Diffusion Prior Interpolation for Flexibility Real-World Face Super-Resolution

Jiarui Yang[1, 2], **Tao Dai**[3, *], **Yufei Zhu**[3], **Naiqi Li**[2], **Jinmin Li**[2], **Shu-Tao Xia**[2]

[1]College of Artificial Intelligence, Nankai University, Tianjin, China
[2]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[3]College of Computer Science and Software Engineering, Shenzhen University, China
1120230264@mail.nankai.edu.cn, daitao.edu@gmail.com
zhuyufei2021@email.szu.edu.cn, ljm22@mails.tsinghua.edu.cn, {linaiqi, xiast}@sz.tsinghua.edu.cn

Abstract

Diffusion models represent the state-of-the-art in generative modeling. Due to their high training costs, many works leverage pre-trained diffusion models' powerful representations for downstream tasks, such as face super-resolution (FSR), through fine-tuning or prior-based methods. However, relying solely on priors without supervised training makes it challenging to meet the pixel-level accuracy requirements of discrimination task. Although prior-based methods can achieve high fidelity and high-quality results, ensuring consistency remains a significant challenge. In this paper, we propose a masking strategy with strong and weak constraints and iterative refinement for real-world FSR, termed Diffusion Prior Interpolation (DPI). We introduce conditions and constraints on consistency by masking different sampling stages based on the structural characteristics of the face. Furthermore, we propose a condition Corrector (CRT) to establish a reciprocal posterior sampling process. DPI can balance consistency and diversity and can be seamlessly integrated into pre-trained models. In extensive experiments conducted on synthetic and real datasets, along with consistency validation in face recognition, DPI demonstrates superiority over SOTA FSR methods.

Introduction

Image super-resolution (SR) is a classic ill-posed problem aimed at enhancing image quality by restoring high-resolution (HR) details from low-resolution (LR) images. In the context of face super-resolution (FSR), this technology is applied in areas such as face recognition (Zou and Yuen 2011) and visual enhancement (Jiang et al. 2021). These applications typically require SR images to exhibit both high consistency and fidelity. Previous SR work has tended to overly pursue distortion-based quantitative metrics (such as PSNR, SSIM) (Chen et al. 2018; Gao et al. 2023). Blau et al. (Blau and Michaeli 2018) mathematically prove that distortion and perceptual quality are at odds with each other, implying that excessive pursuit of PSNR or SSIM indirectly leads to poorer fidelity. Moreover, discriminative model-based SR methods (Gao et al. 2023; Wang et al. 2023a) typically employ end-to-end training, achieving it by minimizing pixel-wise loss between the SR output image and the GT image. It is well-known that such learning objectives favor distortion measures and constrain the SR output to the average of multiple possibilities. This maintains a certain consistency but potentially results in over-smoothing outputs (Sajjadi, Scholkopf, and Hirsch 2017), as shown in Fig. 1(c-d).

In contrast, generative methods such as Variational Autoencoders (VAEs) (Liu, Siu, and Wang 2021), Normalizing Flows (NFs) (Lugmayr et al. 2020), Generative Adversarial Networks (GANs) (Wang et al. 2023c) and Diffusion Models (DMs) (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) have the capability to generate high-fidelity images. Among these, Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020) have recently gained significant attention and research interest due to their impressive generative capabilities. DDPMs exhibit advantages such as stable training and enhanced controllability compared to other generative models (Brock, Donahue, and Simonyan 2018; Kirichenko, Izmailov, and Wilson 2020; Bredell et al. 2023). At present, diffusion-based FSR work can be broadly categorized into those that require task-directed retraining (Saharia et al. 2023; Wei et al. 2023; Li et al. 2022; Whang et al. 2022) and prior-based methods (Choi et al. 2021; Chung et al. 2022; Kawar et al. 2022). Training a conditional DDPM from scratch requires significant computational resources and can limit the prior space to lead to sub-optimal results (Wang et al. 2023b). Introducing conditions to utilize the priors encapsulated in the pre-trained model is an alternating solution. However, adding conditions to a pre-trained model introduces errors and visual artifacts in the intrinsic probability distributions at each time step, leading to the generation of results that deviate from the model's prior manifold (Mei, Nair, and Patel 2022). As shown in Fig. 1(e-f), although these prior-based methods achieve good visual perception, the issue of consistency remains unresolved. This is primarily because prior-based methods are unsupervised and generative in nature. When dealing with low-level tasks requiring pixel-level accuracy, it is challenging to achieve precise discrimination.

We are aware that the sampling process of DDPMs is an iterative one, progressing from coarse to fine. Wang et al. (Wang et al. 2023d) have been demonstrated that there exists a time step that partitions the sampling interval, and beyond this time step, the error between the real posterior distribution and the posterior distribution introduced by con-

*Corresponding author: daitao.edu@gmail.com.

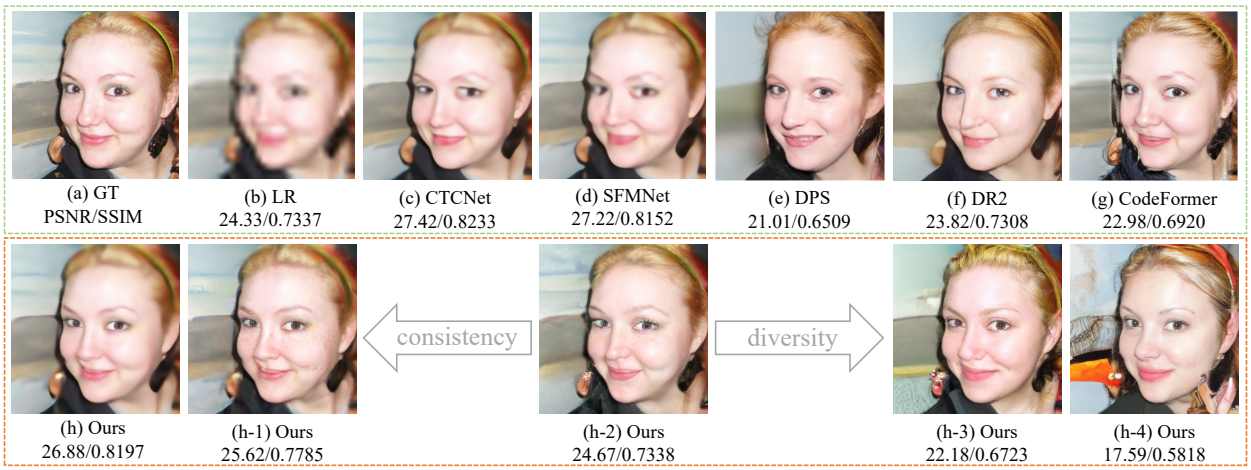


Figure 1: Visualization of FSR results using different types of methods, including those based on discriminative models (c-d) (Gao et al. 2023; Wang et al. 2023a), generative models (e-f) (Chung et al. 2022; Wang et al. 2023d), and prior-based methods (e-h) (Zhou et al. 2022). Our approach combines the strengths of discriminative and generative models, allowing for flexible adjustment of facial diversity and fidelity.

