Blind Face Restoration via Integrating Face Shape and Generative Priors

Feida Zhu Junwei Zhu Wenqing Chu Xinyi Zhang Xiaozhong Ji Chengjie Wang* Ying Tai*
Youtu Lab, Tencent
{feidazhu,junweizhu,wenqingchu,savizhang,xiaozhongji,jasoncjwang,yingtai}@tencent.com
https://github.com/TencentYoutuResearch/BFR-SGPN

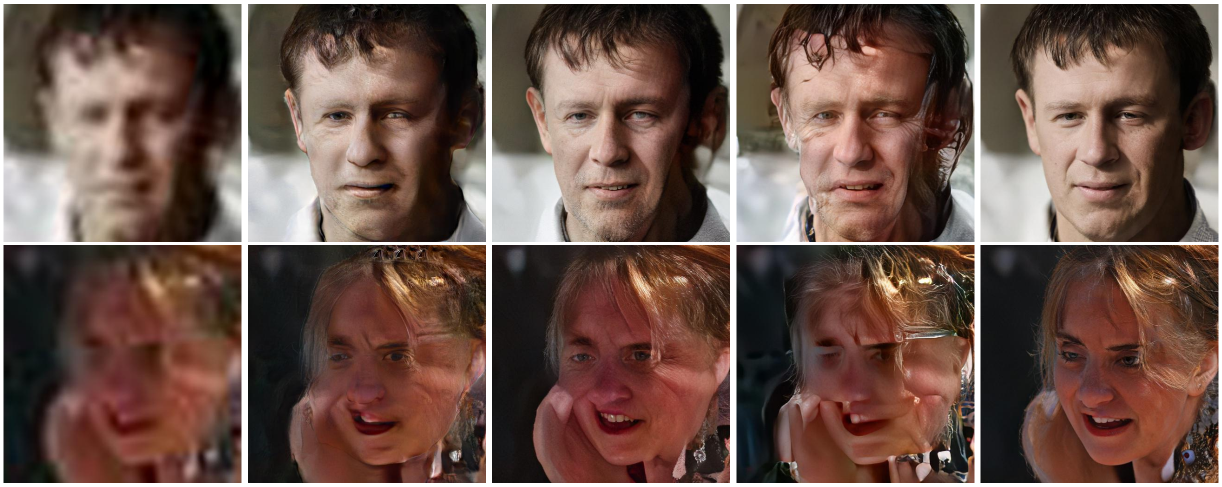


Figure 1. **Comparison with state-of-the-art methods on the real-world low-quality images**. Previous GAN-based methods may overemphasize generation and hallucinate a face with unnatural face shapes or non-realistic face components. The integration of shape and generative prior allows us to achieve realistic restoration results.

Abstract

Blind face restoration, which aims to reconstruct high-quality images from low-quality inputs, can benefit many applications. Although existing generative-based methods achieve significant progress in producing high-quality images, they often fail to restore natural face shapes and high-fidelity facial details from severely-degraded inputs. In this work, we propose to integrate shape and generative priors to guide the challenging blind face restoration. Firstly, we set up a shape restoration module to recover reasonable facial geometry with 3D reconstruction. Secondly, a pretrained facial generator is adopted as decoder to generate photo-realistic high-resolution images. To ensure high-fidelity, hierarchical spatial features extracted from the low-quality inputs and rendered 3D images are inserted into the decoder with our proposed Adaptive Feature Fusion Block (AFFB). Moreover, we introduce hybrid-level losses to jointly train the shape and generative priors together with other network parts such that these two priors better adapt to our blind face restoration task. The proposed Shape and Generative Prior integrated Network (SGPN) can restore high-quality images with clear face shapes and realistic facial details. Experimental results on synthetic and real-world datasets demonstrate SGPN performs favorably against state-of-the-art blind face restoration methods.

1. Introduction

Real-world low-quality face images suffer from unknown degradations during acquisition and Internet transmission. Blind Face Restoration (BFR) has been attracting considerable attention [2, 44, 52] due to its wide applications in real-world scenarios, such as restoring old images and film footage. However, it is still challenging to restore a high-fidelity image with natural facial geometry and realistic facial details from severely degraded face images.

[*] Chengjie Wang and Ying Tai are corresponding authors.

Previous works exploit different kinds of facial priors to help face restoration, *e.g.*, sparse constraints [6, 49, 54], parsing maps [2, 3, 43] and facial landmarks [3]. Additionally, the shape priors [15, 40] are adopted to guide face deblurring and super-resolution. However, they only handle specific image degradation, with over-smoothed results missing details. The resolutions (256 [40] and 128 [15]) are relatively low compared to recent methods. Besides, they first finetune the D3DFR [4] model on paired low- and high-quality images in advance before training their deblurring or super-resolution network. The finetuned model can not produce accurate 3D reconstructions for those LQ images with extreme pose or severe degradation (see Fig. 3).

With the rapid progress of GAN-based high-quality face generation [7, 20, 22], it is observed that the learned convolution weights of generative networks are able to capture a distribution over high-quality images [8, 34]. Such generative prior is adopted to produce visually realistic outputs from extremely low-quality images [8, 32, 34]. GPEN [52] and GFPGAN [44] further improve the fidelity by performing spatial modulation on the features of the embedded face generator. Unfortunately, existing methods [44, 52] often overemphasize generation and hallucinate faces with unnatural facial components on severely-degraded images.

To address the challenges, we introduce a new blind face restoration network designed to achieve a good balance between face shape reconstruction and face detail generation. The proposed Shape and Generative Prior integrated Network (SGPN) consists of two modules: 1) Face shape restoration module. 2) The shape and generative prior integration module. To restore the inherent face structure, we leverage a deep neural network (ResNet50 [13]) to predict the coefficients of 3D morphable face models (3DMMs [1]) from the low-quality input. The rendered 3D image contains natural and sharp face structures. The pretrained generator of StyleGAN2 [23] is adopted as our decoder to generate photo-realistic high-resolution image. We develop a dual-branch encoder to extract hierarchical spatial features from the low-quality inputs and its reconstructed result rendered from the predicted 3DMM coefficients. The spatial features are injected into the decoder progressively with a dedicated Adaptive Feature Fusion Block (AFFB), which learns a explicit weighting mask to adaptively fuse the spatial features from the dual-branch encoder.

