory and low-fidelity results.

To solve this issue, Chan et al. develop GLEAN [12], which is designed for large-factor SR and achieves excellent performance. Firstly, GLEAN extracts latent vectors and the multi-resolution features of the input image through adopting an additional RRDBNet [32]. After that, these features are fused with the intermediate features of the GAN prior. Yang et al. propose GPEN [15] and Wang et al. propose GFPGAN [14], which are based on the GAN prior and obtain superior performance for blind face restoration (BFR) issue. GPEN and GFPGAN also use an additional encoder to extract the multi-resolution features of the input image and further utilize them to condition the intermediate features of the GAN prior. Inspired by the above methods, our CFGPFSR combines GAN and facial priors for FSR. Latent vectors and multi-resolution spatial features are firstly extracted from the input image by using an encoder. Then, we use such cues and face parsing maps extracted from the pre-trained face parsing network (FPN) to co-modulate the intermediate features of the GAN Prior Bank.

3 Materials and Methods

3.1 Overview of CFGPFSR

The proposed CFGPFSR framework is described in detail in this section. The goal of FSR is to estimate the SR result I^{SR}, which is as close as possible to the ground-truth (GT) I^{HR} of the input LR face image I^{LR}. Figure 1 shows the overall framework of CFGPFSR, which consists of an encoder, a GAN Prior Bank and a decoder and can exploit GAN and facial priors to recover facial details by means of only one forward pass. Given a severely downsampled LR image, we obtain the image I^{BIC} of the I^{LR} bicubic upsampled to half the

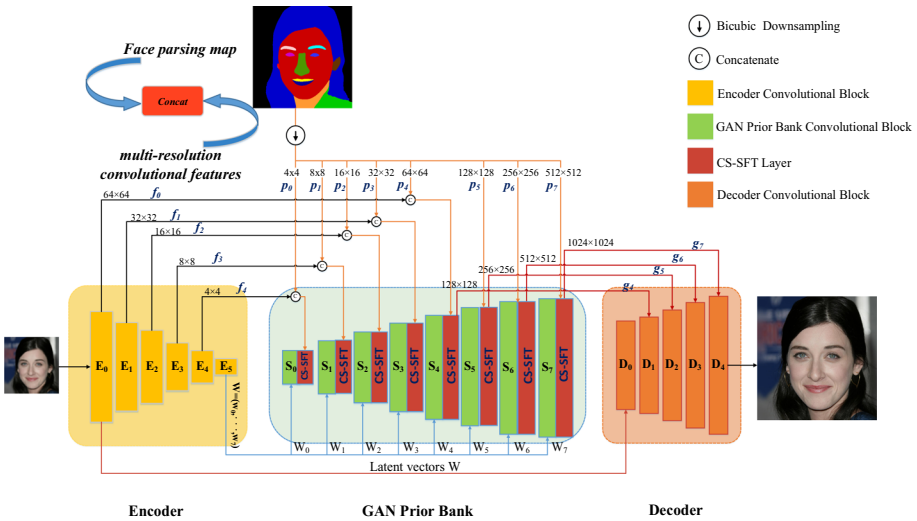

Fig. 1 Overall framework of the proposed CFGPFSR, which consists of an encoder, a GAN Prior Bank and a decoder. The Encoder extracts multi-resolution features f_i and latent vectors w_k. The GAN Prior Bank is modulated by multi-resolution features f_i, the latent vectors w_k and multi-resolution face parsing maps p_k to gererate multi-resolution features g_k. The SR face image is produced by utilizing the decoder to integrate the features f_i and g_k. This example corresponds to the 16 × SR task

size of the I^{HR} and extract the face parsing map P of the I^{BIC} from the pre-trained face parsing network. Firstly, the encoder is used to extract multi-resolution convolutional features and latent vectors from the I^{LR}. In this process, the encoder captures significant high-level cues together with the spatial structure of the I^{LR}. After that, such cues and the face parsing map are utilized through CS-SFT layers to modulate the GAN Prior Bank resulting in another set of convolutional features. Finally, CFGPFSR produces the SR face images by utilizing the decoder, which integrates the convolutional features from the encoder and the GAN Prior Bank. It is worth noting that the SR face images are generated by an additional decoder rather than the GAN Prior Bank, which allows the GAN Prior Bank to be more focused on the generation of details and further strengthen I^{SR} quality. In this work, a pre-trained face StyleGAN2 is employed as the GAN Prior Bank because of its excellent performance.

3.2 Encoder

In order to gain the latent vectors, an RRDBNet [32] are firstly employed to obtain features f_0 from the input, which indicates as E_0. Afterwards, the resolution of the f_0 is decreased progressively by:

$$f_i = E_i(f_{i-1}), i \in \{1, \ldots, N\}, \tag{1}$$

where E_i, i $\in \{1, \cdots, \mathrm{N}\}$, indicates an encoder convolutional block consisting of a convolution with stride-2 and a convolution with stride-1. N denotes the number of stacks. Finally, we obtain latent vectors by using a convolution and a fully connected layer.

$$W = E_{N+1}(f_N), \tag{2}$$

where W denotes a matrix, in which its columns indicate the latent vectors of StyleGAN2.

The input image compressed representation is captured from the latent vectors in W, which can provide high-level information for the GAN Prior Bank. Multi-resolution convolutional features $\{f_i\}$ and latent vectors W are further utilized to condition the GAN Prior Bank.

