GFP-GAN

Towards Real-World Blind Face Restoration with Generative Facial Prior

Xintao Wang Yu Li Honglun Zhang Ying Shan
Applied Research Center (ARC), Tencent PCG
{xintaowang, ianyli, honlanzhang, yingsshan}@tencent.com
https://github.com/TencentARC/GFPGAN

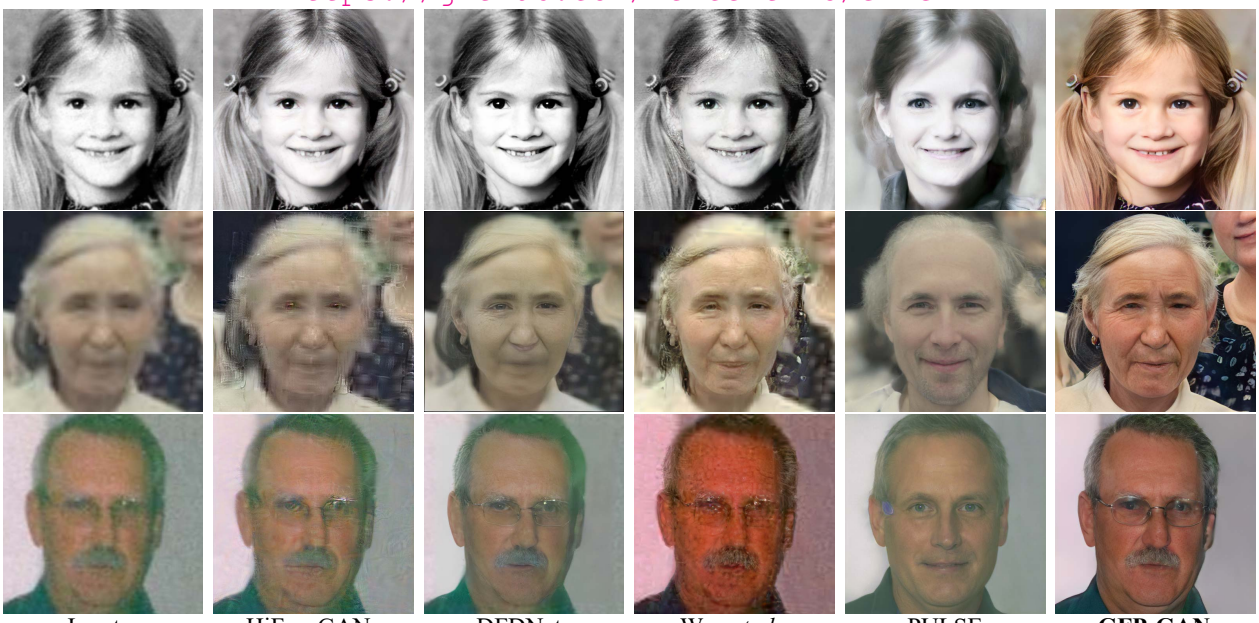


Figure 1: Comparisons with state-of-the-art face restoration methods: HiFaceGAN [67], DFDNet [44], Wan *et al*. [61] and PULSE [52] on the real-world low-quality images. While previous methods struggle to restore faithful facial details or retain face identity, our proposed GFP-GAN achieves a good balance of realness and fidelity with much fewer artifacts. In addition, the powerful generative facial prior allows us to perform restoration and color enhancement jointly. (**Zoom in for best view**)

Abstract

Blind face restoration usually relies on facial priors, such as facial geometry prior or reference prior, to restore realistic and faithful details. However, very low-quality inputs cannot offer accurate geometric prior while high-quality references are inaccessible, limiting the applicability in real-world scenarios. In this work, we propose GFP-GAN that leverages rich and diverse priors encapsulated in a pretrained face GAN for blind face restoration. This Generative Facial Prior (GFP) is incorporated into the face restoration process via spatial feature transform layers, which allow our method to achieve a good balance of realness and fidelity. Thanks to the powerful generative facial prior and delicate designs, our GFP-GAN could jointly restore facial details and enhance colors with just a single forward pass, while GAN inversion methods require image-specific optimization at inference. Extensive experiments show that our method achieves superior performance to prior art on both synthetic and real-world datasets.

1. Introduction

Blind face restoration aims at recovering high-quality faces from the low-quality counterparts suffering from unknown degradation, such as low-resolution [13, 48, 9], noise [71], blur [39, 58], compression artifacts [12], *etc*. When applied to real-world scenarios, it becomes more challenging, due to more complicated degradation, diverse poses and expressions. Previous works [9, 69, 6] typically exploit face-specific priors in face restoration, such as facial landmarks [9], parsing maps [6, 9], facial component heatmaps [69], and show that those *geometry facial priors* are pivotal to recover accurate face shape and details. However, those priors are usually estimated from input images and inevitably degrades with very low-quality inputs in the real world. In addition, despite their semantic guidance, the above priors contain limited texture information for restoring facial details (*e.g*., eye pupil).

Another category of approaches investigates *reference priors*, *i.e*., high-quality guided faces [46, 45, 11] or facial

The hidden secret behind GFP-GAN realistic output is it takes prior from high resolution output of StyleGanv2

component dictionaries [44], to generate realistic results and alleviate the dependency on degraded inputs. However, the inaccessibility of high-resolution references limits its practical applicability, while the limited capacity of dictionaries restricts its diversity and richness of facial details.

In this study, we leverage *Generative Facial Prior* (GFP) for real-world blind face restoration, *i.e.*, the prior implicitly encapsulated in pretrained face Generative Adversarial Network (GAN) [18] models such as StyleGAN [35, 36]. These face GANs are capable of generating faithful faces with a high degree of variability, and thereby providing rich and diverse priors such as geometry, facial textures and colors, making it possible to jointly restore facial details and enhance colors (Fig. 1). However, it is challenging to incorporate such generative priors into the restoration process. Previous attempts typically use GAN inversion [19, 54, 52]. They first 'invert' the degraded image back to a latent code of the pretrained GAN, and then conduct expensive image-specific optimization to reconstruct images. Despite visually realistic outputs, they usually produce images with low fidelity, as the low-dimension latent codes are insufficient to guide accurate restoration.

GAN inversion technique to reconstruct is expension and due to low dimension latent codes it can't accurately restore image.

To address these challenges, we propose the GFP-GAN with delicate designs to achieve a good balance of realness and fidelity in a single forward pass. Specifically, GFP-GAN consists of a degradation removal module and a pretrained face GAN as facial prior. They are connected by a direct latent code mapping, and several Channel-Split Spatial Feature Transform (CS-SFT) layers in a coarse-to-fine manner. The proposed CS-SFT layers perform *spatial* modulation on a split of features and leave the left features to directly pass through for better information preservation, allowing our method to effectively incorporate generative prior while retraining high fidelity. Besides, we introduce facial component loss with local discriminators to further enhance perceptual facial details, while employing identity preserving loss to further improve fidelity.

We summarize the contributions as follows. (1) We leverage rich and diverse generative facial priors for blind face restoration. Those priors contain sufficient facial textures and color information, allowing us to jointly perform face restoration and color enhancement. (2) We propose the GFP-GAN framework with delicate designs of architectures and losses to incorporate generative facial prior. Our GFP-GAN with CS-SFT layers achieves a good balance of fidelity and texture faithfulness in a single forward pass. (3) Extensive experiments show that our method achieves superior performance to prior art on both synthetic and real-world datasets.

