Generative Facial Prior for Large-Factor Blind Face Super-Resolution

Xiaomeng Guo¹², Li Yi², Hang Zou¹² and Yining Gao¹²

¹Communication and Information System, Wuhan Research Institute of Posts and Telecommunications, Wuhan, Hubei, 430000, China
²IAO, Nanjing Fenghuo Tiandi Communication Technology Co., Ltd, Nanjing, Jiangsu, 210000, China

*Corresponding author’s e-mail: dxmgeneral@gmail.com

Abstract. Most existing face super-resolution (SR) methods are developed based on an assumption that the degradation is fixed and known (e.g., bicubic down sampling). However, these methods suffer a severe performance drop in various unknown degradations in real-world applications. Previous methods usually rely on facial priors, such as facial geometry prior or reference prior, to restore realistic face details. Nevertheless, low-quality inputs cannot provide accurate geometric priors while high-quality references are often unavailable, which limits the use of face super-resolution in real-world scenes. In this work, we propose GPLSR which used the rich priors encapsulated in the pre-trained face GAN network to perform blind face super-resolution. This generative facial prior is introduced into the face super-resolution process through channel squeeze-and-excitation spatial feature transformation layer (SE-SFT), which makes our method achieve a good balance between realism and fidelity. Moreover, GPLSR can restore facial details with single forward pass because of powerful generative facial prior information. Extensive experiment shows that when the magnification factor is 16, this method achieves better performance than existing techniques in both synthetic and real datasets.

1. Introduction

Face super-resolution (also known as face illusion) is an image super-resolution problem for specific fields, which aims to improve the resolution of one or a series of low resolution (LR) face images and generate corresponding high resolution (HR) face images. When it is applied to real scenes, it becomes more challenging due to more complex degradation. Due to the serious ambiguity of the image, in this case, the prior information becomes inevitable, especially when restoring texture details.

Previous studies usually use face-specific priors for face restoration, such as facial landmarks, parsing maps, and facial component heatmaps, and show that these geometric facial priors are essential to restore accurate face shapes and details. However, these priors are usually estimated from the input image, and errors inevitably occur as the quality of the input image decreases. In addition, although semantic guidance is provided for the super-resolution process, the semantic information contained in the above-mentioned priors is not sufficient to guide the restoration of facial details.

Another method is to take the information encapsulated in the pre-trained GAN model as the facial priors, such as StyleGAN2[2]. These facial GAN can generate faithful faces with high variability, so as to provide rich and diverse priors such as geometry, facial texture, making it possible to jointly restore facial details. However, it is a challenge to introduce this generative prior into the blind face super-resolution process. Previous attempts usually used Gan inversion [3, 4, 7]. They first "invert" the...
degraded image back to the latent code of the pretrained GAN, and then execute a specific optimization algorithm to reconstruct the image, which is often very time-consuming. Although the outputs have visual realism, they usually produce low fidelity images because low dimensional latent code is not enough to guide accurate super-resolution. GLEAN [1] uses an encoder to obtain potential code and multi-scale features from the image, and then uses multi-scale features to fuse with the features of StyleGAN2 through a convolution, but this leads to that all the prior features (features of StyleGAN2, contributing to realness) are affected by the input feature space (multi-scale features, contributing to fidelity). In particular, when the low-resolution picture is used as the input in the real scene, the "fuzzy" input feature will affect the modulation, and the output inevitably tends to the "fuzzy" result without real and rich facial details.

To solve this problem, we propose a more sophisticated feature fusion based on GLEAN to achieve a good balance between the authenticity and fidelity of the super-resolution results through a forward pass. Specifically, the GPLFSR consists of an encoder, a pre-training facial GAN as a facial prior, and a decoder. The overall construction is similar to GLEAN. They are connected through a direct latent code mapping and multiple spatial feature transformation layers with Squeeze-and-Excitation network (SE-SFT). The proposed SE-SFT layer spatially modulates one set of features and directly transfers another set of features to better preserve information, then through channel attention SE-layers. Our method can effectively combine to generate a prior while retraining high fidelity. In addition, we introduced identity preserving loss to further improve the fidelity of result.

