Panini-Net: GAN Prior based Degradation-Aware Feature Interpolation for Face Restoration

Yinhuai Wang[†], Yujie Hu[†], Jian Zhang[†,‡]

[†]Peking University Shenzhen Graduate School, China
[‡]Peng Cheng Laboratory, China
{yinhuai; hhuyujie}@stu.pku.edu.cn, zhangjian.sz@pku.edu.cn

Abstract

Emerging high-quality face restoration (FR) methods often utilize pre-trained GAN models (*i.e.*, StyleGAN2) as GAN Prior. However, these methods usually struggle to balance realness and fidelity when facing various degradation levels. Besides, there is still a noticeable visual quality gap compared with pre-trained GAN models. In this paper, we propose a novel GAN Prior based degradation-aware feature interpolation network, dubbed Panini-Net, for FR tasks by explicitly learning the abstract representations to distinguish various degradations. Specifically, an unsupervised degradation representation learning (UDRL) strategy is first developed to extract degradation representations (DR) of the input degraded images. Then, a degradation-aware feature interpolation (DAFI) module is proposed to dynamically fuse the two types of informative features (*i.e.*, features from input images and features from GAN Prior) with flexible adaption to various degradations based on DR. Ablation studies reveal the working mechanism of DAFI and its potential for editable FR. Extensive experiments demonstrate that our Panini-Net achieves state-of-the-art performance for multi-degradation face restoration and face super-resolution. The source code is available at https://github.com/jianzhangcs/panini.

Introduction

Face restoration (FR) is typically an ill-posed image inverse problem, especially under high-degradation or multi-degradation (*e.g.*, downsampling, noise, blur, and compression) cases. Traditional deep network-based methods often utilize a single model for end-to-end training (Wang et al. 2018; Yang et al. 2020), which can well grasp the overall structure but lack the richness of details.

Due to the excellent performance of GAN models in recent years (Karras, Laine, and Aila 2019; Karras et al. 2020, 2018), some methods (Menon et al. 2020; Richardson et al. 2021) begin to use pre-trained GAN models as GAN Prior for FR tasks. These methods take advantage of the rich details implicitly encapsulated in GAN Prior by encoding the degraded face image into the latent space of pre-trained GAN. Although high visual qualities are achieved, due to the low dimension of latent space and its poor spatial expression capability, these methods are often unable to fully capture the facial structures of the degraded face image, which usually manifests as identity inconsistency.

To further capture the facial structural information of the degraded face image while preserving the realness contributed by GAN Prior, some methods (Chan et al. 2021; Yang et al. 2021; Wang et al. 2021b) not only encode the degraded face image into the latent space but also fuse external features (*e.g.*, features extracted from the degraded face image) with the GAN Prior features. These methods achieve significant improvements in identity consistency than previous GAN Prior based FR methods. However, they don't provide an explicit design for degradation-aware feature fusion and consequently result in inadequate robustness in visual quality when facing different degradation levels.

Inspired by recent progress in contrastive learning (He et al. 2020; Chen et al. 2020c,a,b) and visual attention (Hu, Shen, and Sun 2018; Woo et al. 2018; Li et al. 2019; Wang et al. 2021a), an unsupervised degradation representation learning (UDRL) strategy has been proposed in this paper to pre-train a degradation representation encoder (DRE). DRE extracts the degradation representations (DR) of input degraded images, regarded as a global condition to guide the restoration process. In addition, we propose a novel degradation-aware feature interpolation (DAFI) module, which can dynamically fuse the features (*i.e.*, features of GAN Prior and features extracted from the degraded face image) according to DR.

Further, we propose a novel network to integrate these designs for FR tasks. As the idea of selecting and fusing different sources of features is interestingly similar to the way of making a panini, we dubbed our network as Panini-Net. For instance, the degraded face image usually contains abundant valid information when suffering from mild degradation. Then, our Panini-Net will increase the fusion proportion of degraded image features while reducing the fusion proportion of GAN Prior features. The degraded face image usually lacks valid information when the degradation is severe. Then, Panini-Net will decrease the fusion proportion of degraded image features while increasing the proportion of GAN Prior features. Besides, the proposed DAFI shows some interesting attributes. In our experiments and ablation studies, DAFI performs better in details compared with nor-

This work was supported in part by Shenzhen Fundamental Research Program (No.GXWD20201231165807007-20200807164903001). *Corresponding author: Jian Zhang.*

mal convolutional feature fusion methods, even if DAFI possesses much fewer parameters. DAFI also shows flexible editability in FR tasks, and this attribute enables Panini-Net to generate multiple high-quality restored images.

The main contributions of this paper are summarized as follows:

- We propose a novel GAN Prior based degradation-aware feature interpolation network for FR, dubbed Panini-Net. It provides a robust solution to balance realness and fidelity considering various degradation levels.
- We propose an unsupervised degradation representation learning strategy to extract discriminative degradation representations for degraded images, which explicitly serve as a global condition for dynamic fusion.
- We propose a novel degradation-aware feature interpolation (DAFI) module, which can dynamically fuse the two types of informative features from input images and GAN Prior based on different degradation representations. We also experimentally show the efficiency and editability of DAFI.
- Experiments on two typical FR tasks, *i.e.*, multi-degradation face restoration, 16× super-resolution, show that Panini-Net achieves state-of-the-art results.

Related Work

Face Restoration. The use of deep neural networks (DNNs) in FR has made great progress in recent years (Dong et al. 2015; Kim, Lee, and Lee 2016; Zhang et al. 2018b,c; Huang et al. 2017; Li et al. 2018; Lin, Zhou, and Chen 2018; Lin et al. 2020). General DNNs based FR methods use paired datasets for end-to-end training. However, they struggle to generate high-quality images. With the development of Generative Adversarial Network (GAN) (Goodfellow et al. 2014; Mirza and Osindero 2014), some image restoration methods begin to introduce adversarial training strategies to improve the visual quality of repaired images (Wang et al. 2018; Yu and Porikli 2017; Yang et al. 2020; Wan et al. 2020; Zhang and Ling 2021). Those methods can restore the overall structure of the image very well but are still inferior to the best generative models (Karras, Laine, and Aila 2019; Karras et al. 2020) in visual quality. The human face has regular structures and details, making it possible to use strong prior information for FR. Commonly used priors include parsing maps (Buhler, Romero, and Timofte 2020; Yu et al. 2018; Chen et al. 2018, 2021; Song et al. 2019), landmarks (Hu et al. 2020; Chen et al. 2018; Song et al. 2019), reference images (Li et al. 2018, 2020), and GAN Prior (Chan et al. 2021; Yang et al. 2021; Wang et al. 2021b). Recent GAN Prior based FR methods achieve unprecedented results.