2. Related Work

Image Restoration typically includes super-resolution [13, 48, 60, 49, 74, 68, 22, 50], denoising [71, 42, 26], deblurring [65, 39, 58] and compression removal [12, 21]. To achieve visually-pleasing results, generative adversarial network [18] is usually employed as loss supervisions to push the solutions closer to the natural manifold [41, 57, 64, 7, 14], while our work attempts to leverage the pretrained face GANs as generative facial priors (GFP).

Face Restoration. Based on general face hallucination [5, 30, 66, 70], two typical face-specific priors: geometry priors and reference priors, are incorporated to further improve the performance. The geometry priors include facial landmarks [9, 37, 77], face parsing maps [58, 6, 9] and facial component heatmaps [69]. However, 1) those priors require estimations from low-quality inputs and inevitably degrades in real-world scenarios. 2) They mainly focus on geometry constraints and may not contain adequate details for restoration. Instead, our employed GFP does not involve an explicit geometry estimation from degraded images, and contains adequate textures inside its pretrained network.

Reference priors [46, 45, 11] usually rely on reference images of the same identity. To overcome this issue, DFDNet [44] suggests to construct a face dictionary of each component (*e.g.*, eyes, mouth) with CNN features to guide the restoration. However, DFDNet mainly focuses on components in the dictionary and thus degrades in the regions beyond its dictionary scope (*e.g.*, hair, ears and face contour), instead, our GFP-GAN could treat faces as a whole to restore. Moreover, the limited size of dictionary restricts its diversity and richness, while the GFP could provide rich and diverse priors including geometry, textures and colors.

Generative Priors of pretrained GANs [34, 35, 36, 3] is previously exploited by GAN inversion [1, 76, 54, 19], whose primary aim is to find the closest latent codes given an input image. PULSE [52] iteratively optimizes the latent code of StyleGAN [35] until the distance between outputs and inputs is below a threshold. mGANprior [19] attempts to optimize multiple codes to improve the reconstruction quality. However, these methods usually produce images with low fidelity, as the low-dimension latent codes are insufficient to guide the restoration. In contrast, our proposed CS-SFT modulation layers enable prior incorporation on multi-resolution spatial features to achieve high fidelity. Besides, expensive iterative optimization is not required in our GFP-GAN during inference.

Channel Split Operation is usually explored to design compact models and improve model representation ability. MobileNet [28] proposes depthwise convolutions and GhostNet [23] splits the convolutional layer into two parts and uses fewer filters to generate intrinsic feature maps. Dual path architecture in DPN [8] enables feature re-usage and new feature exploration for each path, thus improving its representation ability. A similar idea is also employed in super-resolution [75]. Our CS-SFT layers share the similar spirits, but with different operations and purposes. We adopt spatial feature transform [63, 55] on one split and

GFP-GAN uses pretrained GAN models like StyleGAN which generate realistic face image with different textures, colors and variation which is useful attributes to restore faces details and enhance it.

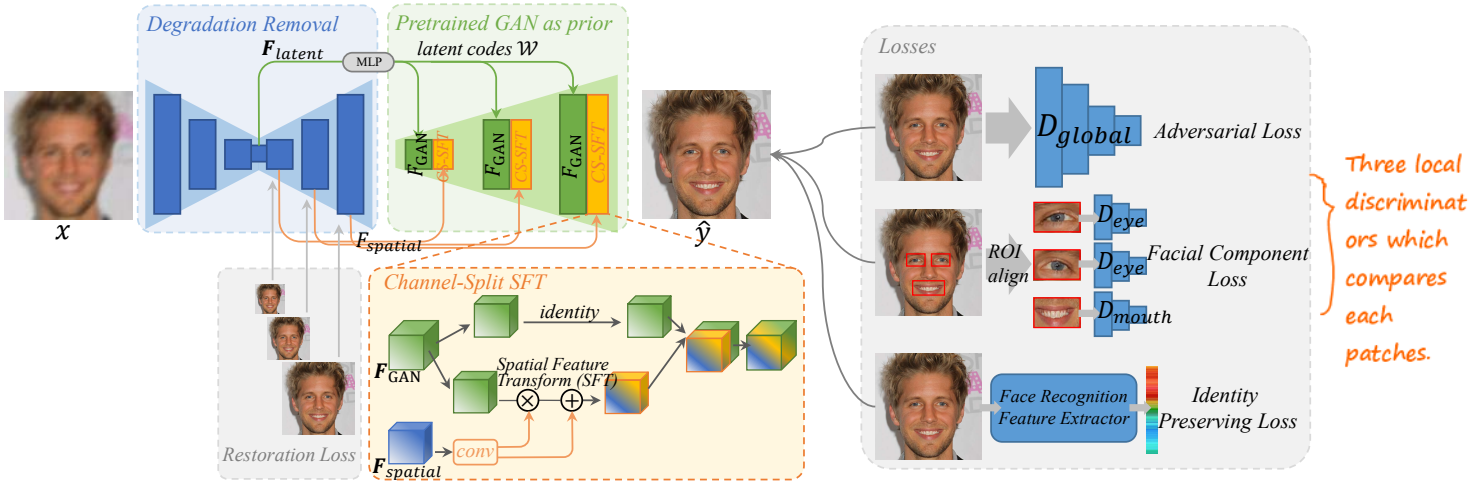


Figure 2: **Overview of GFP-GAN framework**. It consists of a degradation removal module (U-Net) and a pretrained face GAN as facial prior. They are bridged by a latent code mapping and several Channel-Split Spatial Feature Transform (CS-SFT) layers. During training, we employ 1) intermediate restoration losses to remove complex degradation, 2) Facial component loss with discriminators to enhance facial details, and 3) identity preserving loss to retain face identity.

leave the left split as identity to achieve a good balance of realness and fidelity.

Local Component Discriminators. Local discriminator is proposed to focus on local patch distributions [32, 47, 62]. When applied to faces, those discriminative losses are imposed on separate semantic facial regions [43, 20]. Our introduced facial component loss also adopts such designs but with a further style supervision based on the learned discriminative features.

3. Methodology

3.1. Overview of GFP-GAN

We describe GFP-GAN framework in this section. Given an input facial image $\boldsymbol{x}$ suffering from unknown degradation, the aim of blind face restoration is to estimate a high-quality image $\hat{\boldsymbol{y}}$, which is as similar as possible to the ground-truth image $\boldsymbol{y}$, in terms of realness and fidelity.

The overall framework of GFP-GAN is depicted in Fig. 2. GFP-GAN is comprised of a degradation removal module (U-Net) and a pretrained face GAN (such as StyleGAN2 [36]) as prior. They are bridged by a latent code mapping and several Channel-Split Spatial Feature Transform (CS-SFT) layers. Specifically, the degradation removal module is designed to remove complicated degradation, and extract two kinds of features, *i.e.* 1) latent features $\boldsymbol{F}_{latent}$ to map the input image to the closest latent code in StyleGAN2, and 2) multi-resolution spatial features $\boldsymbol{F}_{spatial}$ for modulating the StyleGAN2 features.

After that, $\boldsymbol{F}_{latent}$ is mapped to intermediate latent codes $\mathcal{W}$ by several linear layers. Given the close latent code to the input image, StyleGAN2 could generate *intermediate convolutional features*, denoted by $\boldsymbol{F}_{\text{GAN}}$. These features provide rich facial details captured in the weights of pretrained GAN. Multi-resolution features $\boldsymbol{F}_{spatial}$ are used to spatially modulate the face GAN features $\boldsymbol{F}_{\text{GAN}}$ with the proposed CS-SFT layers in a coarse-to-fine manner, achieving realistic results while preserving high fidelity.