Our contributions are summarized as follows:

(1) We use the rich and diverse generative face priors encapsulated in the pre-trained model to blind face super-resolution. These priors contain rich facial texture and details to enable us to achieve face super-resolution in real scenes.

(2) Based on the existing GLEAN structure, we introduce a more refined feature fusion method and attention mechanism (SE-SFT), and add a new loss. GPLFSR achieves a good balance between realness and fidelity in single forward through SE-SFT layer.

(3) Experiments show that in $16 \times$ SR, our method achieves better performance than the existing technology on synthetic data sets.

2. Related work

2.1. Face Super-Resolution

The earliest face SR algorithm is based on the general SR algorithm to directly learn to map from low-resolution images to high-resolution images. Inspired by the powerful expressiveness of CNN, Zhou et al. [6] first adopted CNN and proposed a dual-channel convolutional neural network (BCCNN) to learn the mapping from LR face images to HR face images. Nie et al [5] adopted Face hallucination via convolution neural network (FHCNN) also applies pixel-wise loss function at every step in cascaded network.

Although these methods have achieved significant results in PSNR, training using only pixel-level constraints usually results in perceptually unconvincing output and severe over-smoothing. In order to solve this problem, GANs are used to approximate natural images, yielding more realistic results. Since Yu et al. [8] developed an original GAN-based facial super-resolution network and achieved good results, GAN-based face super-resolution has become more and more extensive.

In addition, face images have face-specific information, including facial landmarks, facial parsing maps, and face heatmaps. Therefore, researchers have introduced priors based on the uniqueness of the face. Chen et al. [9] predicted landmark heatmaps and parsing maps from LR faces, and then concatenated them with feature maps to adjust finally results. Yu et al. [10] uses the facial component heat map to maintain the facial structure while super-resolving LR faces.

However, their performance largely depends on accurate prior knowledge of the face, which is difficult to obtain from severely degraded face images in real-world, which will lead to unpredictable
failure. And they mainly focus on geometric constraints and cannot guide the restoration of sufficient details. In contrast, the face generation prior we use does not involve explicit geometric estimation of degraded images, and the pre-trained network contains rich information.

2.2. GAN Inversion

Recently, a lot of work has been developed for GAN inversion, that is, to invert a given image back to potential code using a pre-trained GAN model. Some existing methods are to continuously optimize the latent code, for example, PULSE[7] uses pixel constraints between input and output to iteratively optimize the latent code of StyleGAN2. mGANprior [3] optimizes multiple latent codes to increase the expressive ability of the model. DGP [11] further fine-tuned the generator and latent code to narrow the gap between the distribution of training images and test images. These methods not only require expensive iterative optimization, but the low-dimensional characteristics of the underlying coding cannot faithfully retain important spatial information. Therefore, these methods often produce unpredictable results and are not similar to ground truth.

Another kind of method is to train an additional encoder to project the image space back to the latent space [1, 12]. GLEAN[1] adopts a learning-based GAN inversion strategy to train an encoding network which transform from image X to latent vector z by optimizing the super-resolution results, and constrains it with multi-resolution features. GPEN[12] also adopted a similar idea to solve the problem of blind face super-resolution. It encodes the low-quality face image x into the z-space, and then transforms it into the w-space through nonlinear mapping as the input of the GAN model. The disadvantage of this type of method is that the prior features will be excessively affected by multi-scale features. Especially, when degraded images in real world are used as input, the "fuzzy" input features will affect the modulation, so that the output inevitably tends to the "fuzzy" results without real and rich facial details. Therefore, we need a more refined way to fuse prior features and input multi-scale features. So that in this work we adopted multiple spatial feature transformation layers and introduced channel attention mechanism, which makes the generated image achieve a balance between realness and fidelity.