3.3 GAN Prior Bank

A pre-trained face StyleGAN2 is employed as the GAN Prior Bank, which provides priors for generating texture and detail. The latent vectors W from the encoder are fed into the GAN Prior Bank, in which the different latent vectors are employed for each convolutional block of the GAN Prior Bank to enhance expressiveness, rather than using only one latent vector as the input. To be more specific, we have $W = (w_0, \cdots, w_{k-1})$ for k blocks of the GAN Prior Bank, where each $w_k, k \in (0, \cdots, k-1)$ corresponds to one latent vector. After that, the GAN Prior Bank produces intermediate convolutional features for each resolution scale.

$$j_k = \begin{cases} S_0(w_0), \text{ if k = 0,} \\ S_k(w_k, j_{k-1}), \text{ otherwise,} \end{cases} k \in \{0, \cdots, k-1\}, \tag{3}$$

where S_k denotes the style convolutional block of StyleGAN2, and j_k corresponds to the output feature of the k-th style convolutional block.

In order to preserve spatial information from the input face image and better maintain the SR results fidelity, multi-resolution convolutional features $\{f_i\}$ and face parsing maps P are used to further modulate the intermediate features $\{j_k\}$ of the GAN Prior Bank by spatial feature transform (SFT) [33]. The SFT produces affine transformation parameters for spatial

feature modulation, which has demonstrated excellent performance in image generation and image restoration [34, 35] by combining other conditions. Specifically, we first construct a set of multi-resolution face parsing maps $\{p_k\}$, which denotes that P are resized to the same size as $\{j_k\}$ through bicubic interpolation. In addition, we concatenate multi-resolution convolutional features $\{f_i\}$ and multi-resolution face parsing maps $\{p_k\}$ to gain a set of affine transformation parameters (α_k, β_k) for all resolution scale by several convolutional layers. If there are no convolutional features for the corresponding resolution, the (α_k, β_k) are obtained only by using face parsing map $\{p_k\}$. Afterwards, we use (α_k, β_k) to modulate the intermediate features $\{j_k\}$ of the GAN Prior Bank by scaling and shifting. For the above operation, the Spatial Feature Transformation (SFT) method is formulated by:

$$\alpha_k, \beta_k = \begin{cases} \text{Conv}(\text{Concat}[f_{N-k}, p_k]), if k < N, \\ \text{Conv}(p_k), \text{otherwise}, \end{cases} \quad k \in \{0, \ldots, k-1\}, \tag{4}$$

$$g_k = \text{SFT}(j_k|\alpha_k, \beta_k) = \alpha_k \odot j_k + \beta_k, \tag{5}$$

where g_k denotes intermediate features of the GAN Prior Bank that have been modulated through the SFT and Concat$[\cdot, \cdot]$ denotes the concatenation operation.

In order to obtain a better balance between realness and fidelity for the SR results, we further use the Channel-Split Spatial Feature Transform (CS-SFT) layers [14]. As shown in Fig. 2, the intermediate features $\{j_k\}$ of the GAN Prior Bank are split into two parts $\left\{j_{\text{k}}^{\text{split0}}\right\}$ and $\left\{j_{\text{k}}^{\text{split1}}\right\}$ in the channel dimension. After that, a part of the intermediate features $\left\{j_{\text{k}}^{\text{split1}}\right\}$ of the GAN Prior Bank are modulated in the CS-SFT layers by multi-resolution convolutional features $\{f_i\}$ and multi-resolution face parsing maps $\{p_k\}$, which contributes to fidelity. This is because the modulated features carry more identity information. Another part of the features $\left\{j_{\text{k}}^{\text{split0}}\right\}$ are preserved, which contributes to realness. This is because it retains the texture and detail information generated by the GAN Prior Bank. Finally, the two parts of the feature are concatenated in the channel dimension. The Channel-Split Spatial

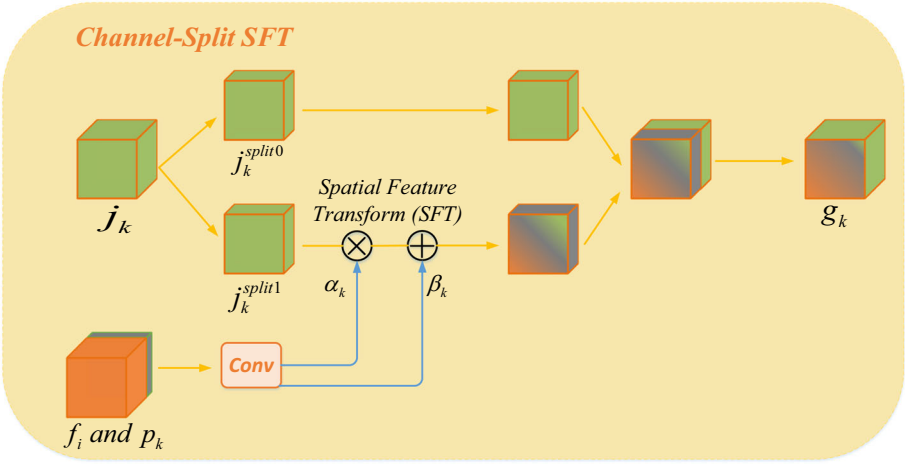

Fig. 2 Channel-Split Spatial Feature Transform (CS-SFT) layers. Half of the intermediate features j_k of the GAN Prior Bank are modulated by affine transformation parameters (α_k, β_k) generated by convolutional features f_i and face parsing maps p_k, and another half are directly passed through

Feature Transform (CS-SFT) method is formulated by:

$$\begin{aligned} g_k =& \mathrm{CS-SFT}(j_k|\alpha_k, \beta_k) \\ =& \mathrm{Concat}\Big[\mathrm{Identity}\Big(j_k^{\mathrm{split0}}\Big), \alpha_k \odot j_k^{\mathrm{split1}} + \beta_k\Big], \end{aligned} \tag{6}$$

where j_k^{split0} and j_k^{split1} are split features from intermediate features of the GAN Prior Bank in channel dimension.

