The following publication T. Yang, P. Ren, X. Xie and L. Zhang, "GAN Prior Embedded Network for Blind Face Restoration in the Wild," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 672-681 is available at https://doi.org/10.1109/CVPR46437.2021.00073.

# GAN Prior Embedded Network for Blind Face Restoration in the Wild

Tao Yang[1], Peiran Ren[1], Xuansong Xie[1], and Lei Zhang[1,2*]
[1]DAMO Academy, Alibaba Group
[2]Department of Computing, The Hong Kong Polytechnic University
yangtao9009@gmail.com, peiran_r@sohu.com, xingtong.xxs@taobao.com, cslzhang@comp.polyu.edu.hk

## Abstract

*Blind face restoration (BFR) from severely degraded face images in the wild is a very challenging problem. Due to the high illness of the problem and the complex unknown degradation, directly training a deep neural network (DNN) usually cannot lead to acceptable results. Existing generative adversarial network (GAN) based methods can produce better results but tend to generate over-smoothed restorations. In this work, we propose a new method by first learning a GAN for high-quality face image generation and embedding it into a U-shaped DNN as a prior decoder, then fine-tuning the GAN prior embedded DNN with a set of synthesized low-quality face images. The GAN blocks are designed to ensure that the latent code and noise input to the GAN can be respectively generated from the deep and shallow features of the DNN, controlling the global face structure, local face details and background of the reconstructed image. The proposed GAN prior embedded network (GPEN) is easy-to-implement, and it can generate visually photo-realistic results. Our experiments demonstrated that the proposed GPEN achieves significantly superior results to state-of-the-art BFR methods both quantitatively and qualitatively, especially for the restoration of severely degraded face images in the wild. The source code and models can be found at* https://github.com/yangxy/GPEN.

## 1. Introduction

Face images are among the most popular types of images in our daily life, while face images are often degraded due to the many factors such as low resolution, blur, noise, compression, etc., or the combination of them. Face image restoration has been attracting significant attentions, aiming at reproducing a clear and realistic face image from the degraded input. Traditional face image restoration methods [50, 3, 2, 36] usually solve an inverse problem based on the degradation model and handcrafted priors, which demonstrate limited performance in practice. Recently, deep neural networks (DNNs) have shown superior results in a variety of computer vision tasks [24, 48, 13, 25, 30], and many DNN based face restoration methods [49, 29, 16] have also been developed and they have demonstrated much better performance than traditional ones.

Though much progress has been made for face restoration, blind face restoration (BFR) remains a challenging research problem because of the unknown and complex degradation of low quality (LQ) face images in the wild. In order to recover a high-quality (HQ) face image with photo-realistic textures from an LQ face image, a number of BFR methods have been proposed by resorting to the spatial transformer networks [49], exemplar images [29, 28, 9], 3D facial priors [16], and facial component dictionaries [27]. Yang *et al.* [47] proposed a collaborative suppression and replenishment (CSR) approach to progressively replenish facial details. These methods exhibit impressive results on artificially degraded faces; however, they fail to tackle real-world LQ face images. The conditional generative adversarial network (cGAN) based methods such as Pix2Pix [18] and Pix2PixHD [43] learn a direct mapping from input image to output image. These methods achieve more realistic results but tend to over-smooth the images (see Figures 5 and 7), which is commonly blamed to the high illness of real-world BFR tasks.

With the rapid advancement of GAN techniques [21, 22], recently some methods have been proposed to reconstruct faces from extremely low resolution inputs [12, 34, 38]. Richardson *et al.* [38] employed an encoder network to generate a series of style vectors before feeding them into a pre-trained generator, achieving a generic image-to-image translation framework. However, such methods can only work on non-blind image super-resolution problems. Furthermore, they kept the pre-trained GAN unchanged in training for the consistency and convenience of face manipulations. This however leads to unstable quality of restored faces when dealing with real-world LQ face images

[*]This work is partially supported by the Hong Kong RGC RIF grant (R5001-18).

with complex background, because it is hard to accurately project a face image with limited resolution to a desired latent code (e.g., a vector of size 512 in StyleGAN [21, 22]).

In this work, we revisit the problem of BFR and target at restoring HQ faces from degraded face observations in the wild. Our idea is to seamlessly integrate the advantages of GAN and DNN. We first pre-train a GAN for HQ face image generation and embed it into a DNN as a decoder prior for face restoration. The GAN prior embedded DNN is then fine-tuned by a set of synthesized LQ-HQ face image pairs, during which the DNN learns to map the input degraded image to a desired latent space so that the GAN prior network can reproduce the desired HQ face images. We carefully design the GAN blocks to make them well suited for a U-shaped DNN, where the deep features are used to generate the latent code for global face reproduction, while the shallow features are used as noise to generate local face details and keep the image background. In this way, our learned model can reconstruct HQ faces with photo-realistic details from even severely degraded face images in the wild, avoiding over-smoothed results caused by the high illness of the BFR problem. Figure 1 shows an example. One can see that our model reconstructs the face images of those great scientists with clear details from the old photo taken in 1927.

The main contributions of this work are summarized as follows:

- We learn and embed a GAN prior network into a DNN, and fine-tune the GAN embedded DNN for effective BFR in the wild. It is worthy to note that previous works only transfer the pre-trained GAN into a network without fine-tuning.
- The GAN blocks are designed so that they can be easily embedded into a U-shaped DNN for fine-tuning. The latent code and noise input of the GAN are respectively generated from the deep and shallow features of the DNN to reconstruct the global structure, local face details and background of the image accordingly.
- Our model sets new state-of-the-art in BFR. It is capable of tackling severely degraded face images taken in real-world scenarios.