3. Methodology

3.1. Overview

In this section, we will describe the GPLFSR framework, which uses the generated face prior encapsulated in the pre-trained generative adversarial network for blind face super-resolution. Given the input low-resolution face image, GPLSR will estimate the high-resolution and high-quality face image, which is as close to the ground truth as possible in terms of authenticity and fidelity.

In this work, our main framework is based on the existing encoder-latent bank-decoder structure, and replaces the original simple splicing by using a more delicate combination with the generative prior, so as to achieve blindness under large magnification factors. The framework of the network is depicted in figure 1, which consists of an encoder module, a pre-trained GAN network (StyleGAN2) and a decoder module. They are bridged by a latent code mapping and multiple channel squeeze excitation spatial feature transform (SE-SFT) layers. Specifically, the encoder module is designed to extract two kinds of features, i.e. 1) Latent features $F_{\text{latent}}$ used to map the input image to the closest latent code in StyleGAN2; 2) Multi-resolution spatial features $F_{\text{spatial}}$, used to modulate StyleGAN2 features.

Then the latent code $F_{\text{latent}}$ is mapped to the intermediate latent code $W$ through the linear layer. StyleGAN2 can generate intermediate convolution features $F_{\text{GAN}}$ through the latent code close to the input image. These features provide rich prior information through the weights of the pre-trained GAN network. The multi-resolution features $F_{\text{spatial}}$ use the SE-SFT layer to spatially modulate the facial GAN features $F_{\text{GAN}}$ from coarse to fine, so as to achieve realistic results while maintaining high fidelity. Then the shallow $F_{\text{GAN}}$ of StyleGAN2 makes a skip connect with the decoder to generate the final result.
Figure 1. Overview of GPLFSR framework. It consists of an encoder module, a pre-trained GAN network (StyleGAN2) and a decoder module. They are bridged by a latent code mapping and multiple channel squeeze excitation spatial feature transform (SE-SFT) layers. Among them, SE-SFT will use multi-resolution features to modulate the segmented prior information from stylegan2. Then connect the modulated part with the unmodulated part and pass the attention mechanism layer. Finally, enter the result into the next layer of stylegan2. This example corresponds to an input size of 64×64 and an output size of 1024×1024.

3.2. Encoder
In order to generate latent code and multi-resolution features, we use an RRDB-Net[13] and a series of convolutions as our encoder, and first use RRDB to extract the feature \( F_0 \) from the input LR image \( X \). Then we use a series of convolution blocks consisting of a stride-2 and a stride-1 to extract the multi-resolution feature \( F_{\text{spatial}} \) while reducing the feature resolution:

\[
F_{\text{spatial}} = E_i(F_{i-1}), \quad i \in \{1, \ldots, N\}
\]

(1)

Where \( E_i \) denotes convolution blocks.

Finally, we use convolution and fully connected layers to generate latent code \( F_{\text{latent}} \)

\[
F_{\text{latent}} = E_{N+1}(F_{N})
\]

(2)

Where \( F_{\text{latent}} \) is a matrix whose columns represent the latent code of StyleGAN2.

The overall formula is as follows:

\[
F_{\text{latent}}, F_{\text{spatial}} = E_{\text{coder}}(X)
\]

(3)

The latent vector \( F_{\text{latent}} \) is a compressed representation of the image, which provides high-level information for the latent generation library. However, due to the low dimensionality of the underlying code, important spatial information may not be faithfully retained. In order to further improve the fidelity...
of the super-resolution results and provide additional guidance for structural repair, we need to input the multi-resolution features $F_{\text{spatial}}$ into the pre-training StyleGAN2.

