# Arbitrary-steps Image Super-resolution via Diffusion Inversion

Zongsheng Yue[1,2], Kang Liao[2], Chen Change Loy[2]

[1]Xi'an Jiaotong University, Xi'an, China
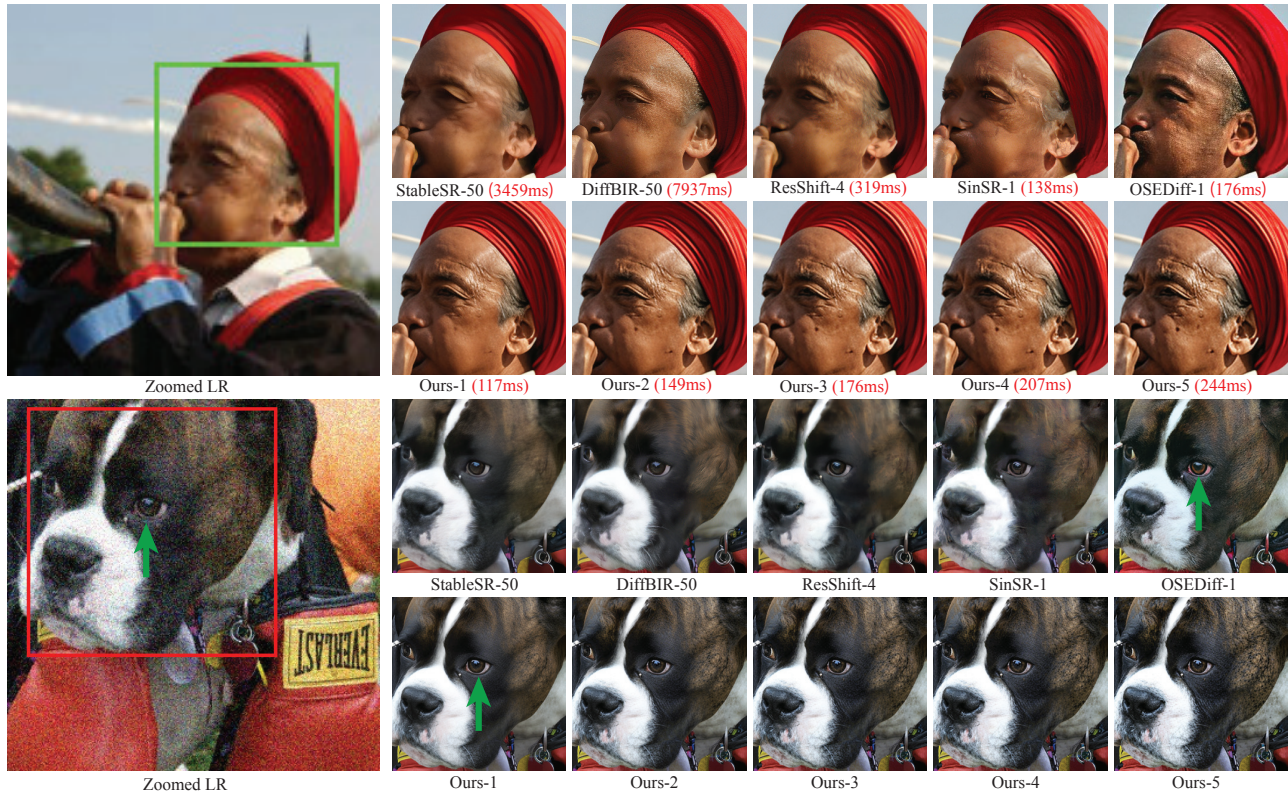[2]S-Lab, Nanyang Technological University, Singapore

Figure 1. Qualitative comparisons of our proposed method to recent state-of-the-art diffusion-based approaches on two real-world examples, where the number of sampling steps is annotated in the format "Method name-Steps". We provide the runtime (in milliseconds) highlighted by red in the sub-caption of the first example , which is tested on ×4 (128 → 512) SR task on an A100 GPU. Our method offers an efficient and flexible sampling mechanism, allowing users to freely adjust the number of sampling steps based on the degradation type or their specific requirements. In the first example, mainly degraded by blurriness, multi-step sampling is preferable to single-step sampling as it progressively recovers finer details. Conversely, in the second example with severe noise, a single sampling step is sufficient to achieve satisfactory results, whereas additional steps may amplify the noise and introduce unwanted artifacts. *(Zoom-in for best view)*

