Blind Face Restoration under Extreme Conditions: Leveraging 3D-2D Prior Fusion for Superior Structural and Texture Recovery

Zhengrui Chen[1], Liying Lu[3], Ziyang Yuan[2], Yiming Zhu[2, 3]
Yu Li[3*†], Chun Yuan[2*], Weihong Deng[1*]
[1]Beijing University of Posts and Telecommunications
[2]Tsinghua University
[3]International Digital Economy Academy (IDEA)

Abstract

Blind face restoration under extreme conditions entails reconstructing high-quality facial images from severely degraded inputs, often characterized by poor quality and extreme facial poses. These challenges lead to errors in facial structure and unnatural artifacts in restored images. This paper demonstrates the efficacy of leveraging 3D priors to address structural deficiencies in 2D priors while preserving texture details. We introduce FREx (Face Restoration under Extreme conditions), which integrates structure-accurate 3D priors and texture-rich 2D priors into pretrained generative networks. To fuse information from 3D and 2D priors, we propose an adaptive weight module that adjusts feature importance based on input image conditions. This approach enables our model to restore accurate facial structures and natural appearances, even when images suffer significant information loss due to degradation and extreme poses. Extensive experiments on synthetic and real-world datasets validate the effectiveness of our approach.

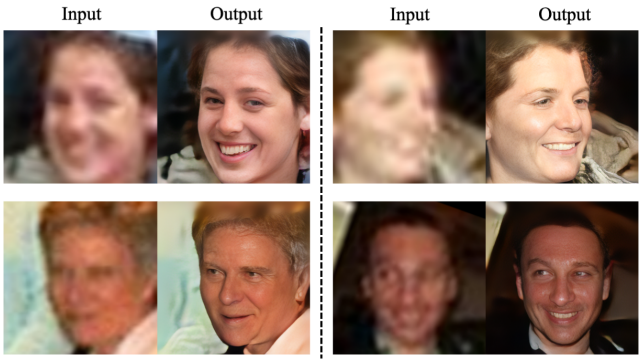

Figure 1: Restoration results of some challenging real-world cases produced by our method. Our method could generate rational facial structures and fine details even when the majority of information is lost in the input images, and our method could handle various head poses.

Introduction

Blind face restoration has demonstrated remarkable potential in restoring high-quality facial images from unknown degradations, including low resolution (Dong et al. 2014; Lim et al. 2017), blur (Kupyn et al. 2018; Shen et al. 2018), noise (Zhang et al. 2017), and compression artifacts (Dong et al. 2015; Li et al. 2014). With the development of deep learning techniques, existing methods (Wang et al. 2021; Zhou et al. 2022) are becoming increasingly effective in enhancing the visual quality of facial images. However, these methods often struggle to restore rational facial structures and details when dealing with images with extremely severe degradations.

In this paper, we present the FREx method, which is designed to handle extreme scenarios such as very severe noise and blur. It is also capable of restoring facial images with large head pose, which is a challenging task in the field of blind face restoration.

Since severely degraded images contain limited useful information, it is crucial to employ prior knowledge to aid in the restoration process. Prior-based methods (Wang et al. 2021; Chan et al. 2021; Yang et al. 2021; Zhou et al. 2022; Gu et al. 2022) have been developed to address this challenge, with previous approaches primarily utilizing 2D priors, such as reference priors (Zhou et al. 2022; Gu et al. 2022) and generative priors (Wang et al. 2021; Chan et al. 2021; Yang et al. 2021).

Reference prior-based methods (Zhou et al. 2022; Gu et al. 2022) typically employ vector quantized codebooks (Van Den Oord, Vinyals et al. 2017) learned by self-reconstruction to predict high-quality facial codes from low-quality inputs. On the other hand, recent generative prior-based methods (Wang et al. 2021; Chan et al. 2021; Yang et al. 2021) encode low-quality images into the latent space and subsequently decode the latent code into high-quality images using a pre-trained generator (Karras et al. 2020). While these methods are effective in generating complex facial details, they often produce erroneous facial structures when dealing with images under extreme facial poses, as they lack 3D information.

To tackle the challenge of restoring severely degraded facial images, we propose to use both 3D and 2D priors simultaneously to achieve superior results. The rationale behind this is twofold: firstly, 3D priors are useful in providing

*Corresponding Authors.
†Project Lead.

robust facial structures, which is essential for further restoration of fine details. Secondly, by leveraging 2D priors on top of these reasonable facial structures, we can better restore rich textures.

To accomplish this, we employ a pretrained 3D module (Chan et al. 2022; Yuan et al. 2023) to provide implicit 3D geometric priors for severely degraded images. Additionally, we use the abundant 2D texture priors encoded in StyleGAN2 (Karras et al. 2020) to restore fine details. Furthermore, to achieve faithful and natural results, we introduce a Adaptive Weighting Block for 3D and 2D information fusion. This module can dynamically control the fusion of 3D and 2D information based on the input image's facial poses, allowing us to optimally utilize the 2D and 3D priors and restore a more authentic result.

Our contributions are summarized as follows:

- We propose a novel FREx framework for the extreme blind face restoration task. Our FREx incorporates 3D structural priors and 2D texture priors into the restoration process, leading to high-quality restoration results under extreme conditions. Some examples are show in Fig. 1.
- We introduce a novel module called the Adaptive Weighting Block, which enhances the aggregation of 3D and 2D information, resulting in high-quality restored images with improved fine details and reduced artifacts.
- Extensive experiments demonstrate that our method achieve superior performance compared to previous methods in blind face restoration under extreme condition task.

Related Work

Image Restoration

Image restoration is a fundamental task in computer vision that involves various tasks such as super-resolution (Dong et al. 2014; Lim et al. 2017; Timofte et al. 2017; Liu et al. 2018; Zhang et al. 2018b; Yu et al. 2021; Guo et al. 2020; Liu et al. 2020), denoising (Zhang et al. 2017; El Helou, Zhou, and Süsstrunk 2020; Lefkimmiatis 2017), deblurring (Shen et al. 2018; Kupyn et al. 2018; Xu et al. 2014), compression removal (Guo and Chao 2016; Dong et al. 2015), low-light recovery (Guo, Li, and Ling 2016; Lv, Li, and Lu 2021; Zhang et al. 2021), and de-weathering (Hao et al. 2019; Li, Tan, and Brown 2015; Li et al. 2016, 2017a,b; Zhang et al. 2023). Recently, deep neural networks have shown remarkable success in these tasks, but most existing methods focus on a specific type of degradation. Therefore, restoring images under extreme conditions remains a challenging and critical task in the field of image restoration.