## 2. Related Work

**Face Image Restoration.** As a specific but important branch of image restoration, face image restoration has been widely studied for many years. In the early stage, Zhang *et al.* [50] presented a joint blind image restoration and recognition method by using sparse representation to handle face recognition from LQ images. Nishiyama *et al.* [36] proposed to improve the recognition performance of blurry faces by using a pre-defined set of blur kernels to restore them. With the unprecedented success of DNNs in solving image restoration tasks such as denoising [13], deblurring [24, 40], inpainting [48, 31] and image super-resolution [25, 30], many DNN based face image restoration methods have also been proposed [7, 23, 33], which advance the traditional methods by a large margin. Considering the fact that facial images have specific structures, it is interesting to investigate whether we can restore a clear face image from severely degraded ones without knowing the degradation model. The so-called blind face restoration (BFR) problem has been attracting intensive research attentions in recent years [29, 9, 27, 47], while it is still a challenging task due to the complex image degradations in the wild.

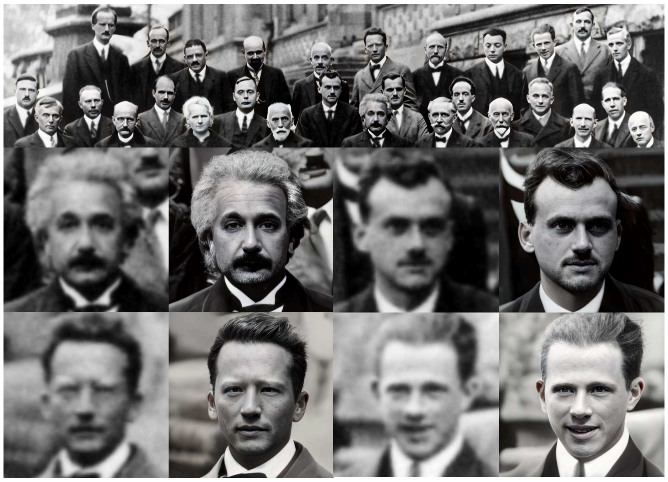


Figure 1: Restored face images from the group photo taken in the Solvay Conference, 1927. Best viewed by zooming to 200% in the screen.

Huang *et al.* [17] presented a wavelet-based approach that can ultra-resolve a very low-resolution (LR) face image. Chen *et al.* [7] learned the facial geometry prior to recover the high-resolution (HR) faces. Ma *et al.* [33] performed face super-resolution with iterative collaboration between two recurrent networks on facial image recovery and landmark estimation, respectively. Li *et al.* [29] used a guiding image and a wrapper subnetwork to cope with appearance variations between the LR input and the HR guiding image. This work was further extended by using an unconstrained HR face image [9], multi-exemplar images [28], and multi-scale component dictionaries [27]. Hu *et al.* [16] explicitly incorporated 3D facial priors to grasp the sharp facial structures. A collaborative suppression and replenishment approach was proposed by Yang *et al.* [47] to progressively replenish facial details. Existing works have generated impressive results on artificially degraded faces, but often failed in real-world scenarios due to the complex unknown degradation. Furthermore, their performance depends heavily on the accurate facial prior knowledge which however is hard to obtain from severely degraded face images in the wild, leading to unpredictable failures.

**Generative Adversarial Network (GAN).** Since the seminal work by Goodfellow *et al.* [11], great progress has been accomplished on learning GAN models [20, 4, 21, 22]. GAN has been widely used for various computer vision applications due to its powerful ability to generate photo-realistic images. Some typical applications include image inpainting [48], super-resolution [25, 44], image colorization [18, 42], texture synthesis [41], etc. Particularly, to provide more user controls for image synthesis, conditional GAN (cGAN) has been proposed [35]. By feeding the generator with different conditional information [37, 18, 52], cGANs succeed in handling various image-to-image translation problems. Isola *et al.* [18] showed that the conditional adversarial networks can be used as a general-purpose solution to image-to-image translation problems. Many following works, such as unsupervised learning [52], disentangled learning [26], few-shot learning [32], high resolution image synthesis [43], multi-domain translation [8], multi-modal translation [53], have been proposed to extend cGAN to different scenarios. The cGAN learns a direct mapping from the input domain to the output one. Unfortunately, the generated results by cGANs are usually over-smoothed in highly ill-posed tasks such as BFR.

**GAN Prior for Image Generation.** Deep generative models are popular in solving many inverse problems, e.g. deblurring [24], image inpainting [48], phase retrieval [14], etc. Recently, many works have been developed for the task of GAN inversion, i.e., reversing a given image back to a latent code with a pre-trained GAN model. Existing methods either optimize the latent code [1] or learn an extra encoder to project the image space back to the latent space [12, 38]. Abdal *et al.* [1] embedded images into an extended latent space of StyleGAN, allowing further semantic image editing operations. Gu *et al.* [12] employed multiple latent codes to generate multiple feature maps to output the final image. These optimization-based methods, however, are slow and improper for real-world applications. To address this issue, Pixel2Style2Pixel (pSp) [38] embeds real images into extended latent space without additional optimization, which can be used in a wide range of image-to-image translation tasks. Menon *et al.* [34] proposed a self-supervised approach that traverses the HR natural image manifold, searching for images that can downscale to the original LR image. GAN inversion is an important step for applying GANs to real-world applications. However, it is difficult to perfectly project the image space back to the latent space. Moreover, it is hard, if not possible, to invert a blindly degraded face into a latent space.

Some works were proposed to transfer GAN priors. Wang *et al.* [46] applied domain adaptation to image generation with GANs. They further proposed a novel knowledge transfer method for generative models by using a knowledge mining network [45]. Frégier and Gouray [10] introduced a novel approach for transfer learning with GAN architecture. These works target at transferring the knowledge from the source domain to different target domains, while in our work, the source and target domains are the same. We embed the GAN prior learned for face generation into a DNN for face restoration, and jointly fine-tune the GAN prior network with the DNN so that the latent code and noise input can be well generated from the degraded face image at different network layers.

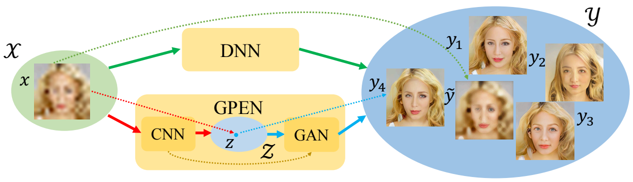


Figure 2: Illustration of the motivation and framework of our GAN prior embedded network (GPEN).