ditioning becomes sufficiently small. This means that during the early sampling phase, the impact of the conditions is more significant because the distribution of conditions is closer to the true distribution. This inspire us to impose strong and weak constraints on the posterior distribution to flexibly regulate the consistency and diversity of FSR, as depicted in the lower part of Fig. 1.

Specifically, we take the LR image as a condition and design various masks tailored to facial features, combining them to form different Condition Masks (CMs). The CMs include a Fixed Condition Mask (FCM) and a Randomly Adaptive Condition Mask (RACM), to mask the sampling process. This is similar to the prior-based inpainting methods (Song et al. 2020; Lugmayr et al. 2022), but our CMs can provide detailed neighborhood information to better utilize the priors. We use an adjustable scalar to divide the sampling process into two stages, where the first stage utilizes the FCM to limit the prior space and ensure consistency in FSR. In the second stage, the RACM, guided by facial structure supervision, ensures consistency while enhancing the fidelity and diversity of the sampling. The trade-off between consistency and fidelity can be realized by adjusting this scalar value. In addition, we introduce a condition Corrector (CRT) to create a reciprocal sampling process where conditions are updated with the assistance of priors, and samples improve with the refinement of conditions. The CRT trades a very small time cost for better performance. Our approach leverages the priors of diffusion to interpolate the condition masks, which we call Diffusion Prior Interpolation (DPI). **Our main contributions are summarized as follows**:

- We propose DPI, which effectively leverages the priors of pre-trained diffusion models for real-world FSR. Extensive experiments on both synthetic and real-world datasets demonstrate that our method outperforms SOTA FSR methods.
- We propose a novel masking strategy tailored for facial features to mask the diffusion sampling process. Our approach ensures structural consistency and detail diversity in the face while supporting flexible sampling.
- By introducing a condition CRT, we establish a reciprocal process between conditions and samples. CRT incurs a small time cost in exchange for improved performance. CRT is not limited to specific networks and endows DPI with scalability.

Related Work

Prior-based Face Super-Resolution

Face images possess distinctive characteristics such as subject-centered focus, prominent foreground-background contrast, and well-defined face structures. Leveraging these prior information in previous work has effectively improved FSR performance (Menon et al. 2020; Chen et al. 2021; Leng and Wang 2022). FSRCH (Lu et al. 2022) introduces a pre-prior guiding method that extracts face priors from HR images and incorporates it into LR inputs, generating LRmix as a new SR input. Wang et al. (Wang et al. 2022a) employ distillation to propagate real face priors learned by a teacher network to guide the learning of a student network for FSR. In order to enhance the ability of face restoration, Wang et al. (Wang et al. 2022b) propose a Restoreformer that utilizes a high-quality dictionary that not only provides priors for the face, nose, and mouth but is generated through a high-quality face network that learns from a large number of undegraded faces. GLEAN (Chan et al. 2022) directly leverages rich and diverse priors encapsulated in a pre-trained face GAN. Zhou et al. (Zhou et al. 2022) introduce CodeFormer, a transformer-based prediction network that utilizes discrete codebooks learned in a compact proxy space through blind face recovery. The primary goal of CodeFormer is to reduce the uncertainty and ambiguity associated with recovery mapping by employing code prediction as a task.

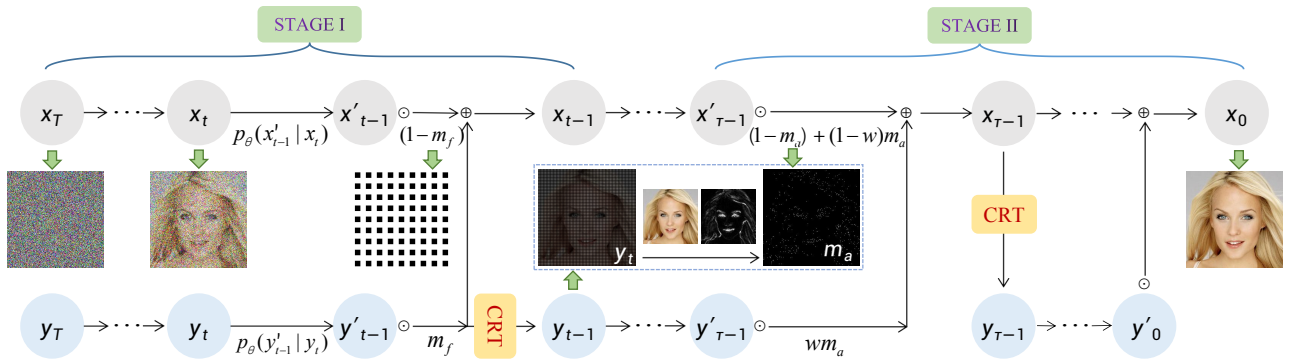


Figure 2: **Graphical model of Diffusion Prior Interpolation.** $\odot$ represents element-wise matrix multiplication. $\boldsymbol{x}_T$ and $\boldsymbol{y}_T$ correspond to the initial random noise and the initial condition respectively. We use the scalar τ to divide the sampling process into two stages. CRT is the Corrector function that applies Eq. 16 to CMs correction. $\boldsymbol{y}_t$ represents the intermediate condition. After posterior sampling, $\boldsymbol{y}'_t$ is multiplied by $\boldsymbol{m}_f$ and $\boldsymbol{m}_a$ to obtain CMs, including the FCM and RACM. Algorithm 1 provides a detailed description of our DPI.