Blind Face Restoration

The blind face restoration task aims to reconstruct high-quality face images from low-quality inputs that suffer from various unknown degradations. In recent years, there has been an increasing amount of literature to address this task. However, it is challenge to restore high-quality results due to the ill-posed nature of this task. Therefore, some existing works utilize different priors to aid the restoration process. In general, there are three main popular ways that employ different types of face-specific prior knowledge: geometric priors, reference priors and generative priors.

Geometric Priors, which include facial landmarks (Chen et al. 2018; Kim et al. 2019; Zhu et al. 2016; Yang et al. 2023), face parsing maps (Shen et al. 2018; Chen et al. 2021, 2018), facial component heatmaps (Yu et al. 2018) and face shapes (Zhu et al. 2022), can assist in localizing facial parts. However, there are two main limitations of these priors: 1) these priors are estimated directly from the low-quality inputs, which can lead to erroneous results and may perform poorly in real-world situations. 2) They primarily concentrate on the geometric constraints and may not provide sufficient texture details for effective restoration.

Reference Priors. Previous reference-based methods (Li et al. 2020b, 2018; Dogan, Gu, and Timofte 2019) usually rely on reference images of the same identity, which limits its potential applications in real cases. Recently, some works investigate face dictionaries (Li et al. 2020a) or vector quantized codebooks (Gu et al. 2022; Zhou et al. 2022; Wang et al. 2022) to predict high-quality facial codes from low-quality inputs. These methods first learn a discrete codebook (Van Den Oord, Vinyals et al. 2017) to store the high-quality visual parts of high-resolution face images, then project the degraded image into discrete codes and find their nearest neighbors in the learned codebook for face generation. However, since most high-quality codes are obtained from front-facing images, it becomes challenging to find corresponding parts in the codebook when the input facial pose is unusual. This can result in distorted high-frequency details or loss of sharpness in the restored image.

Generative Priors. Existing generative prior-based works (Gu, Shen, and Zhou 2020; Menon et al. 2020; Wang et al. 2021; Chan et al. 2021; Wu et al. 2021; Yang et al. 2021) generally encode low-quality images into the latent space and decode the latent codes into high-quality images using a pretrained generator (Karras et al. 2020). Some of them (Wang et al. 2021; Chan et al. 2021; Yang et al. 2021) also employ skip connections to enhance the restoration details. Despite their ability in preserving facial identity, they are highly sensitive to disturbances or perturbations. As the degradation level increases, these techniques are more likely to extract incorrect facial structures, leading to flawed structures and distorted facial features in the outputs. In our method, we propose to combine 2D priors with 3D priors, enabling us to generate results with improved fine structures and details.

Methodology

Overview

The objective of extreme blind face restoration is to restore a high-quality image $\boldsymbol{I}_{\text{out}}$ from a severely degraded input face image $\boldsymbol{I}_{\text{Input}}$, while maintaining authenticity with respect to the ground-truth image $\boldsymbol{I}_{\text{HQ}}$.

To achieve this, we propose a novel framework called FREx, which is depicted in Fig. 2(a). Given the input $\boldsymbol{I}_{\text{Input}}$, a 3D encoder is used to extract its structural information and a pretrained 3D decoder will return the 3D feature $\boldsymbol{F}_{\text{3D}}$ that

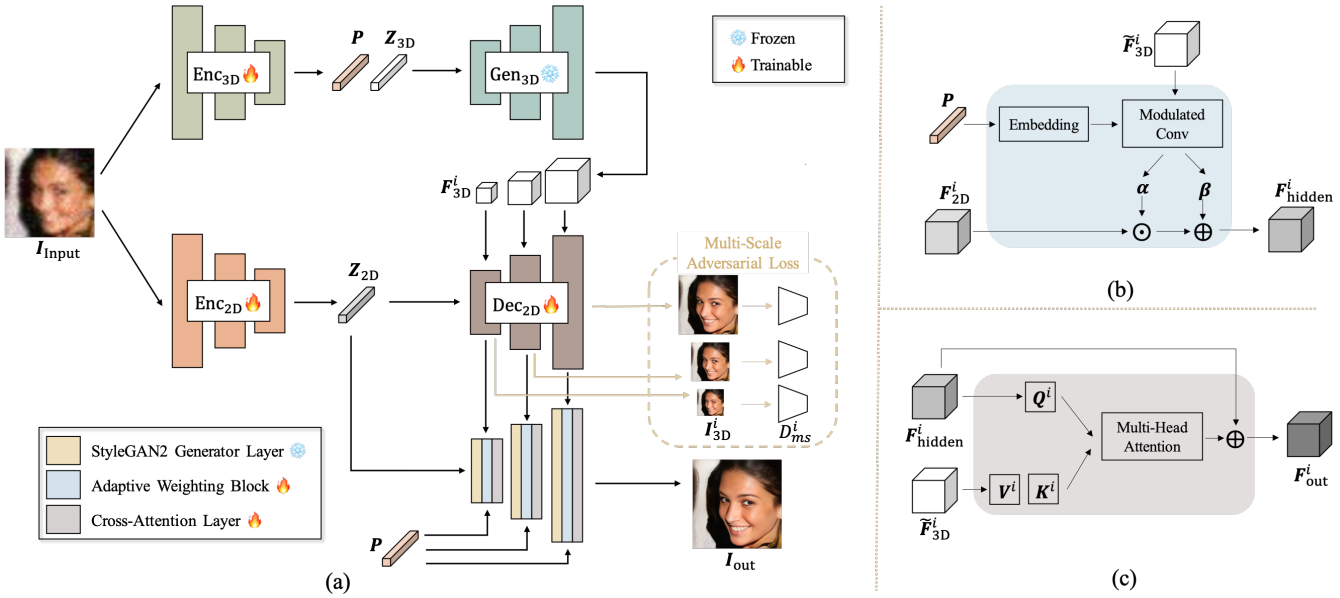

Figure 2: (a) The overview of our framework. (b) The detailed architecture of the proposed Adaptive Weighting Block (AWB). (c) The pipeline of the cross-attention.

contains facial geometry information for further restoration. Afterwards, a 2D encoder is used to extract the remaining information $\boldsymbol{Z}_{2D}$ in the input image, which is used to aid the extracted 3D features to produce more authentic 3D features. At last, we fuse the 3D feature with rich 2D texture priors from the StyleGAN2 (Karras et al. 2020) and produce the final result.