# 3. Proposed Method

## 3.1. Motivation and Framework

BFR is a typical ill-posed inverse problem. Denote by $\mathcal{X}$ the space of degraded LQ faces, and by $\mathcal{Y}$ the space of original HQ face images. Given an input LQ face image $\mathbf{x} \in \mathcal{X}$, BFR aims to find its corresponding clear face image $\mathbf{y} \in \mathcal{Y}$. Most of the DNN based methods learn a mapping function $\Phi$ to achieve this goal, i.e., $\Phi(\mathbf{x}) \rightarrow \mathbf{y}$. However, this is a one-to-many inverse problem, and there are many possible face images (e.g., $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$) in $\mathcal{Y}$ that can match to the input $\mathbf{x}$. Existing methods [5, 29, 9] usually train DNNs to perform mapping between $\mathbf{x}$ and $\mathbf{y}$ using some pixel-wise loss functions. As a result, as we illustrated in Figure 2, the final solution $\Phi(\mathbf{x})$ tends to be the mean of those HQ faces, which is over-smoothed and loses details. This coincides with the visual perception global-first theory [6]. The cGAN methods [18, 43] can partially dilute this issue by adversarial training to reduce the uncertainty in mapping. However, when the degradation is severe, the problem remains and cGANs can hardly generate clear face images with realistic textures and details (see Figure 5 for example).

Different from previous methods [5, 29, 9, 43, 47], we first train a GAN prior network, and then embed it into a DNN as decoder for HQ face image restoration. We call our method GAN prior embedded network (GPEN). As illustrated in Figure 2, the first part of our GPEN is a CNN encoder, which learns to map the input degraded image $\mathbf{x}$ to a desired latent code $\mathbf{z}$ in the latent space $\mathcal{Z}$ of the GAN. The GAN prior network can then reproduce the desired HQ face image via $G(\mathbf{z}) \rightarrow \mathbf{y}$, where G refers to the learned generator of GAN. The generation process is basically a one-to-one mapping, largely alleviating the uncertainty of one-to-many mapping in previous methods. It should be

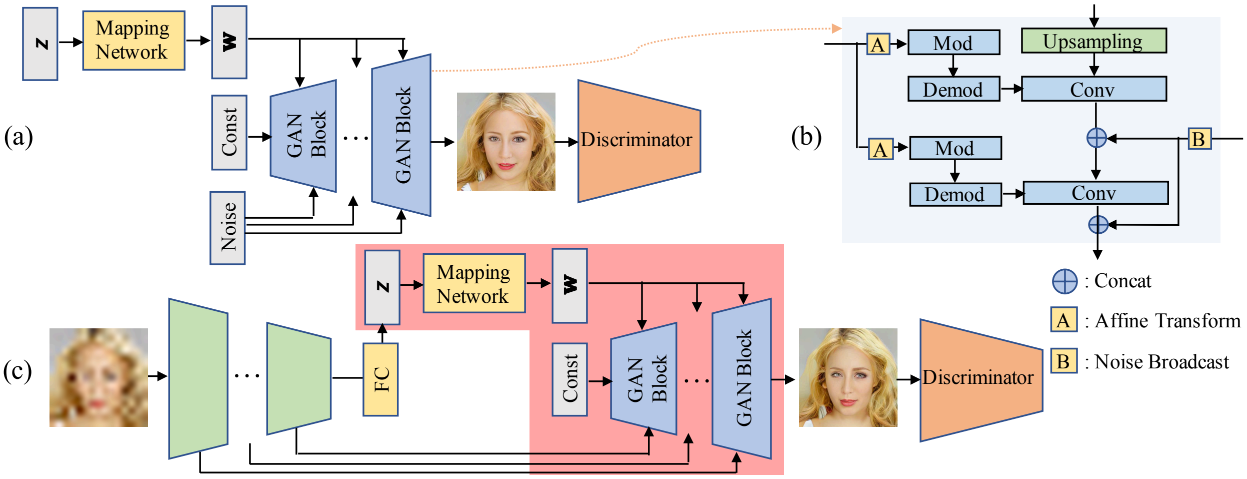


Figure 3: The architecture of GPEN. (a) The GAN prior network; (b) detailed structures of a GAN block; and (c) the full network architecture of GPEN. The definition of "Mod" and "Demod" can be found in [22].

noted that the GAN inversion methods [12, 34, 38] share a similar idea with our GPEN; however, they keep the pre-trained GANs unchanged for consistent and convenient face manipulations. While in GPEN, we carefully design and pre-train the GAN blocks and fine-tune the GAN priors for effective BFR. The architectures of GPEN and GAN blocks are shown in Figure 3 and will be explained in detail in the following sections.

### 3.2. Network Architecture

**The GAN prior network.** U-Net [39] has been successfully and widely used in many image restoration tasks [43, 13] and demonstrated its effectiveness in preserving image details. Therefore, our GPEN overall follows a U-shaped encoder-decoder architecture (see Figure 3(c)). Accordingly, the GAN prior network should be designed to meet two requirements: 1) it is capable of generating HQ face images; and 2) it can be readily embedded into the U-shaped GPEN as a decoder. Inspired by the state-of-the-art GAN architectures, e.g., StyleGAN [21, 22], we use a mapping network to project latent code $\mathbf{z}$ into a less entangled space $\mathbf{w} \in \mathcal{W}$, as illustrated in Figure 3(a). The intermediate code $\mathbf{w}$ is then broadcasted to each GAN block. Since the GAN prior network will be embedded into a U-shaped DNN for finetuning, we need to leave room for the skipped feature maps extracted by the encoder of the U-shaped DNN. We thus provide additional noise inputs to each GAN block.

For the structure of GAN block, there are several options. In this work, we adopt the architecture in StyleGAN v2 (see Figure 3(b)) due to its high capability to generate HQ images. (Alternative GAN architectures such as StyleGAN v1 [21], PGGAN [20] and BigGAN [4] can also be easily adopted into our GPEN.) The number of GAN blocks is equal to the number of skipped feature maps extracted in the U-shaped DNN (and the number of noise inputs), which is related to the resolution of input face image. StyleGAN requires two different noise inputs in each GAN block. To enable the GAN prior network to be readily embedded into the U-shaped GPEN, different from StyleGAN, the noise inputs are reused at the same spatial resolution for all GAN blocks. Furthermore, the noise inputs are concatenated rather than added to the convolutions in StyleGAN. We empirically found that this can bring more details in the restored face image.