The whole networks including shape and generative prior are jointly optimized with a combination of image- and mesh-level objectives. Specifically, the image-level loss favors pixel-level reconstruction and global realness, while the mesh-level loss encourages the face shape recovery. Experiments demonstrate that our method is able to recover realistic facial details and natural face shapes. In addition, our method can be easily generalized to face inpainting. In summary, the contributions of our work are as follows:

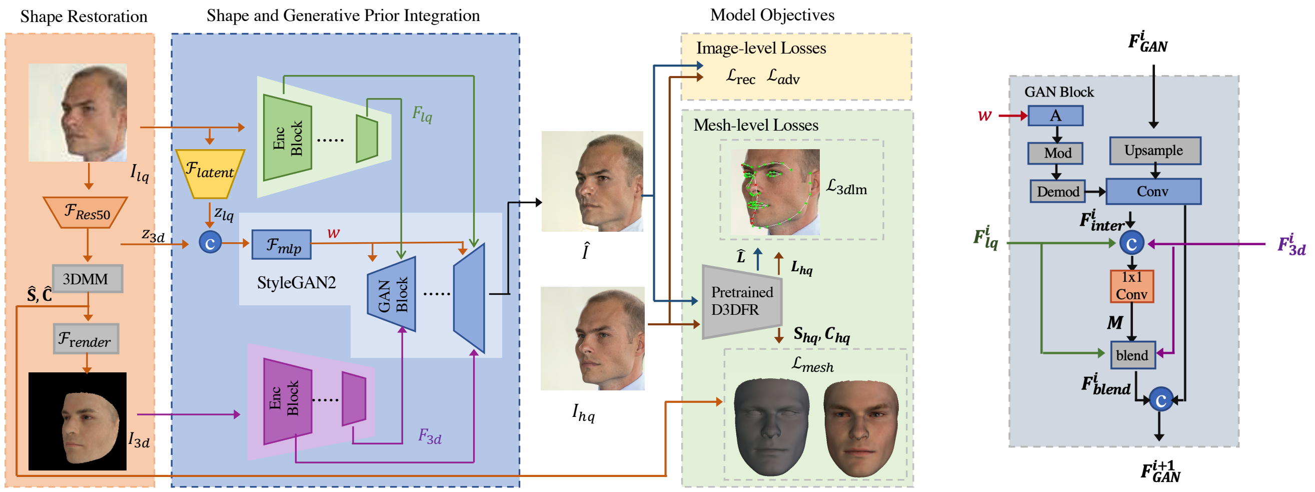
- To combine the merits of face shape and generative prior, we propose a blind face restoration framework to integrate them seamlessly. Our SGPN with adaptive feature fusion block achieves a good balance between face shape reconstruction and face detail generation.
- The face shape and generative priors are jointly optimized with other network parts to better facilitate the blind face restoration task.
- Extensive experiments demonstrate that our method achieves superior performance on both synthetic and real-world low-quality images, along with good generalization ability to face inpainting.

2. Related Work

Image Restoration. Image restoration has been a long-standing research topic. In recent years, deep convolutional neural networks have gained great success in large amounts of image restoration tasks including denoising [10, 55], deblurring [24, 43], super-resolution [3, 48], inpainting [29, 53], and compression artifacts reduction [5, 9]. However, most existing image restoration methods only consider a specific degradation type. It is challenging to restore the real-world low-quality images which contain complex unknown degradations.

Blind Face Restoration. As an important branch of image restoration, BFR has achieved great progress recently. BFR aims to handle severely degraded face images in the wild. One major line is the reference-based approaches. GFRNet [28] and ASFFNet [27] leverage a warped high-quality image to extract high frequency details for guiding the image reconstruction. However, requiring high-quality exemplar images limits the practical applicability. DFDNet [26] proposes to first generate deep dictionaries for facial components from high-quality images and then resort to them for better recovery of fine facial details. Another line is to adopt face generative networks [22, 23] to improve the reconstruction quality. GAN inversion based techniques [8, 32, 34] usually produce images with low fidelity and are time-consuming. GPEN [52] and GFPGAN [44] develop an encoder-decoder based architecture. They improve the fidelity by performing spatial modulation on the features of the embedded face decoder. Unfortunately, these generative prior based methods ignore recovering plausible facial geometry structure and may hallucinate faces with unnatural facial components on severely-degraded images. In contrast, our method explores 3D facial structure information from the low-quality inputs and integrates it with the generative prior to achieve natural face shape reconstruction and realistic facial details generation.

Priors in Face Restoration. There are strong priors in human faces. It is a common practice to exploit facial priors in face restoration. For example, facial semantic priors are



(a) Overall Architecture

(b) Adaptive Feature Fusion Block

Figure 2. **The overall architecture of our proposed SGPN**. (a) Our network contains the shape restoration module, and the shape and generative prior integration module. (b) Adaptive feature fusion block as our basic generative GAN block. We employ image-level reconstruction loss $\mathcal{L}_{rec}$ and adversarial loss $\mathcal{L}_{adv}$ to enhance fidelity and realness. Besides, we employ mesh-level 3D landmark loss $\mathcal{L}_{3dlm}$ and mesh loss $\mathcal{L}_{mesh}$ to enhance shape restoration.