We design CMs based on the unique features of the face, such as facial contours. As a result, our masking strategy effectively ensures consistent sampling of facial structures. Additionally, by leveraging the priors from pre-trained models, DPI enhances the fidelity and quality of FSR.

Conditional Diffusion Models

In the spatial domain, concatenating LR image to train a DDPM from scratch is a simple and efficient SR method, such as SR3 (Saharia et al. 2023) and SRDiff (Li et al. 2022). Whang et al. (Whang et al. 2022) train a DDPM to learn the residual between MSE-estimated images and HR images to enhance the diversity of the MSE-estimated images. RainDiffusion (Wei et al. 2023) employs two DDPMs for generating clear and degraded image pairs and then utilizes an idea similar to SR3 to achieve real-world image restoration. However, these conditional methods necessitate retraining the DDPM, which can be costly. To mitigate this, Brian et al. (Moser et al. 2023) propose training a DDPM in the discrete wavelet domain to learn conditional residual information, effectively reducing model parameters and achieving good SR performance. Additionally, most works introduce conditions into pre-trained DDPMs to circumvent training from scratch. PGDiff (Yang et al. 2024) utilizes the structure and color statistics of reference images to alleviate severe degradation issues. ILVR (Choi et al. 2021) incorporates low-frequency information from reference images into the posterior distribution to control the conditional generation of pre-trained DDPMs. Fei et al. (Fei et al. 2023) propose generating diffusion priors to model the posterior distribution of a pre-trained DDPM through unsupervised sampling. DPS (Chung et al. 2022) applies conditional gradient correction to the posterior for guiding the sampling process. Methods based on pre-trained models are influenced by conditional intensity and controllability; if the condition intensity or control is weak, the fidelity of the results is higher, but consistency is lacking, and vice versa.

Method

Preliminary

DDPMs define a Markovian forward diffusion process with T steps, which continuously transforms the initial state $\boldsymbol{x}_0$ through pre-specified noise scheduler $\{\beta_1, \beta_2, ..., \beta_T\}$ into an isotropic Gaussian distribution $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Each step of the forward process can be represented as a Gaussian transition, denoted as:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t|\sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I}) \tag{1}$$

where $\alpha_t = 1 - \beta_t$. Additionally, we can obtain the state at step t through a single-step transition:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t|\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_t$. The inverse process starts from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and employs a U-Net denoiser (Ronneberger, Fischer, and Brox 2015) with learnable parameters θ to fit the true posterior distribution $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. A judicious noise scheduler is employed to ensure that the inverse process also follows a Gaussian distribution (Bartlett 1978), which can be represented as:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}|\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)) \tag{3}$$

where $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ and $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ are the mean and variance predicted by the denoiser, respectively. Ho et al. (Ho, Jain, and Abbeel 2020) observe that directly predicting the mean is not optimal. Instead, a prevalent methodology involves parameterizing the $\boldsymbol{\mu}_\theta(\hat{\boldsymbol{x}}_0, \boldsymbol{x}_t, t)$ using a simplified loss function $\mathcal{L}_{simple} = ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta||_2^2$ to predict the noise $\boldsymbol{\epsilon}_\theta$. This can be represented by the following formula:

$$\boldsymbol{\mu}_\theta(\hat{\boldsymbol{x}}_0, \boldsymbol{x}_t, t) = \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\boldsymbol{x}}_0 \tag{4}$$

where $\hat{\boldsymbol{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t))$. Furthermore, Dhariwal et al. (Dhariwal and Nichol 2021) propose that incorporating an additional output v from the denoiser to

Algorithm 1: Diffusion Prior Interpolation, given a diffusion model $(\boldsymbol{\mu}_\theta(\cdot), \boldsymbol{\Sigma}_\theta(\cdot))$ and Corrector $CRT(\cdot)$.

Input: $\boldsymbol{y}_T$, τ, s, ω, $\boldsymbol{m}_f$
$\boldsymbol{x}_T \leftarrow$ sample from $\mathcal{N}(\mathbf{0}, \mathbf{I})$
for all t from T to 1 **do**
  $\boldsymbol{\mu}_{x_t}, \boldsymbol{\Sigma}_{x_t} \leftarrow \boldsymbol{\mu}_\theta(\hat{\boldsymbol{x}}_0, \boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$
  $\boldsymbol{x}'_{t-1} \leftarrow$ sample from $\mathcal{N}(\boldsymbol{\mu}_{x_t}, \boldsymbol{\Sigma}_{x_t})$
  $\boldsymbol{y}_t^n = \sqrt{\bar{\alpha}_t}\boldsymbol{y}_t + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta$
  $\boldsymbol{\mu}_{y_t} = \boldsymbol{\mu}_\theta(\boldsymbol{y}_t, \boldsymbol{y}_t^n, t)$
  $\boldsymbol{y}'_{t-1} \leftarrow$ sample from $\mathcal{N}(\boldsymbol{\mu}_{y_t}, \boldsymbol{\Sigma}_{x_t})$
  if $t > \tau$ **then**
    $\boldsymbol{x}_{t-1} = (1-\boldsymbol{m}_f) \odot \boldsymbol{x}'_{t-1} + \boldsymbol{m}_f \odot \boldsymbol{y}'_{t-1}$
  else
    $\boldsymbol{m}_a = Mask_{gen}(\boldsymbol{y}_t, s)$
    $w = t/\omega$
    $\boldsymbol{x}_{t-1} = (1-\boldsymbol{m}_a) \odot \boldsymbol{x}'_{t-1} + w\boldsymbol{m}_a \odot \boldsymbol{y}'_{t-1} + (1-w)\boldsymbol{m}_a \odot \boldsymbol{x}'_{t-1}$
    $\boldsymbol{y}'_{t-1} = \boldsymbol{x}_{t-1}$
  end if
  $\boldsymbol{y}_{t-1} = CRT(\boldsymbol{m}_f \odot \boldsymbol{y}'_{t-1}, \boldsymbol{y}_T, t)$
end for
return $\boldsymbol{x}_0$

parameterize the variance $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ is superior to the conventional approach of using a fixed variance β_t, we get:

$$\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t) = exp(vlog\beta_t + (1-v)log\tilde{\beta}_t) \tag{5}$$

where $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. Both improved algorithms like IDDPM (Nichol and Dhariwal 2021) and accelerated sampling algorithms like DDIM (Song, Meng, and Ermon 2020) follow a similar paradigm. Our proposed DPI can be plug-and-play in this paradigm.

