Accelerating Diffusion Models for Inverse Problems through Shortcut Sampling

Gongye Liu , Haoze Sun , Jiayi Li , Fei Yin , Yujiu Yang
Tsinghua University
{lgy22, shz22, lijy20, yinf20}@mails.tsinghua.edu.cn, yang.yujiu@sz.tsinghua.edu.cn

Abstract

Diffusion models have recently demonstrated an impressive ability to address inverse problems in an unsupervised manner. While existing methods primarily focus on modifying the posterior sampling process, the potential of the forward process remains largely unexplored. In this work, we propose Shortcut Sampling for Diffusion(SSD), a novel approach for solving inverse problems in a zero-shot manner. Instead of initiating from random noise, the core concept of SSD is to find a specific transitional state that bridges the measurement image y and the restored image x. By utilizing the shortcut path of "input - transitional state - output", SSD can achieve precise restoration with fewer steps. Experimentally, we demonstrate SSD's effectiveness on multiple representative IR tasks. Our method achieves competitive results with only 30 NFEs compared to state-of-the-art zero-shot methods(100 NFEs) and outperforms them with 100 NFEs in certain tasks. Code is available at https://github.com/GongyeLiu/SSD.

1 Introduction

Inverse problem is a classic problem in the field of machine learning. Given a low-quality (LQ) input measurement image y and a degradation operator H, inverse problems aim to restore the original high-quality (HQ) image x from $y = Hx + n$. Many image restoration tasks, including super-resolution [Haris *et al.*, 2018; Wang *et al.*, 2018], colorization [Larsson *et al.*, 2016], inpainting [Yeh *et al.*, 2017], deblurring [Guo *et al.*, 2019; Zhang *et al.*, 2018a] and denoising [Wang *et al.*, 2022], can be considered as applications of solving inverse problems. In general, the restored image should exhibit two critical attributes: *Realism* and *Faithfulness*. The former indicates that the restored image should be of high-quality and photo-realistic, while the latter denotes that the restored image should be consistent with the input image in the degenerate subspace.

Recently, diffusion models [Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020; Song *et al.*, 2021c] have demonstrated phenomenal performance in generation tasks [Rombach *et al.*, 2022; Dhariwal and Nichol, 2021; Xiao *et al.*, 2022]. Due to their powerful capability in modeling complex distributions, recent methods [Kawar *et al.*, 2021; Kawar *et al.*, 2022; Chung *et al.*, 2022b; Chung *et al.*, 2023; Wang *et al.*, 2023; Lugmayr *et al.*, 2022; Song *et al.*, 2021b] have sought to utilize pre-trained diffusion models for solving inverse problems in an unsupervised manner. These methods leverage the generative priors to enhance *realism*, and enforce additional consistency constraints to ensure *faithfulness*. Specifically, diffusion models have predefined a forward process expressed as $p(x_t|x_{t-1})$ and a generation process expressed as $p(x_{t-1}|x_t)$. Existing methods are primarily concerned with the generation process, altering the posterior sampling to a conditional sampling $p(x_{t-1}|x_t, y)$ based on the LQ image y.

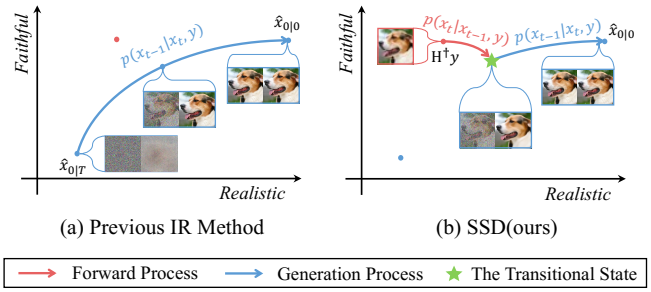

Figure 1: **Visual Depiction of "Shortcut Sampling". (a)** ***Previous IR methods*** initiate from random noise x_T, taking unnecessary steps to generate the layout and structure; **(b)** ***SSD(ours)*** modifies the forward process to obtain a better transitional state, employing a shortcut-sampling path of *"Input-Transitional State-Target"* to restore images with fewer steps.

Despite the successful application of existing methods in solving various inverse problems, their relatively slow sampling speed is a major drawback. These methods overlook the importance of modifying the forward process to achieve an improved initial state, simply initiating from random noise $x_T \sim N(0, I)$. However, as depicted in Fig. 1, since the initial state of pure noise x_T is considerably distant from the target HQ images x_0, prior methods have to travel through a long journey of sampling, typically requiring at least 100-250 neural function evaluations (NFEs), to generate the overall layout, structure, appearance and detailed texture of the restored image, and finally achieve a satisfactory result.

In this work, we propose **Shortcut Sampling for Diffu-**

sion (SSD), a novel approach for solving inverse problems in a zero-shot manner. The primary concept behind SSD is to find a specific transitional state that bridges the gap between the input image y and the target restoration x. For convenience, we use "$\mathscr{E}$" to refer to this specific transitional state. As depicted in Fig. 1, by employing a shortcut path of *"Input - $\mathscr{E}$ - Target"*($H^\dagger y \rightarrow x_t \rightarrow x_0$) instead of the previous *"Noise-Target"*($x_T \rightarrow x_0$), SSD enables precise and fast restoration.

For the forward process denoted as $p(x_t|x_{t-1}, y)$, we observe that the original forward process erodes information from the input image, yielding realistic yet unfaithful results. Conversely, an alternative, DDIM Inversion[Song *et al.*, 2021a], widely adopted for image editing[Hertz *et al.*, 2022; Tumanyan *et al.*, 2023], tends to produce unrealistic outcomes. To address this dilemma, we introduce Distortion Adaptive Inversion (DA Inversion). By incorporating a controllable random disturbance at each forward step, DA Inversion is capable of deriving $\mathscr{E}$ that adheres to the predetermined noise distribution while preserving the majority of the input image's information.