During training, except for the global discriminative loss, we introduce facial component loss with discriminators to enhance the perceptually significant face components, *i.e.*, eyes and mouth. In order to retrain identity, we also employ identity preserving guidance.

3.2. Degradation Removal Module

Real-world blind face restoration faces with complicated and severer degradation, which is typically a mixture of low-resolution, blur, noise and JPEG artifacts. The degradation removal module is designed to explicitly remove the above degradation and extract 'clean' features $\boldsymbol{F}_{latent}$ and $\boldsymbol{F}_{spatial}$, alleviating the burden of subsequent modules. We adopt the U-Net [56] structure as our degradation remove module, as it could 1) increase receptive field for large blur elimination, and 2) generate multi-resolution features. The formulation is as follows:

$$\boldsymbol{F}_{latent}, \boldsymbol{F}_{spatial} = \texttt{U-Net}(\boldsymbol{x}). \tag{1}$$

The latent features $\boldsymbol{F}_{latent}$ is used to map the input image to the closest latent code in StyleGAN2 (Sec. 3.3). The multi-resolution spatial features $\boldsymbol{F}_{spatial}$ are used to modulate the StyleGAN2 features (Sec. 3.4).

In order to have an intermediate supervision for removing degradation, we employ the L1 restoration loss in each resolution scale in the early stage of training. Specifically, we also output images for each resolution scale of the U-Net decoder, and then restrict these outputs to be close to the pyramid of the ground-truth image.

The idea is to use U-NET to reconstruct the image, in which degradation of image can be improve at certain level. Prior taken from StyleGAN2 output where style vectors are generated through compressed latent vectors of U-NET passed to the several layers MLP. Decoding layers of U-NET output give multi-resolution spatial features is used for channel-wise spatial feature transform.

3.3. Generative Facial Prior and Latent Code Mapping

A pre-trained face GAN captures a distribution over faces in its leaned weights of convolutions, namely, generative prior [19, 54]. We leverage such pretrained face GANs to provide diverse and rich facial details for our task. A typical way of deploying generative priors is to map the input image to its closest latent codes Z, and then generate the corresponding output by a pretrained GAN [1, 76, 54, 19]. However, these methods usually require time-consuming iterative optimization for preserving fidelity. Instead of producing a final image directly, we generate *intermediate convolutional features* $\boldsymbol{F}_{\text{GAN}}$ of the closest face, as it contains more details and could be further modulated by input features for better fidelity (see Sec. 3.4).

Specifically, given the encoded vector $\boldsymbol{F}_{latent}$ of the input image (produced by the U-Net, Eq. 1), we first map it to intermediate latent codes $\mathcal{W}$ for better preserving semantic property *i.e.*, the intermediate space transformed from Z with several multi-layer perceptron layers (MLP) [76]. The latent codes $\mathcal{W}$ then pass through each convolution layer in the pre-trained GAN, and generate GAN features for *each resolution scale*.

$$\begin{aligned} \mathcal{W} &= \texttt{MLP}(\boldsymbol{F}_{latent}), \\ \boldsymbol{F}_{\text{GAN}} &= \texttt{StyleGAN}(\mathcal{W}). \end{aligned} \tag{2}$$

Discussion: Joint Restoration and Color Enhancement. Generative models capture diverse and rich priors beyond realistic details and vivid textures. For instance, they also encapsulate *color* priors, which could be employed in our task for joint face restoration and color enhancement. Real-world face images, *e.g.*, old photos, usually have black-and-white color, vintage yellow color, or dim color. Lively color prior in generative facial prior allows us to perform color enhancement including colorization [72]. We believe the generative facial priors also incorporate conventional geometric priors [9, 69], 3D priors [16], *etc*. for restoration and manipulation.

3.4. Channel-Split Spatial Feature Transform

In order to better preserve fidelity, we further use the input spatial features $\boldsymbol{F}_{spatial}$ (produced by the U-Net, Eq. 1) to modulate the GAN features $\boldsymbol{F}_{\text{GAN}}$ from Eq. 2. Preserving spatial information from inputs is crucial for face restoration, as it usually requires local characteristics for fidelity preservation, and adaptive restoration at different spatial locations of a face. Therefore, we employ Spatial Feature Transform (SFT) [63], which generates affine transformation parameters for spatial-wise feature modulation, and has shown its effectiveness in incorporating other conditions in image restoration [63, 44] and image generation [55].

Specifically, at each resolution scale, we generate a pair of affine transformation parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ from input features $\boldsymbol{F}_{spatial}$ by several convolutional layers. After that, the modulation is carried out by scaling and shifting the GAN features $\boldsymbol{F}_{\text{GAN}}$, formulated by:

Learned modulation parameters with single conv layer.

$$\begin{aligned} \boldsymbol{\alpha}, \boldsymbol{\beta} &= \texttt{Conv}(\boldsymbol{F}_{spatial}), \\ \boldsymbol{F}_{output} &= \texttt{SFT}(\boldsymbol{F}_{\text{GAN}} | \boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha} \odot \boldsymbol{F}_{\text{GAN}} + \boldsymbol{\beta}. \end{aligned} \tag{3}$$

To achieve a better balance of realness and fidelity, we further propose channel-split spatial feature transform (CS-SFT) layers, which perform spatial modulation on part of the GAN features by input features $\boldsymbol{F}_{spatial}$ (contributing to fidelity) and leave the left GAN features (contributing to realness) to directly pass through, as shown in Fig. 2:

$$\begin{aligned} \boldsymbol{F}_{output} &= \texttt{CS-SFT}(\boldsymbol{F}_{\text{GAN}} | \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \texttt{Concat}[\texttt{Identity}(\boldsymbol{F}_{\text{GAN}}^{split0}), \boldsymbol{\alpha} \odot \boldsymbol{F}_{\text{GAN}}^{split1} + \boldsymbol{\beta}], \end{aligned} \tag{4}$$

where $\boldsymbol{F}_{\text{GAN}}^{split0}$ and $\boldsymbol{F}_{\text{GAN}}^{split1}$ are split features from $\boldsymbol{F}_{\text{GAN}}$ in channel dimension, and $\texttt{Concat}[\cdot, \cdot]$ denotes the concatenation operation.

Why they split 0 and 1 only?

As a result, CS-SFT enjoys the benefits of directly incorporating prior information and effective modulating by input images, thereby achieving a good balance between texture faithfulness and fidelity. Besides, CS-SFT could also reduce complexity as it requires fewer channels for modulation, similar to GhostNet [23].

We conduct channel-split SFT layers at each resolution scale, and finally generate a restored face $\hat{\boldsymbol{y}}$.

3.5. Model Objectives

Total 4 types of losses are used.

The learning objective of training our GFP-GAN consists of: 1) reconstruction loss that constraints the outputs $\hat{\boldsymbol{y}}$ close to the ground-truth $\boldsymbol{y}$, 2) adversarial loss for restoring realistic textures, 3) proposed facial component loss to further enhance facial details, and 4) identity preserving loss.

Reconstruction Loss. We adopt the widely-used L1 loss and perceptual loss [33, 41] as our reconstruction loss $\mathcal{L}_{rec}$, defined as follows:

$$\mathcal{L}_{rec} = \lambda_{l1} \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_1 + \lambda_{per} \|\phi(\hat{\boldsymbol{y}}) - \phi(\boldsymbol{y})\|_1, \tag{5}$$

Perpetual loss (between feature maps) + L1 Loss (pixel-wise difference)

where ϕ is the pretrained VGG-19 network [59] and we use the $\{\texttt{conv1}, \cdots, \texttt{conv5}\}$ feature maps before activation [64]. λ_{l1} and λ_{per} denote the loss weights of L1 and perceptual loss, respectively.