Diffusion Prior Interpolation

As shown in the upper part of Fig. 2, it represents an unconditional diffusion sampling process. Conditions are introduced through masking, as illustrated in the lower half. The conditions consist of an initial condition $\boldsymbol{y}_T$ and intermediate conditions $\boldsymbol{y}_t$. We design a fixed mask $\boldsymbol{m}_f$ and an adaptive mask $\boldsymbol{m}_a$ for generating Condition Masks (CMs). We will describe the form of CMs in detail in the following subsection.

We start by forward sampling the condition, and we get:

$$\boldsymbol{y}_t^n = \sqrt{\bar{\alpha}_t}\boldsymbol{y}_t + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \tag{6}$$

where $\boldsymbol{y}_t^n$ represents the noisy condition obtained through the reparameterization of Eq. 2 and $\boldsymbol{\epsilon}_\theta$ corresponds to the noise predicted by the pre-trained DDPM. Subsequently, we sample the conditional posterior distribution $\boldsymbol{y}'_{t-1}$ through the following equation:

$$p_\theta(\boldsymbol{y}'_{t-1}|\boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}|\boldsymbol{\mu}_\theta(\boldsymbol{y}_t, \boldsymbol{y}_t^n, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)) \tag{7}$$

where the mean and variance are aligned with the posterior distribution of the pre-trained DDPM ($\boldsymbol{x}'_{t-1}$). We divide the sampling interval by a scalar τ. In the first stage, i.e. for $t \geq \tau$, we mask the $\boldsymbol{x}'_{t-1}$ to constrain the sample space, ensuring consistent sampling. We obtain:

$$\boldsymbol{x}_{t-1} = (1-\boldsymbol{m}_f) \odot \boldsymbol{x}'_{t-1} + \boldsymbol{m}_f \odot \boldsymbol{y}'_{t-1} \tag{8}$$

where $\boldsymbol{m}_f \odot \boldsymbol{y}'_{t-1}$ is the Fixed Condition Mask (FCM) and $\boldsymbol{x}_{t-1}$ denotes the new posterior distribution obtained from this CM masking. $\boldsymbol{x}'_{t-1}$ are injected with conditional information, and in the subsequent sampling, the denoising prior is utilized to interpolate the $\boldsymbol{x}_{t-1}$. In the second stage, fidelity is further enhanced by incorporating a weighted Randomly Adaptive Condition Mask (RACM). When $t < \tau$, the condition is added as follows:

$$\boldsymbol{x}_{t-1} = (1-\boldsymbol{m}_a) \odot \boldsymbol{x}'_{t-1} + w\boldsymbol{m}_a \odot \boldsymbol{y}'_{t-1} + (1-w)\boldsymbol{m}_a \odot \boldsymbol{x}'_{t-1} \tag{9}$$

where $w = t/\omega$ represents a time-dependent weight controlled by the parameter ω and $\boldsymbol{m}_a \odot \boldsymbol{y}'_{t-1}$ is the RACM. w gradually decreases over time steps, leading to a reduction in the intensity of the conditioning.

Condition Masks

In this section, we will mainly discuss the forms of the CMs and how to design both the strongly constrained FCM and the weakly constrained RACM. With the constraints of different CMs, we can flexibly realize consistent sampling and diversity generation.

Given an LR image $\boldsymbol{I}_L$ of size (h, w), the objective of FSR is to upscale it to the size (H, W) of HR image $\boldsymbol{I}_H$ with a scale factor of H/h. To achieve this, we first upsample $\boldsymbol{I}_L$ using Bicubic interpolation to the size $(H/k, W/k)$, which serves as the base condition $\boldsymbol{I}_L^{bc}$. Here, the parameter k controls sparsity. Subsequently, we design different forms of masks for the two stages of posterior sampling. Specifically, in the first stage, a fixed mask $\boldsymbol{m}_f$ is designed as shown in Fig. 2. The specific definition of $\boldsymbol{m}_f$ is as follows:

$$\boldsymbol{m}_f(i,j) = \begin{cases} 1, if \;\; i,j \bmod k = 0 \\ 0, otherwise \end{cases} \tag{10}$$

where $\boldsymbol{m}_f \in R^{H \times W}$. When k is set to 2, the form of $\boldsymbol{m}_f$ resembles a grid mask with a grid size of 1 pixel. We project $\boldsymbol{I}_L^{bc}$ onto $\boldsymbol{m}_f$ to generate an initial condition, denoted as $\boldsymbol{y}_T$:

$$\boldsymbol{y}_T(i,j) = \begin{cases} \boldsymbol{I}_L^{bc}(\left\lfloor \frac{i}{k} \right\rfloor, \left\lfloor \frac{j}{k} \right\rfloor), if \;\; \boldsymbol{m}_f(i,j) = 1 \\ 0, otherwise \end{cases} \tag{11}$$

where $\boldsymbol{I}_L^{bc} \in R^{\frac{H}{k} \times \frac{W}{k}}$ and $\boldsymbol{y}_T \in R^{H \times W}$. The $\boldsymbol{y}_T$ is initially composed of information from the LR image, then is updated to an intermediate condition $\boldsymbol{y}_t$ by CRT during the sampling phase. In the second stage, we backtrack the intermediate condition $\boldsymbol{y}_t$ and we obtain $\boldsymbol{y}_t^b$:

$$\boldsymbol{y}_t^b(i,j) = \boldsymbol{y}_t(ki, kj) \tag{12}$$

where $\boldsymbol{y}_t \in R^{H \times W}$ and $\boldsymbol{y}_t^b \in R^{\frac{H}{k} \times \frac{W}{k}}$. In Fig. 2, we illustrate the visualization of $\boldsymbol{y}_t^b$. Then, we extract an edge map $\nabla^2 \boldsymbol{y}_t^b$ using the first-order Laplacian operator, which exhibits characteristics where pixel values are smaller in low-frequency regions and larger in high-frequency regions. Next, we normalize the edge map to the range of (0-1) to obtain a probability map $\boldsymbol{p}$:

$$\boldsymbol{p} = \frac{\nabla^2 \boldsymbol{y}_t^b - \min \nabla^2 \boldsymbol{y}_t^b}{\max \nabla^2 \boldsymbol{y}_t^b - \min \nabla^2 \boldsymbol{y}_t^b} \tag{13}$$