For the generation process denoted as $p(x_{t-1}|x_t, y)$, we utilize generation priors to produce extra details and texture, and introduce the back projection technique[Tirer and Giryes, 2018; Wang *et al.*, 2023] as additional consistency constraints. In particular, we add a projection step after each denoising step to project the coarse restored image onto the degenerate subspace, obtaining a revised version of the restored image that aligns with the input image in the degenerate subspace. We further propose SSD$^+$ to extend applicability to scenarios with unknown noise and more intricate degradation conditions.

To validate the effectiveness of SSD, we conduct experiments across various inverse problems, including super-resolution, colorization, inpainting, and deblurring on CelebA and ImageNet. Experiments demonstrate that SSD achieves competitive results in comparison to state-of-the-art zero-shot methods (with 100 NFEs), even though it employs only 30 NFEs. Additionally, we observe that SSD, when operated with 100 NFEs, can surpass state-of-the-art methods in certain IR tasks.

2 Related Works

2.1 Diffusion Models

Denoising Diffusion Probabilistic Models. Diffusion models [Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020; Song *et al.*, 2021c] are a family of generative models designed to model complex probability distributions of high-dimensional data. Denoising Diffusion Probabilistic Models(DDPM) [Ho *et al.*, 2020] comprise both a forward process and a generation process. In the forward process, an image x_0 is transformed into Gaussian noise $x_T \sim \mathcal{N}(0, 1)$ by gradually adding random noise over T steps. We can describe each step in the forward process as:

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}z_t, z_t \sim \mathcal{N}(0, 1) \tag{1}$$

where $[x_t]_{t=0}^T$ is the noisy image at time-step t, $[\beta_t]_{t=0}^T$ is the predefined variance schedule, and $[z_t]_{t=0}^T$ is the random gauss noise added at time-step t. Using reparameterization tricks [Kingma and Welling, 2013], The resulting noisy image x_t can be expressed as:

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon, \epsilon \sim \mathcal{N}(0, 1) \tag{2}$$

where $\alpha_t = \prod_{i=1}^t (1-\beta_i)$ and ϵ is the reparameterized noise. The generation process transforms gaussian noise x_T to image x_0, the transition from x_t to x_{t-1} can be expressed as:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)) + \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t \tag{3}$$

and $\epsilon_\theta(x_t, t)$ is a neural network trained to predict the noise ϵ from noisy image x_t at time-step t. The noise approximation model $\epsilon_\theta(x_t, t)$ can be trained by minimize the following objective:

$$\min_\theta \mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim N(0,I)} \|\epsilon - \epsilon_\theta(x_t, t)\|_2^2 \tag{4}$$

Denoising Diffusion Implicit Models. Meanwhile, [Song *et al.*, 2021a] generalize DDPM via a non-Markov diffusion process that shares the same training objective, whose generation process is outlined as follows:

$$x_{t-1} = \sqrt{\alpha_{t-1}}f_\theta(x_t, t) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\epsilon_\theta(x_t, t) + \sigma_t\epsilon_t \tag{5}$$

where $f_\theta(x_t, t)$ is the prediction of x_0 at time-step t:

$$f_\theta(x_t, t) = \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \tag{6}$$

When $\sigma_t = 0$, DDIM samples images through a deterministic generation process, which allows for high-quality sampling with fewer steps:

$$x_{t-1} = \sqrt{\alpha_{t-1}}f_\theta(x_t, t) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(x_t, t) \tag{7}$$

2.2 Solving Inverse Problems in a Zero-shot Way

A general inverse problem aims to restore a high-quality image x from a known degradation operator H and the degraded measurement y with random additional noise n:

$$y = Hx + n \tag{8}$$

Some methods investigate leveraging the generative priors of pre-trained generative models to restore degraded images in a zero-shot way. GAN Inversion aims to find the closest latent vector in the GAN space for an input image. Utilizing the GAN Inversion technique, existing methods[Ulyanov *et al.*, 2018; Pan *et al.*, 2021; Menon *et al.*, 2020] have exhibited remarkable effectiveness in tackling inverse problems.

Compared to GAN, diffusion models provide a forward process that enables the direct acquisition of latent vectors within the Gaussian noise space. By performing generation processes and using consistency constraints at each step, diffusion models can be applied to various IR problems. DDRM[Kawar *et al.*, 2022] applies SVD to decompose the degradation operators and perform diffusion in its spectral space for various IR tasks. DDNM[Wang *et al.*, 2023] uses range-null space decomposition to decompose the restored image as a null-space part and a range-space part,