By using CS-SFT layers, we incorporate the GAN and facial priors into the FSR process so that SR face images not only have a clear face structure, but also achieve a promising balance between fidelity and realness. In contrast to using SFT, CS-SFT also decreases the complexity of our model because it requires modulating fewer channels. In addition, for better fusion of the features from the GAN Prior Bank and the encoder, the features $\{g_k\}$ are output and further passed to the decoder, rather than directly generate the final result from the GAN Prior Bank.

3.4 Decoder

Our CFGPFSR employs an extra progressive fusion decoder to combine the features from the encoder and GAN Prior Bank for generating the SR face images. The RRDBNet features f_0 are taken as the inputs of the decoder. Moreover, the decoder features are gradually fused with the features $\{g_k\}$ from the GAN Prior Bank:

$$d_i = \begin{cases} D_0(f_0) \mathrm{if} i = 0, \\ D_i(d_{i-1}, g_{N-1+i}) \mathrm{otherwise}, \end{cases} \tag{7}$$

where D_i is a 3×3 convolution and d_i denote its output. Moreover, a pixel shuffle layer [2] is existed after each convolutional layer besides the final output layer. The information captured by the encoder can be enhanced by skip-connection between the encoder and decoder, which allows the GAN Prior Bank to be more focused on the generation of texture and detail.

3.5 Objective Functions

The joint optimization of different objective functions makes the model training converge faster and better. In this work, we apply MSE loss, perceptual loss and adversarial loss to train our CFGPFSR.

MSE Loss: To achieve better performance, we apply MSE loss to optimize the SR face image to achieve the fidelity, which can be formulated as:

$$\mathcal{L}_{\mathrm{mse}} = \frac{1}{N} \left\| I^{\mathrm{HR}} - I^{\mathrm{SR}} \right\|_2^2, \tag{8}$$

where N denotes the number of pixels, I^{HR} and I^{SR} respectively correspond to the ground-truth image and the SR result.

Perceptual Loss: For obtaining the SR face images with better perceptual quality, we further utilize perceptual loss [36], which uses a pre-trained VGG16 network [37] to extract high level features from SR and HR images. The perceptual loss formulated as:

$$\mathcal{L}_{\mathrm{percep}} = \frac{1}{N} \left\| f(I^{\mathrm{HR}}) - f(I^{\mathrm{SR}}) \right\|_2^2, \tag{9}$$

where $f(\cdot)$ represents the feature embedding space of the VGG16 network.

Adversarial Loss: For further improving the realness of the SR face images, we also use adversarial loss [38] to enhance the expressiveness of the SR face images. The loss functions of the generator G and discriminator D are formulated as:

$$\mathcal{L}_{\mathrm{g}} = \log\big(1 - D\big(I^{\mathrm{SR}}\big)\big), \tag{10}$$

$$\mathcal{L}_{d} = \log\big(1 - D(I^{\mathrm{SR}})\big) + \log D\big(I^{\mathrm{HR}}\big), \tag{11}$$

where D corresponds to the StyleGAN [31] discriminator.

Overall Loss: Thus, the final objective function of our CFGPFSR is as follows:

$$\mathcal{L}_{\mathrm{overall}} = \mathcal{L}_{\mathrm{mse}} + \alpha_{\mathrm{percep}} \cdot \mathcal{L}_{\mathrm{percep}} + \alpha_{g} \cdot \mathcal{L}_{g}. \tag{12}$$

where α_{percep} and α_{g} are trade-off parameters.

4 Experiments

A pre-trained face StyleGAN2 [13] is adopted as our GAN Prior Bank. During the training of CFGPFSR, we fix the weights of the GAN Prior Bank (from S_0 to S_7) and only update the weights of the encoder (from E_0 to E_5), decoder (from D_0 to D_4) and CS-SFT layers, which contributes to better utilization of the GAN prior and avoids the deviation of the training distribution.

4.1 Datasets and Implementation

Datasets and evaluation metrics. The proposed CFGPFSR is trained on the FFHQ dataset [31]. This dataset consists of 70 k high-quality face images with the resolution of 1024 × 1024 crawled from the internet. A synthetic dataset and a real dataset are constructed as our test datasets. For the synthetic test dataset, we randomly select 1000 images from the CelebA-HQ dataset. For the real test dataset, we collect 50 old photos from the Internet, detect the faces in them, and crop and align them. We use bicubic downsampling in matlab to generate the LR inputs with the resolution of 64 × 64 for the train and test dataset. For evaluation, we use not only pixel-level metrics (PSNR and SSIM) and perceptual metrics (LPIPS [39]), but also non-reference perceptual metrics (FID [40] and NIQE [41]) that are already widely used. Furthermore, the cosine similarity on the ArcFace [42] embedding space is also calculated as an evaluation metric.

Implementation. For training, we set mini-batch size to 1 and the initial learning rates of the generator and discriminator to 5×10^{-5}. The weights α_{percep} and α_{g} are both set to 10^{-2} and the training datas are augmented by means of horizontal flip. In addition, Cosine Annealing Scheme [43] and Adam optimizer [44] (β_1=0.9, $\beta_2 = 0.99$) are adopted to train our model for a total of 500 k iterations. We exercise the PyTorch framework to implement our model and use two gpus (a NVDIA TITAN Xp and a NVDIA TITAN X) for training.