leveraged through concatenating the degraded image and semantic labels as input to the deep neural networks [3, 43]. Furthermore, semantic features generated from a segmentation network are adopted as guidance of the spatially-adaptive normalization operation [35] for image inpainting [57]. However, these semantic priors extracted from severely degraded images may be unreliable [44, 52] and not able to provide detailed structure information. On the other hand, 3D facial prior has been proven to effectively capture the facial structure and applied in various facial editing applications such as face swapping [46], face deblurring [40] and face super-resolution [15, 16, 30]. Specifically, 3D coefficients representing identity, pose and expression are predicted first and fed into deep neural networks as guidance [15, 46]. Unfortunately, 3D facial prior can not provide vivid textures and realistic facial details. In recent years, several attempts [8, 32, 34, 44, 52] have been proposed to utilize generative priors embedded in StyleGAN [23]. Despite visually realistic outputs, they fail in recovering plausible geometry structure for extreme poses or heavy degradations. In summary, existing methods only employ a specific facial prior and thus could not handle the challenging blind face restoration perfectly. In this work, we not only combine 3D face shape and generative prior, but also propose a dedicated fusion module to make them collaborate seamlessly and thus obtain satisfactory performance. There are two reasons for using 3D prior: 1) It is challenging to train a 2D parsing model, which is prone to error for severely degraded images. 2) Although these 2D priors provide global component regions, they can not provide the detailed edges, illumination or expressions.

3. Methodology

First, we describe the overall architecture of SGPN. Then, we introduce the shape restoration module, and the shape and generative prior integration module in detail. Finally, we present the model objectives.

3.1. Overview of SGPN

The overall architecture of SGPN is depicted in Fig. 2a. Given a severely degraded low-quality image I_{lq}, our network first applies a shape restoration module to recover reasonable facial geometry with 3D reconstruction technique. Following the practice of D3DFR [4], we regress 3DMM coefficients with ResNet50, and then transform the coefficients to the face shape $\hat{\mathbf{S}}$ and colored texture $\hat{\mathbf{C}}$. The 3D reconstructions are further projected onto the 2D image plane to obtain a rendered 3D image I_{3d}. The 3DMM coefficients and the rendered 3D image serve as the shape prior.

Besides, we utilize StyleGAN2 [23] as our generative prior. We propose an integration module to combine these two priors. Specifically, we employ a latent encoder to extract the latent vector z_{lq} from LQ image. The 3DMM coefficients z_{3d} and the latent vector z_{lq} are concatenated together to generate intermediate latent code w for StyleGAN2. The intermediate code is then broadcasted to all GAN blocks to modulate the convolutional weights. The LQ image I_{lq} and rendered 3D image I_{3d} go through a dual-branch encoder to generate multi-resolution spatial features F_{lq} and F_{3d}, which will be further concatenated to the features inside the GAN block. We propose an adaptive feature fusion block to blend F_{lq} and F_{3d} adaptively, as shown in Fig. 2b. Details will be elaborated in Sec. 3.3.

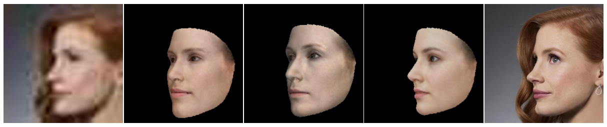


Figure 3. **Our method restores better shape** than the original D3DFR model [4] and the finetuned model [15] from LQ images.

3.2. Shape Restoration Module

ResNet-50 is leveraged to predict 3DMM coefficients, illumination and face pose from the input LQ image I_{lq}.

$$z_{3d} = \mathcal{F}_{Res50}(I_{lq}). \tag{1}$$

The output is a vector $z_{3d} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{p}) \in \mathbb{R}^{257}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{80}, \boldsymbol{\beta} \in \mathbb{R}^{64}, \boldsymbol{\delta} \in \mathbb{R}^{80}, \boldsymbol{\gamma} \in \mathbb{R}^{27}, \mathbf{p} \in \mathbb{R}^{6}$ represent the coefficients of the Basel Face Model (BFM) identity [37], expression [12], BFM texture, illumination with Spherical Harmonics (SH) [38] and pose. For completeness, we first review the 3D face reconstruction procedure.

3DMM Model. With predicted 3DMM coefficients, the 3D face shape $\hat{\mathbf{S}}$ and albedo texture $\hat{\mathbf{T}}$ are constructed as,

$$\hat{\mathbf{S}} = \hat{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \bar{\mathbf{S}} + \mathbf{B}_{id}\boldsymbol{\alpha} + \mathbf{B}_{exp}\boldsymbol{\beta}, \tag{2}$$
$$\hat{\mathbf{T}} = \hat{\mathbf{T}}(\boldsymbol{\delta}) = \bar{\mathbf{T}} + \mathbf{B}_{t}\boldsymbol{\delta}, \tag{3}$$

where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ are the mean face shape and albedo texture. $\mathbf{B}_{id}, \mathbf{B}_{exp}$ and $\mathbf{B}_{t}$ denotes the PCA bases of identity, expression and texture, respectively. Afterwards, Spherical Harmonics lighting is utilized to produce realistic illumination. The real color texture $\hat{\mathbf{C}}$ at vertex i is formulated as,

$$\hat{\mathbf{C}}(i) = \hat{\mathbf{c}}_i(\mathbf{n}_i, \mathbf{t}_i, \boldsymbol{\gamma}) = \mathbf{t}_i \cdot \sum_{b=1}^{27} \gamma_b \Phi_b(\mathbf{n}_i), \tag{4}$$

where $\mathbf{n}_i$ and $\mathbf{t}_i$ are the surface normal and albedo texture at vertex i. Φ_b is the SH basis function. We refer the readers to [4] for more details.

Renderer Model. With a differentiable mesh renderer, the reconstructed 3D face can be projected onto the 2D image plane according to the predicted face pose $\mathbf{p}$,

$$I_{3d} = \mathcal{F}_{render}(\hat{\mathbf{S}}, \hat{\mathbf{C}}, \mathbf{p}). \tag{5}$$

Given a severely degraded LQ image, the pretrained D3DFR can not provide accurate 3D reconstructions. The methods [15, 40] finetune the D3DFR on the paired LQ-HQ images. Instead, we joint train the shape restoration module with other modules such that the shape prior better adapts to our blind face restoration task. Besides, we introduce constraints on the reconstructed 3D mesh directly with mesh-level loss $\mathcal{L}_{mesh}$. Specifically, the shape $\hat{\mathbf{S}}$ and color texture $\hat{\mathbf{C}}$ should be close to $\mathbf{S}_{hq}$ and $\mathbf{C}_{hq}$ reconstructed from HQ images. Details will be elaborated in Sec. 3.4. As shown in Fig. 3, we can see that our method can restore better shape than the original D3DFR model and finetuned model [15] from the input LQ image. We also quantitatively measure the shape prediction accuracy of [4, 15] and our method by the vertex distance between the constructed meshes from LQ images and GT meshes from HQ images. The errors are 0.0183, 0.0121, 0.0058, respectively, demonstrating the superiority of our method.