**Full network architecture.** Once the GAN prior network is trained by using some dataset (e.g., the FFHQ [21] dataset), we embed it into the U-shaped DNN as a decoder, as shown in Figure 3(c). The latent code $\mathbf{z}$ and the noise inputs to the GAN network are replaced by the output of the fully-connected layer (i.e., deeper features) and shallower layers of the encoder of the DNN, respectively, which will control the reconstruction of global face structure, local face details, as well as the background of face image. Since the proposed model is not fully convolutional, LQ face images are first resized to the desired resolution (e.g., $1024^2$) using simple bilinear interpolator before being input to the GPEN. After embedding, the whole GPEN will be fine-tuned so that the encoder part and decoder part can learn to adapt to each other.

### 3.3. Training Strategy

We first pre-train the GAN prior network using a dataset of HQ face images following the training strategies of StyleGAN [21, 22]. The pre-trained GAN model is embedded into the proposed GPEN, and we fine-tune the whole network using a set of synthesized LQ-HQ face image pairs (the image synthesis process will be given in Section 4.2).

To fine-tune the GPEN model, we adopt three loss functions: the adversarial loss $L_A$, the content loss $L_C$, and the feature matching loss $L_F$. $L_A$ is inherited from the GAN prior network:

$$L_A = \min_G \max_D E_{(X)} \log \left(1 + \exp\left(-D\big(G(\tilde{X})\big)\right)\right), \quad (1)$$

where $X$ and $\tilde{X}$ denote the ground-truth HQ image and the degraded LQ one, $G$ is the generator during training, and $D$ is the discriminator. $L_C$ is defined as the $L_1$-norm distance between the final results of the generator and the corresponding ground-truth images. $L_F$ is similar to the perceptual loss [19] but it is based on the discriminator rather than the pre-trained VGG network to fit our task. It is formulated as follows:

$$L_F = \min_G E_{(X)} \left( \sum_{i=0}^{T} \left\| D^i(X) - D^i(G(\tilde{X})) \right\|_2 \right), \quad (2)$$

where $T$ is the total number of intermediate layers used for feature extraction. $D^i(X)$ is the extracted feature at the $i$-th layer of discriminator $D$.

The final loss $L$ is as follows:

$$L = L_A + \alpha L_C + \beta L_F, \quad (3)$$

where $\alpha$ and $\beta$ are balancing parameters. The content loss $L_C$ enforces the fine features and preserves the original color information. By introducing the feature matching loss $L_F$ on the discriminator, the adversarial loss $L_A$ can be better balanced to recover more realistic face images with vivid details. In all the following experiments, we empirically set $\alpha = 1$ and $\beta = 0.02$.

# 4. Experiments

## 4.1. Datasets and Evaluation Metric

The FFHQ dataset [21], which contains $70,000$ HQ face images of resolution $1024^2$, is used to train our GPEN model. We first use it to train the GAN prior network, and then synthesize LQ images from it to fine-tune the whole GPEN. To evaluate our model, we use the CelebA-HQ dataset [20] to simulate LQ face images to quantitatively compare GPEN with other state-of-the-art methods. We also collet $1,000$ real-world LQ faces (will be made publicly available) from internet to qualitatively evaluate the performance of our model in the wild. In the quantitative evaluation, the Peak Signal-to-Noise Ratio (PSNR), the Fréchet Inception Distances (FID) [15] and the Learned Perceptual Image Patch Similarity (LPIPS) [51] indices are used. It is worth mentioning that all these indices can only be used as references for evaluation because they cannot truly reflect the performance of a BFR method, especially for BFR in the wild.

## 4.2. Implementation Details

We first train the GAN prior network using the FFHQ dataset with similar settings to StyleGAN [21, 22]. The pre-trained GAN prior network is embedded into the GPEN to perform fine-tuning. To build LQ-HQ image pairs for fine-tuning, we synthesize degraded faces from the HQ images in FFHQ using the following degradation model:

$$I^d = ((I \otimes \mathbf{k}) \downarrow_s + \mathbf{n}_\sigma)_{JPEG_q}, \quad (4)$$

where $I$, $\mathbf{k}$, $\mathbf{n}_\sigma$, $I^d$ are respectively the input face image, the blur kernel, the Gaussian noise with standard deviation $\sigma$ and the degraded image. $\otimes$, $\downarrow_s$, $JPEG_q$ respectively denote the two-dimensional convolution, the standard $s$-fold downsampler and the JPEG compression operator with a quality factor $q$.

The above degradation model has been used in previous methods [29, 27]. In our implementation, for each image the blur kernel $\mathbf{k}$ is randomly selected from a set of blurring models, including Gaussian blur and motion blur with varying kernel sizes. The additive Gaussian noise $\mathbf{n}_\sigma$ is sampled channel-wise from a normal distribution, and $\sigma$ is chosen from $[0, 25]$. The value of $s$ is randomly and uniformly sampled from $[10, 200]$ (i.e., up to 200 times downscaling) and $q$ is randomly and uniformly sampled from $[5, 50]$ (i.e., up to $95\%$ JPEG compression) per image. By using those severely degraded images to fine-tune the model, the encoder part of our GPEN can learn to generate suitable latent code and noise inputs to the GAN prior decoder network, which is updated simultaneously to tackle severely degraded faces in real-world scenarios.

During model updating, we adopt the Adam optimizer with a batch size of 1. The learning rate ($LR$) varies for different parts of GPEN, including the encoder, the decoder and the discriminator. In our implementation, we let $LR_{\text{encoder}} = 0.002$, and set $LR_{\text{encoder}} : LR_{\text{decoder}} : LR_{\text{discriminator}} = 100 : 10 : 1$. It should be noted that the discriminator part will be removed in the testing stage.