Adversarial Loss. We employ adversarial loss L_{adv} to encourage the GFP-GAN to favor the solutions in the natural image manifold and generate realistic textures. Similar to StyleGAN2 [36], logistic loss [18] is adopted:

$$\mathcal{L}_{adv} = -\lambda_{adv} \mathbb{E}_{\hat{\boldsymbol{y}}} \, \texttt{softplus}(D(\hat{\boldsymbol{y}})) \tag{6}$$

Same as standard GAN loss.

where D denotes the discriminator and λ_{adv} represents the adversarial loss weight.

Facial Component Loss. In order to further enhance the perceptually significant face components, we introduce facial component loss with local discriminators for left eye,

The main purpose to use spatial feature transform is to learn modulation parameters that is adpative to the input, similar to SPADE instead of BatchNORM in GauGAN. The learned parameter pair adaptively influences the outputs by applying an affine transformation spatially to each intermediate feature maps. So, by that it preserves the spatial information from inputs which is important to restore the face. They use SFT in each resolution output of U-NET.

Perpetual loss compares high level differences, like content and style discrepancies between two different images that look similar, like the same photo but shifted by one pixel. It is similar to per-pixel loss function but main difference is it sums all the squared errors between all the pixels and taking the mean. The pre-trained network like VGG-19 in GFP-GAN which helps to define the perceptual loss functions needed to measure the perceptual differences of the predicted and real features maps.

right eye and mouth. As shown in Fig. 2, we first crop interested regions with ROI align [24]. For each region, we train separate and small local discriminators to distinguish whether the restore patches are real, pushing the patches close to the natural facial component distributions.

Inspired by [62], we further incorporate a feature style loss based on the learned discriminators. Different from previous feature matching loss with spatial-wise constraints [62], our feature style loss attempts to match the Gram matrix statistics [15] of real and restored patches. Gram matrix calculates the feature correlations and usually effectively captures texture information [17]. We extract features from multiple layers of the learned local discriminators and learn to match these Gram statistic of intermediate representations from the real and restored patches. Empirically, we found the feature style loss performs better than previous feature matching loss in terms of generating realistic facial details and reducing unpleasant artifacts.

Mainly Gram matrix is used which captures information and is position invariant. So, it helps to effectively compare generated and real.

The facial component loss is defined as follows. The first term is the discriminative loss [18] and the second term is the feature style loss:

$$\mathcal{L}_{comp} = \sum_{\texttt{ROI}} \lambda_{local}\, \mathbb{E}_{\hat{\boldsymbol{y}}_{\texttt{ROI}}}[\log(1 - D_{\texttt{ROI}}(\hat{\boldsymbol{y}}_{\texttt{ROI}}))] + \lambda_{fs}\, \|\texttt{Gram}(\psi(\hat{\boldsymbol{y}}_{\texttt{ROI}})) - \texttt{Gram}(\psi(\boldsymbol{y}_{\texttt{ROI}}))\|_1 \tag{7}$$

where ROI is region of interest from the component collection {left_eye, right_eye, mouth}. $D_{\texttt{ROI}}$ is the local discriminator for each region. ψ denotes the multi-resolution features from the learned discriminators. λ_{local} and λ_{fs} represent the loss weights of local discriminative loss and feature style loss, respectively.

Identity Preserving Loss. We draw inspiration from [31] and apply identity preserving loss in our model. Similar to perceptual loss [33], we define the loss based on the feature embedding of an input face. Specifically, we adopt the pretrained face recognition ArcFace [10] model, which captures the most prominent features for identity discrimination. The identity preserving loss enforces the restored result to have a small distance with the ground truth in the compact deep feature space:

$$\mathcal{L}_{id} = \lambda_{id} \|\eta(\hat{\boldsymbol{y}}) - \eta(\boldsymbol{y})\|_1, \tag{8}$$

where η represents face feature extractor, *i.e.* ArcFace [10] in our implementation. λ_{id} denotes the weight of identity preserving loss.

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

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{adv} + \mathcal{L}_{comp} + \mathcal{L}_{id}. \tag{9}$$

The loss hyper-parameters are set as follows: $\lambda_{l1} = 0.1$, $\lambda_{per} = 1$, $\lambda_{adv} = 0.1$, $\lambda_{local} = 1$, $\lambda_{fs} = 200$ and $\lambda_{id} = 10$.

4. Experiments

Interesting, Go through it. They prove GFP-GAN is more better than others.

4.1. Datasets and Implementation

Training Datasets. We train our GFP-GAN on the FFHQ dataset [35], which consists of 70,000 high-quality images. We resize all the images to 512^2 during training.

Our GFP-GAN is trained on synthetic data that approximate to the real low-quality images and generalize to real-world images during inference. We follow the practice in [46, 44] and adopt the following degradation model to synthesize training data:

$$\boldsymbol{x} = [(\boldsymbol{y} \circledast \boldsymbol{k}_\sigma) \downarrow_r + \boldsymbol{n}_\delta]_{\texttt{JPEG}_q}. \tag{10}$$

The high quality image $\boldsymbol{y}$ is first convolved with Gaussian blur kernel $\boldsymbol{k}_\sigma$ followed by a downsampling operation with a scale factor r. After that, additive white Gaussian noise $\boldsymbol{n}_\delta$ is added to the image and finally it is compressed by JPEG with quality factor q. Similar to [44], for each training pair, we randomly sample σ, r, δ and q from $\{0.2 : 10\}$, $\{1 : 8\}$, $\{0 : 15\}$, $\{60 : 100\}$, respectively. We also add color jittering during training for color enhancement.

Testing Datasets. We construct one synthetic dataset and three different **real** datasets with distinct sources. All these datasets have no overlap with our training dataset. We provide a brief introduction here.

- *CelebA-Test* is the synthetic dataset with 3,000 CelebA-HQ images from its testing partition [51]. The generation way is the same as that during training.
- *LFW-Test.* LFW [29] contains low-quality images **in the wild**. We group all the first image for each identity in the validation partition, forming 1711 testing images.
- *CelebChild-Test* contains 180 child faces of celebrities collected from **the Internet**. They are low-quality and many of them are black-and-white old photos.
- *WebPhoto-Test.* We crawled 188 low-quality photos **in real life from the Internet** and extracted 407 faces to construct the WebPhoto testing dataset. These photos have diverse and complicated degradation. Some of them are old photos with very severe degradation on both details and color.

Implementation. We adopt the pretrained StyleGAN2 [36] with 512^2 outputs as our generative facial prior. The channel multiplier of StyleGAN2 is set to one for compact model size. The UNet for degradation removal consists of seven downsamples and seven upsamples, each with a residual block [25]. For each CS-SFT layer, we use two convolutional layers to generate the affine parameters α and β respectively.

The training mini-batch size is set to 12. We augment the training data with horizontal flip and color jittering. We consider three components: left_eye, right_eye, mouth for face component loss as they are perceptually significant. Each component is cropped by ROI align [24] with face

For perpetual loss, they use pretrained VGG-19 network.
Similar to perpetual loss, they formulate identity preserving loss for feature embedding of face. For that, they use ArcFace pretrained face recognition model which captures most important features for identity discrimination i.e constraint to restored image should have minimum distance to ground truth image.