3.3. Shape and Generative Prior Integration

The facial generative prior network is capable of generating high quality face image. A few attempts [8, 32, 44, 52] have been made to utilize StyleGAN as a facial prior to restore HQ images from LQ images. However, these methods may suffer from producing images with low fidelity [8, 32] and unnatural shapes [44, 52], given severely degraded LQ images. In comparison, the aforementioned shape prior can restore reasonable shapes from LQ images.

As shown in Fig. 2a, we propose a shape and generative prior integration module to take advantage of their merit. We first extract a latent vector z_{lq} from LQ image. The concatenation of 3D coefficients z_{3d} and latent vector z_{lq} will be mapped to the intermediate latent space $w \in \mathcal{W}$ to modulate the convolutional weights of StyleGAN,

$$z_{lq} = \mathcal{F}_{latent}(I_{lq}), \tag{6}$$
$$w = \mathcal{F}_{mlp}(z_{3d}, z_{lq}). \tag{7}$$

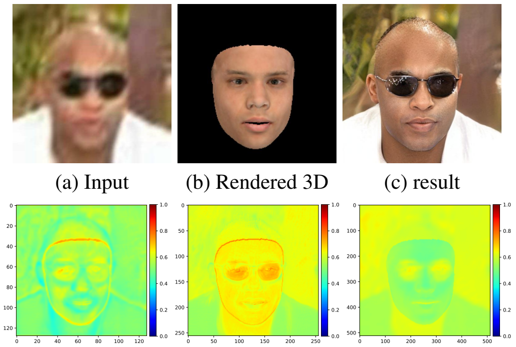
We decrease the mapping network depth from 8 to 2, as recommended by Karras *et al.* [21].

In order to produce high-fidelity and faithful restorations, we condition the generative model on the spatial features F_{lq} and F_{3d} extracted from I_{lq} and I_{3d}. We note that the rendered 3D image can provide sharp face structures. However, the 3DMM can not reconstruct inner mouth areas, eyes or accessories (*i.e.* sunglasses) on the face. One example is shown in Fig. 4. We propose an Adaptive Feature Fusion Block (AFFB) to adaptively fuse F_{lq} and F_{3d}. Fig. 2b shows the detailed structure of AFFB. At resolution scale i, we first generate a spatial mask,

$$F^i_{inter} = \mathbf{StyleConv}(F^i_{GAN}|w), \tag{8}$$
$$M = \mathbf{Conv}_{1\times1}(F^i_{inter}, F^i_{lq}, F^i_{3d}), \tag{9}$$

where the operation **StyleConv** denotes the style convolution in StyleGAN. More details can be found in [23]. The generated spatial mask has the same size with F^i_{lq} and F^i_{3d}. The 1×1 *Conv* is followed by a sigmoid activation such that the mask values are between 0 and 1. The blended feature is formulated as:

$$F^i_{blend} = F^i_{lq} \cdot M + F^i_{3d} \cdot (1 - M), \tag{10}$$



(a) Input (b) Rendered 3D (c) result

(d) Scale 128 (e) Scale 256 (f) Scale 512

Figure 4. **Visualization of learned spatial masks of AFFB**.

We show the visualizations of the spatial mask by calculating the mean across the channel dimension. The visualizations at scale 128, 256 and 512 are presented in Fig. 4. We can see that the sunglass regions have relatively higher activations than the skin regions. It indicates that the LQ spatial feature F_{lq}^i contributes more to the sunglass regions than the 3D spatial feature F_{3d}^i. Since the mask is learned without supervision, the background regions of F_{3d} are hard to be masked out completely for all channels.

3.4. Model Objectives

Recall that our SGPN includes the following trainable parts, 3DMM coefficients prediction model $\mathcal{F}_{Res50}$, latent encoder $\mathcal{F}_{latent}$, mapping network $\mathcal{F}_{mlp}$, spatial feature encoder $\mathcal{F}_{enc}$, and generative blocks $\mathcal{F}_{gan}$.

Our learning objectives can be divided into two categories: 1) Image-level losses, and 2) mesh-level losses.

Image-level Losses: We adopt the widely-used L1 loss as our reconstruction loss:

$$\mathcal{L}_{rec} = \|\hat{I} - I_{hq}\|_1, \tag{11}$$

where $\hat{I}$ and I_{hq} denote the generated result and the HQ image. Adversarial loss is inherited from StyleGAN2,

$$\mathcal{L}_{adv} = \mathbb{E}_{\hat{I}} \log\Big(1 + \exp\big(-\mathrm{D}(\hat{\mathrm{I}})\big)\Big), \tag{12}$$

where D is the discriminator.

Mesh-level Losses: We use pretrained D3DFR to predict 3D mesh from $\hat{I}$ and I_{hq}. The constructed mesh contains ~35.7K vertices, from which we extract 68 pre-defined 3D landmark points [41]. Landmark loss is formulated by:

$$\mathcal{L}_{lm} = \frac{1}{68}\sum_{i=1}^{68} \|\hat{L}(i) - L_{lm}(i)\|_2, \tag{13}$$

where $\hat{L}$ and L_{lm} denote the 3D landmarks predicted from $\hat{I}$ and I_{hq}, respectively.