4.2 Comparison on Synthetic Dataset

We compare our CFGPFSR with several state-of-the-art methods on the synthetic dataset. These methods include PULSE and mGANprior, which are based on GAN inversion, and GPEN and GLEAN, which are based on GAN prior. For the 16 × SR task, we give quantitative and qualitative results. In order to make a fair comparison, all methods are trained on the

Table 1 Quantitative comparison with state-of-the-art methods on the synthetic dataset for 16 × SR task. *Italic* represents the optimal result and **bold** represents the sub-optimal result. Similarity indicates cosine similarity of ArcFace Embeddings. We do not test the NIQE of PULSE and mGANprior because their results difficultly preserve the identity, which could result in our quantitative comparison to be interfered by their NIQE metrics

Method	$16 \times \left(64^2 \rightarrow 1024^2\right)$					
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	NIQE ↓	Similarity ↑
PULSE [10]	21.46	0.5963	0.4588	37.63	-	0.4049
mGANprior [9]	22.11	0.6379	0.4939	65.90	-	0.5870
GPEN [15]	25.75	**0.6744**	0.3691	26.34	5.04	0.9361
GLEAN [12]	**26.50**	0.6715	**0.2814**	*13.94*	**4.42**	*0.9616*
CFGPFSR(Ours)	*26.63*	*0.6798*	*0.2761*	**14.79**	*4.36*	**0.9582**

same dataset by using the code already released by the above methods and tested on the same test dataset. During training, the GAN prior parameters of GLEAN are fixed, while the GAN prior parameters of GPEN are provided with a learning rate of 2×10^{-4}, which is identical to [15] settings.

The quantitative results on the synthetic dataset are provided in Table 1. It is shown that the proposed CFGPFSR obtains the optimal result in PSNR, SSIM, LPIPS and NIQE, and the sub-optimal result in FID and cosine similarity on the ArcFace [42] embedding space. This indicates that CFGPFSR can generate images with higher quality than other contrastive methods. It is noteworthy that our CFGPFSR achieves better performance than GLEAN that only utilizes multi-resolution convolution features to condition the intermediate features of the GAN prior, which shows that rational use of face parsing maps can help improve FSR performance. On the other hand, since the simple latent code exploration strategy makes it difficult to maintain the identity of the generated face images, the PULSE and mGANprior based on GAN inversion obtain noticeably poorer results than the GAN-prior-based methods. It needs to be stated that we do not test the NIQE of PULSE and mGANprior because their outputs difficultly preserve the identity, which could result in our quantitative comparison to be interfered by their NIQE metrics.

Figure 3 provides some qualitative results on the synthetic dataset. It is shown that GAN inversion methods are unable to preserve the identity of the generated images, while the GAN-prior-based methods can generate images with good fidelity. However, the results of GPEN differ significantly from the ground-truths in terms of local texture details, such as: the eye area of the first row result and the tooth area of the fifth row result. Compared with GPEN, the results of GLEAN are closer to the ground-truths. However, its GAN prior is modulated by the multi-resolution convolution features only through a simple concatenation operation, which may not make sufficient utilization of the extracted information, resulting in the performance of the model being affected. The proposed CFGPFSR obtains the closest result to the ground-truths in texture details by using multi-resolution convolutional features and multi-resolution face parsing maps more effectively modulated GAN Prior through the designed CS-SFT layers. In a word, our CFGPFSR can achieve the highest quality of fidelity and realness among all comparison methods.

Comparison with reference-based method. The use of the GAN Prior Bank is reminiscent of the reference-based SR task [34, 45], where an explicit image dictionary is constructed by employing external HR reference images. DFDNet [34] is a representative reference-based

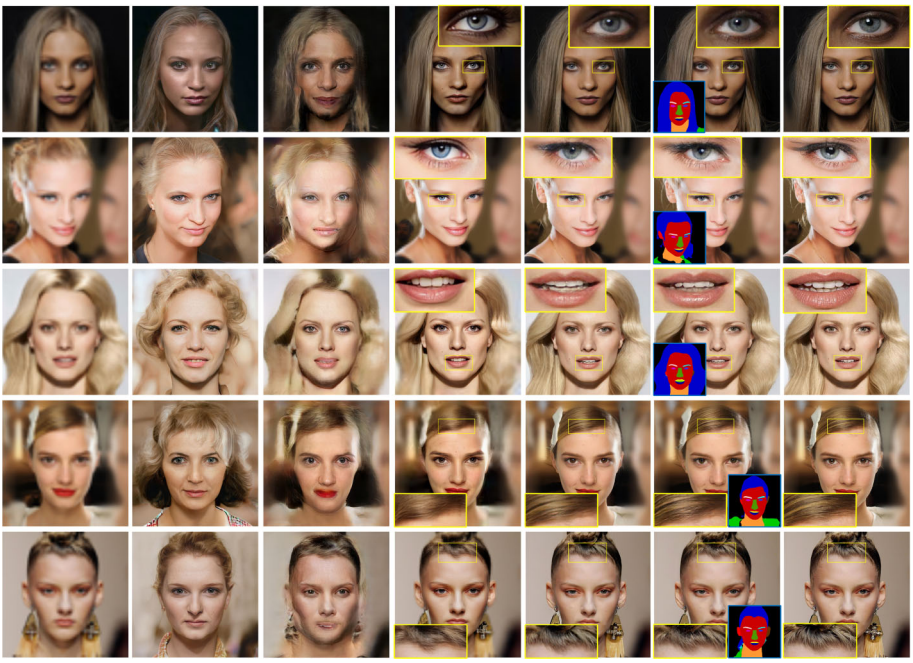

Fig. 3 Qualitative comparison on the synthetic dataset for 16 × SR. The face parsing maps predicted by pre-trained FPN are overlapped at the corner of our results. The resolution of ground-truth (GT) image is 1024 × 1024