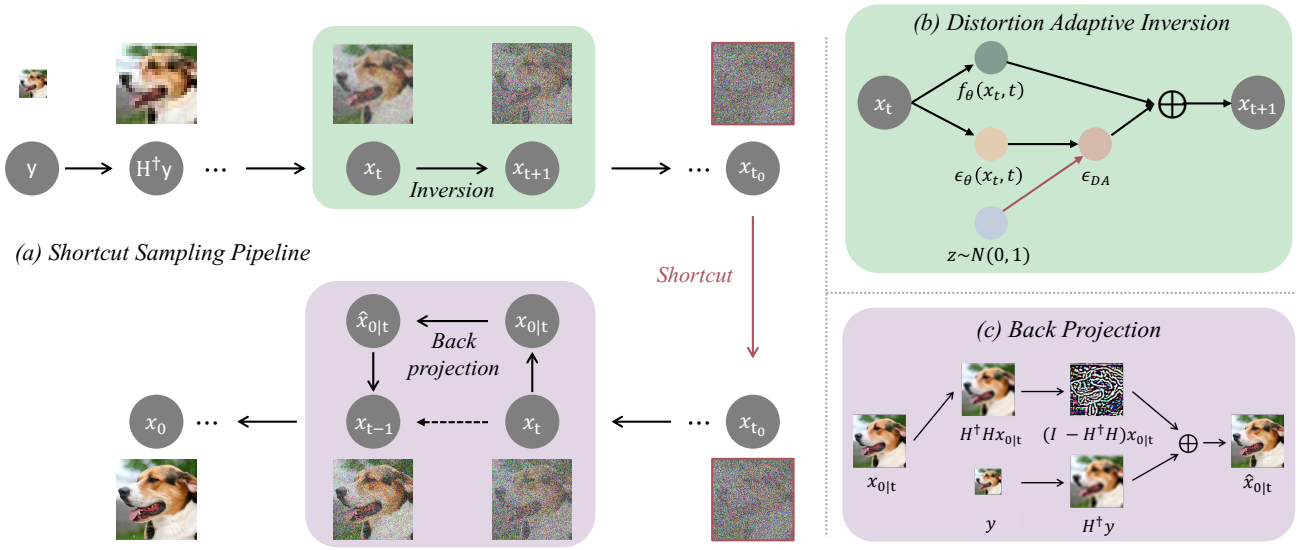

Figure 2: **Overview of the proposed SSD.** We propose a shortcut sampling pipeline, instead of starting from random noise and spending lots of steps to generate the overall layout and structure, we use Distortion Adaptive Inversion to obtain the transitional state, a noisy image that contains most structure information of the input image. Then during the generation process, we iteratively perform the denoising step and the back projection step to generate images with detailed texture while keeping the restored images consistent with the input images.

they keep the range-space part unchanged to force consistency, and obtain the null-space part through iterative refinement. Meanwhile, inspired by guided diffusion[Dhariwal and Nichol, 2021; Liu *et al.*, 2023], some recently developed methods[Chung *et al.*, 2022b; Chung *et al.*, 2023; Song *et al.*, 2022; Mardani *et al.*, 2023; Pokle *et al.*, 2023] adopted gradient-based guidance to generate faithful restoration. MCG[Chung *et al.*, 2022b] applies a gradient-based measurement consistency step at each denoising step to achieve image restoration. These methods typically perform better in terms of perceptual quality, while requiring additional inference consumes compared to non-gradient methods.

3 Method

3.1 Shortcut Sampling

As discussed above, diffusion models comprise two processes: a forward process denoted as $p(x_t|x_{t-1})$, which progressively adds noise to the image until complete Gaussian noise; and a generation process denoted as $p(x_{t-1}|x_t)$, which generates realistic images through iteratively denoising. Previous methods mainly focus on modifying the posterior sampling process to $p(x_{t-1}|x_t, y)$ during the generation process, while ignoring the utilization of the forward process. Instead, these methods typically sample x_T directly from the Gaussian prior as the initial state.

In this work, we propose Shortcut Sampling for Diffusion (SSD), a novel pipeline for solving inverse problems in a zero-shot manner. Different from previous methods that initiate from pure noise, SSD enhances the forward process to obtain an intermediate state $\mathscr{E}$, which serves as a bridge between the measurement image y and the restored image x. Throughout the shortcut sample path of "input-$\mathscr{E}$-output", SSD can achieve efficient and satisfactory restoration results.

For convenience, we denote the transition from the measurement image y to $\mathscr{E}$ as the "inversion process"; and the transition from $\mathscr{E}$ to the restored image x_0 as the "generation process". Given a LQ image y and corresponding degraded operator H, we start from $H^\dagger y$ and apply Distortion Adaptive Inversion (DA Inversion) to derive $\mathscr{E}$ in the inversion process(Sec. 3.2). Subsequently, during the generation process, we iteratively perform the denoising step and the back projection step to generate both faithful and realistic results(Sec. 3.3). Further, due to SSD relies on an accurate estimation of degraded operators to exhibit high performance, we proposed an enhanced version called SSD$^+$ that makes SSD suitable for noisy situations or inaccurate estimation of H.(Sec. 3.4). The overall pipeline is illustrated in Fig. 2.

3.2 Distortion Adaptive Inversion.

We expect to obtain the transitional state by enhancing the forward process. As previously discussed, the transitional state $\mathscr{E}$ should satisfy the following criterias:

Criteria (i): The transitional state should contain information from the input image;

Criteria (ii): The transitional state should retain the capacity for generating a high-quality image.

Why DDIM Inversion Cannot Work Well. To satisfy *Criteria (i)*, a naive approach is to apply DDIM Inversion, capable of providing a deterministic mapping from the input image y to a noisy transitional state[Hertz *et al.*, 2022; Tumanyan *et al.*, 2023; Zhang *et al.*, 2024]. Given the deterministic nature of the DDIM generation process, we can establish the DDIM Inversion process by reversing Eq. (7) in the following manner:

$$x_{t+1} = \sqrt{\alpha_{t+1}} f_\theta(x_t, t) + \sqrt{1-\alpha_{t+1}} \epsilon_\theta(x_t, t) \quad (9)$$