GAN Prior. In recent years, GANs represented by StyleGAN (Karras, Laine, and Aila 2019), which can generate high-resolution images, have inspired a lot of GAN Prior based image editing works (Menon et al. 2020; Richardson et al. 2021; Alaluf, Patashnik, and Cohen-Or 2021; Abdal et al. 2021; Luo et al. 2020; Abdal, Qin, and Wonka 2019, 2020; Nitzan et al. 2020; Tewari et al. 2020; Zhu et al. 2020; Patashnik et al. 2021). These methods often implement image editing by GAN Inversion. Specifically, the input image is first embedded into the latent space of StyleGAN as a latent code. Then the latent code is controlled to achieve image editing, such as super-resolution, inpainting, attribute transformation, *etc*. In PULSE (Menon et al. 2020), the authors optimize the latent code of StyleGAN to solve image super-resolution tasks. In PSP (Richardson et al. 2021), the authors train an encoder to directly predict the latent code instead of iterative optimizing it, thus accelerating the inference time. However, the common flaw for latent code editing is that it can not capture the spatial structures well. Therefore, the identities are usually inconsistent between the input image and the edited one. Even so, the advantages of latent code editing are evident, and the resulting images are of high quality, with realistic details.

Considering the image realness of StyleGAN and the image structural fidelity of the general DNN based methods, recent methods (Chan et al. 2021; Yang et al. 2021; Wang et al. 2021b) try to combine those two merits for FR by fusing external structural features into the intermediate features of pre-trained StyleGAN. GPEN (Yang et al. 2021) utilizes a pre-trained StyleGAN2 generator as the Decoder. Given a degraded image as input, the Encoder predicts a latent code of StyleGAN2 and extracts intermediate features as the input noise of StyleGAN2. GLEAN (Chan et al. 2021) utilizes StyleGAN2 generator as the latent bank and uses an encoder-latent bank-decoder framework for super-resolution. The features connected between each part are fused in a concatenation-convolution way. GFP-GAN (Wang et al. 2021b) uses a degradation removal module to predict latent codes and extracts the intermediate features to modulate the StyleGAN features. The degradation removal module is supervised during training to improve the quality of extracted features. In addition, the eyes and mouth are separately trained by GAN to improve quality. These methods achieve significant improvements in identity consistency than previous GAN Prior based FR methods, but they don't have an explicit design for dynamic feature fusion considering different degradation levels, thus showing insufficient robustness in image visual quality when facing severe degradations.

In Image2StyleGAN++ (Abdal, Qin, and Wonka 2020), the authors find that editing the "activation tensors" of StyleGAN could achieve more precise spatial-wise editing. Luo *et al.* carry out style-mix on the lower layers of latent codes $\mathbf{w}^+$ to migrate the target identity (Luo et al. 2020). Barbershop (Zhu et al. 2021) replaces the output "activations" of a specific layer in StyleGAN2 to achieve better identity migration. These works reveal that the bottom layers of StyleGAN have greater influences on the coarse structures, while the higher layers mainly contribute to the details. These characteristics of StyleGAN inspire us to solve FR tasks by editing the bottom features of StyleGAN to achieve identity consistency while keeping the higher layers untouched to preserve the textures and facial details.

Contrastive Learning. Recently, contrastive learning (He et al. 2020; Chen et al. 2020a,b,c) has made great strides in

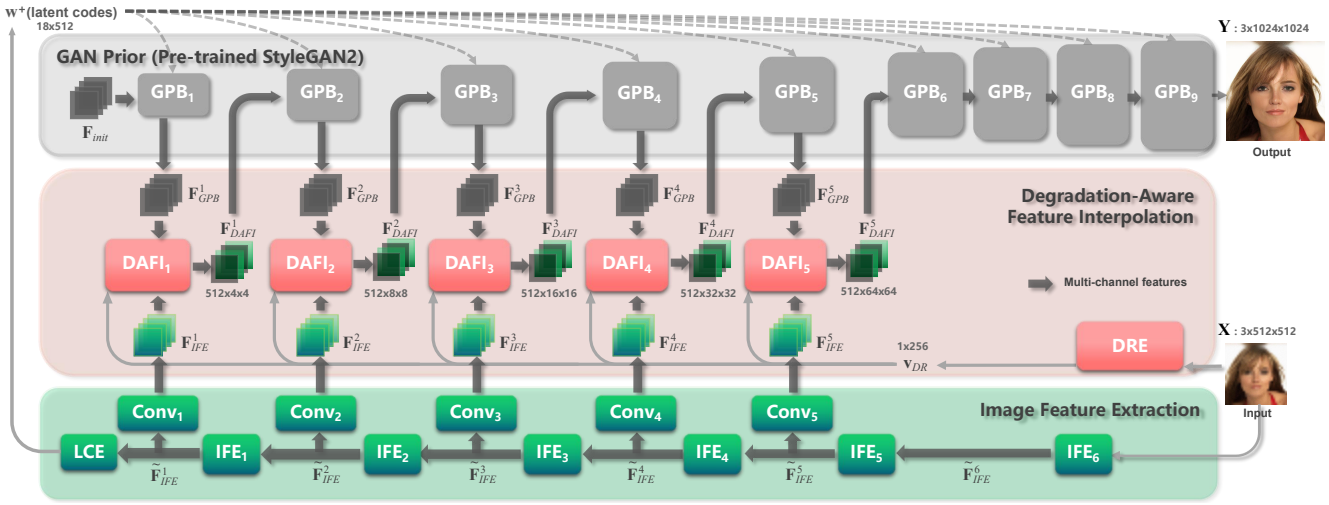

Figure 1: Overview of Panini-Net. It consists of an image feature extraction module, a degradation-aware feature interpolation (DAFI) module, and a pre-trained StyleGAN2 as GAN Prior module (GPM). Given a degraded face image $\mathbf{X}$ as the input, the image feature extraction module extracts features $\mathbf{F}^i_{IFE}, i \in \{1, ..., 5\}$, and predicts latent codes $\mathbf{w}^+ \in \mathbb{R}^{18\times512}$. The latent codes $\mathbf{w}^+$ can coarsely fetch a similar high-quality face from GPM. Then, 5 DAFI blocks (denoted as DAFI_i) are used to progressively interpolate $\mathbf{F}^i_{IFE}$ into $\mathbf{F}^i_{GPB}$ to incorporate the valid structural information of the degraded face image. A pre-trained degradation representations encoder (DRE) encodes the degradation representations as a vector $\mathbf{v}_{DR}$, which can be regarded as a global condition to guide DAFI blocks for restoration.