3D Prior Learning

3D Geometry Prior Extraction. Since a pretrained 3D face GAN is capable of providing features that exhibit sufficient structural information, we leverage an existing state-of-the-art network (Chan et al. 2022) as the decoder of our Geometry Prior Extraction Module, whose detailed architecture can be found in the supplementary materials.

As shown in Fig. 2, the input low-quality image is first encoded into a latent code $\boldsymbol{Z}_{3D}$ by the 3D encoder, which is then passed through the pretrained decoder, producing the 3D features $\boldsymbol{F}^i_{3D} \in \mathbb{R}^{\frac{128}{i}\times\frac{128}{i}\times 32}$. Specifically, in order to extract more structural information from the degraded image, we draw inspiration from (Richardson et al. 2021) and enhance the expressiveness of $\boldsymbol{Z}_{3D}$ by mapping the input image to the $\mathcal{W}+$ space as follows:

$$\boldsymbol{Z}_{3D},\ \overline{\boldsymbol{Z}},\ \boldsymbol{P} = \mathrm{Enc}_{3D}\left(\boldsymbol{I}_{\text{input}}\right), \tag{1}$$

$$\{\boldsymbol{F}^i_{3D}\} = \mathrm{Gen}_{3D}\left(\boldsymbol{Z}_{3D} + \overline{\boldsymbol{Z}}\right),\ \ i = 0, 1, 2\ , \tag{2}$$

where $\boldsymbol{P}$ denotes the camera parameters estimated from the input image and will be used in the following. Enc_{3D} and Gen_{3D} are the 3D encoder and decoder, and $\overline{\boldsymbol{Z}}$ are the offset deviated from $\boldsymbol{Z}_{3D}$.

Faithful 3D Feature Generation. After the above stage, we obtain the 3D features that contain sufficient geometry information. However, there is a possibility of inaccuracy of such information since it is extracted by a pretrained 3D GAN whose network parameters are fixed during training. To guarantee the fidelity and accuracy of the 3D information, we further propose to aid the extracted 3D features with the remaining information from the input image and apply a **multi-scale adversarial loss** to regularize this process. As shown in Fig 2(a), we first map the input image to a latent code $\boldsymbol{Z}_{2D}$ by a 2D encoder as:

$$\boldsymbol{Z}_{2D} = \mathrm{Enc}_{2D}\left(\boldsymbol{I}_{\text{input}}\right)\ , \tag{3}$$

and then upsample it and fuse the upsampled feature with the corresponding 3D feature $\boldsymbol{F}^i_{3D}$ progressively by a 2D decoder, producing a set of enhanced features $\widetilde{\boldsymbol{F}}^i_{3D}$. Afterwards, each $\widetilde{\boldsymbol{F}}^i_{3D}$ is mapped to an RGB image $\boldsymbol{I}^i_{3D}$ by a convolution layer, and the result is fed to the corresponding discriminator to calculate the multi-scale adversarial loss as:

$$\boldsymbol{I}^i_{3D} = \mathrm{Conv}\left(\widetilde{\boldsymbol{F}}^i_{3D}\right) \tag{4}$$

$$\mathcal{L}^i_{ms-adv} = -\lambda^i_{ms-adv}\mathbb{E}_{\boldsymbol{I}^i_{3D}}\,\mathrm{softplus}(D^i_{ms}(\boldsymbol{I}^i_{3D})), \tag{5}$$

where D^i_{ms} denotes the discriminator at the i-th scale, and λ^i_{cs-adv} denotes the corresponding loss weight.

With the help of the information from the original input image and the supervision of our multi-scale adversarial loss, the geometry information contained in the resulted features $\widetilde{\boldsymbol{F}}^i_{3D}$ is learned to be more reliable and authentic.

3D-2D Prior Fusion

As mentioned, the 3D features generated in our approach contain rich geometric information that helps in reconstructing facial images with good structures. However, they may lack sufficient 2D texture information. To overcome this

limitation, we incorporate the StyleGAN2 generator to extract adequate 2D texture priors. Specifically, we use the latent code $\boldsymbol{Z}$2D and upsample it progressively through pretrained StyleGAN2 generator layers to obtain a set of features $\boldsymbol{F}^i_{\text{2D}} \in \mathbb{R}^{\frac{512}{i} \times \frac{512}{i} \times 32}$ that contain rich texture information:

$$\{\boldsymbol{F}^i_{\text{2D}}\} = \text{Gen}_{\text{2D}}\left(\boldsymbol{Z}_{\text{2D}}\right), \quad i = 0, 1, 2\ . \tag{6}$$

The 3D and 2D features are then adaptively fused together by our proposed **Adaptive Weighting Block (AWB)** to get better results. The detailed architecture of the AWB is shown in Fig. 2(b), which is inspired by the Weight Modulation of StyleGAN2. The difference is that our AWB is controlled by the head pose of the input image. This is motivated by the fact that for input images with large head poses, StyleGAN2 may generate results with distorted facial structures and unpleasing artifacts. Thus, the proportion of such information should be decreased. Conversely, for input images with minor head poses, the 2D features can be more effectively utilized to improve the visual quality of the reconstructed facial images. This approach ensures that the fusion process optimally leverages the strengths of each feature type while mitigating the negative effects of potential distortions and artifacts.

Specifically, at the i-th scale, we first use an embedding layer to encode the head pose $\boldsymbol{P}$, which is then combined with the 3D feature $\widetilde{\boldsymbol{F}}^i_{\text{3D}}$ to produce two modulation parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that are used to modulate the 2D feature $\boldsymbol{F}^i_{\text{2D}}$. The pipeline of the AWB is formally formulated as follows:

$$\boldsymbol{\alpha}, \boldsymbol{\beta} = \text{MConv}\left(\widetilde{\boldsymbol{F}}^i_{\text{3D}}, \text{Embed}\left(\boldsymbol{P}\right)\right), \tag{7}$$

$$\boldsymbol{F}^i_{\text{hidden}} = \boldsymbol{\alpha} \odot \boldsymbol{F}^i_{\text{2D}} + \boldsymbol{\beta}, \tag{8}$$

where $MConv$ is the modulated convolution used in StyleGAN2. $\odot$ denotes element-wise product and $+$ denotes element-wise addition. $\boldsymbol{F}^i_{\text{hidden}}$ is the result feature map.