## 4.3. Ablation Study

To better understand the roles of different components of GPEN and the training strategy, in this section we conduct an ablation study by introducing some variants of GPEN and comparing their BFR performance. The first variant is denoted by GPEN-w/o-ft, i.e., the embedded GAN prior network is kept unchanged in the fine-tuning process. The second variant is denoted by GPEN-w/o-noise, which refers to the GPEN model without noise inputs. The third variant is denoted by GPEN-noise-add, i.e., that the noise inputs are added rather than concatenated to the convolutions.

We perform BFR on the CelebA-HQ dataset to evaluate GPEN and its three variants. The LQ images are synthesized by using the degradation model in Eq. (4) and the

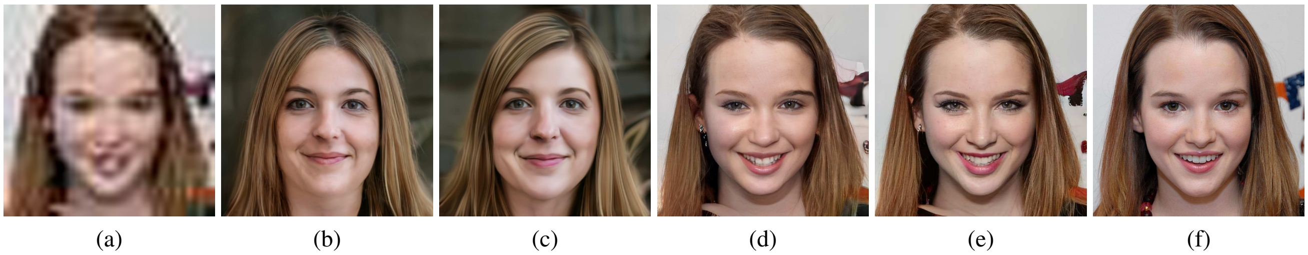


Figure 4: Comparisons of our variants BFR. (a) LQ input; (b) GPEN-w/o-ft; (c) GPEN-w/o-noise; (d) GPEN-noise-add; (e) GPEN; (f) Ground truth.

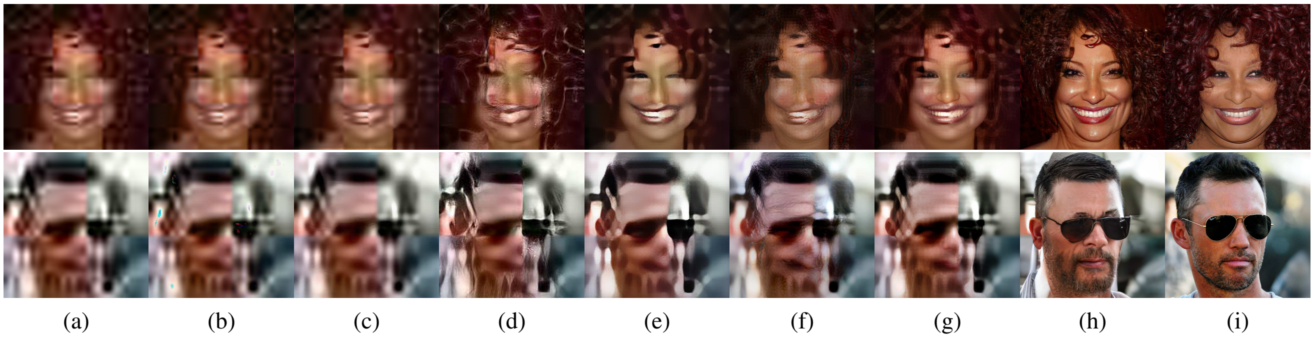


Figure 5: Blind face restoration results on synthsized degraded faces. (a) Degraded faces; (b) Super-FAN [5]; (c) GFRNet [29]; (d) GWAInet [9]; (e) Pix2PixHD [43]; (f) DFDNet [27]; (g) HiFaceGAN [47]; (h) GPEN; (i) Ground truth.

Table 1: Comparison (PSNR, FID and LPIPS) of different variants of GPEN.

| Method | PSNR↑ | FID↓ | LPIPS↓ |
|---|---|---|---|
| GPEN-w/o-ft | 12.55 | 92.71 | 0.653 |
| GPEN-w/o-noise | 13.30 | 95.62 | 0.709 |
| GPEN-noise-add | 20.71 | 34.26 | 0.359 |
| GPEN | **20.80** | **31.72** | **0.346** |

same set of parameters used in Section 4.2. Table 1 lists the PSNR, FID and LPIPS results. One can see that GPEN achieves better quantitative measures than its variants. Figure 4 shows the BSR results of the networks on an image. We can see that GPEN-w/o-ft can generate clean HQ face image; however, the appearance of the face is rather different from the ground-truth, and the background of the image is totally different. This is because without fine-tuning the GAN prior, it is difficult to generate the desired latent code into the latent space $\mathcal{Z}$, which coincides with the findings in many GAN inversion works [1, 38]. By discarding the noise input, the result of GPEN-w/o-noise is blurrier than GPEN-w/o-ft, and there are some artifact generated in the boundary of the image. This implies that the noise input plays an import role in synthesizing localize details. GPEN-noise-add achieves comparable result to GPEN but with slightly less facial details, while it generates some false details in the background of the image. Overall, GPEN shows superior performance to its variants, demonstrating the effectiveness of concatenated U-shaped architecture and our training strategy for the BFR tasks.