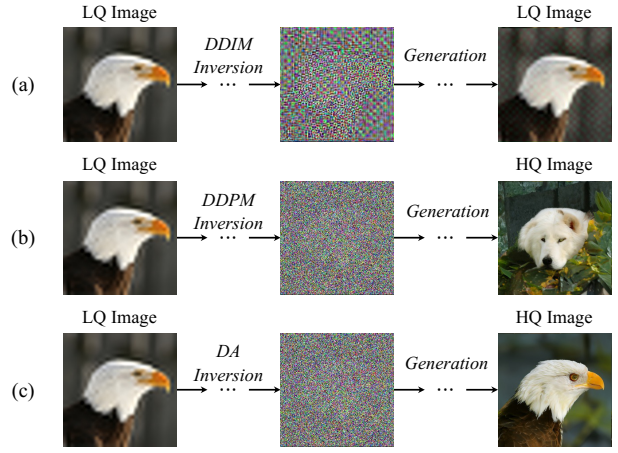

Figure 3: **Comparison of reconstruction results between different Inversion Methods. (a)** *DDIM Inversion* produces faithful but unrealistic results. **(b)** *DDPM Inversion* produces realistic but unfaithful results **(c)** *Distortion Adaptive Inversion(ours)* produces both realistic and faithful results

The transitional state obtained through DDIM Inversion preserves most information of the input images since we can reconstruct it by iteratively executing Eq. (7). However, as depicted in Fig. 3 (a), the application of $\mathscr{E}$ in the generation process produces faithful but unrealistic results, thus violating *Criteria (ii)*.

We attribute the failure to the observation that, the obtained $\mathscr{E}$ deviates from the predefined noise distribution. More specifically, given a low-quality input image y, the predicted noise $\{\epsilon_\theta(x_t, t)\}$ during DDIM Inversion process deviates from the standard normal distribution, resulting in the deviation of $\mathscr{E}$ from the predefined noise distribution. During the generation process, the pre-trained model receives out-of-domain input distributions, thereby generating unrealistic results. We summarize this observation as follows, more details are available in Appendix B.

Assumption 1. *Diffusion Models rely on in-domain noise distribution to generate high-quality images. When facing low-quality input images, the distribution of predicted noise $\epsilon_\theta(x_t, t)$ during the DDIM Inversion process exhibits a greater deviation from the standard normal distribution.*

Why Original Forward Process Cannot Work Well. Another extreme scenario is the original forward process, which can be regarded as a special inversion technique termed DDPM Inversion. In DDPM Inversion, the predicted noise is replaced with randomly sampled noise from Gaussian distribution. We can redefine the forward process in Eq. 1 in a similar form of DDIM Inversion in Eq. 9:

$$\begin{aligned} x_{t+1} = &\sqrt{\alpha_{t+1}} f_\theta(x_t, t) + \sqrt{1 - \alpha_{t+1} - \beta_{t+1}} \epsilon_\theta(x_t, t) \\ &+ \sqrt{\beta_{t+1}} z, \qquad z \sim \mathcal{N}(0, 1) \end{aligned} \tag{10}$$

As shown in Fig. 3(b), DDPM Inversion converts the input image y into pure noise, thereby violating **Criterion (i)** and producing results that are realistic yet lack faithfulness.

Distortion Adaptive Inversion. Since a deterministic inversion process like DDIM Inversion produces unrealistic results, while a stochastic inversion process like DDPM Inversion yields unfaithful results. To resolve this dilemma, we propose a novel inversion approach called Distortion Adaptive Inversion(DA Inversion). The definition of DA Inversion is stated as follows:

Definition 1 (Distortion Adaptive Inversion)**.** *We define the iterative process of DA Inversion as:*

$$\begin{aligned} x_{t+1} = &\sqrt{\alpha_{t+1}} f_\theta(x_t, t) + \sqrt{1 - \alpha_{t+1} - \eta\beta_{t+1}} \epsilon_\theta(x_t, t) \\ &+ \sqrt{\eta\beta_{t+1}} z \qquad z \sim \mathcal{N}(0, 1) \end{aligned} \tag{11}$$

where η control the proportion of random disturbances and $0 < \eta < 1$.

For ease of exposition, the predicted noise in DA Inversion at each timestep can be rephrased as

$$\begin{aligned} \epsilon_{DA} = &\frac{1}{\sqrt{1 - \alpha_{t+1}}} (\sqrt{1 - \alpha_{t+1} - \eta\beta_{t+1}} \epsilon_\theta(x_t, t) \\ &+ \sqrt{\eta\beta_{t+1}} z), \quad z \sim \mathcal{N}(0, 1) \end{aligned} \tag{12}$$

By adding controllable random perturbations in each inverse step, DA Inversion has the capability to generate high-quality images while preserving the essential information including layout and structure, which is shown in Fig. 3 (c).

For **Criterion (i)**, since the random perturbation only replaces a portion of the predicted noise, the $\mathscr{E}$ obtained through DA Inversion actually preserves a substantial amount of information from the input image. For **Criterion (ii)**, we have verified that incorporating random disturbances can bring the predicted noise closer to $\mathcal{N}(0, 1)$. Proofs are available in Appendix A.

Theorem 1. *Assuming $\epsilon_\theta(x_t, t) \sim \mathcal{N}(\mu, \sigma^2)$, We have:*

$$\epsilon_{DA} \sim \mathcal{N}(\mu_{\epsilon_{DA}}, \sigma^2_{\epsilon_{DA}}) \tag{13}$$

$$\mu_{\epsilon_{DA}} = \frac{\sqrt{1 - \alpha_{t+1} - \eta\beta_{t+1}}}{\sqrt{1 - \alpha_{t+1}}} \mu \tag{14}$$

$$\sigma^2_{\epsilon_{DA}} = 1 + \frac{1 - \alpha_{t+1} - \eta\beta_{t+1}}{1 - \alpha_{t+1}} (\sigma^2 - 1) \tag{15}$$

thus:

$$\begin{aligned} \|\mu_{\epsilon_{DA}}\| &< \|\mu\| \\ \|\sigma^2_{\epsilon_{DA}} - 1\| &< \|\sigma^2 - 1\| \end{aligned} \tag{16}$$

which indicates that after adding random disturbance, ϵ_{DA} becomes closer to $\mathcal{N}(0, 1)$.