unsupervised representational learning. Some methods begin to use contrastive learning for image generation tasks (Park et al. 2020; Liu et al. 2021). In DASR (Wang et al. 2021a), the authors use the MoCo framework to pre-train a degradation encoder to learn degradation representations, then use degradation representations to guide the network for super-resolution tasks. In the pre-training of the degradation encoder, two patches are randomly selected from the same degraded image, and one patch is taken as a positive example for the other. These two patches share the same degradation functions, while the contents may also be similar. Therefore, the degradation encoder may learn not only the degradation representations but also the content representations. Moreover, for FR tasks, a portrait photo is often shot with a large aperture. The accompanying depth of field (DOF) may blur the background. If the patch of DASR is happened to locate in the background, the degradation encoder may be hard to differentiate whether the background blur is caused by DOF or by degradation. Therefore, we design an unsupervised degradation representation learning (UDRL) strategy to focus on learning the overall degradation representations of the degraded face image and encourage learning the degradation rather than the contents.

Visual Attention. Visual attention is usually explored to improve the performance of CNN, which not only tells where to focus but also improves the representation of interests. In (Hu, Shen, and Sun 2018), the authors propose the Squeeze-and-Excitation (SE) block, which adaptively adjusts channel-wise feature weights by explicitly modeling interdependencies between channels. In (Woo et al. 2018), the proposed Convolutional Block Attention Module (CBAM) explicitly incorporates both channel-wise and spatial-wise attention. In (Li et al. 2019), the Selective Kernel (SK) unit adaptively adjusts the receptive field size based on the region of interest. These methods provide robust baselines for universal visual tasks. However, they all extract attention from local features which need to be processed, while in FR models, these features may hardly contain valid degradation information. Besides, they can only reinforce features from a single source, while our urgent demand is to dynamically fuse two sources of features (*i.e.*, features from GAN Prior and features extracted from degraded face image). To overcome these deficiencies, our proposed DAFI module extracts the global condition directly from the degraded face image and enforces adaptive feature fusion for features from distinct sources.

Proposed Panini-Net

An overview of Panini-Net is depicted in Fig. 1. Specifically, Panini-Net consists of an image feature extraction module, a pre-trained face GAN model (*e.g.*, StyleGAN2) as the GAN Prior module, and a degradation-aware feature interpolation module. Given a degraded image $\mathbf{X}$, its degradation representation (DR) is encoded by the pre-trained degradation representation encoder (DRE) as $\mathbf{v}_{DR}$. The image feature extraction module extracts features $\mathbf{F}_{IFE}$ from the degraded face image and generates latent codes $\mathbf{w}^+$. $\mathbf{w}^+$ can coarsely fetch a similar high-quality face from GAN Prior module. Then the degradation-aware feature interpolation module progressively interpolates features $\mathbf{F}_{IFE}$ into GAN Prior features to rectify the facial structures of the fetched similar face, and finally, obtain a high-quality face image $\mathbf{Y}$ with realness and identity consistency.

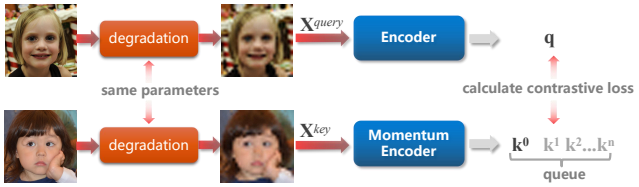

Figure 2: Overview of the unsupervised degradation representation learning strategy for degradation representation encoder (DRE). For each iteration, we randomly generate a set of new degradation parameters, and operate them on two different new HQ images to generate positive example pairs. Let the history images in queue be negative examples, to encourage learning the degradation rather than the contents.

Image Feature Extraction Module

As shown in Fig. 1, the image feature extraction module is designed to extract features $\mathbf{F}_{IFE}$ from the degraded face image and generate latent codes $\mathbf{w}^+$. Given an input image $\mathbf{X}$, we use image feature extractor (IFE) to extract preliminary features $\tilde{\mathbf{F}}^i_{IFE} \in \mathbb{R}^{C_i \times H_i \times W_i}$ progressively:

$$\tilde{\mathbf{F}}^i_{IFE} = \begin{cases} \mathcal{H}^i_{IFE}(\mathbf{X}), & i = 6; \\ \mathcal{H}^i_{IFE}(\tilde{\mathbf{F}}^{i+1}_{IFE}), & 1 \le i < 6, \end{cases} \tag{1}$$

where $\mathcal{H}^6_{IFE}(\cdot)$ is a dense block, while $\mathcal{H}^i_{IFE}(\cdot)$, $i \in \{1, ..., 5\}$, is just a convolution layer.

In order to avoid coupling of adjacent features, we add additional convolution branch to further extract features $\mathbf{F}^i_{IFE} \in \mathbb{R}^{C_i \times H_i \times W_i}$ as the final features for fusion:

$$\mathbf{F}^i_{IFE} = \mathcal{H}^i_{Conv}(\tilde{\mathbf{F}}^i_{IFE}), \quad i \in \{1, ..., 5\}, \tag{2}$$

where $\mathcal{H}^i_{Conv}(\cdot)$ is just a convolution layer, which has been experimentally verified useful.