## Abstract

*This study presents a new image super-resolution (SR) technique based on diffusion inversion, aiming at harnessing the rich image priors encapsulated in large pre-trained diffusion models to improve SR performance. We design a Partial noise Prediction strategy to construct an intermediate state of the diffusion model, which serves as the starting sampling point. Central to our approach is a deep noise predictor to estimate the optimal noise maps for the forward diffusion process. Once trained, this noise predictor can be used to initialize the sampling process partially along the diffusion trajectory, generating the desirable high-resolution result. Compared to existing approaches, our method offers a flexible and efficient sampling mechanism that supports an arbitrary number of sampling steps, ranging from one to five. Even with a single sampling step, our method demonstrates superior or comparable performance to recent state-of-the-art approaches. The code and model are publicly available at https://github.com/zsyOAOA/InvSR.*

## 1. Introduction

Image super-resolution (SR) is a fundamental yet challenging problem in computer vision, aiming to restore a high-resolution (HR) image from a given low-resolution (LR) observation. The main challenge of SR arises from the complexity and often unknown nature of the degradation model in real-world scenarios, making SR an ill-posed problem. Recent breakthroughs in diffusion models [16, 44, 47], particularly large-scale text-to-image (T2I) models, have demonstrated remarkable success in generating high-quality images. Owing to the strong generative capability of these T2I models, recent studies have begun to use them as a reliable prior to alleviate the ill-posedness of SR. This work follows this research line, further exploring the potential of diffusion priors in SR.

The prevailing SR approaches leveraging diffusion priors usually attempt to modify the intermediate features of the diffusion network, either through optimization [7, 23, 55] or fine-tuning [30, 55, 59, 63], to better align them with the given LQ observations. In this work, we propose a new technique based on diffusion inversion to harness diffusion priors. Unlike existing approaches, it attempts to find an optimal noise map as the input of the diffusion model, without any modification to the diffusion network itself, thereby maximizing the utility of diffusion prior.

While considerable advances have been made in generative adversarial networks (GANs) [14] inversion for various applications [61, 74], including SR [4, 15, 40], extending these principles to diffusion models presents unique challenges, particularly for SR tasks that demand high fidelity preservation. In particular, the multi-step stochastic sampling process of diffusion models makes inversion nontrivial. The straightforward inversion approach to optimize the distinct noise maps at each diffusion step is expensive and complex. Additionally, the iterative inference mechanism would accumulate prediction errors and randomness at each step, which can significantly compromise fidelity. Therefore, recent diffusion inversion methods have mainly focused on tasks with lower fidelity requirements, such as image editing [13, 36].

In this work, we reformulate diffusion inversion for the more challenging task of SR. To enable diffusion inversion for SR, we introduce a deep neural network called noise predictor to estimate the noise map from a given LR image. In addition, a Partial noise Prediction (PnP) strategy is devised to construct an intermediate state for the diffusion model, serving as the starting point for sampling. This is made possible by *adding noise onto the LR image* according to the diffusion model's forward process, *where the noise predictor predicts the added noise instead of random sampling*. This approach is driven by the following key motivations:

- **Rationality**. LR and HR images differ only in high-frequency details. With the addition of appropriate noise,

the LR image becomes indistinguishable from its HR counterpart. Thus, the noisy LR can serve as a proxy for deriving the inversion trajectory during reverse diffusion.
- **Complexity**. Rather than predicting noise maps for all diffusion steps, the PnP strategy simplifies the inversion task by limiting predictions to the starting step, thereby reducing the overall complexity of the inversion process.
- **Flexibility**. The noise predictor can be trained to predict noise maps for multiple predefined starting steps. During inference, we can freely select a starting step from them and then use any existing sampling algorithm with an arbitrary number of steps, offering favorable flexibility in controlling the sampling process.
- **Fidelity**. The starting steps during training are carefully selected to have a high signal-to-noise ratio (SNR), ensuring robust fidelity preservation for SR. In practice, we enforce an SNR threshold greater than 1.44, corresponding to the timestep of 250 in Stable Diffusion [42].
- **Efficiency**. As the sampling process begins from a step earlier than 250 (SNR larger than 1.44), the PnP strategy effectively *reduces the number of sampling steps to fewer than five* when combined with off-the-shelf accelerated sampling algorithms [22, 45]. This addresses the common inefficiency issue in diffusion-based SR approaches.