Predicting an accurate 3D reconstruction from a LQ image is non-trivial. We introduce mesh loss $\mathcal{L}_{mesh}$ to better adapt the shape prior to our BFR task. We re-use the pretrained D3DFR to predict the shape $\mathbf{S}_{hq}$ and color texture $\mathbf{C}_{hq}$ from I_{hq}. The mesh loss enforces $\hat{\mathbf{S}}$ and $\hat{\mathbf{C}}$ to be close to $\mathbf{S}_{hq}$ and $\mathbf{C}_{hq}$ at all vertices,

$$\mathcal{L}_{mesh} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{\mathbf{S}}(i) - \mathbf{S}_{hq}(i)\|_2 + \|\hat{\mathbf{C}}(i) - \mathbf{C}_{hq}(i)\|_2, \tag{14}$$

where i denotes the vertex index. The overall loss $\mathcal{L}$ is:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{lm}\mathcal{L}_{lm} + \lambda_{vt}\mathcal{L}_{vt}. \tag{15}$$

where $\lambda_{adv} = 1$, $\lambda_{lm} = 100$ and $\lambda_{vt} = 100$.

4. Experiments

4.1. Datasets and Implementation

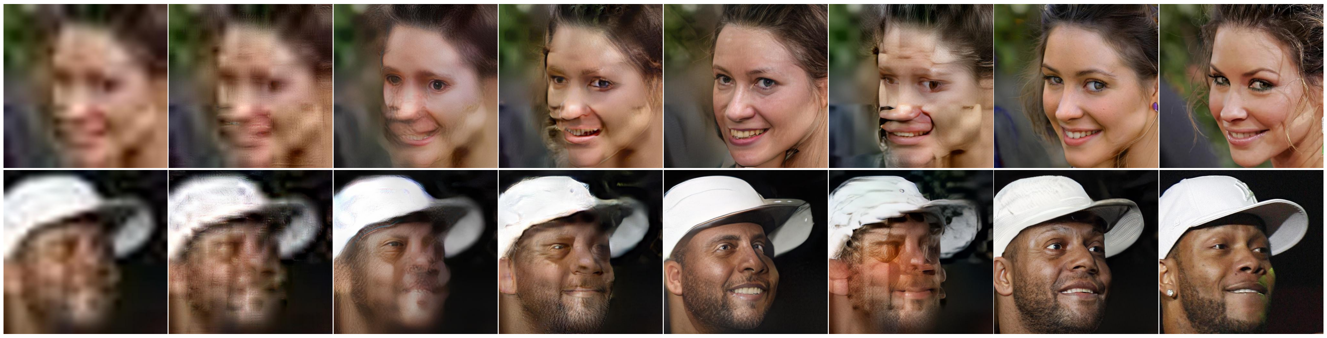
Training Datasets. We utilize FFHQ dataset [22], which consists of $70,000$ high-quality images, to train our SGPN. All images are resized to 512^2 during training. To build training data, the low quality images are synthesized from the HQ images with the following degradation model [28]:

$$I_{LQ} = ((I_{HQ} \otimes k)_{\downarrow_r} + n_\sigma)_{JPEG_q}. \tag{16}$$

The high quality image is first convolved with blur kernel k, which includes Gaussian blur with standard deviation $\varrho \in \{0 : 0.1 : 5\}$ and 32 motion blur kernels from [26]. Downsampling scale r, additive white Gaussian noise intensity σ, and JPEG compression quality factor q are randomly sampled from $\{1 : 20\}$, $\{0 : 10\}$ and $\{30 : 70\}$, respectively. The real image degradations usually become more complicated after several Internet transmissions. Inspired by [45], we repeat the degradation process one or two times randomly to obtain the final LQ image.

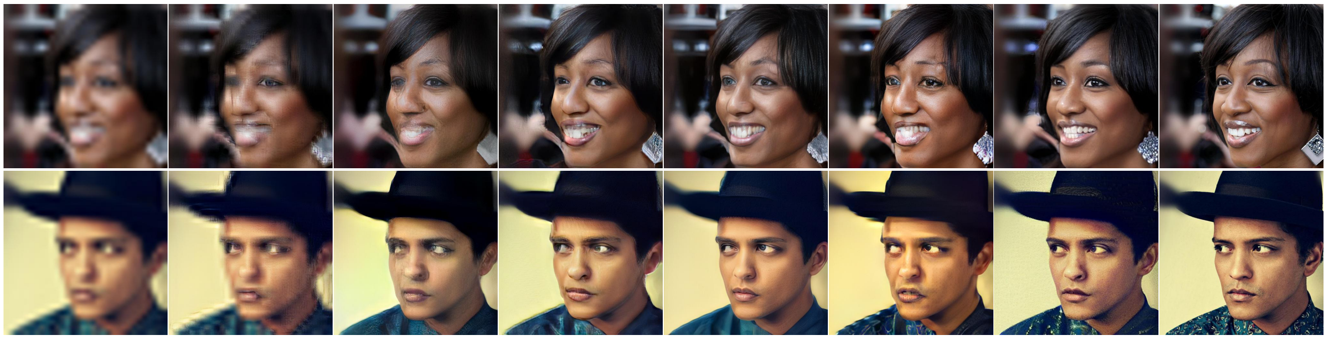
Implementation Details. We pretrained the D3DFR [4] and StyleGAN2 with 512^2 resolution [23] as our face shape and generative prior. The spatial feature encoder consists of seven down-sample convolutional layers. In practice, the encoders for I_{lq} and I_{3d} share the same weights. During training, the training data is augmented with horizontal flip. We adopt the Adam optimizer with a batch size of 32 for a total of 400K iterations. The learning rate was set to 0.002 for all trainable parameters. We implement our models with PyTorch [36] framework and the differentiable renderer with PyTorch3D [39]. The training time was 2 days with 8 Tesla V100 GPUs.

Inference Speed. The network parameters and inference speed on Tesla V100 of our method and other SOTA methods are reported in Tab. 1. Our inference time includes three



LQ input HiFaceGAN [50] DFDNet [26] PSFRGAN [2] GPEN [52] GFPGAN [44] SGPN GT

Figure 5. **Visual comparison of Blind Face Restorations (BFR)**. Our SGPN is able to restore reasonable face shapes and details.