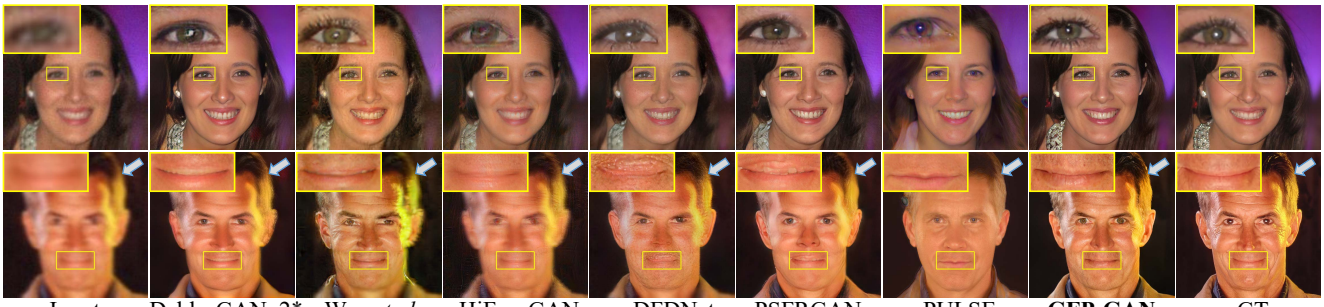


Figure 3: Qualitative comparison on the **CelebA-Test** for blind face restoration. Our GFP-GAN produces faithful details in eyes, mouth and hair. **Zoom in for best view.**

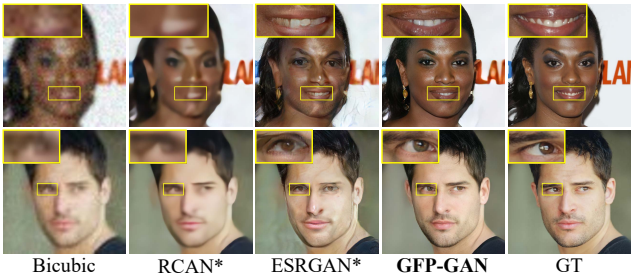


Figure 4: Comparison on the **CelebA-Test** for $\times 4$ face super-resolution. Our GFP-GAN restores realistic teeth and faithful eye gaze direction. **Zoom in for best view.**

landmarks provided in the origin training dataset. We train our model with Adam optimizer [38] for a total of 800k iterations. The learning rate was set to 2×10^{-3} and then decayed by a factor of 2 at the 700k-th, 750k-th iterations. We implement our models with the PyTorch framework and train them using four NVIDIA Tesla P40 GPUs.

4.2. Comparisons with State-of-the-art Methods

We compare our GFP-GAN with several state-of-the-art face restoration methods: HiFaceGAN [67], DFDNet [44], PSFRGAN [6], Super-FAN [4] and Wan *et al*. [61]. GAN inversion methods for face restoration: PULSE [52] and mGANprior [19] are also included for comparison. We also compare our GFP-GAN with image restoration methods: RCAN [74], ESRGAN [64] and DeblurGANv2 [40], and we finetune them on our face training set for fair comparisons. We adopt their *official* codes except for Super-FAN, for which we use a re-implementation.

For the evaluation, we employ the widely-used non-reference perceptual metrics: FID [27] and NIQE [53]. We also adopt pixel-wise metrics (PSNR and SSIM) and the perceptual metric (LPIPS [73]) for the CelebA-Test with Ground-Truth (GT). We measure the identity distance with angels in the ArcFace [10] feature embedding, where smaller values indicate closer identity to the GT.

Synthetic CelebA-Test. The comparisons are conducted under two settings: 1) blind face restoration whose inputs and outputs have the same resolution. 2) $4\times$ face super-resolution. Note that our method could take upsampled images as inputs for face super-resolution.

Table 1: Quantitative comparison on **CelebA-Test** for blind face restoration. **Red** and blue indicates the best and the second best performance. '*' denotes finetuning on our training set. Deg. represents the identity distance.

Methods	LPIPS↓	FID↓	NIQE ↓	Deg.↓	PSNR↑	SSIM↑
Input	0.4866	143.98	13.440	47.94	25.35	0.6848
DeblurGANv2* [40]	0.4001	52.69	4.917	39.64	**25.91**	**0.6952**
Wan *et al*. [61]	0.4826	67.58	5.356	43.00	24.71	0.6320
HiFaceGAN [67]	0.4770	66.09	4.916	42.18	24.92	0.6195
DFDNet [44]	0.4341	59.08	4.341	40.31	23.68	0.6622
PSFRGAN [6]	0.4240	47.59	5.123	39.69	24.71	0.6557
mGANprior [19]	0.4584	82.27	6.422	55.45	24.30	0.6758
PULSE [52]	0.4851	67.56	5.305	69.55	21.61	0.6200
GFP-GAN (ours)	**0.3646**	**42.62**	**4.077**	**34.60**	25.08	0.6777
GT	0	43.43	4.292	0	∞	1

The quantitative results for each setting are shown in Table. 1 and Table. 2. On both settings, GFP-GAN achieves the lowest LPIPS, indicating that our results is perceptually close to the ground-truth. GFP-GAN also obtain the lowest FID and NIQE, showing that the outputs have a close distance to the real face distribution and natural image distribution, respectively. Besides the perceptual performance, our method also retains better identity, indicated by the smallest degree in the face feature embedding. Note that 1) the lower FID and NIQE of our method than GT does not indicate that our performance is better than GT, as those 'perceptual' metrics are well correlated with the human-opinion-scores on a coarse scale, but *not always well correlated on a finer scale* [2]; 2) the pixel-wise metrics PSNR and SSIM are not correlation well with the subjective evaluation of human observers [2, 41] and our model is not good at these two metrics.

Qualitative results are presented in Fig. 3 and Fig. 4. 1) Thanks to the powerful generative facial prior, our GFP-GAN recovers faithful details in the eyes (pupils and eyelashes), teeth, *etc*. 2) Our method treats faces as whole in restoration and could also generate realistic hair, while previous methods that rely on component dictionaries (DFDNet) or parsing maps (PSFRGAN) fail to produce faithful hair textures (2nd row, Fig. 3). 3) GFP-GAN is capable of retaining fidelity, *e.g*., it produces natural closed mouth

Figure 5: Qualitative comparisons on three **real-world** datasets. **Zoom in for best view.**

Table 2: Quantitative comparison on **CelebA-Test** for 4× face super-resolution. **Red** and blue indicates the best and the second best performance. '*' denotes finetuning on our training set. Deg. represents the identity distance.

Methods	LPIPS↓	FID↓	NIQE ↓	Deg.↓	PSNR↑	SSIM↑
Bicubic	0.4834	148.87	10.767	49.60	25.377	0.6985
RCAN* [74]	0.4159	93.66	9.907	38.45	**27.24**	**0.7533**
ESRGAN* [64]	0.4127	49.20	4.099	51.21	23.74	0.6319
Super-FAN [4]	0.4791	139.49	10.828	49.14	25.28	0.7033
GFP-GAN (ours)	**0.3653**	**42.36**	**4.078**	**34.67**	25.04	0.6744
GT	0	43.43	4.292	0	∞	1

Table 3: Quantitative comparison on the *real-world* **LFW**, **CelebChild**, **WebPhoto**. **Red** and blue indicates the best and the second best performance. '*' denotes finetuning on our training set. Deg. represents the identity distance.