Unlike most existing diffusion-based methods that rely on fixed sampling steps, our flexible sampling mechanism offers a versatile solution for handling varying degrees of degradation in SR. In SR, it is common to encounter different types and intensities of corruption. Intuitively, the number of sampling steps should adapt to the specific degradation conditions. For example, as shown in Fig. 1, multi-step sampling is preferable to single-step sampling in the first case, as it effectively reduces blurriness and restores finer details. In contrast, for the second example with severe noise, a single sampling step achieves satisfactory results, while additional steps may amplify the noise and introduce unwanted artifacts. Our method uniquely allows users to adjust sampling to suit different degradation types.

The main contributions of this work are twofold. First, we propose a novel SR approach based on diffusion inversion, which effectively leverages the diffusion prior by integrating an auxiliary noise predictor while keeping the entire diffusion backbone fixed. Second, our method introduces a flexible and efficient sampling mechanism that allows for arbitrary sampling steps, ranging from one to five. Remarkably, even when the steps are reduced to just one, our approach still achieves superior or comparable performance to recent dedicated one-step diffusion methods.

## 2. Related Work

**Diffusion Prior for SR.** Existing diffusion prior-based SR approaches can be broadly categorized into two classes. The first class of methods involves re-optimizing the in-

termediate results of the diffusion model to ensure consistency with the given LR images via pre-defined or estimated degradation models. Representative works include DDRM [23], CCDF [7], and DDNM [55], among others [6, 8, 11, 37, 46, 62, 66]. While effective, these methods are limited by their computational complexity, as they require solving an optimization problem at each diffusion step, leading to slow inference. Furthermore, they often rely on manually defined assumed degradation models and thus cannot handle the blind SR problem in real-world scenarios. The second class directly fine-tunes a pre-trained large T2I model for the SR task. StableSR [52] pioneers this paradigm by incorporating spatial feature transform layers [53] to guide the T2I model toward generating HR outputs. Subsequent works follow by proposing various fine-tuning strategies to exploit diffusion priors, including DiffBIR [30], SeeSR [59], PASD [63], S3Diff [70], and so on [27, 39, 48, 58, 60, 65]. These methods have achieved impressive performance, validating the effectiveness of diffusion priors for SR.

**Diffusion Inversion.** Diffusion inversion focuses on determining the optimal noise map set that, when processed through the diffusion model, reconstructs a given image. DDIM [45] first addressed this by generalizing the diffusion model via a class of non-Markovian processes, thereby establishing a deterministic generation process. Subsequent approaches, such as those by Rinon *et al.* [12] and Mokady *et al.* [36], proposed optimizing the text embedding to better align with the desired textual guidance. Recent efforts have further refined the optimization strategies for both the textual and visual prompts [35, 38], as well as for intermediate noise maps [13, 19, 20, 33, 50, 71], leading to notable enhancements in inversion quality. Despite these advances, existing methods mainly focus on image editing and cannot meet the high-fidelity requirements of SR.

In this work, we tailor the diffusion inversion technique for SR. While Chihaoui *et al.* [5] have recently explored diffusion inversion for image restoration, their method relies on solving an optimization problem at each inversion step, significantly limiting its inference efficiency. In contrast, our approach introduces a noise prediction module that, once trained, enables efficient inversion without requiring iterative optimization during inference. This leads to substantial improvements in both the efficiency and practicality of diffusion inversion for SR tasks.