Apart from the proposed AWB, we also perform cross-attention (Vaswani et al. 2017; Liang et al. 2021) between $\boldsymbol{F}^i_{\text{hidden}}$ and $\widetilde{\boldsymbol{F}}^i_{\text{3D}}$ to improve the fusion quality. The pipeline is shown in Fig. 2(c). Specifically, $\boldsymbol{F}^i_{\text{hidden}}$ is mapped into the query $\boldsymbol{Q^i}$, and $\widetilde{\boldsymbol{F}}^i_{\text{3D}}$ is mapped into the key $\boldsymbol{K^i}$ and the value $\boldsymbol{V^i}$ through linear layers. Then the cross-attention is computed as:

$$\boldsymbol{F}^i_{\text{out}} = \text{MH-Attn}\left(\boldsymbol{Q^i K^i}\right)\boldsymbol{V^i} + \boldsymbol{F}^i_{\text{hidden}}, \tag{9}$$

where MH-Attn denotes the the multi-head attention. At last, we feed $\boldsymbol{F}^0_{\text{out}}$ to a convolutional layer to produce the final output image $\boldsymbol{I}_{\text{out}}$. As a result, by fusing both 3D and 2D priors, our method is able to generate results with reliable structures and vivid details.

Model Objectives

The learning objectives of training our FREx consists of: 1) 3D geometry loss that constrains the rendered results of 3D GAN close to the ground-truth $\boldsymbol{I}_{\text{HQ}}$, 2) reconstruction and adversarial loss for authentic and realistic restoration, 3) proposed face structure loss for better supervision of facial structures, 4) proposed multi-scale adversarial loss for multi-scale fidelity, mentioned in Equation 5

3D Geometry Loss. To retain face structure better, we use reconstruction loss following (Richardson et al. 2021), including pixel-wise $\mathcal{L}_2$ loss, perceptual loss (Zhang et al. 2018a) and the ID loss (Wang et al. 2021) which focus on identity similarity with a pretrained ArcFace network (Deng et al. 2019a). Moreover, to constrain the $\boldsymbol{Z}_{\text{3D}}$ to the $\mathcal{W}+$ space, we adopt the adversarial loss with $\boldsymbol{R}_1$ regularization. Finally, to predict a precise camera parameters $\boldsymbol{P}$, we use $\mathcal{L}_1$ loss with the ground-truth $\boldsymbol{P}_{gt}$ estimated by (Deng et al. 2019b). The total 3D geometry loss can be written as:

$$\begin{aligned}\mathcal{L}_{\text{3D-rec}} &= \lambda_{l1}\|\boldsymbol{I}_{\text{3D-out}} - \boldsymbol{I}_{\text{HQ}}\|_2 \\ &+ \lambda_{\text{3D-per}}\|\phi(\boldsymbol{I}_{\text{3D-out}}) - \phi(\boldsymbol{I}_{\text{HQ}})\|_1 \\ &+ \lambda_{3D-ID}\,\text{ID}\left(\boldsymbol{I}_{\text{3D-out}} - \boldsymbol{I}_{\text{HQ}}\right),\end{aligned} \tag{10}$$

$$\mathcal{L}_{Z-adv} = -\lambda_{Z-adv}\mathbb{E}_{\boldsymbol{Z}_{\text{3D}}}\,\text{softplus}(D_{\text{3D}}(\boldsymbol{Z}_{\text{3D}})), \tag{11}$$

$$\mathcal{L}_P = \lambda_P\|\boldsymbol{P} - \boldsymbol{P}_{\text{gt}}\|_1, \tag{12}$$

$$\mathcal{L}_{3D} = \mathcal{L}_{\text{3D-rec}} + \mathcal{L}_{Z-adv} + \mathcal{L}_P. \tag{13}$$

Reconstruction and GAN Loss. We adopt the widely-used L1 loss, the perceptual loss (Zhang et al. 2018a) and the ID loss (Wang et al. 2021) as our reconstruction loss. We also employ the adversarial loss used in (Chan et al. 2022), which enforce the consistency between $\boldsymbol{I}_{\text{Input}}$ and $\boldsymbol{I}_{\text{HQ}}$ through a dual-discriminator. The reconstruction and GAN loss are defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{2D-rec}} &= \lambda_{l1}\|\boldsymbol{I}_{\text{out}} - \boldsymbol{I}_{\text{HQ}}\|_1 \\ &+ \lambda_{\text{per}}\|\phi(\boldsymbol{I}_{\text{out}}) - \phi(\boldsymbol{I}_{\text{HQ}})\|_1 \\ &+ \lambda_{ID}\,\text{ID}\left(\boldsymbol{I}_{\text{out}} - \boldsymbol{I}_{\text{HQ}}\right),\end{aligned} \tag{14}$$

$$\mathcal{L}_{2D-adv} = -\lambda_{adv}\mathbb{E}_{\boldsymbol{I}_{\text{out}}}\,\text{softplus}(D_{\text{dual}}(\boldsymbol{I}_{\text{out}}, \boldsymbol{I}_{\text{LQ}})), \tag{15}$$

$$\mathcal{L}_{2D} = \mathcal{L}_{\text{2D-rec}} + \mathcal{L}_{2D-adv}, \tag{16}$$

where $\boldsymbol{I}_{\text{out}}$ is the restoration output of FREx, D_{dual} denotes the dual-discriminator and λ denotes the loss weight.

Facial Structure Loss. In order to further enhance the structure accuracy of our restored result, we introduce the facial structure loss. We use the pretrained ArcFace (Deng et al. 2019a) model to extract high-level perceptual facial features of the output and ground truth images. Specifically, we extract features gradually from hidden layers and calculate the L1 loss as:

$$\mathcal{L}_{\text{struct}} = \lambda_{\text{struct}}\|\phi_{Arc}(\boldsymbol{I}_{\text{out}}) - \phi_{Arc}(\boldsymbol{I}_{\text{HQ}})\|_1, \tag{17}$$

where ϕ_{Arc} denotes the pretrained ArcFace network and we use its $\{\text{layer1}, \ldots, \text{layer4}\}$ feature maps before activation. λ_{struct} denotes the loss weight of the facial structure loss.