Methods	×8 LPIPS↓	×8 FID↓	×8 PSNR↑	×16 LPIPS↓	×16 FID↓	×16 PSNR↑
Dataset	**CelebA1000**					
Bicubic	0.2263	80.16	24.95	0.2989	158.38	21.65
CodeFormer	0.1460	23.12	23.89	0.2162	34.15	20.75
DDNM	0.1468	29.12	24.88	0.2074	32.38	21.56
DDRM	0.1589	36.34	24.81	0.2088	42.48	21.61
SR3	0.1435	56.69	25.48	0.2574	76.45	21.71
ILVR	0.1446	25.53	24.21	0.2001	33.49	20.95
DR2	0.1957	52.52	22.31	0.2243	53.18	20.57
DPS	0.1893	23.48	21.53	0.2303	25.83	19.20
DiffFace	0.1553	24.09	24.49	0.2167	31.80	20.32
PGDiff	0.1439	22.49	24.53	0.2003	27.68	21.24
DPI(Ours)	0.1401	22.86	24.97	0.1927	27.33	21.68

Table 1: Quantitative comparisons on **CelebA1000** testset. **Red** and blue indicates the best and the second best.

For each pixel value $\boldsymbol{p}(i,j)$, we use it as a probability and then project it to $\boldsymbol{m}_f$ to generate a randomly adaptive mask $\boldsymbol{m}_a$ as follows:

$$\boldsymbol{m}_a(i,j) = \begin{cases} 0, & \text{if } \boldsymbol{m}_f(i,j) = 0, \\ 1, & \text{otherwise, with } \boldsymbol{P}\left(\left\lfloor \frac{i}{k} \right\rfloor, \left\lfloor \frac{j}{k} \right\rfloor\right)^s. \end{cases} \quad (14)$$

where $\boldsymbol{P}$ represents probability, $\boldsymbol{m}_a \in R^{H\times W}$ and s adjusts the probability distribution. We refer to the aforementioned process as $Mask_{gen}(\boldsymbol{y}_t, s)$. It's important to note that a different $\boldsymbol{m}_a$ is generated at each time step. We generate the RACM by masking the conditional posterior distribution $\boldsymbol{y}'_{t-1}$ using $\boldsymbol{m}_a$. The RACM exhibits high sparsity and features similar to the edge guidance in ControlNet (Zhang, Rao, and Agrawala 2023). This amplifies the prior space, enabling the generation of more texture details, thereby enhancing the fidelity and diversity of the results.

Condition Corrector

We propose a condition Corrector (CRT), to pull back the conditions to the prior space and establish a reciprocal sampling, as illustrated in Fig. 3. CRT is a small neural network conditioned on the initial condition $\boldsymbol{y}_T$, designed to denoise the posterior distribution of the ground truth (GT) and predict GT conditions. The loss function of CRT is defined as follows:

$$\mathcal{L}_{prior} = ||\boldsymbol{I}_G \odot \boldsymbol{m}_f - CRT(\boldsymbol{m}_f \odot \boldsymbol{I}'_{G_{t-1}}, \boldsymbol{y}_T, t)||_2^2 \quad (15)$$

where $CRT(\cdot)$ stands for the CRT function, $\boldsymbol{I}_G$ represents the GT image and $\boldsymbol{I}'_{G_{t-1}}$ is obtained by Eq. 3. Building upon the assumptions presented in (Wang et al. 2023d), it is assumed that there exists a time step γ such that for $t > \gamma$, the distance between $q(\boldsymbol{x}_t|\boldsymbol{x})$ and $q(\boldsymbol{y}_t|\boldsymbol{y})$ becomes sufficiently small. At this point, we can get the approximate estimate:

$$\begin{aligned} \boldsymbol{y}_{t-1} &= CRT(\boldsymbol{m}_f \odot \boldsymbol{y}'_{t-1}, \boldsymbol{y}_T, t) \\ &\approx CRT(\boldsymbol{m}_f \odot \boldsymbol{I}'_{G_{t-1}}, \boldsymbol{y}_T, t) \end{aligned} \quad (16)$$

However, the CRT is essentially a discriminative model that favors the average output and the gap between $q(\boldsymbol{x}_t|\boldsymbol{x})$ and $q(\boldsymbol{y}_t|\boldsymbol{y})$ progressively increases for $t < \gamma$. For this reason, we need to adapt the CRT to the intermediate conditions as well as mitigate the gap. Ultimately, the objective function for training the CRT is as follows:

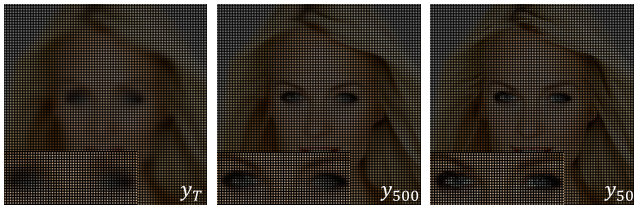


Figure 3: **Condition Refinement**. We present the initial condition $\boldsymbol{y}_T$ along with the refined conditions in the intermediate stages ($\boldsymbol{y}_{500}, \boldsymbol{y}_{50}$). Please zoom in for a better viewing.

$$\mathcal{L}_{gap} = ||\boldsymbol{I}_G \odot \boldsymbol{m}_f - CRT(\boldsymbol{m}_f \odot \hat{\boldsymbol{x}}_{crt}, \boldsymbol{y}_T, t)||_2^2 \quad (17)$$

$$\mathcal{L}_{crt} = \Omega\mathcal{L}_{prior} + (1-\Omega)\mathcal{L}_{gap} \quad (18)$$

where $\hat{\boldsymbol{x}}_{crt}$ represents the intermediate condition output of CRT and Ω is a weight that decreases over t. The structure of CRT can be found in the Appendix. In the first stage, due to the sufficiently heavy input noise, CRT can only extract information from $\boldsymbol{y}_T$. The output of CRT during this stage is smooth. In the second stage, as the noise decreases and RACM amplifies the prior space, CRT can utilize prior information to update conditions.