# 3. Methodology

In this section, we present the proposed diffusion inversion technique for SR. To maintain consistency with the notations used in diffusion models, we denote the LR image as $\boldsymbol{y}_0$ and the corresponding HR image as $\boldsymbol{x}_0$.

## 3.1. Motivation

The diffusion model [16, 44] was first introduced as a probabilistic generative model inspired by nonequilibrium thermodynamics. Subsequently, Song *et al.* [47] reformulated it within the framework of stochastic differential equations (SDEs). In this paper, we propose a general diffusion inversion technique that is applicable to both the probabilistic and SDE-based diffusion formulations. To facilitate understanding, we employ the probabilistic framework of the Denoising Diffusion Probabilistic Model (DDPM) [16] throughout our presentation.

The DDPM framework [16] is indeed a Markov chain of length $T$, where the forward process is characterized by a Gaussian transition kernel:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t\boldsymbol{I}), \qquad (1)$$

where $\beta_t$ is a pre-defined hyper-parameter controlling variance schedule. Notably, this transition kernel allows the derivation of the marginal distribution $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$, *i.e.*,

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

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, $\alpha_s = 1 - \beta_s$. The reverse process aims to generate a high-quality image from an initial random noise map $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, which can be expressed as:

$$\boldsymbol{x}_{t-1} = g_\theta(\boldsymbol{x}_t, t) + \sigma_t\boldsymbol{z}_{t-1}, \ t = T, \cdots, 1, \qquad (3)$$

where

$$g_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right), \quad (4)$$

$\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ is a pre-trained denoising network parameterized by $\theta$. The noise term $\boldsymbol{z}_t$ satisfies $\boldsymbol{z}_0 = \boldsymbol{0}$ and $\boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ for $t = 1, \cdots, T-1$.

Equation (3) indicates that the synthesized image $\boldsymbol{x}_0$ is fully determined by the set of noise maps $\mathcal{M} = \boldsymbol{x}_T \cup \{\boldsymbol{z}_t\}_{t=1}^{T-1}$. In the context of SR, our goal is to generate an HR image $\boldsymbol{x}_0$ conditioned on an LR image $\boldsymbol{y}_0$. To this end, we propose diffusion inversion to find an optimal set of noise maps $\mathcal{M}^*$ that reconstruct the target HR image $\boldsymbol{x}_0$ via the reverse process of Eq. (3). In the following sections, we detail how to achieve this goal by training a noise predictor.

## 3.2. Diffusion Inversion

To achieve diffusion inversion, we introduce a noise prediction network with parameter $w$, denoted as $f_w$, which takes the LR image $\boldsymbol{y}_0$ and the timestep $t$ as input and outputs the desired noise maps $\mathcal{M}^*$. Unlike the strategy [5] of directly optimizing $\mathcal{M}^*$ for each testing image, we train such a noise predictor to enable fast sampling during inference, thereby significantly improving the inference efficiency. To ensure the output of $f_w$ conforms to Gaussian distribution, we adopt the reparameterization trick of VAE [25], which predicts the mean and variance parameters of Gaussian distribution rather than directly estimating the noise map.

### 3.2.1. Problem Simplification

Training this noise predictor is inherently challenging. The noise map set $\mathcal{M}$ consists of $T$ noise maps (typically $T = 1000$ in most current diffusion models), corresponding to each step of the diffusion process. Naturally, it is non-trivial to simultaneously estimate such a large number of noise maps using a single, compact network. What's worse, the iterative sampling paradigm of diffusion models can gradually accumulate prediction errors, which may adversely affect the final SR performance.

To address these challenges, we design a Partial Noise Prediction (PnP) strategy. Specifically, let's consider diffusion inversion in the context of SR, where the observed LR image $\boldsymbol{y}_0$ only slightly deviates from the target HR image $\boldsymbol{x}_0$ in most cases, primarily in high-frequency components. This observation inspires us to initiate the sampling process from an intermediate timestep $N$ ($N < T$), effectively reducing the number of noise maps in $\mathcal{M}$ from $T$ to $N$, $i.e.$, $\mathcal{M} = \{\boldsymbol{z}_t\}_{t=1}^{N}$. Furthermore, given the high-fidelity requirements of SR, we constrain $\boldsymbol{x}_N$ to have a relatively high SNR, implying mild noise corruption. This constraint encourages the selection of a smaller $N$, and in practice, we set $N \leq 250$, corresponding to an SNR threshold of $1.44$ in the widely used Stable Diffusion [42].