Dataset	**LFW-Test**		**CelebChild**		**WebPhoto**	
Methods	FID↓	NIQE ↓	FID↓	NIQE ↓	FID↓	NIQE ↓
Input	137.56	11.214	144.42	9.170	170.11	12.755
DeblurGANv2* [40]	57.28	4.309	110.51	4.453	100.58	4.666
Wan *et al.* [61]	73.19	5.034	115.70	4.849	100.40	5.705
HiFaceGAN [67]	64.50	4.510	113.00	4.855	116.12	4.885
DFDNet [44]	62.57	4.026	111.55	4.414	100.68	5.293
PSFRGAN [6]	51.89	5.096	107.40	4.804	88.45	5.582
mGANprior [19]	73.00	6.051	126.54	6.841	120.75	7.226
PULSE [52]	64.86	5.097	**102.74**	5.225	**86.45**	5.146
GFP-GAN (ours)	**49.96**	**3.882**	111.78	**4.349**	87.35	**4.144**

without forced addition of teeth as PSFRGAN does (2nd row, Fig. 3). And in Fig. 4, GFP-GAN also restores reasonable eye gaze direction.

Real-World LFW, CelebChild and WedPhoto-Test. To test the generalization ability, we evaluate our model on three different real-world datasets. The quantitative results are show in Table. 3. Our GFP-GAN achieves superior performance on all the three real-world datasets, showing its remarkable generalization capability. Although PULSE [52] could also obtain high perceptual quality (lower FID scores), it could not retain the face identity as shown in Fig 5.

The qualitative comparisons are shown in Fig. 5. GFP-GAN could jointly conduct face restoration and color enhancement for real-life photos with the powerful generative prior. Our method could produce plausible and realistic faces on complicated real-world degradation while other methods fail to recover faithful facial details or produces ar-

Table 4: Ablation study results on **CelebA-Test** under blind face restoration.

Configuration	LPIPS↓	FID↓	NIQE ↓	Deg.↓
Our GFP-GAN with SC-SFT	**0.3646**	**42.62**	**4.077**	**34.60**
a) No spatial modulation	0.550 (↑)	60.44 (↑)	4.183 (↑)	74.76 (↑)
b) Use SFT	0.387 (↑)	47.65 (↑)	4.146(↑)	34.38 (↓)
c) w/o GFP	0.379 (↑)	48.47 (↑)	4.153 (↑)	35.04 (↑)
d) − Pyramid Restoration Loss	0.369 (↑)	45.17 (↑)	4.284 (↑)	35.50 (↑)

tifacts (especially in WebPhoto-Test in Fig 5). Besides the common facial components like eyes and teeth, GFP-GAN also perform better in hair and ears, as the GFP prior takes the whole face into consideration rather than separate parts. With SC-SFT layers, our model is capable of achieving high fidelity. As shown in the last row of Fig. 5, most previous methods fail to recover the closed eyes, while ours could successfully restore them with fewer artifacts.

4.3. Ablation Studies

CS-SFT layers. As shown in Table. 4 [configuration a)] and Fig. 6, when we remove spatial modulation layers, *i.e.*, only keep the latent code mapping without spatial information, the restored faces could not retain face identity even with identity-preserving loss (high LIPS score and large Deg.). Thus, the multi-resolution spatial features used in CS-SFT layers is vital to preserve fidelity. When we switch CS-SFT layers to simple SFT layers [configuration b) in Table. 4], we observe that 1) the perceptual quality degrades on all metrics and 2) it preserves stronger identity (smaller Deg.), as the input image features impose influence on all the modulated features and the outputs bias to the degraded inputs, thus leading to lower perceptual quality. By contrast, CS-SFT layers provide a good balance of realness and fidelity by modulating a split of features.

Pretrained GAN as GFP. Pretrained GAN provides rich and diverse features for restoration. A performance drop is observed if we do not use the generative facial prior, as shown in Table. 4 [configuration c)] and Fig. 6.

Pyramid Restoration Loss. Pyramid restoration loss is employed in the degradation removal module and strengthens the restoration ability for complicated degradation in the real world. Without this intermediate supervision, the multi-resolution spatial features for subsequent modulations may still have degradation, resulting in inferior performance, as shown in Table. 4 [configuration d)] and Fig. 6.

Facial Component Loss. We compare the results of 1) removing all the facial component loss, 2) only keeping the component discriminators, 3) adding extra feature matching loss as in [62], and 4) adopting extra feature style loss based on Gram statistics [15]. It is shown in Fig 7 that component discriminators with feature style loss could better capture the eye distribution and restore the plausible details.

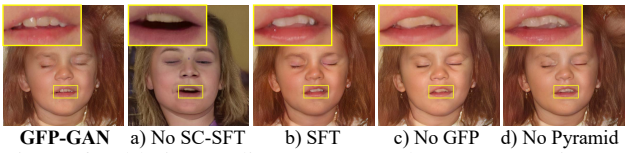


Figure 6: Ablation studies on CS-SFT layers, GFP prior and pyramid restoration loss. **Zoom in for best view.**

Figure 7: Ablation studies on facial component loss. In the figure, *D*, *fm*, *fs* denotes component discriminator, feature

Figure 8: Results on dark-skinned faces.

Figure 9: Limitations of our model. The results of PSFR-GAN [6] are also presented.

4.4. Discussion and Limitations

Training bias. Our method performs well on most dark-skinned faces and various population groups (Fig. 8), as our method uses both the pretrained GAN and input image features for modulation. Beside, we employ reconstruction loss and identity preserving loss to restrict the outputs to retain fidelity with inputs. However, when input images are gray-scale, the face color may have a bias (last example in Fig. 8), as the inputs do not contain sufficient color information. Thus, a diverse *and* balanced dataset is in need.

Limitations. As shown in Fig. 9, when the degradation of real images is severe, the restored facial details by GFP-GAN are twisted with artifacts. Our method also produces unnatural results for very large poses. This is because the synthetic degradation and training data distribution are different from those in real-world. One possible way is to learn those distributions from real data instead of merely using synthetic data, which is left as future work.

5. Conclusion

We have proposed the GFP-GAN framework that leverages the rich and diverse generative facial prior for the challenging blind face restoration task. This prior is incorporated into the restoration process with channel-split spatial feature transform layers, allowing us to achieve a good balance of realness and fidelity. Extensive comparisons demonstrate the superior capability of GFP-GAN in joint face restoration and color enhancement for real-world images, outperforming prior art.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019. 2, 4

[2] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *ECCVW*, 2018. 6

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[4] 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. 6, 7

[5] Qingxing Cao, Liang Lin, Yukai Shi, Xiaodan Liang, and Guanbin Li. Attention-aware face hallucination via deep reinforcement learning. In *CVPR*, 2017. 2

[6] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K. Wong. Progressive semantic-aware style transformation for blind face restoration. *arXiv:2009.08709*, 2020. 1, 2, 6, 7, 8

[7] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *CVPR*, 2018. 2

[8] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In *NeurIPS*, 2017. 2

[9] 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. 1, 2, 4

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 5, 6

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

[12] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *ICCV*, 2015. 1, 2

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 1, 2

[14] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. Deep generative adversarial compression artifact removal. In *ICCV*, 2017. 2

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 5, 8

[16] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *CVPR*, 2019. 4

[17] Muhammad Waleed Gondal, Bernhard Schölkopf, and Michael Hirsch. The unreasonable effectiveness of texture transfer for single image super-resolution. In *ECCV*, 2018. 5

[18] 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, 4, 5

[19] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, 2020. 2, 4, 6, 7

[20] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *ICCV*, 2019. 3

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

[22] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. Closed-loop matters: Dual regression networks for single image super-resolution. In *CVPR*, 2020. 2

[23] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *CVPR*, 2020. 2, 4

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 5

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[26] Majed El Helou, Ruofan Zhou, and Sabine Süsstrunk. Stochastic frequency masking to improve super-resolution and denoising networks. In *ECCV*, 2020. 2