In practice, inspired by [Meng *et al.*, 2021; Kim *et al.*, 2022; Chung *et al.*, 2022a], rather than performing the inversion process until the last time-step T, we find that we can achieve acceleration by performing until time-step $t_0 < T$.

3.3 Back Projection

Although $\mathscr{E}$ obtained from DA Inversion carries information about the input image, and the generation process started from which can produce images with high quality, the result may not entirely align with the input LQ image in the degenerate subspace. Following [Tirer and Giryes, 2018; Kawar *et al.*, 2022; Wang *et al.*, 2023], we introduce the back

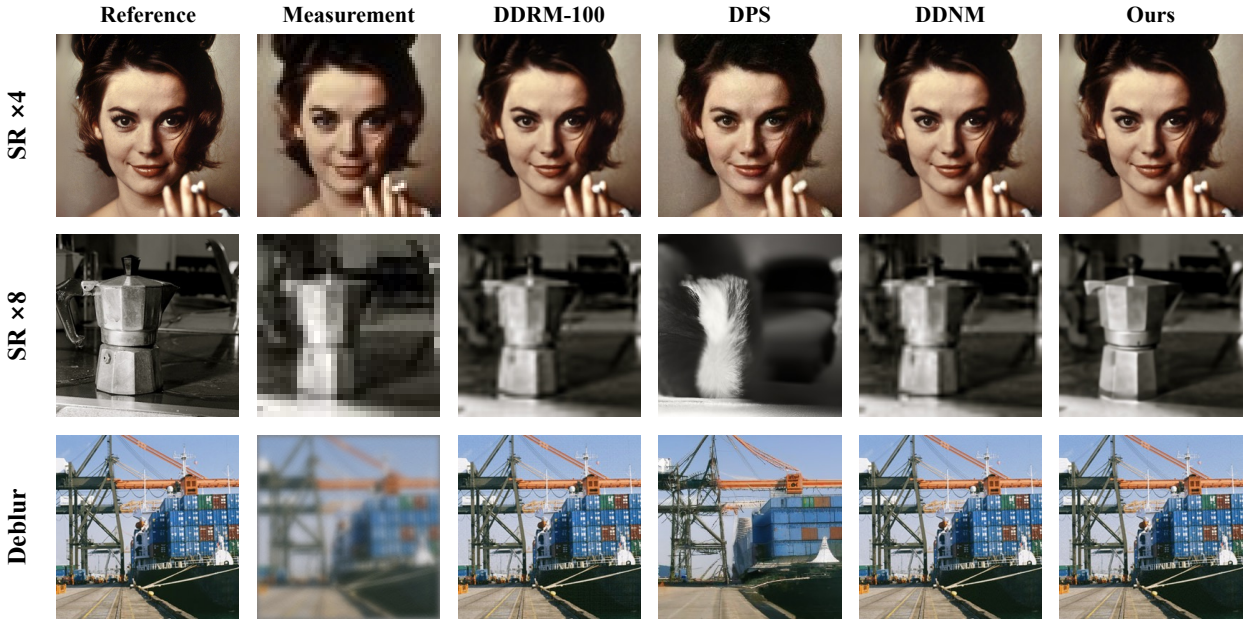

Figure 4: Qualitative results of different zero-shot IR methods on CelebA and Imagenet Dataset.

projection technique as consistency constraints to address this problem.

For details, given a measurement image y and corresponding degraded operator H, we can project the coarse restored image x onto the affine subspace $\{H\mathbb{R}^n = y\}$ by the following operation to force consistency:

$$x' = (I - H^\dagger H)x + H^\dagger y \tag{17}$$

where $H^\dagger$ is the pseudo-inverse of H, details about calculating H and $H^\dagger$ is available in Appendix E.

The refined restored image x' consists of two parts: the former part, denoted as $(I - H^\dagger H)x$, represents the residual between x and the image obtained after projection-in and projection-back, which can be interpreted as the enhancement details of x. The latter part, denoted as $H^\dagger y$, can be regarded as the preservation of input image y. After the back-projection step, we have:

$$Hx' \equiv H[(I - H^\dagger H)x + H^\dagger y] \equiv y \tag{18}$$

which indicates that x' entirely aligns with the input measurement image y in the degenerate subspace.

At each timestep, we initiate the process by performing a denoising step to obtain the predicted x_0 in Eq. 6. Subsequently, we process with a back projection step to refine the results of x_0 and derive x_{t-1} through Eq. 5. The complete transformations from x_t to x_{t-1} can be expressed as follows:

$$\begin{aligned} x_{0|t} =& \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \\ \hat{x}_{0|t} =& (I - H^\dagger H)x_{0|t} + H^\dagger y \\ x_{t-1} =& \sqrt{\alpha_{t-1}}\hat{x}_{0|t} + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\epsilon_\theta(x_t, t) + \sigma_t\epsilon_t \end{aligned} \tag{19}$$

3.4 Expand SSD to Noisy IR Tasks

Although SSD is effective in addressing various noiseless inverse problems, it tends to exhibit poor performance when faced with noisy tasks. This limitation can be primarily attributed to the utilization of back projection. The success of back projection hinges on a precise estimation of the degraded operator H. When applied to blind image restoration or noisy IR tasks, back-projection tends to result in disappointing restorations because of the inability to satisfy Eq. 18.