Table 2: Comparison (PSNR, FID and LPIPS) of different BFR methods. *

| Method | PSNR↑ | FID↓ | LPIPS↓ |
|---|---|---|---|
| Pix2PixHD [43] | 20.45 | 76.89 | 0.494 |
| Super-FAN [5] | 21.56 | 136.83 | 0.616 |
| GFRNet [29] | **21.70** | 134.92 | 0.597 |
| GWAInet [9] | 19.84 | 135.84 | 0.569 |
| HiFaceGAN [47] | 21.33 | 56.67 | 0.392 |
| GPEN | 20.80 | **31.72** | **0.346** |

## 4.4. Experiments on Synthetic Images

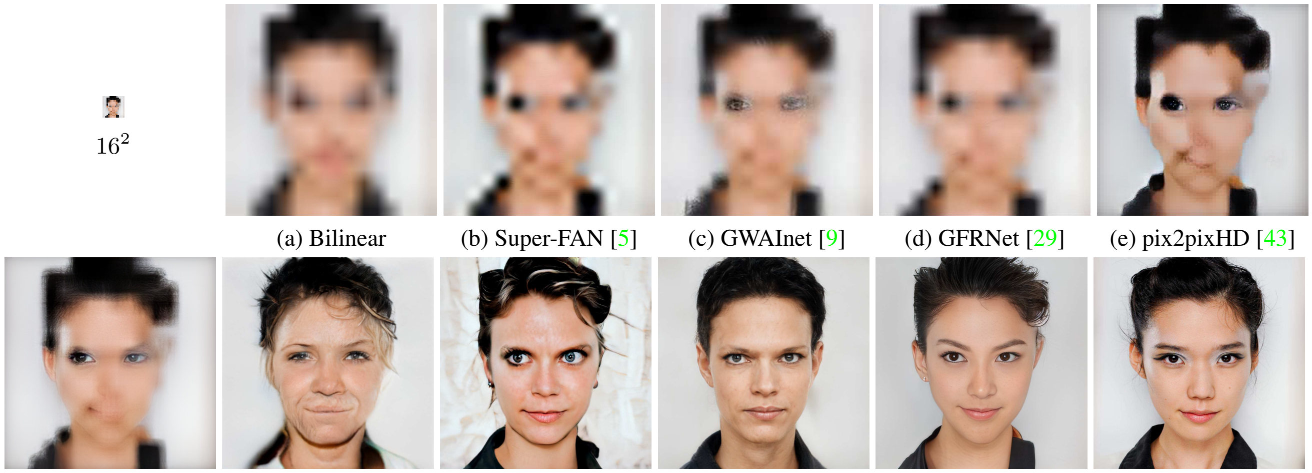
To quantitatively compare GPEN with other state-of-the-arts, we first perform experiments on synthetic images. Considering that many face restoration methods [12, 34, 38] are actually designed for FSR instead of BFR, we perform experiments on BFR and FSR separately, where different competing methods are used for fair comparison.

**Blind Face Restoration.** By using the degradation model in Eq. (4) and the same set of parameters used in Section 4.2, we synthesized a set of LQ face images on the CelebA-HQ dataset for evaluation. We compare GPEN with the latest BFR methods, including Pix2PixHD [43], Super-FAN [5], GFRNet [29], GWAInet [9], DFDNet [27], HiFaceGAN [47]. The models trained by the original authors are used in the experiments. We do not compare with those FSR methods [12, 34, 38] in this experiment because they assume a very simple degradation model (e.g., bicu-

*Note that the results of DFDNet [27] are not reported because it fails to recover many face images in this experiment.

Table 3: Comparison (PSNR, FID and LPIPS) of various FSR methods. Since mGANprior [12] and PULSE [34] are very time-consuming, we only used the first $1,000$ images of CelebA-HQ dataset to compute their measures. "-" means that the result is not available.

| **Method** | **PSNR↑** | | | | | | **FID↓** | | | | | | **LPIPS↓** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 8× | 16× | 32× | 64× | 128× | 256× | 8× | 16× | 32× | 64× | 128× | 256× | 8× | 16× | 32× | 64× | 128× | 256× |
| Bilinear | **28.73** | **26.13** | **22.81** | **20.49** | **17.75** | **15.17** | 89.29 | 183.50 | 206.03 | 342.63 | 528.17 | 495.03 | 0.471 | 0.567 | 0.659 | 0.713 | 0.765 | 0.812 |
| Super-FAN [5] | - | 20.95 | - | - | - | - | - | 92.65 | - | - | - | - | - | 0.453 | - | - | - | - |
| GFRNet [29] | 28.08 | 24.73 | 21.39 | - | - | - | 47.38 | 70.49 | 132.88 | - | - | - | 0.324 | 0.423 | 0.578 | - | - | - |
| GWAInet [9] | 25.79 | - | - | - | - | - | 56.81 | - | - | - | - | - | 0.339 | - | - | - | - | - |
| DFDNet [27] | 25.37 | 23.11 | - | - | - | - | 29.97 | 35.46 | - | - | - | - | 0.212 | 0.274 | - | - | - | - |
| HiFaceGAN [47] | 26.36 | 24.66 | 22.42 | 19.83 | - | - | **29.95** | 36.26 | 47.17 | 88.28 | - | - | 0.211 | 0.266 | 0.349 | 0.460 | - | - |
| mGANprior [12] | 21.44 | 21.29 | 20.53 | 18.09 | 15.45 | 13.39 | 104.20 | 100.84 | 95.82 | 108.05 | 113.73 | 113.28 | 0.521 | 0.518 | 0.472 | 0.519 | 0.558 | 0.582 |
| PULSE [34] | 24.32 | 22.54 | 19.98 | 16.09 | 13.39 | 11.49 | 65.89 | 65.33 | 81.23 | 87.45 | 102.48 | 101.35 | 0.421 | 0.425 | 0.405 | 0.492 | 0.544 | 0.579 |
| pSp [38] | 18.99 | 18.73 | 18.62 | 18.02 | 16.18 | 14.57 | 40.97 | 43.37 | 75.92 | 74.46 | 88.44 | 123.85 | 0.415 | 0.424 | 0.441 | 0.458 | 0.504 | 0.581 |
| GPEN | 24.66 | 23.27 | 21.23 | 19.02 | 15.74 | 13.66 | 30.49 | **31.37** | **31.60** | **32.56** | **46.08** | **82.72** | **0.210** | **0.261** | **0.317** | **0.381** | **0.503** | **0.564** |



(a) Bilinear (b) Super-FAN [5] (c) GWAInet [9] (d) GFRNet [29] (e) pix2pixHD [43]

(f) HiFaceGAN [47] (g) mGANprior [12] (h) PULSE [34] (i) pSp [38] (j) GPEN (k) Ground truth

Figure 6: Face super-resolution results by state-of-the-art methods. The input image has a resolution of $16^2$.

bic downsampling) and cannot handle this challenging BFR task. The PSNR, FID and LPIPS results are listed in Table 2. One can see that our GPEN achieves comparable PSNR index to other competing methods, but it achieves significantly better results on FID and LPIPS indices, which are better measures than PSNR for the face image perceptual quality.