[27] 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

[28] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*, 2017. 2

[29] 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, University of Massachusetts, Amherst, 2007. 5

[30] 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

[31] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *CVPR*, 2017. 5

[32] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 3

[33] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 4, 5

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

[35] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2018. 2, 5

[36] 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

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

[38] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[39] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. 1, 2

[40] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, 2019. 6, 7

[41] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2, 4, 6

[42] Stamatios Lefkimmiatis. Non-local color image denoising with convolutional neural networks. In *CVPR*, 2017. 2

[43] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *ACM MM*, 2018. 3

[44] 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, 4, 5, 6, 7

[45] 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

[46] 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, 5

[47] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *CVPR*, 2017. 3

[48] 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

[49] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *NeurIPS*, 2018. 2

[50] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, 2020. 2

[51] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[52] 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, 2, 6, 7

[53] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[54] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. In *ECCV*, 2020. 2, 4

[55] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 2, 4

[56] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015. 3

[57] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ECCV*, 2017. 2

[58] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *CVPR*, 2018. 1, 2

[59] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4

[60] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 2

[61] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *CVPR*, 2020. 1, 6, 7

[62] 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. 3, 5, 8

[63] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 2, 4

[64] 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. 2, 4, 6, 7

[65] Li Xu, Jimmy S Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. In *NeurIPS*, 2014. 2

[66] 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

[67] Lingbo Yang, Chang Liu, Pan Wang, Shanshe Wang, Peiran Ren, Siwei Ma, and Wen Gao. Hifacegan: Face renovation via collaborative suppression and replenishment. *ACM Multimedia*, 2020. 1, 6, 7

[68] Ke Yu, Xintao Wang, Chao Dong, Xiaoou Tang, and Chen Change Loy. Path-restore: Learning network path selection for image restoration. *arXiv:1904.10343*, 2019. 2

[69] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *ECCV*, pages 217–233, 2018. 1, 2, 4

[70] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *CVPR*, 2018. 2

[71] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE TIP*, 26(7):3142–3155, 2017. 1, 2

[72] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 4

[73] 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

[74] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 2, 6, 7

[75] Xiaole Zhao, Yulun Zhang, Tao Zhang, and Xueming Zou. Channel splitting network for single mr image super-resolution. *IEEE Transactions on Image Processing*, 28(11):5649–5662, 2019. 2

[76] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *ECCV*, 2020. 2, 4

[77] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *ECCV*, 2016. 2

GFP-GAN Summary

Prepared By: Sushant Gautam
Part of 30 Day GAN Paper Reading
https://github.com/sushant097/30-Days-GANs-Paper-Reading

Proposed GFP-GAN that take Generative Facial Prior (GFP) from pretrained GAN for blind face restoration. The hidden secret behind its sucess is it takes prior from high resolution output of StyleGanv2

Previous works try to restore facial details, where prior taken from low quality input and thus that latent vectors would not able to produce high quality output from generator.

GFP-GAN uses various techniques, ideas to restore face which seems realistic. U-NET, Generative Facial prior from StyleGAN2, and Channel-Split Spatial Feature Transormation (CS-SPT) are core techniques that make GFP-GAN output realistic than other previous work.

Important points of GFP-GAN:

1. It uses U-NET architecture for removal of complicated and severer degradation (mixture of low-resolution, blur, noise and JPEG artifacts)
2. Use several layers of MLP to convert latent vectors of U-NET to latent codes W which is input to StyleGANv2 to generate output of various styles.
3. Use CS-SPT which basically is to learn modulation parameters (α , β) that is adpative to the input and use in each layer of U-NET Decoder for each resolution output.
4. Local Component Discriminators to focused on local patch distributions.
4. Use Four types of loss:
    a) Reconstruction loss that constraints the outputs y^ close to the ground-truth y
    b) Adversarial loss for restoring realistic textures,
    c) Facial component loss to further enhance facial details, and
    d) Identity preserving loss.

GFP-GAN Architecture:

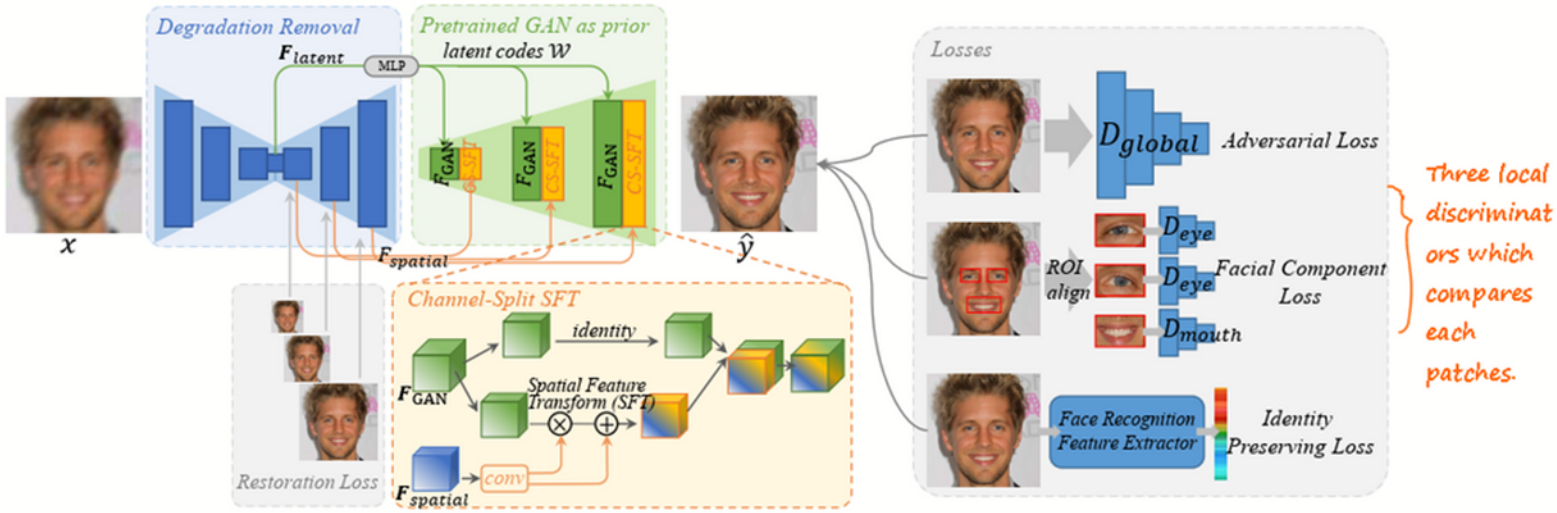


Figure 2: **Overview of GFP-GAN framework**. It consists of a degradation removal module (U-Net) and a pretrained face GAN as facial prior. They are bridged by a latent code mapping and several Channel-Split Spatial Feature Transform (CS-SFT) layers. During training, we employ 1) intermediate restoration losses to remove complex degradation, 2) Facial component loss with discriminators to enhance facial details, and 3) identity preserving loss to retain face identity.

Detail On Architecture:

Degradation Removal:

Why they use U-NET?

They uses U-NET model to remove complicated image degradation in some level.
The idea is U-NET compress input low quality image to form latent codes, and again it try to reconstruct it through upsampling.
If U-NET is trained then, it weights and biased adjusted such that it can reconstruct input image easily that improve quality of input image in some level.