In addition, we further compress the set of the noise maps $\mathcal{M} = \{\boldsymbol{z}_t\}_{t=1}^{N}$ by integrating existing diffusion acceleration algorithms [22, 45]. The common idea of these algorithms is to skip certain steps during inference, which is selected based on specific rules [29], $e.g.$, "linspace" and "trailing". Combining with this skipping strategy, the noise map set is simplified as follows:

$$\mathcal{M} = \{\boldsymbol{z}_{\kappa_i}\}_{i=1}^{M}, \qquad (5)$$

where $\{\kappa_1, \cdots, \kappa_M\} \subseteq \{1, \cdots, N\}$. In practice, we set $M \leq 5$, thus largely reducing the prediction burden on the noise predictor and improving the sampling efficiency.

### 3.2.2. Inversion Trajectory

Given the set of noise maps $\mathcal{M} = \{\boldsymbol{z}_{\kappa_i}\}_{i=1}^{M}$ and the noise prediction network $f_w$, our goal is to restore the HR image $\boldsymbol{x}_0$ from a given LR observation $\boldsymbol{y}_0$, following an inversion trajectory defined by:

$$\boldsymbol{x}_{\kappa_{i-1}} = g_\theta(\boldsymbol{x}_{\kappa_i}, \kappa_i) + \sigma_{\kappa_i} f_w(\boldsymbol{y}_0, \kappa_{i-1}), \qquad (6)$$

where $\kappa_0 = 0$, and $g_\theta(\cdot, \cdot)$ is defined in Eq. (4). The key to initiating this inversion trajectory is constructing the starting state $\boldsymbol{x}_{\kappa_M}$ from the LR image $\boldsymbol{y}_0$.

The marginal distribution $q(\boldsymbol{x}_{\kappa_M}|\boldsymbol{x}_0)$, as defined in Eq. (2), suggests to achieve $\boldsymbol{x}_{\kappa_M}$ as follows:

$$\boldsymbol{x}_{\kappa_M} = \sqrt{\bar{\alpha}_{\kappa_M}}\boldsymbol{x}_0 + \sqrt{1 - \bar{\alpha}_{\kappa_M}}\boldsymbol{\xi}, \; \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \qquad (7)$$

In the context of SR, since the HR image $\boldsymbol{x}_0$ is not accessible during testing, we thus construct an analogous formulation for $\boldsymbol{x}_{\kappa_M}$ directly from the LR image $\boldsymbol{y}_0$ using the noise predictor $f_w(\cdot)$, namely

$$\boldsymbol{x}_{\kappa_M} = \sqrt{\bar{\alpha}_{\kappa_M}}\boldsymbol{y}_0 + \sqrt{1 - \bar{\alpha}_{\kappa_M}}f_w(\boldsymbol{y}_0, \kappa_M). \qquad (8)$$

This design is inspired by the observation that the LR image $\boldsymbol{y}_0$ and the HR image $\boldsymbol{x}_0$ become increasingly indistinguishable when perturbed by Gaussian noise with an appropriate magnitude. Therefore, we aim to seek an optimal noise map $f_w(\boldsymbol{y}_0, \kappa_M)$ to perturb $\boldsymbol{y}_0$ in such a way that the pre-trained diffusion model can generate the corresponding $\boldsymbol{x}_0$ from $\boldsymbol{x}_{\kappa_M}$ that is defined in Eq. (8).

To summarize, we establish an inversion trajectory by combining Eqs. (6) and (8), which can be used to solve the SR problem via iterative generation along this trajectory.