Figure 5 compares the BFR results on some degraded face images by the competing methods. One can see that the competing methods fail to produce reasonable face reconstructions. They tend to generate over-smoothed face images with distorted facial structures. However, our GPEN generate visually photo-realistic face images with clear hair, eye, eyebrow, tooth and mustache details. Even the background can also be partially constructed. This clearly validates the advantages of our GPEN model and the training strategy. More visual comparison results can be found in the supplementary file.

**Face Super-Resolution.** FSR aims to generate an HR image from the input LR version. It can be considered as a special case of BFR, where the image degradation process is specified (i.e., bicubic downsampling). To validate the generality of our GPEN, we still use our model trained for BFR to perform the FSR task, and compare it with those state-of-the-art methods designed for FSR, including Super-FAN [5], GFRNet [29], GWAInet [9], DFDNet [27], HiFaceGAN [47], mGANprior [12], PULSE [34], and pSp [38]. The zooming factor ranges from 8× to 256×, and the LR face images are simulated on the CelebA-HQ dataset.

The quantitative results are presented in Table 3. One can see that the naïve bilinear interpolator achieves the best PSNR index, though it cannot restore any facial details. This actually validates that PSNR is not a suitable index to measure FSR quality. GPEN achieves the best FID and LPIPS scores under almost all the zooming factors. Figure 6 presents a visual comparison example for zooming factor 64×. More visual comparison results can be found in the supplementary.

## 4.5. Experiments on Images in the Wild

Finally, we perform experiments on real-world LQ face images, which suffer from complex unknown degradations. We collected $1,000$ LQ face images from internet for testing. The BFR methods Pix2PixHD [43], Super-FAN [5], GFRNet [29], GWAInet [9], DFDNet [27] and HiFaceGAN [47] are used in the comparison. Figure 7 shows the BFR

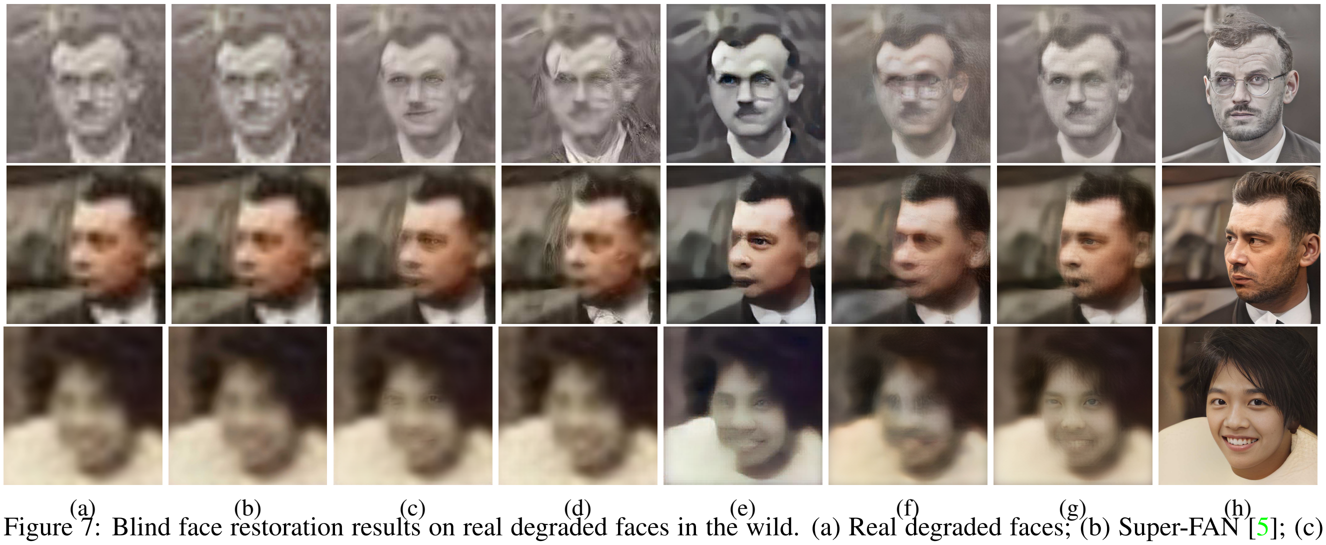


Figure 7: Blind face restoration results on real degraded faces in the wild. (a) Real degraded faces; (b) Super-FAN [5]; (c) GFRNet [29]; (d) GWAInet [9]; (e) Pix2PixHD [43]; (f) DFDNet [27]; (g) HiFaceGAN [47]; (h) GPEN.

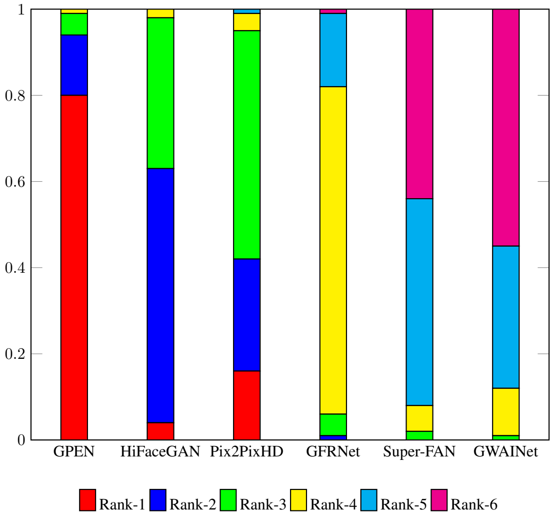


Figure 8: User study results of different BFR methods.

results on three images. One can see that the competing methods fail to restore the facial details. This is mainly because they are trained on synthesized data but have limited generalization capability to the images in the wild. Our method manages to overcome this difficulty by the carefully designed GAN prior embedding and fine-tuning strategies. It not only preserves well the global structure of the face, but also generates realistic details on the face components (e.g., hair, eye, mouth, etc.). Our GPEN can also be successfully used to renovate old photos, as we demonstrated in Figure 1. Please refer to the supplementary material for more results.