Bicubic HiFaceGAN [50] DFDNet [26] PSFRGAN [2] GPEN [52] GFPGAN [44] SGPN GT

Figure 6. **Visual comparison of Face Super-Resolution (FSR)**. The ground truth HQ image is firstly downscaled and then upscaled to the original resolution with bicubic interpolation to synthesize LQ images. The scale factor is $16\times$.

Table 1. **Parameters and inference speed** on Tesla V100.

Method	DFDNet	PSFRGAN	GPEN	GFPGAN	Ours		
					$\mathcal{F}_{Res50}$	$\mathcal{F}_{render}$	$\mathcal{F}_G$
Params (M)	228.99	63.89	25.02	59.96	22.92	-	28.39
Inference Time (s)	0.8327	0.0759	0.0542	0.0556	0.0094	0.0256	0.0570

Table 2. **Quantitative comparison on CelebAHQ-Test for BFR**. Red and blue indicates the best and the second best result.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	Shape ↓
Bilinear	24.20	0.6690	0.6051	129.22	19.1081
HiFaceGAN [50]	23.54	0.5990	0.4345	75.63	13.6619
DFDNet [26]	23.07	0.5775	0.3556	25.08	1.6315
PSFRGAN [2]	23.07	0.6035	0.3232	24.29	0.3842
GPEN [52]	22.93	0.6048	0.2929	13.34	0.2891
GFPGAN [44]	22.06	0.5894	0.3119	20.21	0.6346
Ours	23.10	0.6146	0.2698	7.21	0.1667
GT	∞	1	0	6.51	0

parts: 3DMM coefficients prediction $\mathcal{F}_{Res50}$, the rendering process $\mathcal{F}_{render}$ and image generation $\mathcal{F}_G$. The whole procedures can be completed in $0.092s$, comparable to other SOTAs but with *better face shape and details restoration*.

4.2. Experiments on Synthetic Dataset

We use the CelebAHQ [19] test partition to simulate LQ images, which contains $3,000$ images. The widely used PSNR, SSIM [47] and LPIPS scores [56] are adopted to evaluate the restoration quality. FID score [14] is also reported. It should be noted that there is a domain gap between FFHQ dataset and CelebAHQ dataset. Therefore, we use the remaining CelebAHQ train partition as the reference data to evaluate FID score. We also report the shape error. For fair comparison, we use another 3D face reconstruction method RingNet [42] to regress the FLAME coefficients [25] from the result image and GT. The shape error is computed by L2 distance of the regressed coefficients.

We quantitatively compare our SGPN with state-of-the-art face restoration methods, including HiFaceGAN [50], PSFRGAN [2], DFDNet [26], GPEN [52] and GFPGAN [44]. Their official released models are adopted in the experiments. The comparisons are conducted in two tasks, *i.e.*, blind face restoration and face super-resolution.

Blind Face Restoration. Following the degradation model illustrated in Sec. 4.1, testing LQ images are synthesized for evaluation. The quantitative results are shown in Tab. 2. It can be seen that our SGPN achieves significantly better results on LPIPS and FID scores, showing that the outputs are closer to the original HQ images distribution. Our model achieves comparable PSNR and SSIM scores to other competing methods. It should be noted that the PSNR and SSIM

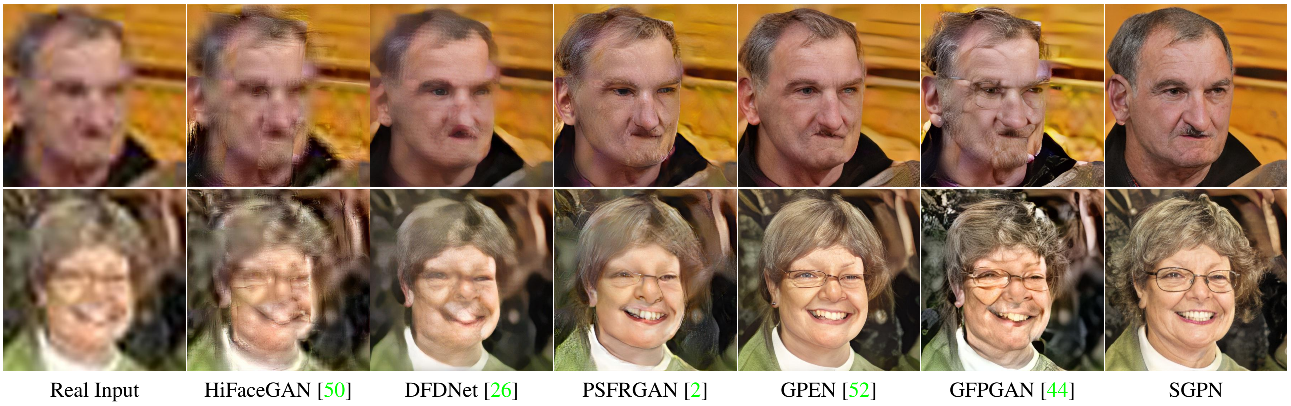


Figure 7. **Visual comparison of real face restorations**. Our SGPN is able to restore natural face shapes and details.

Table 3. **Quantitative comparison on CelebAHQ-Test for face super-resolution**.