Experiments

Experimental Setup

We employ a pre-trained DDPM from DPS (Chung et al. 2022). The model is pre-trained on 49k face images from FFHQ (Karras, Laine, and Aila 2019) at a resolution of 256×256. For evaluation, we utilize synthetic datasets FFHQ1000 and CelebA1000 (Liu et al. 2015), along with real-world datasets LFW (Huang et al. 2008), WebPhoto (Wang et al. 2021), and WIDER (Yang et al. 2016), serving as our testsets. Following the settings of the latest diffusion-based methods (Wang et al. 2023d; Chung et al. 2022), we compare them on synthetic datasets at scales of $\times 4$, $\times 8$, and $\times 16$. Additionally, for each of these three scales, the parameters (τ, s, ω) is set to (100, 1.4, 500), (300, 1.2, 750), and (500, 1, 1000) respectively. For real-world datasets, we adhere to the experimental settings in CodeFormer (Zhou et al. 2022), with fixed parameters set to (500, 1, 1000). The sparsity parameter k for CMs is set to 2 for all experiments.

Comparison with Previous Work

Synthetic Datasets: Diffusion-based FSR works typically use Bicubic settings of different scales as benchmarks. For this purpose, we train CRT at different scales to compare them (Wang, Yu, and Zhang 2022; Kawar et al. 2022; Saharia et al. 2023; Choi et al. 2021; Wang et al. 2023d; Chung et al. 2022; Kim et al. 2022; Yang et al. 2024). Table 1 demonstrates the outstanding performance of DPI across all scales, showcasing SOTA perceptual metrics compared to other diffusion-based methods. Fig. 4 illustrates DPI's superior performance in visual consistency.

Figure 4: From top to bottom are the FSR results of DPI and DDPM-based methods on **CelebA1000** testset at $\times 8$, and $\times 16$ scales, respectively. DPI_{20} refers that the DDIM (Song, Meng, and Ermon 2020) algorithm samples only 20 steps. Please zoom in for best view.

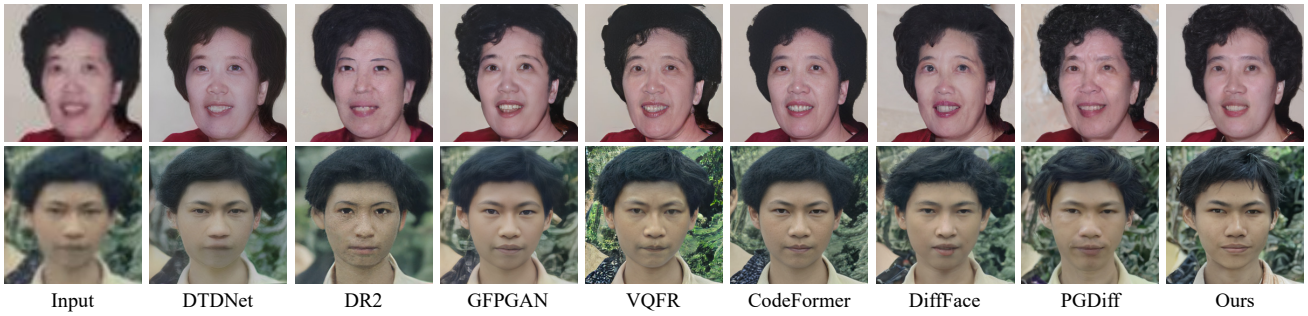


Figure 5: Qualitative comparisons on **WebPhoto** dataset. Please zoom in for best view.

Real-world Datasets: We adhere to the general degradation model (Zhou et al. 2022) represented as follows:

$$\boldsymbol{I}_L = \{[(\boldsymbol{I}_H \otimes k_{s,\sigma})_{\downarrow r} + n_\delta]_{\text{JEPG}_q}\}_{\uparrow r} \tag{19}$$

where $k_{s,\sigma}$ denotes a Gaussian blur kernel with a kernel size of s, n_δ represents Gaussian noise, JPEG_q signifies compression with quality q, and r denotes the sampling scale. CRT is trained on this degradation model to address real-world issues. Our DPI achieves SOTA performance in no-reference metrics on three real-world datasets, as shown in Table 2.

Face Recognition Results: We employ the open-source face recognition framework, DeepFace (Serengil and Ozpinar 2020) with a threshold set at 0.4 to compare the accuracy (ACC) and consistency (CS) of face recognition between GT and SR images. Specifically, we feed the SR results and GT images into the DeepFace model to determine if they correspond to the same person. ACC represents the proportion of accurate recognition in the testset, while CS represents the distance between SR and GT features. Table 3 presents the experimental results on the CelebA1000 testset for different upscaling factors. Especially in the $\times 16$ task, we are far ahead of other methods. The results of face recognition proves that DPI has higher consistency.

Severe Degradation Blind FSR: For severe blind FSR, Eq. 19 is applied to sample parameters r, s, σ, q, δ from $\{8:16\}$, $\{1:17\}$, $\{3:20\}$, $\{40:50\}$, $\{30:90\}$ to construct the testset. The hyperparameters of DPI are consistent with real-world tasks. As illustrated in Fig. 6, it can be seen that DPI can ensure the consistency of the contour even under severe degradation, which reflects the strong robustness.

Figure 6: Qualitative comparison on heavy degradation.

Ablation Studies

In this section, we conduct ablation experiments to analyze the effects of strong and weak constraints, masking strategies, and hyperparameters on consistency and diversity. All ablation studies are performed on the CelebA1000 testset at a scale of $\times 8$. Please refer to the visual ablation studies in Appendix for further insight.