The overall model objective constitutes above losses, where the $\mathcal{L}_{ms-adv}$ is mentioned in Equation 5:

$$\mathcal{L}_{total} = \mathcal{L}_{3D} + \mathcal{L}_{2D} + \sum_i \mathcal{L}^i_{ms-adv} + \mathcal{L}_{struct}. \tag{18}$$

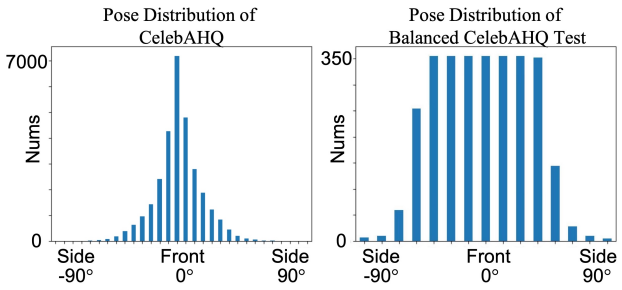

Figure 3: Comparison of pose distribution between original CelebAHQ dataset and our Balanced CelebAHQ Test dataset. Since the pose is more evenly distributed within our dataset, it provides a better evaluation of the model's capabilities across varying poses.

Experiments

Datasets

Training Dataset. We train our FREx on the FFHQ dataset (Karras, Laine, and Aila 2019), which contains 70,000 high-quality images. To simulate the extreme conditions, we synthesize training images using the following degradation model:

$$\boldsymbol{I}_{\mathrm{LQ}} = [(\boldsymbol{I}_{\mathrm{HQ}} \circledast \boldsymbol{k}) \downarrow_r + \boldsymbol{n}_\delta]_{\mathrm{JPEG}_q}. \tag{19}$$

The high-quality image is first convolved with a blur kernel k, which includes Gaussian blur with standard deviation $\sigma \in \{0 : 20\}$ and motion blur with kernel size $s \in \{21 : 10 : 71\}$, followed by a downsampling operation of sacle $r \in \{0 : 16\}$. After that, additive Gaussian noise $\delta \in \{0 : 50\}$ is added to the images, and then JPEG compression with quality factor $q \in \{50 : 100\}$ is applied. Finally, we resize the low-quality image back to 512×512. To approximate complex and extreme conditions, we randomly apply multiple degradation processes to the images.

Testing Datasets. We evaluate our model on our synthetic Balanced CelebAHQ Test dataset. The original CelebAHQ dataset (Karras et al. 2017) exhibits a imbalanced distribution of facial poses. Therefore, to better evaluate the model's capabilities across varied facial poses, we select a more evenly distributed subset consisting of of 3,000 facial images from the CelebAHQ dataset, and employ the same synthesizing approach used during training. A comparison of the pose distributions of the Balanced CelebAHQ Test dataset and the original dataset can be observed in Fig. 3. Moreover, to evaluate the model's performance under realistic extreme degradation scenarios, we collect 300 low-quality face images from CelebA (Liu et al. 2015) and WIDERFACE (Yang et al. 2016) to serve as the real-world test set.

Experimental Settings and Metrics

Implementation Details. We adopt the pretrained 3D face GAN EG3D (Chan et al. 2022) as the decoder for 3D geometry prior extraction, and the pretrained StyleGAN2 for 2D texture prior extraction. The 2D encoder consists of seven downsampling layers with skip connections and the 3D encoder consists of 14 layers to extract $\boldsymbol{P}$, $\boldsymbol{Z}_{3\mathrm{D}}$ and $\overline{\boldsymbol{Z}}$. We augment the training data using random horizontal flipping and EG3D to balance pose distribution. We use the Adam (Kingma and Ba 2014) optimizer with a batch size of 4 for a total of 250k iterations. The learning rate is set to 0.002 and decayed by a factor of 2 at the 200k-th iteration.

Metrics. To evaluate on the balanced CelebAHQ Test with ground truths, we adopt PSNR, SSIM, and LPIPS (Zhang et al. 2018a) as metrics. We also employ the non-reference perceptual metrics FID (Heusel et al. 2017) and NIQE (Mittal, Soundararajan, and Bovik 2012). Additionally, we evaluate identity preservation through the cosine similarity of features derived from the ArcFace network (Deng et al. 2019a), denoted as Deg.

Comparison with State-of-the-art Methods

We compare our FREx with existing state-of-the-art methods, including reference-prior-based methods: VQFR (Gu et al. 2022), CodeFormer (Zhou et al. 2022), RestoreFormer (Wang et al. 2022), and generative-prior-based methods: GPEN (Yang et al. 2021), GLEAN (Chan et al. 2021), GFP-GAN (Wang et al. 2021). Their official released models are adopted in the experiments. More results can be found in the supplementary materials.

Evaluation on Synthetic Dataset. The quantitative comparison on the Balanced CelebAHQ Test dataset is shown in Table. **??**. Our FREx achieves the highest scores for LPIPS, FID, PSNR, and SSIM, demonstrating that the restored faces are both perceptually and visually superior to the results obtained by other methods. FREx also obtains the best Deg. score, showing that our method could retain identity better.

Additionally, we present the qualitative comparison in Fig. 4. Although the degradation is too large for human to recognize and the facial pose is extreme, our FREx can still produce visually pleasing restoration results.

Evaluation on Real-world Dataset. We also shown the visual illustrations of real-world cases in Fig. 5, where the input images are from CelebA and WIDERFACE mentioned above. It can be observed that our method produces superior results compared to other existing state-of-the-art face restoration methods. Specifically, as shown in the first column of Fig. 5, our method could restore much finer beard and wrinkle. Meanwhile, our method also restore better facial structures compared to other methods.