To solve this problem, we proposed SSD$^+$, an enhanced version that makes SSD suitable for noisy situations or inaccurate estimation of H. Earlier studies[Meng *et al.*, 2021; Hertz *et al.*, 2022; Tumanyan *et al.*, 2023] have indicated that diffusion models typically generate the overall layout and color in the early stage, generate the structure and appearance in the middle stage, and generate the texture details in the final stage. We notice that SSD employs shortcut sampling to skip the early stage and ensure the preservation of the overall layout. However, during the middle final stage, the utilization of back projection with an inaccurate H has the potential to deteriorate the fine-textured details, leading to suboptimal outcomes. In SSD$^+$, rather than performing back projection throughout the generation process, we restrict its use to the middle stage, where it still plays a crucial role in maintaining structure consistency. During the final stage of generation, we rely exclusively on diffusion priors to ensure texture details without compromising the integrity of the original structure.

4 Experiments

4.1 Experimental Setup

Pretrained Models and Datasets. To evaluate the performance of SSD, we conduct experiments on two datasets with different distribution characters: CelebA 256×256 [Karras

CelebA Method	**SR** $\times$ **4** PSNR↑ / FID↓ / LPIPS↓	**SR** $\times$ **8** PSNR↑ / FID↓ / LPIPS↓	**Colorization** FID↓ / LPIPS↓	**Deblur (gauss)** PSNR↑ / FID↓ / LPIPS↓	NFEs↓	Time(s/image)↓
$\mathbf{H}^{\dagger}\mathbf{y}$	28.02 / 128.22 / 0.301	24.77 / 153.86 / 0.460	43.99 / 0.197	19.96 / 116.28 / 0.564	0	N/A
DDRM-100	28.84 / 40.52 / 0.214	26.47 / 45.22 / 0.273	25.88 / 0.156	36.17 / 15.32 / 0.119	100	8.26
DPS	24.71 / 34.69 / 0.304	22.38 / **41.01** / 0.348	N/A	24.89 / 32.64 / 0.288	250	47.34
DDNM-100	**28.85** / 35.13 / 0.206	**26.53** / 44.22 / 0.272	23.65 / 0.138	**38.70** / 4.48 / 0.062	100	8.05
SSD-100 (ours)	28.84 / **32.41** / **0.202**	26.44 / 42.42 / **0.267**	**23.62** / **0.138**	38.62 / **4.36** / **0.060**	100	8.08
DDRM-30	28.62 / 46.72 / 0.221	26.28 / 49.32 / 0.281	27.69 / 0.214	36.05 / 15.71 / 0.122	30	2.57
DDNM-30	**28.76** / 41.36 / 0.213	**26.41** / 48.25 / 0.277	25.25 / 0.184	37.40 / 6.65 / 0.084	30	2.47
SSD-30 (ours)	28.71 / **36.77** / **0.208**	26.32 / **44.97** / **0.271**	**24.11** / **0.159**	**38.34** / **4.98** / **0.065**	30	2.48

ImageNet Method	**SR** $\times$ **4** PSNR↑ / FID↓ / LPIPS↓	**SR** $\times$ **8** PSNR↑ / FID↓ / LPIPS↓	**Colorization** FID↓ / LPIPS↓	**Deblur (gauss)** PSNR↑ / FID↓ / LPIPS↓	NFEs↓	Time(s/image)↓
$\mathbf{H}^{\dagger}\mathbf{y}$	26.26 / 106.01 / 0.322	22.86 / 124.89 / 0.4690	27.40 / 0.231	19.33 / 102.33 / 0.553	0	N/A
DDRM-100	27.40 / 43.27 / 0.260	23.74 / 83.08 / 0.420	36.44 / 0.224	36.48 / 11.81 / 0.121	100	16.91
DPS	20.34 / 72.33 / 0.485	18.38 / **76.89** / 0.538	N/A	24.89 / 32.64 / 0.288	250	148.71
DDNM-100	27.44 / 39.42 / 0.251	**23.80** / 80.09 / 0.412	36.46 / 0.219	**40.48** / 3.33 / 0.041	100	16.56
SSD-100 (ours)	**27.45** / **37.69** / **0.248**	23.76 / 82.11 / **0.409**	**35.40** / **0.215**	40.32 / **3.07** / **0.039**	100	16.63
DDRM-30	27.17 / 46.14 / 0.269	23.50 / 84.53 / 0.426	36.48 / 0.237	35.90 / 13.35 / 0.130	30	5.19
DDNM-30	**27.22** / 40.12 / 0.256	**23.53** / **74.60** / 0.414	36.46 / 0.229	37.67 / 6.91 / 0.081	30	4.98
SSD-30 (ours)	27.13 / **38.24** / **0.251**	23.44 / 76.35 / **0.411**	**36.22** / **0.223**	**39.23** / **4.64** / **0.053**	30	5.01

Table 1: Quantitative evaluation on the **CelebA**(*top*) and **ImageNet**(*bottom*) datasets for various typical IR tasks.

et al., 2017] for face images and ImageNet 256×256 [Deng *et al.*, 2009] for natural images, both containing 1k validation images independent of the training dataset. For CelebA 256×256, we use the denoising network VP-SDE[Song *et al.*, 2021c; Meng *et al.*, 2021][1]. For ImageNet 256×256, we use the denoising network guided-diffusion [Dhariwal and Nichol, 2021][2].

Degradation Operators. We conduct experiments on several typical IR tasks, including Super-Resolution($\times$ 4, $\times$ 8), Colorization, Inpainting, and Deblurring. Details of degradation operators can be found in Appendix E.