Finally, a latent code encoder (LCE), which is composed of a convolution layer and a fully-connected layer, is used to predict the latent codes $\mathbf{w}^+ \in \mathbb{R}^{18 \times 512}$, expressed as

$$\mathbf{w}^+ = \mathcal{H}_{LCE}(\tilde{\mathbf{F}}^1_{IFE}). \tag{3}$$

Note that in our Panini-Net, as illustrated in Fig. 1, $\tilde{\mathbf{F}}^i_{IFE}$, $\mathbf{F}^i_{IFE}$, $\mathbf{F}^i_{GPB}$, $\mathbf{F}^i_{DAFI}$ have the same size of $C_i \times H_i \times W_i$.

GAN Prior Module

A pre-trained StyleGAN2 generator (Karras et al. 2020) is utilized as the GAN Prior module (GPM) in our Panini-Net. To be concrete, as shown in Fig. 1, GPM starts with a learned constant features $\mathbf{F}_{init}$, then progressively generates the result by passing $\mathbf{F}_{init}$ through a series of GAN Prior blocks (GPB_i, $i \in \{1, ..., 5\}$). Each GPB_i includes an up-sample operation (except for GPB_1) and outputs the feature $\mathbf{F}^i_{GPB}$. The latent codes $\mathbf{w}^+$ can coarsely fetch a similar high-quality face from GPM. The features $\mathbf{F}^i_{GPB}$, $i \in \{1, ..., 5\}$, are selected for dynamic fusion with our proposed degradation-aware feature interpolation module to rectify the facial structures. To preserve the delicate facial details, we leave the rest part of GPM untouched. It is worth noting that we omit details in StyleGAN2 (or GPM, GPB) for simplicity. For more details about StyleGAN2 please refer to (Karras, Laine, and Aila 2019; Karras et al. 2020).

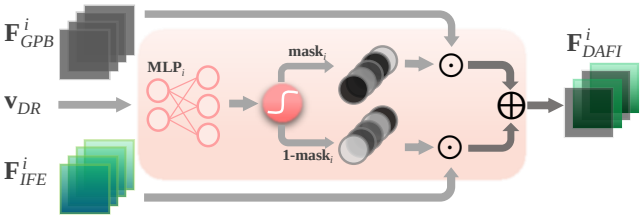

Figure 3: Overview of the degradation-aware feature interpolation (DAFI) block. $\mathbf{v}_{DR}$ serves a condition to generate an adaptive degradation-aware channel-wise $\mathbf{mask}$ through an MLP followed by Softmax. Then, the $\mathbf{mask}$ is used for dynamic interpolation.

Degradation-Aware Feature Interpolation Module

The degradation-aware feature interpolation (DAFI) module is comprised of a degradation representation encoder (DRE) and several DAFI blocks (denoted as DAFI_i, $i \in \{1, ..., 5\}$).

The DRE is designed to extract the degradation representations (DR) of the degraded face images. We adopt an unsupervised degradation representation learning (UDRL) strategy to pre-train DRE. Specifically, as illustrated in Fig. 2, two degraded images $\mathbf{x}^{query}$ and $\mathbf{x}^{key}$ are obtained by applying the same degradation function on two different images. Then we take $\mathbf{x}^{key}$ as the positive example of $\mathbf{x}^{query}$ and use MoCo (Chen et al. 2020c) framework to conduct contrastive training. We take $\mathbf{x}^{query}$ as the input for Encoder to generate vector $\mathbf{q}$, and take $\mathbf{x}^{key}$ as the input for the Momentum Encoder to generate vector $\mathbf{k}^0$. The history generated by Momentum Encoder can form a queue $\mathbf{k}^0, \mathbf{k}^1, ..., \mathbf{k}^n$. For each $\mathbf{q}$, as $\mathbf{k}^0$ is generated from $\mathbf{x}^{key}$ which shares the same degradation mode with $\mathbf{x}^{query}$, vector $\mathbf{k}^0$ should be similar to $\mathbf{q}$, while the rest vectors in the queue should be different from $\mathbf{q}$. An InfoNCE loss is used as training objective to encourage the encoding of $\mathbf{q}$ to approach $\mathbf{k}^0$ and stay away from $\mathbf{k}^1, \mathbf{k}^2, ..., \mathbf{k}^n$, which is formulated as:

$$\mathcal{L}_{DR} = \sum_{i=1}^{B} -\log \frac{\exp(\mathbf{q}_i \cdot \mathbf{k}^0_i / \tau)}{\sum_{j=1}^{n} \exp(\mathbf{q}_i \cdot \mathbf{k}^j_i / \tau)}. \tag{4}$$

We first pre-train the DRE under MoCo framework with our UDRL strategy, then fine-tune Panini-Net with fixed DRE. Given a degraded image $\mathbf{X}$, its degradation representation (DR) is efficiently encoded as a vector $\mathbf{v}_{DR} \in \mathbb{R}^{1 \times 256}$:

$$\mathbf{v}_{DR} = \mathcal{H}_{DRE}(\mathbf{X}). \tag{5}$$

For each feature $\mathbf{F}^i_{GPB}$ generated by GPB_i, $i \in \{1, ..., 5\}$, a dedicated DAFI block is applied to incorporate the valid facial structural information of its counterpart $\mathbf{F}^i_{IFE}$.

As Fig. 3 shows, given $\mathbf{v}_{DR}$ as the global condition, each DAFI block first applies a dedicated MLP followed by Softmax operation to generate an adaptive degradation-aware channel-wise $\mathbf{mask}_i \in \mathbb{R}^{1 \times C_i}$, $i \in \{1, ..., 5\}$, that is

$$(\mathbf{mask}_i, \mathbf{1} - \mathbf{mask}_i) = \text{Softmax}(\mathcal{H}^i_{MLP}(\mathbf{v}_{DR})). \tag{6}$$

Obviously, the $\mathbf{mask}_i$ is a vector whose dimension is equal to the channel number of features to be interpolated, and each $\mathbf{mask}$ element represents an interpolation weight