### 3.2.3. Model Training

Given a pre-trained large-scale diffusion model $\boldsymbol{\epsilon}_\theta(\cdot)$, an estimation of the HR image $\boldsymbol{x}_0$ can be obtained from $\boldsymbol{x}_{\kappa_i}$ by taking a reverse diffusion step:

$$\hat{\boldsymbol{x}}_{0 \leftarrow \kappa_i} = \frac{1}{\sqrt{\bar{\alpha}_{\kappa_i}}}\left[\boldsymbol{x}_{\kappa_i} - \sqrt{1 - \bar{\alpha}_{\kappa_i}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_{\kappa_i}, \kappa_i)\right], \qquad (9)$$

where $\boldsymbol{x}_{\kappa_i}$ is defined by Eq. (8) for $i = M$ and Eq. (6) for $i < M$. It is thus possible to train the noise predictor $f_w(\cdot)$ by minimizing the distance between $\hat{\boldsymbol{x}}_{0 \leftarrow \kappa_i}$ and $\boldsymbol{x}_0$.

However, directly training with this objective is computationally impractical. Specifically, as shown in Eq. (6), calculating $\boldsymbol{x}_{\kappa_i}$ ($i < M$) necessitates recurrent application of the large-scale diffusion model $\boldsymbol{\epsilon}_\theta$, which leads to prohibitive GPU memory usage. To circumvent this, we adopt an alternative version for $\boldsymbol{x}_{\kappa_i}$ based on the marginal distribution in Eq. (2), $i.e.$,

$$\boldsymbol{x}_{\kappa_i} = \sqrt{\bar{\alpha}_{\kappa_i}}\boldsymbol{x}_0 + \sqrt{1 - \bar{\alpha}_{\kappa_i}}f_w(\boldsymbol{y}_0, \kappa_i), \; i < M. \qquad (10)$$

This modification also aligns better with the training process of the employed diffusion model, allowing for more effective leveraging of the prior knowledge embedded in it. We now detail the training procedure step by step:

**Gaussian Constraint**. The pre-trained diffusion model is a powerful denoiser tailored for Gaussian noise with zero mean and varying variances. Hence, it is reasonable to enforce the predicted noise map by $f_w$ to obey a Gaussian distribution. For the initial state $\boldsymbol{x}_{\kappa_M}$, it is observed that the predicted noise map $f_w(\boldsymbol{y}_0, \kappa_M)$ exhibits a mean shift, which is evident when comparing Eqs. (7) and (8), due to the substitution of $\boldsymbol{y}_0$ for $\boldsymbol{x}_0$. Moreover, the visualization presented in Figs. 2 and 3 further validates this observation, illustrating that the predicted noise map is clearly correlated with the LR image. Therefore, we do not consider the Gaussian constraint for $\boldsymbol{x}_{\kappa_M}$.
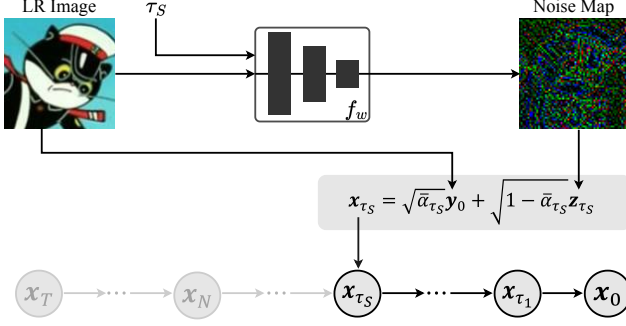
Figure 2. Inference flow of our proposed method, wherein $\{\tau_i\}_{i=1}^{S}$ denotes the inversion timesteps. Note that the predicted noise map $z_{\tau_S}$ exhibits an obvious correlation with the LR image, indicating the non-zero mean property of its statistical distribution.

Conversely, for the intermediate state $x_{\kappa_i}$ as defined in Eq. (10), the predicted noise map $f_w(y_{\kappa_i}, \kappa_i)$ should be enforced to follow $\mathcal{N}(\mathbf{0}, \mathbf{I})$. This naturally raises an interesting question: Is it necessary to predict the noise map using $f_w$ instead of random sampling?