Evaluation. We use PSNR, FID[Heusel *et al.*, 2017], and LPIPS[Zhang *et al.*, 2018b] as the main metrics to quantitatively evaluate the performance of image restoration. Especially due to the inability of PSNR to capture colorization performance, we use FID and LPIPS for colorization. Additionally, we adopt Neural Function Evaluations(NFEs) as the metrics of sampling speed, which is a commonly employed benchmark in diffusion model-based methods. For most methods, NFEs are equivalent to the sampling steps of the generation process. Since SSD introduces an additional inversion process, the NFEs of SSD are calculated by summing the steps involved in both the inversion process and the generation process. All of our experiments are conducted on a single NVIDIA RTX 2080Ti GPU.

Comparison Methods. We compare the restoration performance of the proposed method with recent State-Of-The-Art zero-shot image restoration methods using pre-trained diffusion models: DDRM[Kawar *et al.*, 2022], DPS[Chung *et al.*, 2023] and DDNM[Wang *et al.*, 2023]. For a fair comparison, all methods above use the same pre-trained denoising networks and degradation operator.

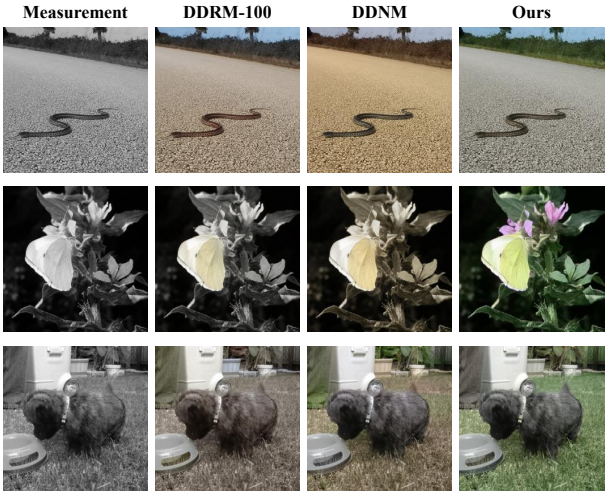

Figure 5: Colorization results of different zero-shot IR methods on ImageNet Dataset

4.2 Noiseless Image Restoration Results

We compare SSD (with 30 and 100 steps) with previous methods mentioned in Sec 4.1. The quantitative evaluation results shown in Table 1 illustrate that the proposed method achieves competitive results compared to state-of-the-art methods. When setting NFEs to 30, SSD-30 outperforms

[1]https://github.com/ermongroup/SDEdit
[2]https://github.com/openai/guided-diffusion

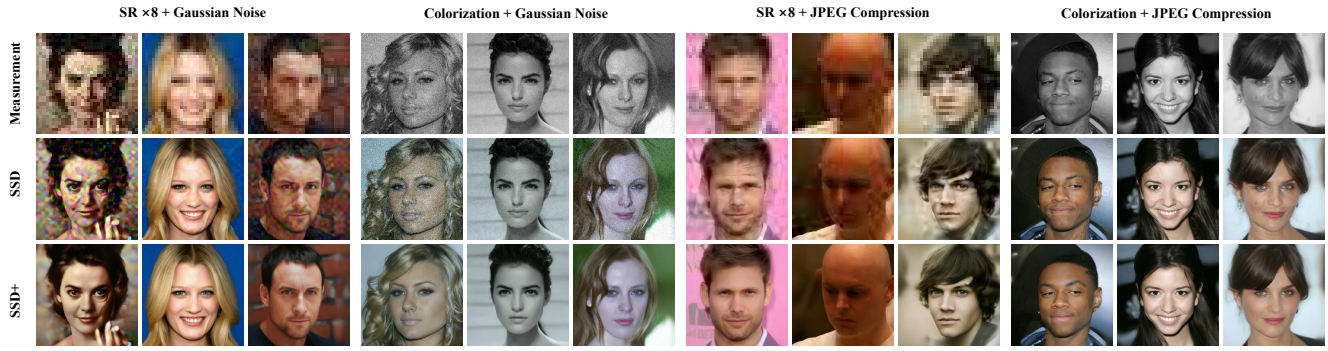

Figure 6: Qualitative results on CelebA of solving inverse problems with additional Gaussian noise and JEPG compression.

CelebA	8× SR + Noise	Colorization + Noise	8× SR + JPEG	Colorization + JPEG
Method	PSNR↑ / LPIPS↓ / FID↓	LPIPS↓ / FID↓	PSNR↑ / LPIPS↓ / FID↓	LPIPS↓ / FID↓
SSD	22.51 / 0.528 / 85.92	0.533 / 68.24	23.23 / 0.414 / 80.74	**0.301** / 48.54
SSD$^+$	**24.60 / 0.299 / 43.84**	**0.373/ 45.02**	**24.23 / 0.301 / 46.32**	0.372 / **45.02**

Table 2: Quantitative evaluation on CelebA of solving inverse problems with additional **Gaussian noise**(*left*) and **JEPG compression**(*right*). **Red** indicates the best performance.

other methods known for fast sampling such as DDRM-30 and achieves better perception-oriented metris (i.e., FID, LPIPS) than SOTA methods (DDNM with 100 NFEs). When setting NFEs to 100, SSD-100 achieves SOTA performance in many IR tasks, including SR $\times 4$ and colorization. As shown in Fig. 4, 5, SSD generates high-quality restoration results in all tested datasets and tasks.