Ablation Studies

Ablation on 3D Geometry Prior. To verify the effectiveness of 3D geometry prior, we remove the 3D geometry prior extraction part and replace the $\boldsymbol{F}_{3\mathrm{D}}$ with the upsampled $\boldsymbol{Z}_{2\mathrm{D}}$. As shown in Table. 2, the absence of 3D geometry prior leads to a drop in LPIPS, FID, and PSNR, suggesting that 3D geometry prior contributes to produce authentic face images. Besides, the superiority of 3D geometry prior is also demonstrated in column 2 and column 7 of Fig. 6. We also show the results directly generate by the pretrained 3D encoder in Fig. 7. It can be observed that the 3D encoder have the capability to restore accurate facial structures, given the

Methods	LPIPS ↓	FID ↓	NIQE ↓	Deg. ↓	PSNR ↑	SSIM ↑
Bilinear	0.6475	261.9	14.86	99.67	20.24	<u>0.5639</u>
GLEAN (Chan et al. 2021)	0.5211	75.37	5.423	96.90	<u>20.68</u>	0.5606
GPEN (Yang et al. 2021)	0.5207	78.49	4.2611	<u>94.28</u>	19.93	0.5175
GFP-GAN (Wang et al. 2021)	0.5501	87.47	<u>3.7898</u>	96.36	19.08	0.4606
RestoreFormer (Wang et al. 2022)	0.5553	111.08	3.9674	97.12	19.62	0.4599
CodeFormer (Zhou et al. 2022)	<u>0.4981</u>	<u>44.83</u>	4.2115	93.88	20.08	0.4765
VQFR (Gu et al. 2022)	0.5478	92.75	**3.6376**	96.53	18.83	0.3910
FREx (ours)	**0.4671**	**39.58**	4.9161	**91.75**	**20.95**	**0.5890**
GT	0.0	8.27	4.0620	32.82	∞	1.0

Table 1: Quantitative comparison on Balanced CelebAHQ Test for face restoration under extreme conditions. Bold and underline indicates the best and the second best performance.

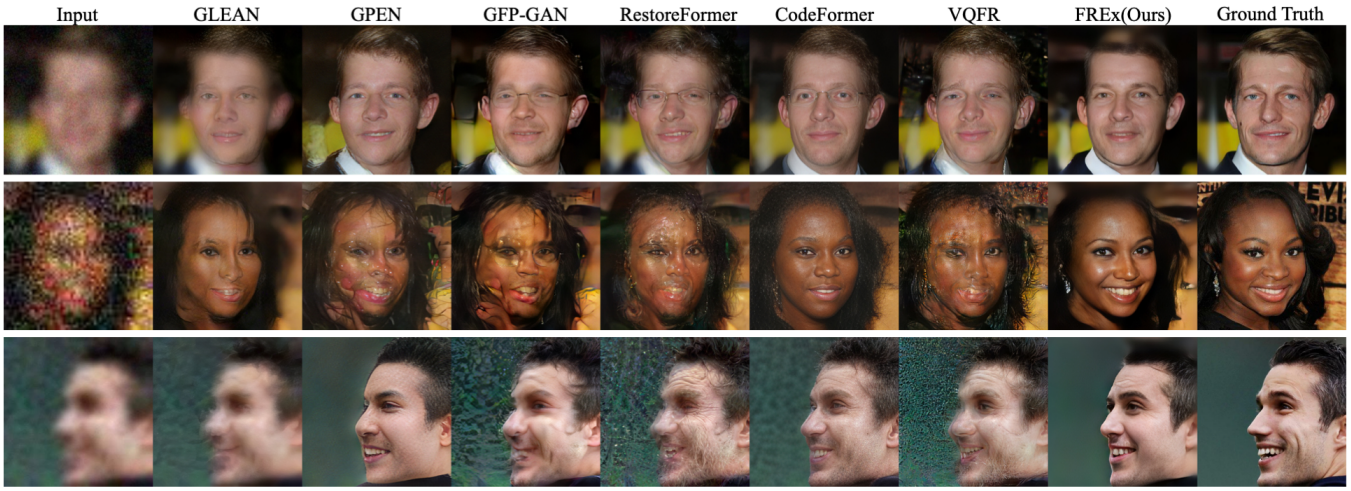

Figure 4: Qualitative comparison on the Balanced CelebAHQ Test. Even though the input images are severely degraded and the facial pose is large, our FREx restores natural and authentic face images.

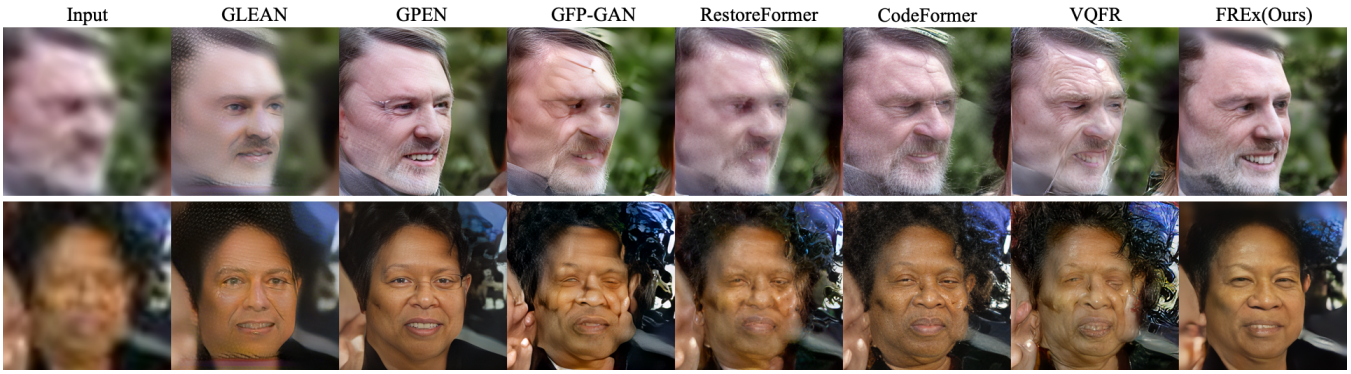

Figure 5: Qualitative comparison on the real-world dataset. Our method demonstrates its robustness even when dealing with challenging real-world cases.

fact that it has been encoded with rich 3D priors. However, the restored results is oversmoothed, lacking of fine details. By further leveraging 2D priors on top of these reasonable facial structures, our FREx method could generate results with both accurate structures and sufficient details.