3.3. Decoder

Different from the direct use of StyleGAN2 to generate images in the past, we choose to use StyleGAN2 to generate features, and then use an additional decoder to generate super-resolution images. The decoder takes the features of the first layer of RRDB-Net[13] as input, and the remaining layers gradually fuse the shallow intermediate multi-resolution features of StyleGAN2:

$$d_i = \begin{cases} D_0(F_0) & i = 0 \\ D_i(d_{i-1}, F_{\text{GAN}}(N-2+i)) & \text{otherwise} \end{cases}$$

(4)

Where $D_i$ denote $3 \times 3$ convolution and $d_i$ is its output. In addition to the final output layer, there is a pixel shuffle[14] layer behind each convolution. The skip connection between the encoder and the decoder can enhance the information captured by the encoder, so that StyleGAN2 can pay more attention to the generation of textures and details to obtain better results.

3.4. Generative Facial Prior and Latent Code Mapping

The current face GAN [2,15] can generate realistic faces with a high degree of diversity and variability. We used the powerful face generation capabilities of pre-trained StyleGAN2[2] to provide rich facial details for our super-resolution work. However, this kind of generation prior cannot directly guide the super-resolution process, because the prior is encapsulated into the GAN network by mapping the random latent code $Z$ to the face image.

Specifically, after the latent space vector $F_{\text{latent}}$ of the image is obtained by the encoder, we map it to the intermediate latent code $W$ through several multi-layer perceptron layers (MLP) in order to better retain the semantic properties. Then we pass the intermediate latent vector $F_{\text{latent}}$ into the generative model. Each block of the generator does not use a single intermediate latent vector $W$ as input, but uses different latent vectors to improve expressiveness. More specifically, we have $W = (w_0, ..., w_{k-1})$ for $k$ blocks, where each $w_i$ corresponds to a latent vector. We found that this modification resulted in fewer artifacts output. This modification also appeared in previous work [3, 16].

The overall formula is as follows:

$$W = \text{MLP}(F_{\text{latent}})$$

(5)

$$F_{\text{GAN}} = \text{StyleGAN}(W)$$

(6)

3.5. Channel Squeeze Excitation Spatial Feature Transform (SE-SFT)

In order to preserve authenticity and fidelity, the key to the problem is to effectively combine the GAN features $F_{\text{GAN}}$ and the spatial features $F_{\text{spatial}}$ extracted from the input image. An effective method is Spatial Feature Transformation (SFT), which generates affine transformation parameters for spatial feature modulation, and shows its effectiveness in combination with other conditions in image restoration and image generation. In our task, a pair of affine transformation parameters $(\alpha, \beta)$ are generated from spatial input features $F_{\text{spatial}}$ by several convolutional layers. After that, the modulation is performed by scaling and shifting the prior feature $F_{\text{GAN}}$. The overall formula is as follows:

$$\hat{\alpha}, \beta = \text{Conv}(F_{\text{spatial}})$$

(7)

$$F_{\text{output}} = \text{SFT}(F_{\text{GAN}} | \alpha, \beta) = \hat{\alpha} \odot F_{\text{GAN}} + \beta$$

(8)
Although it can effectively merge the input face information, it cannot balance the realness and fidelity well, because all the prior features $F_{\text{prior}}$ (contributing to realness) are affected by the input features $F_{\text{spatial}}$ (contributing to fidelity). In particular, when the low-resolution picture is used as the input in the real scene, the "fuzzy" input feature will affect the modulation, and the output inevitably tends to the "fuzzy" result without real and rich facial details.

To solve this problem, we propose channel squeeze-excitation spatial feature transform (SE-SFT) layers inspired by DPN[27], which divides features into two parts according to the channel, spatially modulates only one part, and passes the other part directly to better preserve information. In this way, we retain more prior information, which improves the realness of the image. In order to further balance the realness and fidelity, we introduced channel attention mechanism, that is, we stitch the features and pass the squeeze and excitation (SE) module. The SE module will provide the learned weights for each channel. The multiplication of the data in the channel and the weights will further control the proportion of the original information and the prior information in the results, so as to achieve a balance between realness and fidelity. As shown in equation (9):

$$F_{\text{output}} = SE - SFT(F_{\text{GAN}}(\alpha, \beta) = SE(\text{Concat}(\text{Identity}(F_{\text{split0}}_{\text{GAN}}), \alpha \odot F_{\text{split1}}_{\text{GAN}} + \beta)))$$ (9)

Where $F_{\text{split0}}_{\text{GAN}}$ and $F_{\text{split1}}_{\text{GAN}}$ are $F_{\text{GAN}}$ divided by channel dimension, Concat [...] represents the concatenation operation, and SE is the channel attention mechanism.