This method uses pre-trained StyleGANv2 as facial prior which needs style vectors to generate output. So, U-NET latent code (compressed one) is passed through several layers of MLP which convert latent codes to W (similar concept of StyleGAN).
Also, output of each upsample layer of U-NET is spatial output which is passed through the CS-SPT that learn modulation parameters (α , β) through a single convolution layer i.e adaptive to the input image which not wash away useful details of input image in compare to BatchNorm or Instance Norm. This technique is similar to SPADE of GauGAN - (Check my previous Annotated Paper).

Others concept is cleared by above architecture itself.

Model Objectives:

Altogether GFP-GAN uses 4 losses

1. Reconstruction loss:

   It constraints the outputs y^ close to the ground-truth y .
   Perpetual loss (between feature maps of pretrained VGG-19) + L1 Loss (pixel-wise difference).

$$\mathcal{L}_{rec} = \lambda_{l1}\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_1 + \lambda_{per}\|\phi(\hat{\boldsymbol{y}}) - \phi(\boldsymbol{y})\|_1,$$

where ϕ is the pretrained VGG-19 network. λl1 and λper denote the loss weights of L1 and perceptual loss

2. Adversarial Loss:

   Standard GAN Loss that constraints GFP-GAN to generate realistic textures in a natural way.

$$\mathcal{L}_{adv} = -\lambda_{adv}\mathbb{E}_{\hat{\boldsymbol{y}}}\,\texttt{softplus}(D(\hat{\boldsymbol{y}})) \qquad (6)$$

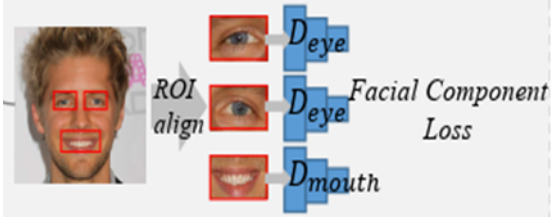
where D denotes the discriminator and λ_{adv} represents the adversarial loss weight.

Ladv = -λadv * Ey^ [log(1-sigmoid(D(G(z)))) + log(sigmoid(D(y)))]

3. Facial Component Loss:

It further enhance facial details like eye orientation, mouth.

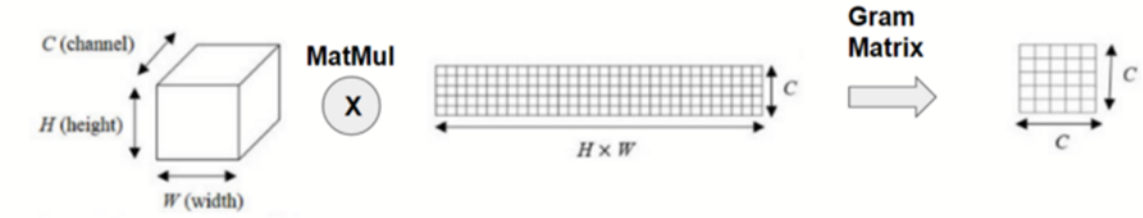
Use 3 local discriminator for left eye, right eye and mouth.



First crop interested regions with ROI align. Then, for each region, they train separate small local discriminators to distinguish whether the restore patches are real, pushing the patches close to the natural facial component distributions.

So, main concept is use of small patch at a time even of only one ROI (eg:eye) which enables discriminators to further capture small details and enforce generator to produce output having those details. - Use of PatchGAN in discriminator.

They use Gram Matrix for that which captures information and is position invariant. So, it helps to effectively compare generated and real image features.



Input : C Features of H x W

Reshape (C, HxW)

See application of Gram Matrix in paper called Neural Style Transfer.

Also, my explanation of it on Medium:

https://pub.towardsai.net/basic-intuition-on-neural-style-transfer-idea-c5ac179d1530

The gram matrix is a correlation operation i.e. dot product of feature maps at a layer that summarizes the activations that co-occur i.e it captures locality. What's being measured is whether, at a particular pixel position, generated image feature tends to cooccur with real image feature.

Discriminative Loss

$$\mathcal{L}_{comp} = \sum_{\text{ROI}} \lambda_{local}\, \mathbb{E}_{\hat{\boldsymbol{y}}_{\text{ROI}}}[\log(1 - D_{\text{ROI}}(\hat{\boldsymbol{y}}_{\text{ROI}}))] + \lambda_{fs}\, \|\texttt{Gram}(\psi(\hat{\boldsymbol{y}}_{\text{ROI}})) - \texttt{Gram}(\psi(\boldsymbol{y}_{\text{ROI}}))\|_1 \quad (7)$$

Feature Loss

where ROI is region of interest from the component collection {left_eye, right_eye, mouth}.

4. Identity Preserving Loss:

It constraint to preserve the facial identity.

For perpetual loss, they use pretrained VGG-19 network.

Similar to perpetual loss, they formulate identity preserving loss for feature embedding of face. For that, they use ArcFace pretrained face recognition model which captures most important features for identity discrimination i.e constraint to restored image should have minimum distance to ground truth image.

$$\mathcal{L}_{id} = \lambda_{id}\|\eta(\hat{\boldsymbol{y}}) - \eta(\boldsymbol{y})\|_1, \quad (8)$$

where η represents face feature extractor, i.e . ArcFace. λid denotes the weight of identity preserving loss

Total Objective Loss of GFP-GAN

Sum of all four losses.

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{adv} + \mathcal{L}_{comp} + \mathcal{L}_{id}. \quad (9)$$

The loss hyper-parameters are set as follows: $\lambda_{l1} = 0.1$, $\lambda_{per} = 1$, $\lambda_{adv} = 0.1$, $\lambda_{local} = 1$, $\lambda_{fs} = 200$ and $\lambda_{id} = 10$.

Output:

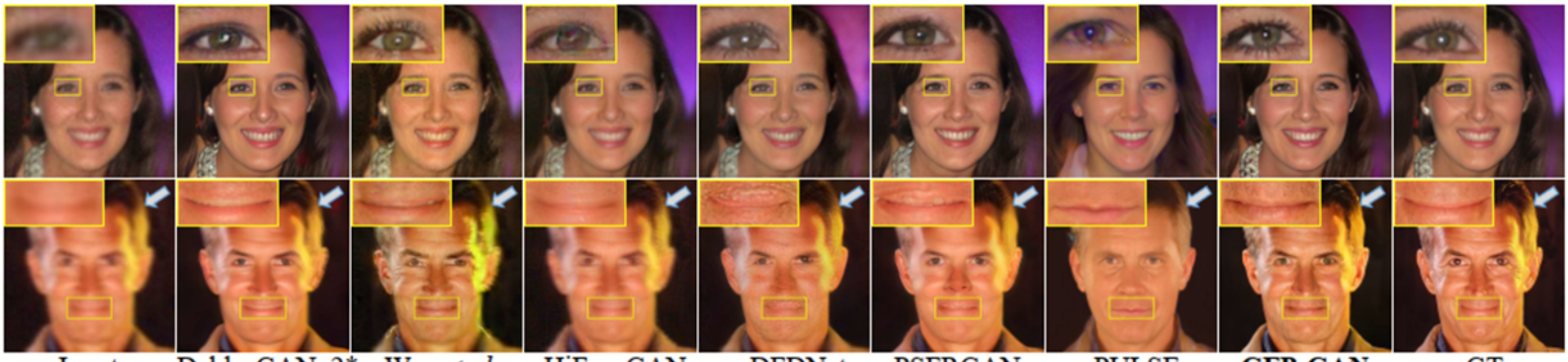


Figure 3: Qualitative comparison on the **CelebA-Test** for blind face restoration. Our GFP-GAN produces faithful details in eyes, mouth and hair. **Zoom in for best view.**

GFP-GAN outperform all other methods. You can easily see in above output.