method. It pre-constructed a dictionary of face components (such as eyes, noses) and then performed face restoration, showing significant results. In contrast to reference-based methods, our CFGPFSR uses a GAN-based dictionary with a pre-trained face StyleGAN2 as condition rather than constructing an image dictionary, which does not depend on any specific components or images. We compare our CFGPFSR with DFDNet on the synthetic dataset to assess the effectiveness of GAN Prior Bank, which is shown in Fig. 4. DFDNet cannot generate faithful results for parts of the dictionary that are not, its outputs differ significantly from the ground-truth images in the skin and hair regions. Unlike DFDNet, our CFGPFSR produce coherent and pleasing results, which are not limited to the enhancement of the expressiveness of specific components.

Robustness to Poses. Our CFGPFSR also demonstrates robustness to diverse poses, which is shown in Fig. 5. We compare the performance of our CFGPFSR with PULSE on some facial images with different poses, which we filter from the synthetic dataset. Guided by the multi-resolution convolution features and multi-resolution face parsing maps, CFGPFSR is still capable of reconstructing non-aligned face images with realistic. On the contrary, the reconstruction results of PULSE tend to align face images and its results are only able to approximate the ground-truths at low resolution.

Fig. 4 Comparison with DFDNet. DFDNet cannot generate faithful results for parts of the dictionary that are not (*e.g.* skin, hair)

4.3 Comparison on Real World Images

To further demonstrate the effectiveness of the proposed CFGPFSR, we compare CFGPFSR with methods in Table 1 on the real dataset. Table 2 provides the quantitative results. Since real images do not have the ground-truth images, we only evaluate non-reference perceptual metrics (NIQE and FID) for various methods. It can be known from Table 2 that our CFGPFSR obtains the best performance in NIQE, and the second best result in FID that is close to GLEAN. The FID of the GAN prior methods is still much smaller than that of the GAN inversion methods.

Figure 6 provides some qualitative results of various methods on the real dataset. As we can see, PULSE and mGANprior are still difficult to preserve the identity and construct images with high fidelity. Compared with GPEN and GLEAN, the proposed CFGPFSR generates more natural facial images with less artifacts. This is because our CFGPFSR combines facial and GAN priors for FSR, in which the GAN Prior Bank provides rich and varied priors for reconstructing faces and multi-resolution convolutional features and multi-resolution face parsing maps obtained from the input image are used to further modulate GAN Prior Bank through CS-SFT layers. This allows the network to be progressively guided to generate finer texture details for each semantic region and to improve the identity of the image. Therefore, the generated face images obtain a promising balance in realness and fidelity. In addition, the face parsing map results in the last column indicate that the pre-trained FPN also performs well for real inputs.

4.4 Ablation Study

To explore the effectiveness of the proposed CFGPFSR, an ablation study is further implemented on the synthetic dataset. Our base model is denoted as the baseline that only uses the latent vectors $\{w_k\}$ to condition the GAN Prior Bank. Furthermore, we also denote a model called CS-SFT_feature on the baseline, which further modulates the intermediate features

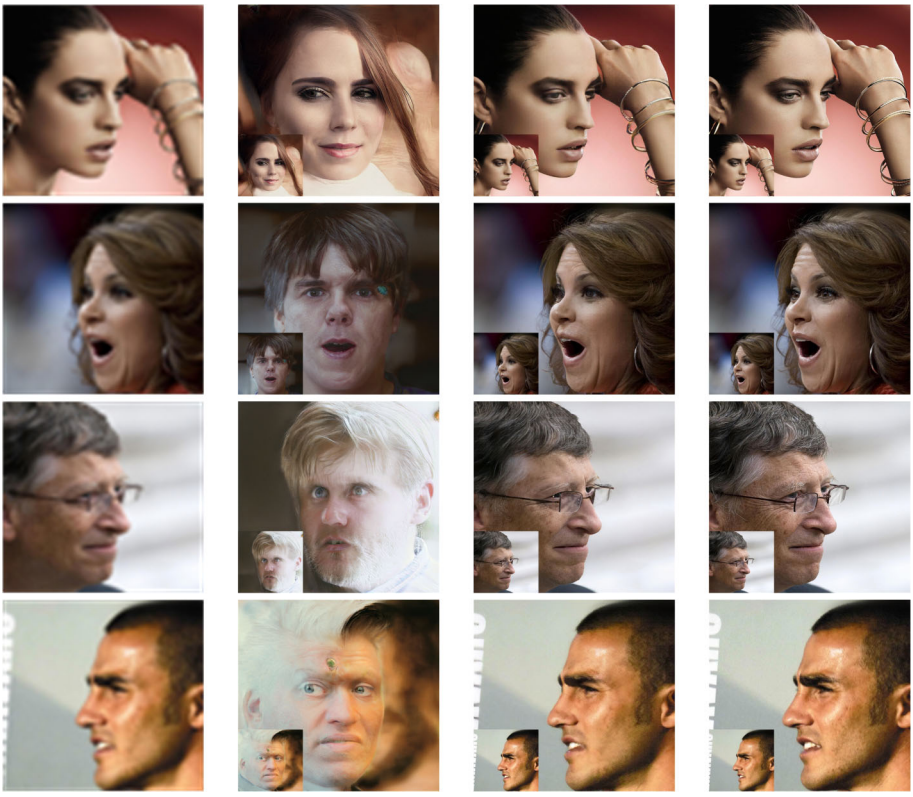

Fig. 5 Reconstruction results with diverse poses. The results of PULSE are noticeably different from the GT at high resolution and are close to the ground-truths only at low resolution *(bottom left)*. Unlike PULSE, although CFGPFSR is trained on aligned face images, it is still capable of reconstructing realistic non-aligned face images