Ablation on Adaptive Weighting Block. Adaptive weighting block (AWB) aims to control the fusion of 2D and 3D information by the face pose. By the control of face pose, the proportion of 3D information should be increased if the pose is large to produce accurate face structure and vice versa. In comparison, upon removing AWB, the model fails to effectively utilize 2D and 3D information, resulting in worse

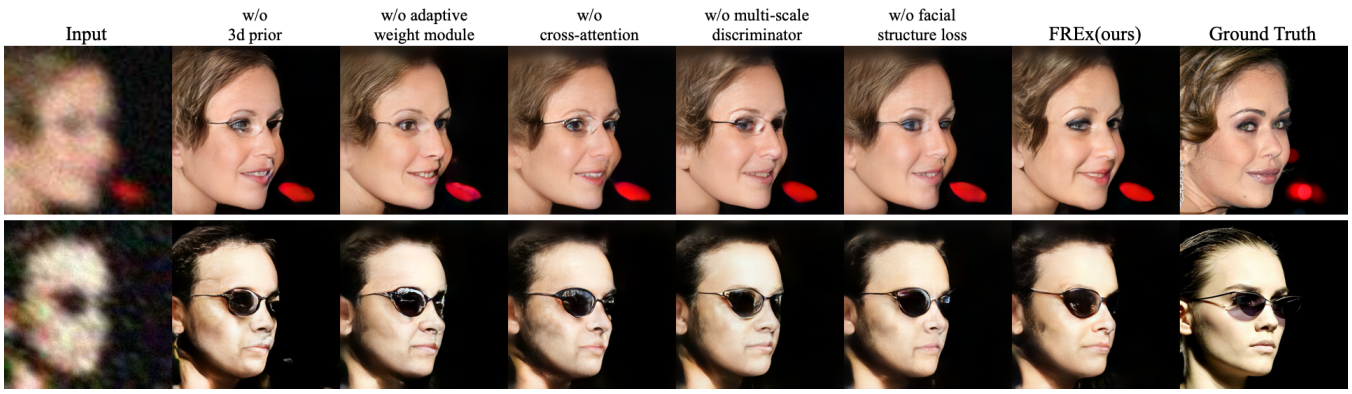

Figure 6: Visual illustrations of the ablation studies on different modules of our method. Our final model produces the best results.

Configuration	LPIPS ↓	FID ↓	PSNR ↑
Our FREx	**0.4671**	**39.58**	**20.95**
w/o 3D geometry prior	0.4703(↑)	41.01(↑)	20.63(↓)
w/o adaptive weight block	0.4730(↑)	48.35(↑)	20.89(↓)
w/o 2D-3D cross-attention	0.4714(↑)	48.34(↑)	20.87(↓)
w/o multi-scale discriminator	0.4698(↑)	45.56(↑)	20.90(↓)
w/o facial structure loss	0.4712(↑)	49.42(↑)	20.87(↓)

Table 2: Ablation studies of different settings on the Balanced CelebAHQ Test.

Figure 7: Results generated by our full FREx method and the 3D decoder.

performance shown in Table. 2 and Fig. 6.

Ablation on 3D-2D Cross-Attention. We employ cross-attention between $\boldsymbol{F}_{\text{hidden}}$ and $\widetilde{\boldsymbol{F}}_{\text{3D}}$ to improve the fusion quality. After removing cross-attention, a decline in performance is noticed, as shown in Table. 2 and Fig. 6

Ablation on Multi-Scale Adversarial Losses. The multi-scale discriminators calculate the adversarial loss for RGB images from each layer during the faithful 3D feature generation process. This approach helps to reduce artifacts and enhance the fidelity of the resulting 3D features. As shown in Table. 2 and Fig. 6, a performance drop is observed without this intermediate restriction.

Ablation on Facial Structure Loss. The Facial structure loss focuses more on the perceptual aspect of the facial images. Without this supervision, the output images are less similar in perceptual distance with the ground truth images and are less natural, as shown in Table. 2 and Fig. 6

Figure 8: Limitations of our model. Our method sometimes generates blurry results on top of heads due to face misalignment in the pretrained 3D decoder (Chan et al. 2022).

Limitations

Our method occasionally produces blurry results on the top of heads, as illustrated in Fig 8. This is caused by face misalignment in the pretrained 3D decoder (Chan et al. 2022), which can only generate a specific range of facial features, resulting in missing structures at the top of the head, as shown in the last column of the figure. Retraining the decoder could address this issue, which we leave for the future.

Conclusion

In this paper, we introduce a novel method called FREx for restoring facial images with extreme unknown degradations. To tackle this challenging task, we propose to incorporate 3D structural priors and 2D texture priors into the restoration process, where pretrained 2D and 3D GANs are utilized for distilling rich priors. We further propose a multi-scale adversarial loss to regularize the extracted 3D features to be more reliable. A novel module called Adaptive Weighting Block is further proposed to fuse the 2D and 3D information in an adaptive manner based on the head poses, endowing our network the capability to utilize different information more effectively. By integrating 3D and 2D information, our model can restore structurally accurate and visually pleasing results even when the majority of information is lost in the inputs or the head poses are extreme.

Acknowledgments

This project was supported in part by Shenzhen Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone, under Grant No. HTHZQSWS-KCCYB-2023052. This work was supported in part by the National Natural Science Foundation of China under Grant No.62276030 and 62236003. This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant (JCYJ20190809172201639, WDZC20200820200655001), Shenzhen Key Laboratory (ZDSYS20210623092001004), and Beijing Key Lab of Networked Multimedia.

References

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 16123–16133.

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 *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 *CVPR*, 11896–11905.

Chen, Y.; Tai, Y.; Liu, X.; Shen, C.; and Yang, J. 2018. Fsrnet: End-to-end learning face super-resolution with facial priors. In *CVPR*, 2492–2501.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019a. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 4690–4699.

Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; and Tong, X. 2019b. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR workshops*, 0–0.

Dogan, B.; Gu, S.; and Timofte, R. 2019. Exemplar guided face image super-resolution without facial landmarks. In *CVPR workshops*, 0–0.

Dong, C.; Deng, Y.; Loy, C. C.; and Tang, X. 2015. Compression artifacts reduction by a deep convolutional network. In *ICCV*, 576–584.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *ECCV*, 184–199.

El Helou, M.; Zhou, R.; and Süsstrunk, S. 2020. Stochastic frequency masking to improve super-resolution and denoising networks. In *ECCV*, 749–766.

Gu, J.; Shen, Y.; and Zhou, B. 2020. Image processing using multi-code gan prior. In *CVPR*, 3012–3021.

Gu, Y.; Wang, X.; Xie, L.; Dong, C.; Li, G.; Shan, Y.; and Cheng, M.-M. 2022. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *ECCV*, 126–143.

Guo, J.; and Chao, H. 2016. Building dual-domain representations for compression artifacts reduction. In *ECCV*, 628–644.

Guo, X.; Li, Y.; and Ling, H. 2016. LIME: Low-light Image Enhancement via Illumination Map Estimation. *TIP*.