Therefore, SE-SFT has the advantages of directly fusing the prior information and effectively modulating by input images, thus achieving a good balance between realness and fidelity. As a result, the dual-path structure of SE-SFT can reuse features and new features exploration for each path, thereby improving its representation ability, and finally use the channel attention mechanism for fine fusion, so as to achieve the balance of the overall effect.

### 3.6. Training

Different from the existing works [5,8,13], we have added identity preserving loss[17] in addition to standard $L_2$ loss, perceptual loss and adversarial loss for training.

**Standard Loss:** MSE $L_2$ loss is used to guide the fidelity of the output images.

$$L_{\text{mse}} = \frac{1}{N}||\hat{y} - y||_2^2$$ (10)

We further combine perceptual loss [18] to improve the perceptual quality:

$$L_{\text{percep}} = \frac{1}{N}||f(\hat{y}) - f(y)||_2^2$$ (11)

Where $f(.)$ denotes the feature embedding space of the pretrained VGG16[19] network, we use {conv1, ⋯, conv5} feature map before activation.

**Adversarial Loss:** Similar to StyleGAN2, we use the adversarial loss as follows:

$$L_{\text{gen}} = \lambda_{adv} \log(1 - D(\hat{y}))$$ (12)

Where $D(.)$ represents the discriminator of stylegan2.
Figure 2. Qualitative comparison on the CelebA-HQ for blind face super-resolution. Our GPLFSR has achieved excellent results superior to other methods in terms of image realness and fidelity.

**Identity Preserving Loss.** We derive inspiration from previous work and apply identity preserving loss to our model. Specifically, we use the pre-trained face recognition ArcFace[10] model, which captures the most salient features of identity recognition. The loss of identity consistency makes the restoration result have greater fidelity:

$$ L_{id} = \lambda_{id} || \eta(\hat{y}) - \eta(y) ||_1 $$

(13)

Where $\eta$ represents the facial feature extractor, that is, ArcFace [20] in our implementation. $\lambda_{id}$ represents the weight of the identity preserving loss.

The overall model goal is a combination of the above losses:

$$ L_g = L_{mse} + L_{percep} + L_{gen} + L_{id} $$

(14)

4. Experiments

4.1. Dataset

**Training data:** We trained GPLFSR based on the FFHQ[2] dataset, which contains 70,000 high-quality images. We use FFHQ to obtain synthetic data through the degradation model as training data, which is used to fit real low-quality images and promote to real-world images in the prediction process. The degradation model is based on existing methods [21], as shown below:

$$ \chi = [(y \otimes k_{\sigma}) \downarrow_{\gamma} + n_{\delta}]_{JPEG} $$

(15)

The original image $y$ is first convoluted with a Gaussian blur kernel $k_{\sigma}$, and then subjected to a down-sampling operation with a random scale factor $\gamma$. Then, add additive white Gaussian noise $n_{\delta}$ to the image, and finally use the quality factor $q$ to perform JPEG compression on the image.