First, the proposed PnP strategy requires the timestep $\kappa_i$ to satisfy a high SNR constraint, indicating that $x_{\kappa_i}$ is corrupted by only mild Gaussian noise. Second, the pre-trained large-scale diffusion model, specifically designed for Gaussian denoising, performs robustly, especially for timesteps $\kappa_i$ with low noise levels. Thus, introducing an extra noise predictor model $f_w$ for the intermediate timesteps, even conditioned on the additional LR observation, does not yield significant performance gains. This is also empirically illustrated by an ablation study provided in supp. Third, predicting the noise map both for the initial and intermediate state would increase the prediction burden on $f_w$ and make the training process more challenging. Considering these reasons together, we discard the noise prediction for the intermediate states, which further reduces the noise map set to $\mathcal{M} = \{z_{\kappa_M}\}$, leading to a more elegant diffusion inversion technique, as detailed in Alg. 1.

**Arbitrary-step Inversion.** As analyzed above, noise map prediction is only required for the starting state, as defined in Eq.(8), to initialize the reverse sampling process. To further enhance the flexibility of the sampling process, the noise predictor is trained to estimate the noise maps for multiple pre-selected steps via time embedding. Once trained, the starting timestep can be freely chosen during inference, which results in a fidelity-realism trade-off, as analyzed in Sec. 4.2. Note that the total number of sampling steps is determined by both the selected starting timestep and the skipping stride of the accelerated sampling algorithm for diffusion models. For clarity, a detailed inference procedure is provided in Fig. 2 and Alg. 1. Given practical efficiency considerations, we focused on the number of sampling steps

---

**Algorithm 1** Inference

**Require:** LR image $y_0$, noise predictor $f_w$, pre-trained diffusion model $\epsilon_\theta$, inversion timesteps $\{\tau_i\}_{i=1}^{S} \subset \{\kappa_i\}_{i=1}^{M}$
1: $x_{\tau_S} = \sqrt{\bar{\alpha}_{\tau_S}} y_0 + \sqrt{1 - \bar{\alpha}_{\tau_S}} f_w(y_0, \tau_S)$
2: **for** $i = S, \cdots, 1$ **do**
3:     $z_{\tau_i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $\tau_i > 1$ else $z_{\tau_i} = \mathbf{0}$
4:     $x_{\tau_{i-1}} = g_\theta(x_{\tau_i}, \tau_i) + \sigma_{\tau_i} z_{\tau_i}$, where $g_\theta$ is defined in Eq. (4)
5: **end for**
6: **return** $x_0$

---

ranging from one to five in this study.

**Loss Function.** To train the noise predictor, we adopt an $L_2$ loss $\mathcal{L}_2$, a LPIPS [73] loss $\mathcal{L}_l$, and a GAN [14] loss $\mathcal{L}_g$ following recent SR approaches [4, 54]. Let's denote the set of pre-selected starting timesteps for training as $\mathcal{S} \subseteq \{\kappa_1, \cdots, \kappa_M\}$, the overall loss function is defined as:

$$\sum_{t \in \mathcal{S}} \mathcal{L}_2(\hat{x}_{0 \leftarrow t}, x_0) + \lambda_l \mathcal{L}_l(\hat{x}_{0 \leftarrow t}, x_0) + \lambda_g \mathcal{L}_g(\hat{x}_{0 \leftarrow t}, x_0), \quad (11)$$

where $\hat{x}_{0 \leftarrow t}$ is defined in Eq. (9), $\lambda_l$ and $\lambda_g$ are hyper-parameters. The adversarial loss $\mathcal{L}_g$ is implemented using a hinge loss, with a discriminator architecture based on diffusion UNet [16] enhanced with a multi-in, multi-out strategy [64]. For the base model $\epsilon_\theta$, we use SD-Turbo [43], which operates in the latent space of VQGAN [10]. We thus compute the whole loss in the latent space, significantly reducing the required GPU memory. To facilitate training, the LPIPS loss is also fine-tuned in the latent space.

## 4. Experiments

In this section, we first provide an analysis of the proposed method and then conduct extensive experiments to evaluate its performance on one synthetic and two real-world datasets. Our investigation focuses mainly on the $\times 4$ SR task following previous works [54, 72]. To ease the presentation, we refer to our method as *InvSR*, standing for Diffusion *Inv*ersion-based *S*uper-*R*esolution.