Since the commonly used quantitative metrics like PSNR and SSIM do not strongly correlate with human visual perception to image quality, we conduct a user study as a subjective assessment on the performance of our method and the competing methods. The BFR results of GPEN, Pix2PixHD [43], Super-FAN [5], GFRNet [29], GWAInet [9], DFDNet [27] and HiFaceGAN [47] on 113 real-world LQ face images collected from internet are presented in a random order to 17 volunteers for subjective evaluation. The volunteers are asked to rank the six BFR outputs of each input image according to their perceptual quality. Finally, we collect $1,915$ votes, and the statistics are presented in Figure 8. As can be seen, our GPEN method receives much more rank-1 votes than the other state-of-the arts.

## 5. Conclusion and Discussion

We proposed a simple yet effective GAN prior embedded network, namely GPEN, for BFR in the wild. By embedding a pre-trained GAN into a U-shaped DNN as a decoder, and fine-tuning the whole network with artificially degraded face images, our model learned to generate high quality face images from severely degraded ones. Our extensive experiments on synthetic data and real-world images demonstrated that GPEN outperforms the latest state-of-the-arts significantly, restoring clear facial details while retaining properly the image background. The proposed method can also be applied to other tasks such as face inpainting and face colorization. Some preliminary results were provided in the supplementary material.

The proposed GPEN does not allow multiple HQ images to be generated from a single LQ image in its current form. StyleGAN controls the synthesis via style mixing; however, such an operation may lead to inconsistent image background in GPEN. In the future, we will extend GPEN to allow multiple HQ outputs for a given LQ image. For example, we can use an extra HQ face image as a reference so that different HQ outputs can be generated by GPEN for different reference images.

## References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019. 3, 6

[2] S. Bakerand and T. Kanade. Hallucinating faces. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2000. 1

[3] Thirimachos Bourlai, Arun Ross, and Anil K. Jain. Restoring degraded face images: A case study in matching faxed, printed, and scanned photos. *IEEE Transactions on Information Forensics and Security*, 6(2):371–384, 2011. 1

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 3, 4

[5] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *CVPR*, 2018. 3, 6, 7, 8

[6] Lin Chen. The topological approach to perceptual organization. *Visual Cognition*, 12(4):553–637, 2005. 3

[7] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *CVPR*, 2018. 2

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 3

[9] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *CVPRW*, 2019. 1, 2, 3, 6, 7, 8

[10] Yaël Frégier and Jean-Baptiste Gouray. Mind2mind : transfer learning for gans. *arXiv*, 2019. 3

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, page 2672–2680, 2014. 3

[12] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. *ArXiv*, 2019. 1, 3, 4, 6, 7

[13] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. *CVPR*, 2019. 1, 2, 4

[14] Paul Hand, Oscar Leong, and Vladislav Voroninski. Phase retrieval under a generative prior. In *NIPS*, 2018. 3

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 5

[16] Xiaobin Hu, Wenqi Ren, John Lamaster, Xiaochun Cao, Xiaoming Li, Zechao Li, Bjoern Menze, and Wei Liu. Face super-resolution guided by 3d facial priors. In *ECCV*, 2020. 1, 2

[17] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *ICCV*, 2017. 2

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 1, 3

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[20] Tero Karras, Timo Aila, and Samuli Laine. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 3, 4, 5

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *ArXiv*, 2018. 1, 2, 3, 4, 5

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *ArXiv*, 2019. 1, 2, 3, 4, 5

[23] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark. *ArXiv*, 2019. 2

[24] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. *ArXiv*, 2017. 1, 2, 3

[25] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 1, 2, 3

[26] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 3

[27] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*, 2020. 1, 2, 5, 6, 7, 8

[28] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *CVPR*, 2020. 1, 2

[29] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *ECCV*, 2018. 1, 2, 3, 5, 6, 7, 8

[30] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 1, 2

[31] Guilin Liu, Fitsum A. Reda, Kevin Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. *ArXiv*, 2018. 2

[32] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsueprvised image-to-image translation. *Arxiv*, 2019. 3

[33] Cheng Ma, Zhenyu Jiang, Yongming Rao, Jiwen Lu, and Jie Zhou. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In *CVPR*, 2020. 2

[34] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020. 1, 3, 4, 6, 7

[35] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *ArXiv*, 2014. 3

[36] Masashi Nishiyama, Hidenori Takeshima, Jamie Shotton, Tatsuo Kozakaya, and Osamu Yamaguchi. Facial deblur inference to improve recognition of blurred faces. In *CVPR*, 2009. 1, 2

[37] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 3

[38] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *Arxiv*, 2020. 1, 3, 4, 6, 7

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet: a convolutional network for biomedical image segmentation. *Arxiv*, 2015. 4

[40] Ziyi Shen, Wei-sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *CVPR*, 2018. 2

[41] Ron Slossberg, Gil Shamai, and Ron Kimmel. High quality facial surface and texture synthesis via generative adversarial networks. In *ICCV*, 2018. 3

[42] Patricia L. Suarez, Angel D. Sappa, and Boris X. Vintimilla. Infrared image colorization based on a triplet dcgan architecture. In *CVPRW*, 2017. 3

[43] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 1, 3, 4, 6, 7, 8

[44] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018. 3

[45] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *CVPR*, 2020. 3

[46] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *ECCV*, 2018. 3

[47] Lingbo Yang, Chang Liu, Pan Wang, Shanshe Wang, Peiran Ren, Siweia Ma, and Gao Wen. Hifacegan: Face renovation via collaborative suppression and replenishment. *Arxiv*, 2020. 1, 2, 3, 6, 7, 8

[48] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. *ArXiv*, 2018. 1, 2, 3

[49] Xin Yu and Fatih Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *CVPR*, 2017. 1

[50] Haichao Zhang, Jianchao Yang, Yanning Zhang, Nasser M. Nasrabadi, and Thomas S. Huang. Close the loop: Joint blind image restoration and recognition with sparse representation prior. In *ICCV*, 2011. 1, 2

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3

[53] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NIPS*, 2017. 3