Table 2 Quantitative comparison with state-of-the-art methods on the real dataset. *Italic* represents the optimal result and **bold** represents the sub-optimal result. Similarly to Table 1, we do not test the NIQE of PULSE and mGANprior

	PULSE	mGANprior	GPEN	GLEAN	Ours
NIQE ↓	-	-	5.69	**5.34**	*5.16*
FID ↓	114.61	119.01	56.31	*44.77*	**44.89**

of the GAN Prior Bank using multi-resolution convolution features $\{f_i\}$ through the CS-SFT layer. Our CFGPFSR is conditioned jointly by the latent vectors $\{w_k\}$, multi-resolution convolution features $\{f_i\}$ and multi-resolution face parsing maps $\{p_k\}$. All variants in our ablation study are trained for the same number of stages.

As shown in Table 3, the performance of the model is improved with the gradual addition of the multi-resolution convolution features and the multi-resolution face parsing maps. Our

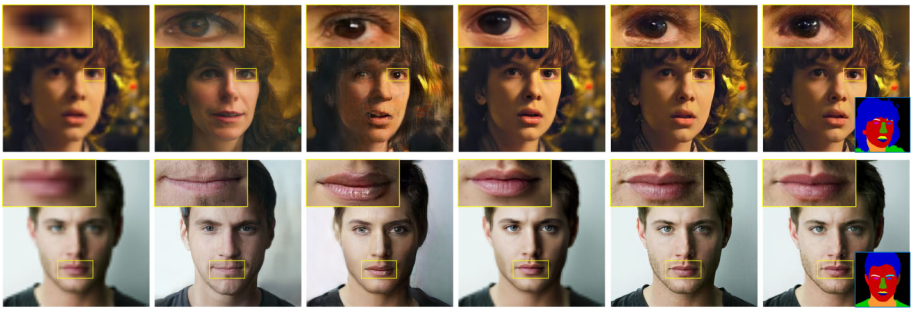

Fig. 6 Qualitative comparison on the real dataset. Our CFGPFSR could generates more natural facial images with less artifacts. The face parsing map results in the last column indicate that the pre-trained FPN also performs well for real inputs

Table 3 Quantitative comparison with different variants on the synthetic dataset. *Italic* represents the optimal result and **bold** represents the sub optimal result

Model	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	NIQE ↓
Baseline	26.41	0.6651	0.2813	16.08	4.57
CS-SFT_feature	**26.54**	**0.6747**	**0.2789**	**15.23**	**4.46**
CFGPFSR	*26.63*	*0.6798*	*0.2761*	*14.79*	*4.36*

CFGPFSR achieves the optimal performance and CS-SFT_feature achieves the sub-optimal performance. In addition, some qualitative results of each variant are presented in Fig. 7. We can see that the baseline is difficult to reconstruct face images with faithful local detail. This is because the spatial information is inadequate preservation, and the finer detail information is difficultly recovered by guided only by low-dimensional vectors. The details become more accurate with the progressive addition of multi-resolution convolution features and multi-resolution face parsing maps, which can modulate the intermediate features of the GAN Prior Bank via the designed CS-SFT layers, such as the eye regions in the first and second rows. The experiments demonstrate the effectiveness of the proposed CFGPFSR, which modulates the GAN Prior Bank intermediate features to optimize the output by using multi-resolution convolution features and multi-resolution face parsing maps via CS-SFT layers.

4.5 Limitation

Thought the proposed method is capable of recovering SR images of faces at a large upsampling factor, it also suffers from some limitations. The establishment of our network architecture is based on the traditional model with massive face samples. Hence, it has the same problems as the existing advanced face SR methods, which can significantly degrade the performance of face SR methods when the test image and the training dataset have distributional gaps.

As shown in Fig. 8, if our training dataset lacks face images with different rotational angles, the performance of the proposed CFGPFSR can be degraded when the input face