Effectiveness of Condition Correction and Refinement: The conditions undergo pixel correction and iterative refinement to provide accurate guidance. Initially, the degraded pixels' impact on performance is analyzed by not using the CRT to repair the conditions. Subsequently, the corrected condition are used solely as conditions at each time step without participating in the iterative refinement. As shown in the the second and third columns of Table 4, degraded pixels lead to

Datasets / Degradation / Methods	LFW *mild* FID↓	IS↑	MUSIQ↑	CLIPIQA↑	WebPhoto *medium* FID↓	IS↑	MUSIQ↑	CLIPIQA↑	WIDER *heavy* FID↓	IS↑	MUSIQ↑	CLIPIQA↑
Input	138.41	2.85	29.41	0.4432	163.24	3.28	21.04	0.3351	192.17	2.88	15.72	0.2976
DFDNet	87.83	3.52	65.92	0.7313	125.22	3.60	61.53	0.6526	135.78	3.27	56.78	0.6249
GFPGAN	/	/	/	/	133.73	3.45	63.78	0.6742	99.51	3.48	63.07	0.6611
VQFR	69.02	3.53	64.44	0.6964	92.81	3.46	63.92	0.6933	/	/	/	/
CodeFormer	68.72	3.56	67.30	0.7034	85.93	3.56	65.87	0.7050	51.61	3.34	64.48	0.7132
DR2	80.94	3.45	59.23	0.6095	99.27	3.53	54.53	0.5469	71.85	3.01	55.06	0.5739
DiffFace	66.40	3.55	63.57	0.6813	89.81	3.58	59.61	0.6624	50.36	3.41	58.56	0.6764
PGDiff	62.31	3.60	65.97	0.7137	85.83	3.61	62.87	0.6950	47.63	3.47	64.56	0.7137
DPI	65.91	3.64	66.13	0.7098	81.77	3.67	64.65	0.7011	49.79	3.50	64.90	0.7150

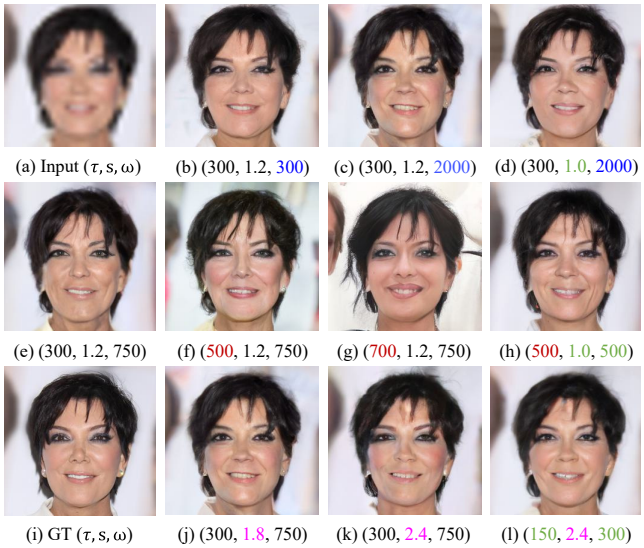
Table 2: Quantitative comparisons on the **real-world** datasets. Red and blue indicate the best and the second best.

Methods	×4 CS↓	×4 ACC↑	×8 CS↓	×8 ACC↑	×16 CS↓	×16 ACC↑
Bicubic	0.1061	90.2	0.3276	78.3	0.6475	60.0
CodeFormer	0.0991	94.8	0.2253	92.2	0.3950	53.5
DDRM	0.1466	92.8	0.2781	87.7	0.4155	47.5
SR3	0.0831	93.7	0.2264	91.2	0.4015	49.9
ILVR	0.0898	91.5	0.2363	90.8	0.3998	53.3
DR2	0.2455	89.4	0.3264	76.9	0.3955	52.8
DPS	0.2470	92.2	0.3603	66.6	0.4550	31.8
PGDiff	0.0870	93.1	0.2103	90.8	0.4064	59.8
DPI	0.0807	95.3	0.2100	92.5	0.3449	72.8

Table 3: Qualitative comparisons of **face recognition** accuracy and consistency on CelebA1000 testset. Red and blue indicate the best and the second best performance.

w/ CRT	✓		✓	✓	✓	✓
w/ Refine	✓	✓		✓	✓	✓
w/ Stage II	✓	✓	✓		✓	✓
w/ Weight	✓	✓	✓			✓
w/ $k = 2$	✓	✓	✓	✓	✓	
SSIM↑	0.6881	0.6933	0.6873	**0.7278**	0.6362	0.6335
LPIPS↓	**0.1401**	0.2242	0.1507	0.1486	0.1679	0.1592
FID↓	**22.86**	73.01	23.65	53.60	33.78	28.09

Table 4: Ablation studies



(a) Input (τ, s, ω) (b) (300, 1.2, 300) (c) (300, 1.2, 2000) (d) (300, 1.0, 2000)
(e) (300, 1.2, 750) (f) (500, 1.2, 750) (g) (700, 1.2, 750) (h) (500, 1.0, 500)
(i) GT (τ, s, ω) (j) (300, 1.8, 750) (k) (300, 2.4, 750) (l) (150, 2.4, 300)

Figure 7: Visual comparisons of the impact of hyperparameters on consistency and diversity.

poor metrics and significantly affect image quality, while the absence of refinement only marginally impacts performance.

Balancing Consistency and Diversity: Our proposed method allows the user to tune the consistency and diversity of FSR by simply adjusting the scalar values, as shown in Fig. 7. In our default settings (Fig. 7 (e)), the parameters (τ, s, ω) are set to (300, 1.2, 750). We will elaborate on the impact of these scalars on the results: 1) By adjusting τ, we can flexibly partition the sampling stages. Increasing the range of the second stage significantly enhances diversity, while conversely improving consistency. As seen in the fourth column of Table 4, removing the second stage ($\tau = 0$) initially yields the best SSIM, but the FID metric noticeably increases, indicating enhanced consistency but reduced fidelity. The second row of Fig. 7 (e-g) clearly demonstrates that increasing τ leads to greater diversity. 2) Under the premise of ensuring consistency (Fig. 7 (e, j, k)), we fine-tune the diversity by changing the sparsity. 3) In Eq. 9, ω is used to reduce conditional intensity for better utilization of prior information. A larger ω implies a smaller weight attached to the condition. From the comparison in Fig. 7 (b) and (c), reducing conditions intensity enhances fidelity and diversity, such as details in hair. The ablation study in the sixth column of Table 4 also confirms this. Furthermore, the comparison of Fig. 7 (f) and Fig. 7 (h) reflects the control of s and ω over consistency and diversity.