**Test data:** Test data: We choose Celeba-HQ[15] as the test data. Celeba-HQ is a synthetic data set containing 30,000 Celeba-HQ images from its test partition. The synthesis method is the same as during training.
Table 1. Quantitative comparison on CelebA-HQ for blind face super-resolution. Bold indicates the best performance. Deg. represents the identity distance

| Methods          | LPIPS↓ | FID↓  | NIQE↓ | Deg.↓ | PSNR↑ | SSIM↑ |
|------------------|--------|-------|-------|-------|-------|-------|
| Input            | 0.4863 | 143.98| 13.440| 47.94 | 25.35 | 0.6848|
| HiFaceGAN[22]    | 0.4795 | 66.13 | 4.954 | 42.18 | 24.92 | 0.6013|
| DFDNet[21]       | 0.4396 | 60.01 | 4.354 | 40.35 | 23.66 | 0.6602|
| PSFRGAN[23]      | 0.4240 | 47.59 | 5.123 | 39.69 | 24.71 | 0.6557|
| mGANPrior[3]     | 0.4683 | 82.35 | 6.422 | 55.45 | 24.30 | 0.6758|
| PULSE[7]         | 0.4451 | 56.56 | 4.344 | 69.55 | 21.61 | 0.6200|
| GLEAN[1]         | 0.4803 | 93.59 | 9.784 | 49.50 | 24.80 | 0.6753|
| GPLFS            | 0.3908 | 45.37 | 4.137 | 33.92 | 24.08 | 0.6677|

4.2. Implementation
We use the 10242 output pre-trained StyleGAN2 as our generated face prior. The channel multiplier of StyleGAN2 is set to 1, for compact size. In order to achieve a compact size, the channel multiplier of stylegan2 is set to 1. For each SE-SFT layer, we use two convolutional layers to generate affine parameters α and β respectively.

Similar to GLEAN[1], in order to use the generated prior, we keep the weight of the StyleGAN2 constant throughout the training process. We use cosine degradation scheme and Adam Optimizer in training. The number of iterations is 300, and the initial learning rate is 10^{-4}. The batch size of the face is 4, and we use Nvidia 3090 GPU to train our model.

4.3. Comparison with the Latest Method
We compare the network with the latest methods designed for face super-resolution, including mGANprior [3], PULSE [7] and GLEAN[1]. Among them, blind face repair methods DFDNet[21], HiFaceGAN[22], PSFRGAN[23], are also included. The magnification factor is 16 × , and the input is 64².

We use popular non-reference perception indicators: FID [24] and NIQE [25]. Pixel-level indicators (PSNR and SSIM) and perceptual indicators (LPIPS [26]). And we are measuring the distance of identity based on ArcFace [20] through feature embedding, the smaller the value, the closer the identity is to GT.

**Qualitative comparison:** The qualitative comparison of 16×SR is shown in figure 2. Thanks to powerful generative facial priors and the sophisticated combination with priors, GPLFSR can achieve good results on complex degraded synthetic images, with more real details and fewer artifacts. However, the traditional network based on bicubic downsampling degradation greatly reduces the effect in blind super-resolution. Such as GLEAN, ESRGAN+, the results still have fuzzy characteristics. However, other super-resolution models based on GAN inversion cannot maintain good fidelity under the guidance of low-dimensional vectors and LR space constraints. In specifically, PULSE [7] and mGANprior [3] cannot recover face images with the same identity, and in addition, artifacts can also be observed in their output. In addition, methods that rely on component dictionary (DFD-Net) or parsing maps (PSFRGAN) cannot generate real texture details. They tend to generate excessively smooth facial images with distorted facial structures. This obviously confirms the superiority of our work.

**Quantitative comparison:** Table 1 shows the quantitative results of various methods. It can be seen that GPLFSR has reached the lowest level on LPIPS, FID, and NIQE, which shows that the images we generate are closer to the real images in perception, which is also consistent with the conclusions of our qualitative observations. In addition to perceived performance, the network also demonstrated the best performance in terms of identity consistency. This shows that our network has achieved a balance between realness and fidelity. However, the pixel-level indicators PSNR and SSIM are often inconsistent with the subjective evaluation of human observers, so our model does not perform well on these two indicators.
5. Conclusion
We proposed GPLFSR for blind face super-resolution. Based on the existing encoder-latent bank-decoder super-resolution network, we introduced a more refined prior combination method channel squeeze excitation spatial feature transform (SE-SFT) layers to make better use of the the prior information in pre-trained network. Through SE-SFT, we have achieved a balance between the authenticity and fidelity. In addition, in order to improve network performance, we introduce identity consistency loss. Experiments show that our method is better than the existing methods.