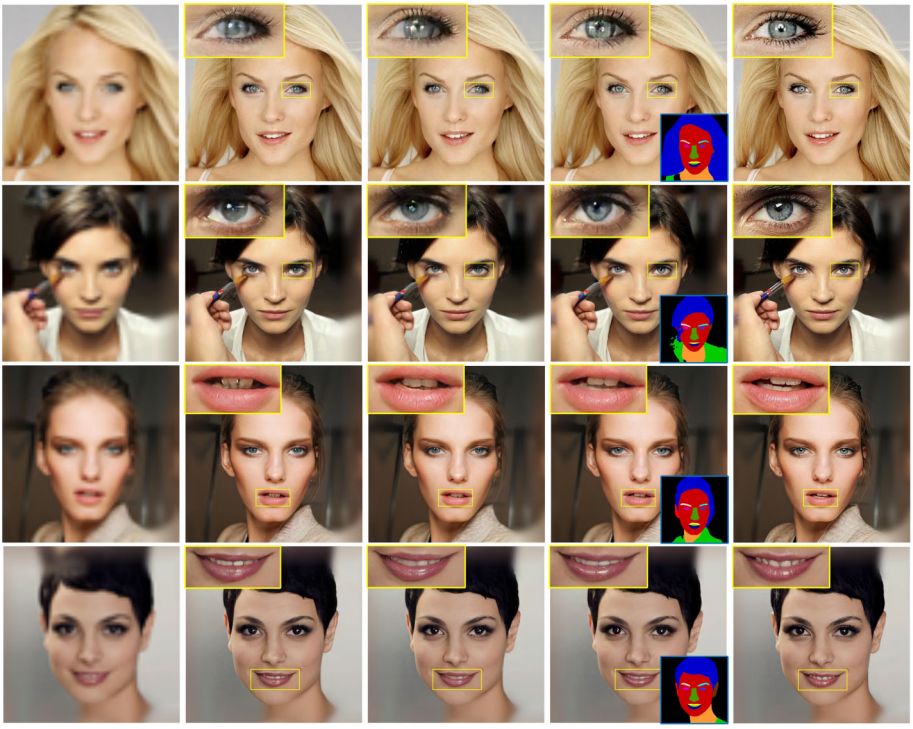

Fig. 7 Qualitative comparison with different variants on the synthetic dataset. With the addition of multi-resolution convolution features and multi-resolution face parsing maps, the generated details are more accurate

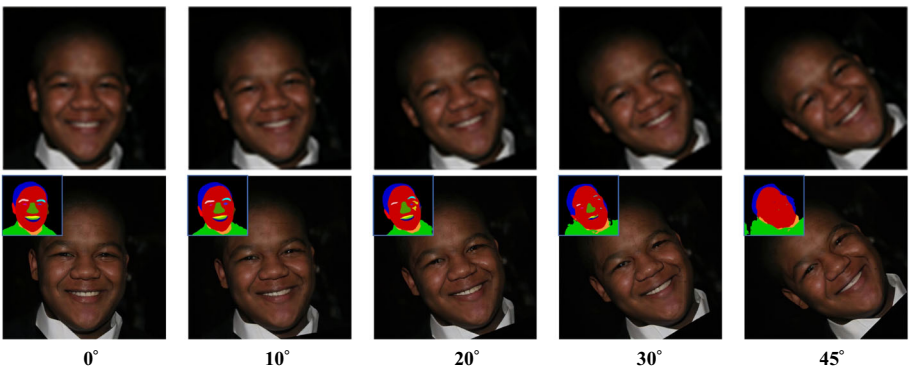

Fig. 8 Reconstruction results after rotating the face LR image by several angles. The first line represents the rotated LR input, the second line represents the result after reconstruction by the proposed CFGPFSR, and the third line indicates the angle of rotation

Fig. 9 An example of failure of the proposed method on a occluded facial LR image. The masked portion of the reconstructed face image fails to produce a pleasing visual effect

images are rotated by several angles (10°, 20°, 30° and 45°). Along with the gradual increase in the rotation angle of the input image, the recovered facial details become progressively unrealistic, such as the teeth and eye regions in the second row, and the extracted face parsing maps become progressively inaccurate. Meanwhile, when the input facial image is masked, the extracted face paring map may be inaccurate and the reconstructed SR face image may show artifacts in the masked portion, which cannot produce a pleasing visual effect, as shown in Fig. 9.

In the future, we will be concerned with developing a more robust network that breaks through the limitations of the proposed method.

5 Conclusion

In this paper, we propose a novel generative FSR framework combining facial and GAN priors named CFGPFSR. Firstly, we pre-train a face StyleGAN2 with rich and diverse priors and a face parsing network (FPN) that can generate decent parsing maps. After that, we exploit the prior informations encapsulated in the face StyleGAN2 (GAN prior) and face parsing maps extracted from the FPN (facial prior) for face super-resolution. To further enhance FSR performance, GAN and facial priors are introduced into the FSR process through Channel-Split Spatial Feature Transform (CS-SFT) layers so that the SR results can be more natural and a good balance is achieved between fidelity and realness. The proposed CFGPFSR can jointly recover facial details utilizing only a single forward pass without expensive image optimization at runtime. Both quantitative and qualitative experiments on synthetic and real test datasets demonstrate the effectiveness of the proposed CFGPFSR. The performance of our method is somewhat influenced by the face and GAN priors. In the future work, we will explore more powerful facial and GAN priors and develop more effective feature fusion methods to improve FSR performance and generate more realistic facial super-resolution results. Meanwhile, we will attempt to construct a dataset considering the face rotation problem and utilize state-of-the-art techniques for solving face masking, which can overcome the limitations of the proposed method.

6 Data Availability Statement

All data can be obtained for scientific research.

Acknowledgements This work was partly supported by NSFC of China (U1504610, 61971339).

Author contributions Jinbo Liu: Conceptualization, Methodology, Software, Investigation, Writing - Original Draft. Zhonghua Liu: Resources, Writing - Review & Editing. Weihua Ou: Writing: Review & Editing. Kaibing Zhang: Writing: Review & Editing. Yong Liu: Writing: Review & Editing.

Funding NSFC of China,U1504610,61971339

Declarations

Competing interests The authors declare no competing interests.