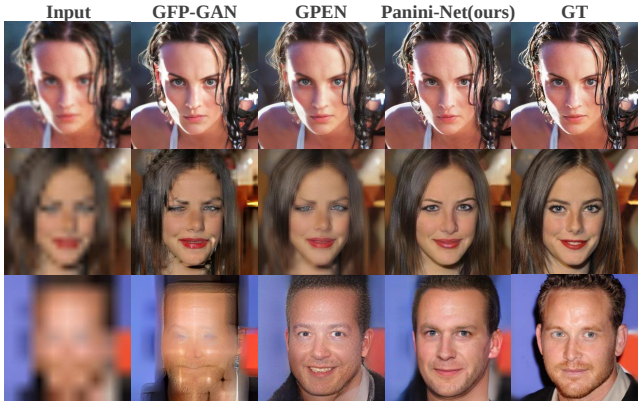

Figure 4: Experiment with multi-degradation face restoration. Given input images with different degradation levels, GFP-GAN performs poorly when degradation is severe, and GPEN struggle to generate realistic details, while our Panini-Net can achieve both highest realness and identity consistency. **Zoom in for best view**.

of a certain channel. Then the degradation-aware interpolation between image feature and GAN prior is formulated as:

$$\mathbf{F}^i_{DAFI} = \mathbf{F}^i_{GPB} \odot \mathbf{mask}_i + \mathbf{F}^i_{IFE} \odot (1 - \mathbf{mask}_i), \quad (7)$$

where $\odot$ denotes channel-wise multiplication.

Next, we replace $\mathbf{F}^i_{GPB}$ with $\mathbf{F}^i_{DAFI}$ as the input for the $(i+1)$-th GAN prior block GPB_{i+1}, $i \in \{1, ..., 5\}$, *i.e.*,

$$\mathbf{F}^i_{GPB} = \begin{cases} \mathcal{H}^i_{GPB}(\mathbf{F}_{init}), & i = 1 \\ \mathcal{H}^i_{GPB}(\mathbf{F}^{i-1}_{DAFI}), & 1 < i < 7 \\ \mathcal{H}^i_{GPB}(\mathbf{F}^{i-1}_{GPB}). & 7 \leq i < 9 \end{cases} \quad (8)$$

The final result $\mathbf{Y}$ is generated by the last block GPB_9 as

$$\mathbf{Y} = \mathcal{H}^9_{GPB}(\mathbf{F}^8_{GPB}). \quad (9)$$

Training and Loss

We first pre-train the DRE by UDRL strategy, then load the pre-trained StyleGAN2 (Karras et al. 2020) generator as our GPM. Finally, we fine-tune the whole Panini-Net. To provide stable degradation representations, we fix the parameters of DRE during fine-tuning. The rest parameters of Panini-Net are learnable. We use standard L1 loss, VGG perceptual loss (Johnson, Alahi, and Fei-Fei 2016), and vanilla adversarial loss (Goodfellow et al. 2014) as objectives for fine-tuning. The pre-trained discriminator of StyleGAN2 is also used for adversarial training. Details about the training and loss functions can be found in the appendix.

Experimental Results

To verify the utility of our Panini-Net on different FR tasks, we conduct extensive experiments on two typical FR tasks: multi-degradation face restoration and face super-resolution.

Face Restoration with Multiple Degradations

We train our Panini-Net on the FFHQ dataset (Karras, Laine, and Aila 2019), which contains 70000 high-quality (HQ) face images. We follow the practice in (Yang et al. 2021; Wang et al. 2021b) that applies a widely used degradation function to synthesize degraded low-quality (LQ) images:

$$\mathbf{X} = [(\mathbf{Y} \circledast \mathbf{k}_\sigma)\downarrow_r + \mathbf{N}_\delta]_{\mathbf{JPEG}_q}. \quad (10)$$

Specifically, HQ image $\mathbf{Y}$ is first convolved by Gaussian blur kernel $\mathbf{k}$ followed by a downsampling operation (with a symmetric upsampling to keep the size invariant), then an additive noise is applied on it. Finally, the image is compressed by JPEG operation with quality factor q. For each HQ image $\mathbf{Y}$, the blur kernel size σ is randomly selected. r, δ, q are randomly chosen from [10:200], [0:25], [5:50] respectively. By this means, we generate HQ-LQ image pairs for training. LQ and HQ images are of size 512×512 and 1024×1024, respectively. We adopt Adam optimizer and Cosine Annealing Scheme to fine-tune Panini-Net, with batch size being 8 and iterations number being 600K.

We randomly choose 1000 HQ face images in CelebA-HQ (Karras et al. 2018) dataset as ground truth (GT), then use the above degradation function Eq. (10) to process these images as input. We compare Panini-Net with recent state-of-the-art GAN Prior based FR methods, including official pre-trained GFP-GAN (Wang et al. 2021b) and GPEN (Yang et al. 2021) models. It is worth noting that these competing methods all use StyleGAN2 as GAN Prior and share the same degradation formulation (may be different in parameters) for synthesizing LQ images. We use PSNR, FID (Heusel et al. 2017), LPIPS (Zhang et al. 2018a) as the metrics for quantitative comparison. Subjective and visual results are shown in Table 1 and Fig. 4, which clearly shows the evident advantage of Panini-Net in the balance between image realness and identity consistency. The images restored by Panini-Net are comparable with the ground truth in visual quality even when the degradation is severe.

Table 1: Quantitative comparison on multi-degradation face restoration. Our Panini-Net achieves the highest PSNR and competitive LPIPS. A significant gain in FID is also obtained, which statistically reflects the improvement of image realness contributed by our method.