Method	PSNR ↑			SSIM ↑			LPIPS ↓			FID ↓			Shape ↓		
	8x	16x	32x	8x	16x	32x	8x	16x	32x	8x	16x	32x	8x	16x	32x
Bilinear	26.93	23.93	21.03	0.7325	0.6710	0.6241	0.4677	0.5850	0.6461	109.71	135.96	190.77	0.0944	0.9132	76.9403
HiFaceGAN [50]	26.24	23.45	20.93	0.6918	0.6141	0.5821	0.2275	0.4218	0.5995	13.27	55.66	187.19	0.0524	0.7395	71.2096
DFD [26]	24.62	22.81	20.72	0.6103	0.5569	0.5654	0.2518	0.3507	0.5495	11.57	26.83	90.76	0.0570	0.4158	15.5578
PSFRGAN [2]	25.06	22.91	20.25	0.6667	0.6037	0.5382	0.2437	0.3069	0.3967	14.02	19.65	43.68	0.0704	0.2211	0.8777
GPEN [52]	25.07	23.09	20.39	0.6666	0.6135	0.5539	0.2230	0.2845	0.3648	12.82	16.36	29.78	0.0607	0.1803	0.9005
GFPGAN [44]	24.72	21.96	19.47	0.6684	0.5922	0.5254	0.2164	0.2730	0.3800	11.02	15.22	32.49	0.0584	0.1801	0.7570
Ours	25.37	23.35	20.61	0.6805	0.6286	0.5740	0.2062	0.2610	0.3480	9.59	10.45	16.88	0.0513	0.1474	0.5293
GT	∞	∞	∞	1	1	1	0	0	0	6.51	6.51	6.51	0	0	0

are not correlated well with human perceptions when there are severe degradations, because the BFR methods aim to hallucinate realistic face details (*e.g*., clear eyes and teeth) that do not exist in the LQ images.

Fig. 5 shows the qualitative comparison between our SGPN and other state-of-the-art methods. Most competing methods fail to restore realistic faces from severely degraded images. Among them, GPEN generates better results. However, it produces distorted face shape (first row) and eyes (second row). In comparison, our SGPN restores reasonable face shapes and visual-realistic facial details.

Face Super-Resolution. Following the common practice in SR tasks, the LQ images are synthesized with bicubic downsampling. FSR experiments are conducted under three scale factors, $8\times$, $16\times$ and $32\times$, respectively. The low-resolution images are resized back to the original resolution with bicubic interpolation before passing through the face restoration models. The quantitative results are listed in Tab. 3. We can see that the bicubic interpolation already achieves the best PSNR and SSIM. However, it cannot restore any meaningful facial details as shown in Fig. 6. SGPN achieves the best LPIPS and FID scores under all three scale factors. Fig. 6 presents the visual comparison for scale factor $16\times$. Our SGPN manages to generate better face shapes at large poses thanks to the carefully designed combination of shape prior and generative prior.

Table 4. **Quantitative comparison on real face restoration**.

Method	DFDNet [26]	PSFRGAN [2]	GPEN [52]	GFPGAN [44]	Ours
FID ↓	30.17	29.65	26.64	24.07	22.94
NIQE ↓	4.394	4.021	3.860	3.712	3.644

4.3. Experiments on Images in the Wild

The ultimate goal of BFR methods is to restore low-quality faces in the wild. We collected LQ images from CelebA [31], WIDERFACE [51] and LFW [17] for testing, forming $1,247$ test images. FID [14] and NIQE [33] are adopted as the non-reference perceptual metrics. Since our test images are mainly from CelebA, we use the whole CelebAHQ dataset as reference to calculate FID. The quantitative comparisons are shown in Tab. 4. Our SGPN achieves superior performance on both FID and NIQE.

Fig. 7 presents the visual comparisons. From the first row, we can see that our SGPN can restore better face details than the competing methods. Other methods can not generate natural mouth shape in this challenging case. In the second row, GPEN generates unnatural smiles and can not hallucinate complete glasses. In comparison, shape prior can provide reasonable face shape information. Then, our generative network can put more effort into restoring facial details including glasses on the face. As a result, SGPN can restore visually better smiles and clearer glasses.

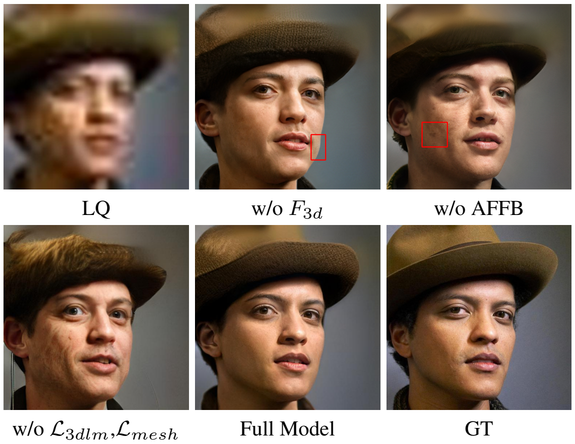


Figure 8. **Visual comparison between variants of SGPN**.

Table 5. **Comparison of different variants of SGPN**.

method	**A** w/o $\mathcal{F}_{3d}$	**B** w/o AFFB	**C** w/o $\mathcal{L}_{3dlm}$,$\mathcal{L}_{mesh}$	Full Model
FID ↓	24.36	23.42	24.63	**22.94**

Table 6. **Quantitative comparison on face inpainting**.

method	Mask Ratio				
	10%	20%	30%	40%	50%
GPEN [52]	31.05/0.0143	27.33/0.1825	25.84/0.2017	24.27/0.1985	23.30/0.2696
CTSDG [11]	32.28/0.0113	30.43/0.0443	28.28/0.1478	26.44/0.1682	25.80/0.2769
SGPN	**34.61/0.0061**	**31.89/0.0360**	**29.07/0.0631**	**27.40/0.1393**	**26.19/0.1422**

4.4. Ablation Studies

To evaluate the effectiveness of our proposed SGPN, we conduct experiments on three variants of our method. Variant **A** (w/o $\mathcal{F}_{3d}$) represents removing the encoder branch for 3D images. Only spatial feature F_{lq} is concatenated the features in the GAN block. Variant **B** (w/o AFFB) represents removing the adaptive feature fusion block. The encoded spatial features F_{lq} and F_{3d} are directly added rather than adaptively fused. Variant **C** (w/o $\mathcal{L}_{3dlm}$,$\mathcal{L}_{mesh}$) denotes removing mesh-level losses during training. The finetuned D3DFR model [15] is used to construct 3D images.