References
[1] Chan, K. C., Wang, X., Xu, X., Gu, J., & Loy, C. C. (2021). Glean: Generative latent bank for large-factor image super-resolution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14245-14254.
[2] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110-8119.
[3] 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. pp. 3012-3021.
[4] Pan X, Zhan X, Dai B, et al. (2020). Exploiting deep generative prior for versatile image restoration and manipulation[C] In: European Conference on Computer Vision. Springer, Cham. pp. 262-277.
[5] Nie, H., Lu, Y., Ikram, J.: (2016) Face hallucination via convolution neural network. In: Proceeding of the IEEE 28th International Conference on Tools with Artificial Intelligence. pp. 485–489.
[6] Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q. (2015) Learning face hallucination in the wild. In: Proceeding of the Associationfor the Advancement of Artificial Intelligence. pp. 3871–3877.
[7] Menon, S., Damian, A., Hu, S., Ravi, N., & Rudin, C. (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. pp. 2437-2445.
[8] Yu, X., & Porikli, F. (2016). Ultra-resolving face images by discriminative generative networks. In European conference on computer vision. Springer, Cham. pp. 318-333.
[9] Chen, Y., Tai, Y., Liu, X., Shen, C., & Yang, J. (2018). Fsrgan: End-to-end learning face super-resolution with facial priors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2492-2501.
[10] Yu, X., Fernando, B., Ghanem, B., Porikli, F., & Hartley, R. (2018). Face super-resolution guided by facial component heatmaps. In Proceedings of the European conference on computer vision (ECCV). pp. 217-233.
[11] Pan X, Zhan X, Dai B, et al. (2020)Exploiting deep generative prior for versatile image restoration and manipulation[C]//European Conference on Computer Vision. Springer, Cham. pp 262-277.
[12] Yang, T., Ren, P., Xie, X., & Zhang, L. (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. pp. 672-681.
[13] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., ... & Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0-0.
[14] Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., ... & Wang, Z. (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. pp. 1874-1883.
[15] Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation. In International Conference on Learning Representations. pp. 1224-1236.

[16] Xu, Y., Shen, Y., Zhu, J., Yang, C., & Zhou, B. (2021). Generative hierarchical features from synthesizing images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4432-4442.

[17] Huang, R., Zhang, S., Li, T., & He, R. (2017). Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In Proceedings of the IEEE international conference on computer vision. pp. 2439-2448.

[18] Johnson, J., Alahi, A., & Fei-Fei, L. (2016, October). Perceptual losses for real-time style transfer and super-resolution. In European conference on computer vision. pp. 694-711.

[19] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[20] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690-4699.

[21] Li, X., Chen, C., Zhou, S., Lin, X., Zuo, W., & Zhang, L. (2020, August). Blind face restoration via deep multi-scale component dictionaries. In European Conference on Computer Vision. Springer, Cham. pp. 399-415.

[22] Yang, L., Wang, S., Ma, S., Gao, W., Liu, C., Wang, P., & Ren, P. (2020). Hifacegan: Face renovation via collaborative suppression and replenishment. In Proceedings of the 28th ACM International Conference on Multimedia. pp. 1551-1560.

[23] Chen, C., Li, X., Yang, L., Lin, X., Zhang, L., & Wong, K. Y. K. (2021). Progressive semantic-aware style transformation for blind face restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11896-11905.

[24] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, p.30.

[25] Mittal, A., Soundararajan, R., & Bovik, A. C. (2012). Making a “completely blind” image quality analyzer. IEEE Signal processing letters, 20(3), 209-212.

[26] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586-595.

[27] Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., & Feng, J. (2017). Dual path networks. arXiv preprint arXiv:1707.01629.