Method	PSNR↑	FID↓	LPIPS↓
GFP-GAN (CVPR'21)	17.22	34.61	**0.4150**
GPEN (CVPR'21)	17.91	32.03	0.4355
Panini-Net (ours)	**18.01**	**24.66**	0.4470

Face Super-Resolution

Downsampling with a fixed ratio can be considered as a constant degradation. Thus we don't need a DRE to extract the degradation representations. Specifically, we simplified Panini-Net for the $16\times$ SR task: (1) We remove DRE and represent $\mathbf{v}_{DR}$ with a learnable constant vector. (2) Some relevant convolutions are adjusted to fit the new size of the input image. In this experiment, we use the FFHQ dataset as the GT and use $16\times$ bilinear interpolation as a downsampling operation to generate low-resolution (LR) images, thus forming GT-LR pairs for training. The size of LR images is

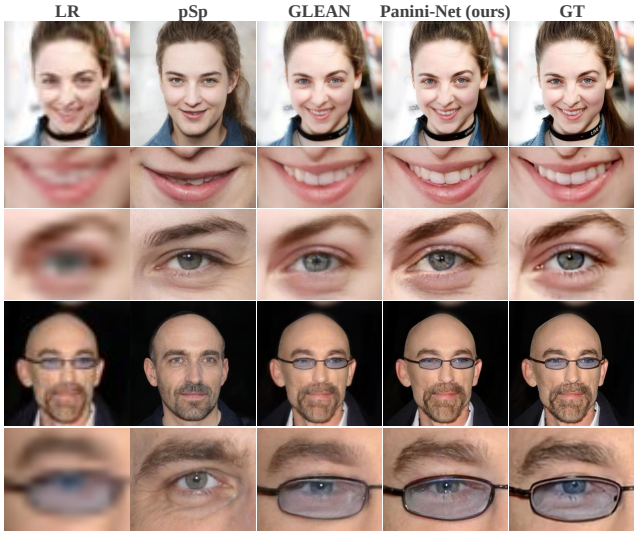

Figure 5: Experiment with 16× SR. Despite pSp can achieve high visual quality, it performs poorly in maintaining the identity of the input face image. GLEAN can achieve good results, while Panini-Net provides comparable even superior performance, especially in details (*e.g.*, eyes and tooth). Since Panini-Net has much fewer parameters than GLEAN, and they share similar training frameworks, we attribute its superiority to the proposed DAFI. **Zoom in for best view**.

Table 2: Quantitative comparison on 16× face SR. Our Panini-Net outperforms other methods in FID and LPIPS, and achieves competitive PSNR. As is shown in Fig. 5, the quality of our results outperforms the GT sometimes, but metrics like PSNR can not well reflect that, as they only calculates the consistency with the GT, without any consideration for other reasonable multiple solutions.

Method	PSNR↑	FID↓	LPIPS↓
pSp (CVPR'21)	12.90	47.65	0.6529
GLEAN (CVPR'21)	**21.66**	19.76	0.4013
Panini-Net (ours)	21.19	**16.77**	**0.3886**

64×64, and 1024×1024 for GT, corresponding to the input and output size of the modified Panini-Net. We use the same way to generate a test dataset from CelebA-HQ and compare Panini-Net with recent SOTA GAN Prior based SR methods on our test dataset, including pSp (Richardson et al. 2021), and GLEAN (Chan et al. 2021), which also use StyleGAN2 as GAN Prior. Results are shown in Fig. 5 and Table 2. Although with fewer parameters, our Panini-Net is still competitive even outperforms these methods considering its superiority in the balance between realness and fidelity.

Ablation Studies

A: Comparison with Different Feature Fusion

GAN Prior based FR methods (Chan et al. 2021; Wang et al. 2021b; Yang et al. 2021) usually achieve identity consistency by fusing external features into features of the GAN Prior model. Concatenation followed by convolution (denoted as Cat-Conv in this paper) is a typical feature fusion method that is also adopted in GLEAN (Chan et al. 2021). Given two informative features from different sources, it first concatenates them by channel, then uses a convolution layer to decrease the channel number. Here we simply replace DAFI in Panini-Net with Cat-Conv. To exclude the effect of DRE, We train Panini-Net with Cat-Conv on the 16× SR task (without DRE) with the same training settings as Panini-Net. Results are shown in Fig. 6 and Table 3.

Figure 6: Ablation experiment **A**. Panini-Net with Cat-Conv is a modified version of Panini-Net, which removes DRE and replaces DAFI with concatenation followed by convolution operations (similar feature fusion operation in GLEAN). Cat-Conv costs more parameters and undermines the visual quality, especially in details (*e.g.*, tooth and eyes). We argue that's because the interpolation operation in DAFI can better preserve the details encapsulated in GAN Prior features, and the global condition guidance can help DAFI better handle feature fusion. **Zoom in for best view**.

Table 3: Quantitative comparison on different feature fusion. Concatenation will double the feature channels. Considering features with 512 channels, Cat-Conv will be costly in parameters. Instead, DAFI uses interpolation for feature fusion with fewer parameters and performs better in visual quality and quantitative metrics.

Method	PSNR↑	FID↓
Panini-Net with Cat-Conv	21.17	18.41
Panini-Net with DAFI (ours)	**21.19**	**16.77**

Although Cat-Conv consumes more parameters than DAFI, it causes performance decline when applied on Panini-Net, especially in detailed visual quality (*e.g.*, tooth and eyes). We argue that's because the interpolation operation in DAFI can better preserve the details encapsulated in GAN Prior features, and the global condition guidance can help DAFI better handle feature fusion.

B: Dissection of Panini-Net

To study the correlations between degradation levels and interpolation ratios, we fix σ, δ, q in Eq. (10), while choosing downsampling rate r as 16, 32, 64, 128, respectively. Then we get four sets of degradation functions, the only difference between these degradation functions is the

Table 4: GAN Prior feature usage ratio in Panini-Net with different degradation levels. Panini-Net can dynamically increase the usage of GAN Prior when the degradation becomes severe.

Downsampling Rate r	16	32	64	128
$\mathbf{F}^5_{GPB}$ usage ratio θ	0.43	0.54	0.65	0.73

downsampling rate r. We use these four functions to process a single HQ image to get four degraded face images $\{\mathbf{X}_{\downarrow 16}, \mathbf{X}_{\downarrow 32}, \mathbf{X}_{\downarrow 64}, \mathbf{X}_{\downarrow 128}\}$ that share the same content but are different in degradation levels. We feed these four images into Panini-Net respectively. As Panini-Net uses progressive interpolation layer by layer, the earlier DAFI$_i$, $i \in \{1, ..., 5\}$, has weaker influences, while DAFI$_5$ has determinate affection on restoration. Hence, for each degraded face image, we only record the interpolation mask of DAFI$_5$. For simplicity, we sum the entire elements of the $\mathbf{mask}_5 \in \mathbb{R}^{1\times 512}$ and divide it with the dimension to get a ratio θ in range (0,1), defined as

$$\theta = \sum_{j=1}^{512} \mathbf{mask}_5[j]/512. \tag{11}$$

Obviously, θ can coarsely reflect the usage ratio of features $\mathbf{F}^5_{GPB}$. The records are shown in Table 4. In DAFI$_5$, θ shows obvious positive correlations with degradation levels: when degradation becomes severe, θ will increase. That means Panini-Net tends to utilize more GAN Prior information when degradation becomes severe, which is in accord with our expectations.