Guo, Y.; Chen, J.; Wang, J.; Chen, Q.; Cao, J.; Deng, Z.; Xu, Y.; and Tan, M. 2020. Closed-loop matters: Dual regression networks for single image super-resolution. In *CVPR*, 5407–5416.

Hao, Z.; You, S.; Li, Y.; Li, K.; and Lu, F. 2019. Learning from synthetic photorealistic raindrop for single image raindrop removal. In *ICCV Workshops*, 0–0.

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. *NeurIPS*, 30.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *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 *CVPR*, 8110–8119.

Kim, D.; Kim, M.; Kwon, G.; and Kim, D.-S. 2019. Progressive face super-resolution via attention to facial landmark. *arXiv preprint arXiv:1908.08239*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kupyn, O.; Budzan, V.; Mykhailych, M.; Mishkin, D.; and Matas, J. 2018. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 8183–8192.

Lefkimmiatis, S. 2017. Non-local color image denoising with convolutional neural networks. In *CVPR*, 3587–3596.

Li, K.; Li, Y.; You, S.; and Barnes, N. 2017a. Photo-realistic Simulation of Road Scene for Data-Driven Methods in Bad Weather. In *CVPR Workshops*, 491–500.

Li, X.; Chen, C.; Zhou, S.; Lin, X.; Zuo, W.; and Zhang, L. 2020a. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*, 399–415.

Li, X.; Li, W.; Ren, D.; Zhang, H.; Wang, M.; and Zuo, W. 2020b. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *CVPR*, 2706–2715.

Li, X.; Liu, M.; Ye, Y.; Zuo, W.; Lin, L.; and Yang, R. 2018. Learning warped guidance for blind face restoration. In *ECCV*, 272–289.

Li, Y.; Guo, F.; Tan, R. T.; and Brown, M. S. 2014. A Contrast Enhancement Framework with JPEG Artifacts Suppression. In *ECCV*, 174–188.

Li, Y.; Tan, R. T.; and Brown, M. S. 2015. Nighttime haze removal with glow and multiple light colors. In *ICCV*, 226–234.

Li, Y.; Tan, R. T.; Guo, X.; Lu, J.; and Brown, M. S. 2016. Rain streak removal using layer priors. In *CVPR*, 2736–2744.

Li, Y.; You, S.; Brown, M. S.; and Tan, R. T. 2017b. Haze visibility enhancement: A survey and quantitative benchmarking. *CVIU*, 165: 1–16.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *ICCV*, 1833–1844.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *CVPR workshops*, 136–144.

Liu, D.; Wen, B.; Fan, Y.; Loy, C. C.; and Huang, T. S. 2018. Non-local recurrent network for image restoration. *NeurIPS*, 31.

Liu, J.; Zhang, W.; Tang, Y.; Tang, J.; and Wu, G. 2020. Residual feature aggregation network for image super-resolution. In *CVPR*, 2359–2368.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*, 3730–3738.

Lv, F.; Li, Y.; and Lu, F. 2021. Attention Guided Low-light Image Enhancement with a Large Scale Low-light Simulation Dataset. *IJCV*, 129(7): 2175–2193.

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 *CVPR*, 2437–2445.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3): 209–212.

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. arXiv:2008.00951.

Shen, Z.; Lai, W.-S.; Xu, T.; Kautz, J.; and Yang, M.-H. 2018. Deep semantic face deblurring. In *CVPR*, 8260–8269.

Timofte, R.; Agustsson, E.; Van Gool, L.; Yang, M.-H.; and Zhang, L. 2017. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPR workshops*, 114–125.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *NeurIPS*, 30.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Wang, X.; Li, Y.; Zhang, H.; and Shan, Y. 2021. Towards Real-World Blind Face Restoration with Generative Facial Prior. In *CVPR*, 9168–9178.

Wang, Z.; Zhang, J.; Chen, R.; Wang, W.; and Luo, P. 2022. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *CVPR*, 17512–17521.

Wu, Y.; Wang, X.; Li, Y.; Zhang, H.; Zhao, X.; and Shan, Y. 2021. Towards Vivid and Diverse Image Colorization with Generative Color Prior. In *ICCV*, 14377–14386.

Xu, L.; Ren, J. S.; Liu, C.; and Jia, J. 2014. Deep convolutional neural network for image deconvolution. *NeurIPS*, 27.

Yang, S.; Luo, P.; Loy, C.-C.; and Tang, X. 2016. Wider face: A face detection benchmark. In *CVPR*, 5525–5533.

Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2021. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, 672–681.

Yang, Z.; Zeng, A.; Yuan, C.; and Li, Y. 2023. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, 4210–4220.

Yu, K.; Wang, X.; Dong, C.; Tang, X.; and Loy, C. C. 2021. Path-restore: Learning network path selection for image restoration. *PAMI*, 44(10): 7078–7092.

Yu, X.; Fernando, B.; Ghanem, B.; Porikli, F.; and Hartley, R. 2018. Face super-resolution guided by facial component heatmaps. In *ECCV*, 217–233.

Yuan, Z.; Zhu, Y.; Li, Y.; Liu, H.; and Yuan, C. 2023. Make Encoder Great Again in 3D GAN Inversion through Geometry and Occlusion-Aware Encoding. In *ICCV*.

Zhang, F.; Li, Y.; You, S.; and Fu, Y. 2021. Learning Temporal Consistency for Low Light Video Enhancement From Single Images. In *CVPR*, 4967–4976.

Zhang, F.; You, S.; Li, Y.; and Fu, Y. 2023. Learning Rain Location Prior for Nighttime Deraining. In *ICCV*, 13148–13157.

Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; and Zhang, L. 2017. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *TIP*, 26(7): 3142–3155.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *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 *ECCV*, 286–301.

Zhou, S.; Chan, K.; Li, C.; and Loy, C. C. 2022. Towards robust blind face restoration with codebook lookup transformer. *NeurIPS*, 35: 30599–30611.

Zhu, F.; Zhu, J.; Chu, W.; Zhang, X.; Ji, X.; Wang, C.; and Tai, Y. 2022. Blind face restoration via integrating face shape and generative priors. In *CVPR*, 7662–7671.

Zhu, S.; Liu, S.; Loy, C. C.; and Tang, X. 2016. Deep cascaded bi-network for face hallucination. In *ECCV*, 614–630.