Conclusion

We propose DPI, which effectively leverages the prior knowledge of pre-trained models for face super-resolution. By implementing a masking strategy tailored to facial features, we achieve a balance between consistency and diversity during the sampling process. Additionally, we introduce CRT to establish a reciprocal sampling process, where samples and conditions are iteratively refined. Extensive experiments demonstrate the superior performance of DPI and its ability to ensure consistency.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China, under Grant (62302309, 62171248), Shenzhen Science and Technology Program (JCYJ20220818101014030, JCYJ20220818101012025).

References

Bartlett, M. S. 1978. *An introduction to stochastic processes: with special reference to methods and applications*. CUP Archive.

Blau, Y.; and Michaeli, T. 2018. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6228–6237.

Bredell, G.; Flouris, K.; Chaitanya, K.; Erdil, E.; and Konukoglu, E. 2023. Explicitly Minimizing the Blur Error of Variational Autoencoders. *arXiv preprint arXiv:2304.05939*.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

Chan, K. C.; Xu, X.; Wang, X.; Gu, J.; and Loy, C. C. 2022. GLEAN: Generative latent bank for image super-resolution and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3154–3168.

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*, 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2492–2501.

Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14367–14376.

Chung, H.; Kim, J.; Mccann, M. T.; Klasky, M. L.; and Ye, J. C. 2022. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *The Eleventh International Conference on Learning Representations*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Fei, B.; Lyu, Z.; Pan, L.; Zhang, J.; Yang, W.; Luo, T.; Zhang, B.; and Dai, B. 2023. Generative Diffusion Prior for Unified Image Restoration and Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9935–9946.

Gao, G.; Xu, Z.; Li, J.; Yang, J.; Zeng, T.; and Qi, G.-J. 2023. Ctcnet: A cnn-transformer cooperation network for face image super-resolution. *IEEE Transactions on Image Processing*, 32: 1978–1991.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, G. B.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

Jiang, J.; Wang, C.; Liu, X.; and Ma, J. 2021. Deep learning-based face super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 55(1): 1–36.

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*, 4401–4410.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35: 23593–23606.

Kim, K.; Kim, Y.; Cho, S.; Seo, J.; Nam, J.; Lee, K.; Kim, S.; and Lee, K. 2022. Diffface: Diffusion-based face swapping with facial guidance. *arXiv preprint arXiv:2212.13344*.

Kirichenko, P.; Izmailov, P.; and Wilson, A. G. 2020. Why normalizing flows fail to detect out-of-distribution data. *Advances in neural information processing systems*, 33: 20578–20589.

Leng, J.; and Wang, Y. 2022. RCNet: Recurrent collaboration network guided by facial priors for face super-resolution. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 01–06. IEEE.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479: 47–59.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Liu, Z.-S.; Siu, W.-C.; and Wang, L.-W. 2021. Variational autoencoder for reference based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 516–525.

Lu, T.; Wang, Y.; Zhang, Y.; Jiang, J.; Wang, Z.; and Xiong, Z. 2022. Rethinking prior-guided face super-resolution: a new paradigm with facial component prior. *IEEE Transactions on Neural Networks and Learning Systems*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Lugmayr, A.; Danelljan, M.; Van Gool, L.; and Timofte, R. 2020. Srflow: Learning the super-resolution space with normalizing flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, 715–732. Springer.

Mei, K.; Nair, N. G.; and Patel, V. M. 2022. Bi-Noising Diffusion: Towards Conditional Diffusion Models with Generative Restoration Priors. *arXiv preprint arXiv:2212.07352*.

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*, 2437–2445.

Moser, B.; Frolov, S.; Raue, F.; Palacio, S.; and Dengel, A. 2023. DWA: Differential Wavelet Amplifier for Image Super-Resolution. *arXiv preprint arXiv:2304.01994*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2023. Image Super-Resolution via Iterative Refinement. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(04): 4713–4726.

Sajjadi, M. S.; Scholkopf, B.; and Hirsch, M. 2017. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE international conference on computer vision*, 4491–4500.

Serengil, S. I.; and Ozpinar, A. 2020. Lightface: A hybrid deep face recognition framework. In *2020 innovations in intelligent systems and applications conference (ASYU)*, 1–5. IEEE.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Wang, C.; Jiang, J.; Zhong, Z.; and Liu, X. 2022a. Propagating facial prior knowledge for multitask learning in face super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7317–7331.

Wang, C.; Jiang, J.; Zhong, Z.; and Liu, X. 2023a. Spatial-Frequency Mutual Learning for Face Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22356–22366.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2023b. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *arXiv preprint arXiv:2305.07015*.

Wang, X.; Li, Y.; Zhang, H.; and Shan, Y. 2021. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9168–9178.

Wang, Y.; Hu, Y.; Yu, J.; and Zhang, J. 2023c. Gan prior based null-space learning for consistent super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2724–2732.

Wang, Y.; Yu, J.; and Zhang, J. 2022. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*.

Wang, Z.; Zhang, J.; Chen, R.; Wang, W.; and Luo, P. 2022b. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17512–17521.

Wang, Z.; Zhang, Z.; Zhang, X.; Zheng, H.; Zhou, M.; Zhang, Y.; and Wang, Y. 2023d. DR2: Diffusion-based Robust Degradation Remover for Blind Face Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1704–1713.

Wei, M.; Shen, Y.; Wang, Y.; Xie, H.; and Wang, F. L. 2023. RainDiffusion: When Unsupervised Learning Meets Diffusion Models for Real-world Image Deraining. *arXiv preprint arXiv:2301.09430*.

Whang, J.; Delbracio, M.; Talebi, H.; Saharia, C.; Dimakis, A. G.; and Milanfar, P. 2022. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16293–16303.

Yang, P.; Zhou, S.; Tao, Q.; and Loy, C. C. 2024. PGDiff: Guiding diffusion models for versatile face restoration via partial guidance. *Advances in Neural Information Processing Systems*, 36.

Yang, S.; Luo, P.; Loy, C.-C.; and Tang, X. 2016. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5525–5533.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhou, S.; Chan, K.; Li, C.; and Loy, C. C. 2022. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35: 30599–30611.

Zou, W. W.; and Yuen, P. C. 2011. Very low resolution face recognition problem. *IEEE Transactions on image processing*, 21(1): 327–340.