Further, we expect to see what exactly does features $\mathbf{F}_{GPB}$ and $\mathbf{F}_{IFE}$ contribute to the final result when facing different degradation levels. We feed $\{\mathbf{X}_{\downarrow 16}, \mathbf{X}_{\downarrow 32}, \mathbf{X}_{\downarrow 64}, \mathbf{X}_{\downarrow 128}\}$ into Panini-Net respectively. At the first round, we get the standard results $\mathbf{Y}$, the operation of DAFI is the same as Eq. (7). At the second round, we manually set $\mathbf{F}_{GPB}$ as zero during inference to get results $\mathbf{Y}_{IFE}$, the operation of DAFI can be formulated as:

$$\mathbf{F}^i_{DAFI} = \mathbf{F}^i_{IFE} \odot (\mathbf{1} - \mathbf{mask}_i). \tag{12}$$

At the third round, we manually set $\mathbf{F}_{IFE}$ as zero during inference to get results $\mathbf{Y}_{GPB}$, the operation of DAFI can be formulated as:

$$\mathbf{F}^i_{DAFI} = \mathbf{F}^i_{GPB} \odot \mathbf{mask}_i. \tag{13}$$

Results are shown in Fig. 7. When the degradation becomes severe, the valid contents of $\mathbf{Y}_{IFE}$ will decline, and $\mathbf{Y}_{GPB}$ will become more complete, which means Panini-Net can dynamically increase the use of GAN Prior to offset the decline of valid contents in the degraded face image. Besides, when the degradation is mild, the content of $\mathbf{Y}_{IFE}$ is mainly about the coarse structures of the input image, while the content of $\mathbf{Y}_{GPB}$ is mainly about the refinement of the eyes, nose, mouth, and textures. It indicates that Panini-Net learned how to extract valid information from GAN Prior and degraded face images. Another interesting discovery is that the masked features $\mathbf{F}^i_{GPB} \odot \mathbf{mask}_i$ and $\mathbf{F}^i_{IFE} \odot (\mathbf{1} - \mathbf{mask}_i)$ can generate reasonable results respectively, which implies the potential of DAFI for editing.

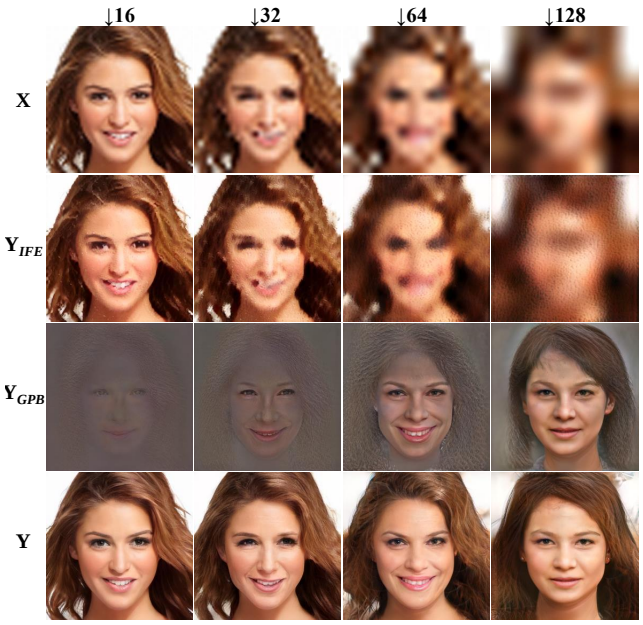

Figure 7: Ablation experiment **B**. Given four degraded face images $\mathbf{X}$ with downsampling rate 16, 32, 64, 128 respectively, $\mathbf{Y}$ are the standard results restored from $\mathbf{X}$ by Panini-Net. To solely use $\mathbf{F}_{GPB}$ for generation, we manually set $\mathbf{F}_{IFE}$ as $\mathbf{0}$ to generate the results $\mathbf{Y}_{GPB}$. To solely use $\mathbf{F}_{IFE}$ for generation, we manually set $\mathbf{F}_{GPB}$ as $\mathbf{0}$ to generate the results $\mathbf{Y}_{IFE}$. These results accord with our expectations for Panini-Net: (1) Dynamically fuse the two types of informative features (*i.e.*, $\mathbf{F}_{IFE}$ and $\mathbf{F}_{GPB}$) with flexible adaption to various degradations. (2) Discriminatively utilize the rich details in GAN Prior and the overall structures of the degraded face image $\mathbf{X}$. **Zoom in for best view**.

Conclusions

In this paper, we present a novel feature fusion framework for face restoration, dubbed Panini-Net, which can dynamically fuse the features depending on degradation levels. The proposed DAFI provides a concise and efficient way to fuse external features into GAN prior. Besides, there is still a lot of room to explore the interpolation forms (*e.g.* spatial-wise interpolation) and the way of mask generation. As we showed on the $16\times$ SR experiment, a simple modification on the way of mask generation can make Panini-Net competent to a new task, and we believe Panini-Net can also perform well in other supervised image-to-image tasks with limited modifications. On the other hand, the characteristics of interpolation operation enable us to explore the editability of the restored results. We can generate multiple high-quality solutions simply by adding a bias on mask elements (see appendix for more details). This merit brings more possibilities for practical applications and is worthy of study.

References

Abdal, R.; Qin, Y.; and Wonka, P. 2019. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4432–4441.

Abdal, R.; Qin, Y.; and Wonka, P. 2020. Image2StyleGAN++: How to Edit the Embedded Images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8296–8305.

Abdal, R.; Zhu, P.; Mitra, N. J.; and Wonka, P. 2021. StyleFlow: Attribute-conditioned Exploration of StyleGAN-Generated Images using Conditional Continuous Normalizing Flows. *ACM Transactions on Graphics*, 40(3): 1–21.