The FID metrics on the real test images of **A**, **B**, **C** and our full model are listed in Tab. 5. Fig. 8 shows one BFR example. Without the shape prior, the edge of the face generated by model **A** protrudes unnaturally. On the other hand, model **B** generates unwanted freckles on the cheek. Model **C** produces apparently worse results. Our full model has none of the above flaws. From the above observations, it can be inferred that the shape prior helps regularize face structures. The AFFB block allows our model to adaptively condition on F_{lq} and F_{3d} to synthesize realistic textures. The mesh-level loss enforces the shape prior to better adapt to our blind face restoration task.

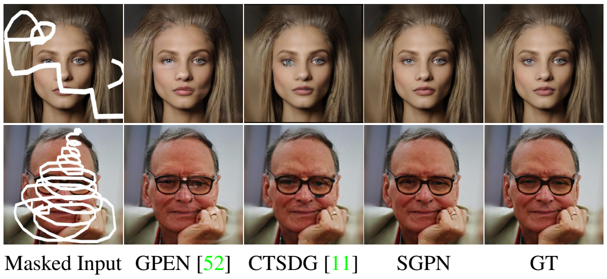


Figure 9. **Visual comparison with SOTA inpainting methods**.

4.5. Extension to Face Inpainting

Besides blind face restoration, our method can be easily generalized to face inpainting. The shape prior can be used to guide face inpainting. We use the public-available QD-IMD [18] masks to draw irregular holes on the FFHQ dataset to synthesize training pairs. We compare with state-of-the-art face inpainting methods, including GPEN [52] and CTSDG [11]. We conduct testing with different mask ratios on the CelebAHQ dataset. Larger mask ratio means that more pixels are erased. The quantitative comparisons are shown in Tab. 6. Our SGPN achieves better performance at all mask ratios. Fig. 9 presents the qualitative comparisons. We can see that there are still visible strokes in GPEN result. On the other hand, CTSDG performance suffers from low resolution. In comparison, our model retrieves better high-quality faces.

5. Conclusion

In this paper, we propose a novel approach for blind face restoration through integrating face shape and generative priors. The shape restoration module first predicts the parameters of 3DMMs from the low-quality observation and then renders a new facial image which exhibits accurate facial structure information. After that, the shape and generative prior integration module combines the priors seamlessly with adaptive feature fusion block. Moreover, the face shape and generative priors are jointly optimized with other network parts such that these two priors better adapt to our blind face restoration task. Extensive experiments on both synthetic and real-world benchmarks demonstrate that the proposed SGPN is superior to existing face restorations methods in terms of face shape and texture recovery.

Limitations. Our model relies on 3DMM model [4] to restore the face shape. Better 3D face model and 3D face reconstruction network may further improve the restoration quality. Besides, SGPN mainly focuses on facial part and may overlook the background region restoration.

Societal Impact. The identity information is actually difficult to restore from completely degraded images. The restored faces might not have the same identity as the severely degraded inputs although 3D prior is used.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the annual conference on Computer graphics and interactive techniques*, 1999. 2

[2] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. In *CVPR*, 2021. 1, 2, 6, 7

[3] 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, 3

[4] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPRW*, 2019. 2, 3, 4, 5, 8

[5] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *ICCV*, 2015. 2

[6] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on image processing*, 2011. 2

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[8] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, 2020. 2, 3, 4

[9] Jun Guo and Hongyang Chao. Building dual-domain representations for compression artifacts reduction. In *ECCV*, 2016. 2

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

[11] Xiefan Guo, Hongyu Yang, and Di Huang. Image inpainting via conditional texture and structure dual generation. In *ICCV*, 2021. 8

[12] Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 4

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[14] 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 *NeurIPS*, 2017. 6, 7

[15] 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. 2, 3, 4, 8

[16] Xiaobin Hu, Wenqi Ren, Jiaolong Yang, Xiaochun Cao, David P Wipf, Bjoern Menze, Xin Tong, and Hongbin Zha. Face restoration via plug-and-play 3d facial priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[17] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 7

[18] Karim Iskakov. Semi-parametric image inpainting. *arXiv:1807.02855*, 2018. 8

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2017. 6

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 2

[21] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *arXiv:2106.12423*, 2021. 4

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 5

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2, 3, 4, 5

[24] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. 2

[25] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2017. 6

[26] 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. 2, 5, 6, 7

[27] 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. 2

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

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

[30] Jing Liu, Weikang Wang, Jiexiao Yu, Chunping Zhang, and Yuting Su. 3dfp-fcgan: Face completion generative adversarial network with 3d facial prior. *Journal of Visual Communication and Image Representation*, 82:103380, 2022. 3

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 7

[32] 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. 2, 3, 4

[33] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 2012. 7

[34] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 3

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

[36] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NeurIPSW*, 2017. 5

[37] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *Proceedings of the IEEE international conference on advanced video and signal based surveillance*, 2009. 4

[38] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the annual conference on Computer graphics and interactive techniques*, 2001. 4

[39] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 5

[40] Wenqi Ren, Jiaolong Yang, Senyou Deng, David Wipf, Xiaochun Cao, and Xin Tong. Face video deblurring using 3d facial priors. In *ICCV*, 2019. 2, 3, 4

[41] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, 2013. 5

[42] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *CVPR*, 2019. 6

[43] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *CVPR*, 2018. 2, 3

[44] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 1, 2, 3, 4, 6, 7

[45] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021. 5

[46] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. In *IJCAI*, 2021. 3

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 2004. 6

[48] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *ICCV*, 2017. 2

[49] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 2010. 2

[50] Lingbo Yang, Shanshe Wang, Siwei Ma, Wen Gao, Chang Liu, Pan Wang, and Peiran Ren. Hifacegan: Face renovation via collaborative suppression and replenishment. In *ACMMM*, 2020. 6, 7

[51] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[52] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, 2021. 1, 2, 3, 4, 6, 7, 8

[53] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 2

[54] 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. 2

[55] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 2017. 2

[56] 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. 6

[57] Wendong Zhang, Junwei Zhu, Ying Tai, Yunbo Wang, Wenqing Chu, Bingbing Ni, Chengjie Wang, and Xiaokang Yang. Context-aware image inpainting with learned semantic priors. In *IJCAI*, 2021. 3