Conflicts of Interest The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Dun Y, Da Z, Yang S et al (2021) Kernel-attended residual network for single image super-resolution. Knowl-Based Syst 213:106663
2. Shi W, Caballero J, Huszár F, et al. (2016) 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 (CVPR), pp. 1874–1883
3. Chen Y, Tai Y, Liu X, et al. (2018) Fsrnet: End-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 2492–2501
4. Yu X, Fernando B, Ghanem B, et al. (2018) Face super-resolution guided by facial component heatmaps. In: European conference on computer vision (ECCV), pp. 217–233
5. Li M, Zhang Z et al (2021) Learning face image super-resolution through facial semantic attribute transformation and self-attentive structure enhancement. IEEE Trans Multimedia 23:468–483
6. MaC, Jiang Z, Rao Y, et al. (2020) 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 (CVPR), pp. 5569–5578
7. Xin J, Wang N, Jiang X et al (2020) Facial attribute capsules for noise face super resolution. Proc Assoc Adv Artif Intell 34(7):12476–12483
8. Yu X, Fernando B, Hartley R et al (2020) Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes. IEEE Trans Pattern Anal Mach Intell 42(11):2926–2943
9. Gu J, Shen Y, Zhou B (2020) Image processing using multi-code gan prior. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 3009–3018
10. Menon S, Damian A, Hu S, et al. (2020) 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 (CVPR), pp. 2434–2442
11. Pan X, Zhan X, et al. (2020) Exploiting deep generative prior for versatile image restoration and manipulation. In: European conference on computer vision (ECCV), pp. 23–28

12. Chan KCK, Wang X, Xu X, et al. (2021) Glean: Generative latent bank for large-factor image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 14245–14254
13. Karras T, Laine S, Aittala M, et al. (2020) Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 8110–8119
14. Wang X, Li Y, et al. (2021) Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 9168–9178
15. Yang T, Ren P, Xie X, et al. (2021) Gan prior embedded network for blind face restoration in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 672–681
16. Baker S, Kanade T (2000) Hallucinating faces. In: Proceedings fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580), pp. 83–88
17. Wang X, Tang X (2005) Hallucinating face by eigentransformation. IEEE Trans Syst Man Cyber Part C 35(3):425–434
18. Liu C, Shum H-Y, Freeman WT (2007) Face hallucination: Theory and practice. Int J Comput Vis (IJCV) 75(1):115–134
19. Ma X, Zhang J, Qi C (2010) Hallucinating face by position-patch. Pattern Recogn (PR) 43(6):2224–2236
20. Ma X, Zhang J, Qi C (2010) Hallucinating face by position-patch. Pattern Recogn 43(6):2224–2236
21. SongY, Zhang J, He S, et al. (2017) Learning to hallucinate face images via component generation and enhancement. In: International joint conference on artificial intelligence (IJCAI), pp. 4537–4543
22. Jing Y, Yang Y, Wang X, et al. (2021) Amalgamating knowledge from heterogeneous graph neural networks. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 15704–15713
23. Q. Cao, L. Lin, Y. Shi, et al. Attention-aware face hallucination via deep reinforcement learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1656–1664
24. Chen X, Wang X, Yao Lu et al (2020) Rbpnet: An asymptotic residual back-projection network for super-resolution of very low-resolution face image. Neurocomputing 376:119–127
25. Jiang K, Wang Z, Yi P et al (2022) Dual-path deep fusion network for face image hallucination. IEEE Trans Neural Netw Learn Syst 33(1):378–391
26. Chen C, Gong D, Wang H et al (2021) Learning spatial attention for face super-resolution. IEEE Trans Image Process 30:1219–1231
27. Jing Y, Yang Y, Wang X, et al. (2021) Meta-aggregator: learning to aggregate for 1-bit graph neural networks. In: IEEE/CVF international conference on computer vision (ICCV), pp. 5281–5290
28. Jing Y, Yuan C, Ju L, et al. (2023) Deep graph reprogramming. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 24345–24354
29. Gao G, Tang L, Fei Wu et al (2023) Jdsr-gan: constructing an efficient joint learning network for masked face super-resolution. IEEE Trans Multimed 25:1505–1512
30. Wang H, Qian Hu, Chengdong Wu et al (2021) Dclnet: Dual closed-loop networks for face super-resolution. Knowl-Based Syst 222:106987
31. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 4401–4410
32. Wang X, Yu K, Wu S, et al. (2018) Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV) workshops, Springer, pp. 63–79
33. Wang X, Yu K, Dong C, et al. (2018) Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 606–615
34. Li X, Chen C, Zhou S, et al. (2020) Blind face restoration via deep multi-scale component dictionaries. In: European conference on computer vision (ECCV), pp. 399–415
35. Park T, Liu M-Y, et al. (2019) Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 2332–2341
36. Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision (ECCV), Springer, pp. 694–711
37. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations (ICLR), pp. 1–14
38. Goodfellow I, Pouget-Abadie J, Mirza M, et al. (2014) Generative adversarial nets. In: Proceedings of the 27th international conference on neural information processing systems, pp. 2672–2680
39. Zhang R, Isola P, Efros AA, et al. (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 586–595

40. Heusel M, Ramsauer H, Unterthiner T, et al. (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st international conference on neural information processing systems, Vol. 30. pp. 6629–6640
41. Mittal A, Soundararajan R, Bovik AC (2012) Making a "completely blind" image quality analyzer. IEEE Signal Process Lett 20(3):209–212
42. Deng J, Guo J, Xue N, et al. (2019) Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 4690–4699
43. Loshchilov I, Hutter F (2016) SGDR: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983
44. Kingma D, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
45. Dehkordi RA et al (2020) Single image super-resolution based on sparse representation using dictionaries trained with input image patches. IET Image Process 14:1587–1593