Alaluf, Y.; Patashnik, O.; and Cohen-Or, D. 2021. Only a Matter of Style: Age Transformation Using a Style-Based Regression Model. *ACM Transactions on Graphics*, 40(4).

Buhler, M. C.; Romero, A.; and Timofte, R. 2020. DeepSEE: Deep Disentangled Semantic Explorative Extreme Super-Resolution. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 624–642.

Chan, K. C.; Wang, X.; Xu, X.; Gu, J.; and Loy, C. C. 2021. GLEAN: Generative Latent Bank for Large-Factor Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14245–14254.

Chen, C.; Li, X.; Yang, L.; Lin, X.; Zhang, L.; and Wong, K.-Y. K. 2021. Progressive Semantic-Aware Style Transformation for Blind Face Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11896–11905.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning (ICML)*, 1597–1607.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. E. 2020b. Big Self-Supervised Models are Strong Semi-Supervised Learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 22243–22255.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020c. Improved Baselines with Momentum Contrastive Learning. arXiv:2003.04297v1.

Chen, Y.; Tai, Y.; Liu, X.; Shen, C.; and Yang, J. 2018. FSRNet: End-to-End Learning Face Super-Resolution With Facial Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2492–2501.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2015. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2): 295–307.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. *Advances in Neural Information Processing Systems (NeurIPS)*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9729–9738.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7132–7141.

Hu, X.; Ren, W.; LaMaster, J.; Cao, X.; Li, X.; Li, Z.; Menze, B.; and Liu, W. 2020. Face Super-Resolution Guided by 3D Facial Priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 763–780.

Huang, H.; He, R.; Sun, Z.; and Tan, T. 2017. Wavelet-SRNet: A Wavelet-Based CNN for Multi-Scale Face Super Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1689–1697.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 694–711.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations (ICLR)*.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8110–8119.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1646–1654.

Li, X.; Chen, C.; Zhou, S.; Lin, X.; Zuo, W.; and Zhang, L. 2020. Blind Face Restoration via Deep Multi-scale Component Dictionaries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 399–415.

Li, X.; Liu, M.; Ye, Y.; Zuo, W.; Lin, L.; and Yang, R. 2018. Learning Warped Guidance for Blind Face Restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 272–289.

Li, X.; Wang, W.; Hu, X.; and Yang, J. 2019. Selective Kernel Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 510–519.

Lin, J.; Zhou, T.; and Chen, Z. 2018. Multi-Scale Face Restoration With Sequential Gating Ensemble Network.

Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 32(1).

Lin, S.; Zhang, J.; Pan, J.; Liu, Y.; Wang, Y.; Chen, J.; and Ren, J. 2020. Learning to Deblur Face Images via Sketch Synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 34(07): 11523–11530.

Liu, R.; Ge, Y.; Choi, C. L.; Wang, X.; and Li, H. 2021. DivCo: Diverse Conditional Image Synthesis via Contrastive Generative Adversarial Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16377–16386.

Luo, X.; Zhang, X.; Yoo, P.; Martin-Brualla, R.; Lawrence, J.; and M. Seitz, S. 2020. Time-Travel Rephotography. arXiv:2012.12261v1.

Menon, S.; Damian, A.; Hu, S.; Ravi, N.; and Rudin, C. 2020. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2437–2445.

Mirza, M.; and Osindero, S. 2014. Conditional Generative Adversarial Nets. arXiv:1411.1784.

Nitzan, Y.; Bermano, A.; Li, Y.; and Cohen-Or, D. 2020. Face identity disentanglement via latent space mapping. *ACM Transactions on Graphics*, 39: 1–14.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive Learning for Unpaired Image-to-Image Translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 319–345.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2085–2094.

Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2021. Encoding in Style: A StyleGAN Encoder for Image-to-Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2287–2296.

Song, L.; Cao, J.; Song, L.; Hu, Y.; and He, R. 2019. Geometry-Aware Face Completion and Editing. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 33(01): 2506–2513.

Tewari, A.; Elgharib, M.; Bharaj, G.; Bernard, F.; Seidel, H.-P.; Pérez, P.; Zollhofer, M.; and Theobalt, C. 2020. StyleRig: Rigging StyleGAN for 3D Control Over Portrait Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6142–6151.

Wan, Z.; Zhang, B.; Chen, D.; Zhang, P.; Chen, D.; Liao, J.; and Wen, F. 2020. Bringing Old Photos Back to Life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2747–2757.

Wang, L.; Wang, Y.; Dong, X.; Xu, Q.; Yang, J.; An, W.; and Guo, Y. 2021a. Unsupervised Degradation Representation Learning for Blind Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10581–10590.

Wang, X.; Li, Y.; Zhang, H.; and Shan, Y. 2021b. Towards Real-World Blind Face Restoration with Generative Facial Prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9168–9178.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Loy, C. C. 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 63–79.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19.

Yang, L.; Wang, S.; Ma, S.; Gao, W.; Liu, C.; Wang, P.; and Ren, P. 2020. HiFaceGAN: Face Renovation via Collaborative Suppression and Replenishment. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, 1551–1560.

Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2021. GAN Prior Embedded Network for Blind Face Restoration in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 672–681.

Yu, X.; Fernando, B.; Ghanem, B.; Porikli, F.; and Hartley, R. 2018. Face Super-resolution Guided by Facial Component Heatmaps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 217–233.

Yu, X.; and Porikli, F. 2017. Hallucinating Very Low-Resolution Unaligned and Noisy Face Images by Transformative Discriminative Autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3760–3768.

Zhang, M.; and Ling, Q. 2021. Supervised Pixel-Wise GAN for Face Super-Resolution. *IEEE Transactions on Multimedia*, 23: 1938–1950.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 586–595.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018b. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 286–301.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2018c. Residual Dense Network for Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2472–2481.

Zhu, P.; Abdal, R.; Femiani, J.; and Wonka, P. 2021. Barbershop: GAN-based Image Compositing using Segmentation Masks. arXiv:2106.01505v1.

Zhu, P.; Abdal, R.; Qin, Y.; and Wonka, P. 2020. SEAN: Image Synthesis With Semantic Region-Adaptive Normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5104–5113.