4.3 Noisy Image Restoration Results

To illustrate the robustness of SSD$^+$ in the face of noisy situations and complex degradation, we evaluate SSD and SSD$^+$ on diverse inverse problems with gaussian noise and JPEG compression[Shin and Song, 2017]. For gaussian noise, we add gaussian noise $z \sim \mathcal{N}(0, \sigma^2)$ to the degraded image y, where σ represents the intensity of noise and is randomly distributed in $[0.0, 0.2]$. For JPEG compression, we perform JPEG compression[Wang *et al.*, 2021] with a quality factor of 60 after the degraded operator H is applied. The quantitative result is shown in Tab. 2. Qualitative results are available in Fig. 6. Compared to SSD, SSD$^+$ exhibits enhanced robustness when encountering more intricate degradation, producing satisfactory restoration results.

4.4 Sampling Speed

We further investigate the performance of various methods concerning FID in relation to the change in NFEs, as illustrated in Fig. 7. We conduct experiments with SR $\times 4$ task on the CelebA dataset. DPS [Chung *et al.*, 2023] is excluded from the comparison as it fails to generate reasonable results at low NFEs(< 100). The results indicate that SSD surpasses all other methods at both high and low NFEs, with a greater advantage when the NFEs is relatively small.

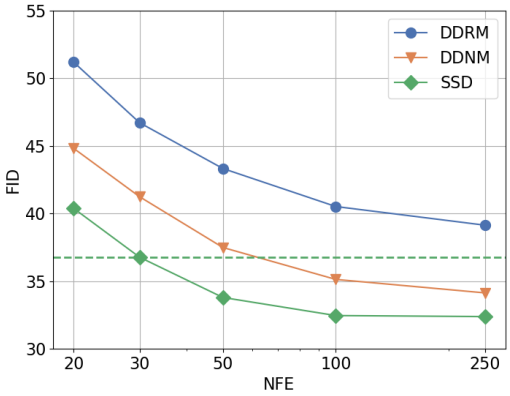

Figure 7: Comparison of the performance of various methods affected by NFEs with SR $\times$ 4 task on CelebA dataset

5 Conclusion

In this paper, we propose SSD, a novel framework for solving inverse problems in a zero-shot manner. We have departed from the conventional "Noise-Target" paradigm and instead proposed a shortcut sampling pathway of "Input-Embryo-Target". This novel approach enables us to achieve satisfactory results with reduced steps. We further propose SSD$^+$, an enhanced version of SSD tailored to excel in scenarios where degradation estimation is less accurate or in the presence of noise. We hope the proposed pipeline may inspire future work on inverse problems to solve them in a more efficient manner.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (Grant No. 61991451)

and the Shenzhen Science and Technology Program (JSGG20220831093004008).

References

[Chung *et al.*, 2022a] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-Closer-Diffuse-Faster: Accelerating Conditional Diffusion Models for Inverse Problems through Stochastic Contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[Chung *et al.*, 2022b] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[Chung *et al.*, 2023] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *In International Conference on Learning Representations (ICLR)*, 2023.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009.

[Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[Guo *et al.*, 2019] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1712–1722, 2019.

[Haris *et al.*, 2018] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1664–1673, 2018.

[Hertz *et al.*, 2022] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems (NeurIPS)*, volume 30, 2017.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[Karras *et al.*, 2017] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[Kawar *et al.*, 2021] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems*, 34:21757–21769, 2021.

[Kawar *et al.*, 2022] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

[Kim *et al.*, 2022] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2426–2435, 2022.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Larsson *et al.*, 2016] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 577–593. Springer, 2016.

[Liu *et al.*, 2023] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 289–299, 2023.

[Lugmayr *et al.*, 2022] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, June 2022.

[Mardani *et al.*, 2023] Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*, 2023.

[Meng *et al.*, 2021] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[Menon *et al.*, 2020] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. 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)*, pages 2437–2445, 2020.

[Pan *et al.*, 2021] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and

manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7474–7489, 2021.

[Pokle *et al.*, 2023] Ashwini Pokle, Matthew J Muckley, Ricky TQ Chen, and Brian Karrer. Training-free linear image inversion via flows. *arXiv preprint arXiv:2310.04432*, 2023.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[Shin and Song, 2017] Richard Shin and Dawn Song. Jpeg-resistant adversarial images. In *NeurIPS Workshop on Machine Learning and Computer Security*, page 3, 2017.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265. PMLR, 2015.

[Song *et al.*, 2021a] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *In International Conference on Learning Representations (ICLR)*, 2021.

[Song *et al.*, 2021b] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *In International Conference on Learning Representations (ICLR)*, 2021.

[Song *et al.*, 2021c] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *In International Conference on Learning Representations (ICLR)*, 2021.

[Song *et al.*, 2022] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2022.

[Tirer and Giryes, 2018] Tom Tirer and Raja Giryes. Image restoration by iterative denoising and backward projections. *IEEE Transactions on Image Processing*, 28(3):1220–1234, 2018.

[Tumanyan *et al.*, 2023] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, June 2023.

[Ulyanov *et al.*, 2018] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9446–9454, 2018.

[Wang *et al.*, 2018] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[Wang *et al.*, 2021] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1905–1914, October 2021.

[Wang *et al.*, 2022] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17683–17693, 2022.

[Wang *et al.*, 2023] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *In International Conference on Learning Representations (ICLR)*, 2023.

[Xiao *et al.*, 2022] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[Yeh *et al.*, 2017] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5485–5493, 2017.

[Zhang *et al.*, 2018a] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.

[Zhang *et al.*, 2018b] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.

[Zhang *et al.*, 2024] Yuechen Zhang, Jinbo Xing, Eric Lo, and Jiaya Jia. Real-world image variation by aligning diffusion inversion chain. *Advances in Neural Information Processing Systems*, 36, 2024.