### 4.1. Experimental Setup

**Training Details.** Following the setup of recent works [58, 59], we trained the noise predictor on the LSDIR [28] dataset and a subset of 20k face images from the FFHQ [21] dataset. At each iteration, we randomly cropped an image patch with a resolution of $512 \times 512$ from the source image and synthesized the LR image using the pipeline of RealESRGAN [54]. The text prompt was fixed as a general description[1] in both the training and testing phases. To optimize the network parameters, we adopted the Adam [26]

---

[1]"Cinematic, high-contrast, photo-realistic, 8k, ultra HD, meticulous detailing, hyper sharpness, perfect without deformations."

Table 1. Quantitative results of *InvSR* with various numbers of sampling steps ranging from one to five on the *ImageNet-Test* dataset.

| #Steps | Index of the sampled timesteps | Metrics | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | PSNR↑ | SSIM↑ | LPIPS↓ | NIQE↓ | PI↓ | CLIPIQA↑ | MUSIQ↑ |
| 5 | {250, 200, 150, 100, 50} | 22.70 | 0.6412 | 0.2844 | 4.8757 | 3.4744 | 0.6733 | 69.8427 |
| 3 | {250, 150, 50} | 22.92 | 0.6478 | 0.2762 | 4.7980 | 3.4002 | 0.6823 | 70.4688 |
| | {200, 100, 50} | 23.41 | 0.6609 | 0.2648 | 4.5089 | 3.2074 | 0.6851 | 70.7024 |
| | {150, 100, 50} | 23.84 | 0.6713 | 0.2575 | 4.2719 | 3.0527 | 0.6823 | 70.4569 |
| 1 | {250} | 23.84 | 0.6713 | 0.2575 | 4.5287 | 3.1748 | 0.7132 | 72.5773 |
| | {200} | 24.14 | 0.6789 | 0.2517 | 4.3815 | 3.0866 | 0.7093 | 72.2909 |
| | {150} | 24.42 | 0.6851 | 0.2469 | 4.2194 | 2.9979 | 0.7019 | 71.7100 |
| | {100} | 24.66 | 0.6891 | 0.2450 | 4.0606 | 2.8951 | 0.6912 | 70.8251 |

algorithm with default settings of PyTorch [41]. The training process takes over 100k iterations with a batch size of 64 and a fixed learning rate of $5e^{-5}$. The hyper-parameters $\lambda_l$ and $\lambda_g$ in the loss function were set to 2.0 and 0.1, respectively. The architecture of the noise predictor was based on the encoder of VQGAN [10], containing two down-sampling blocks, each equipped with a self-attention layer [49].

In the training stage, we randomly select a starting timestep from $\mathcal{S} = \{250, 200, 150, 100\}$ to train the noise predictor at each iteration. During inference, five inversion steps, *i.e.*, $\mathcal{M} = \{250, 200, 150, 100, 50\}$, are used throughout our experiments.

**Testing Datasets**. To evaluate the performance of *InvSR*, we constructed a synthetic dataset named *ImageNet-Test*, comprising 3,000 images from the validation set of ImageNet [9]. The LR and HR images, with resolutions of $128 \times 128$ and $512 \times 512$, respectively, were synthesized using the degradation settings of ResShift [68]. Notably, we selected the HR images from ImageNet rather than the commonly used datasets in SR, such as Set5 [1], Set14 [69], and Urban100 [17], mainly because these datasets only contain very few source images, which fails to thoroughly assess various methods under complicated degradation types.

We further conducted experiments on two real-world datasets to validate the effectiveness of *InvSR*. The first dataset is *RealSR* [3], which contains 100 real images captured by Canon 5D3 and Nikon D810 cameras. The second dataset, *RealSet80* [68], comprises 80 LR images widely used in existing literature [18, 30–32, 54, 67].

**Compared Methods**. We evaluate the effectiveness of *InvSR* in comparison to nine recent methods, including two GAN-based methods, namely BSRGAN [72] and RealESRGAN [54], as well as seven diffusion-based methods, including LDM [42], StableSR [52], DiffBIR [30], SeeSR [59], ResShift [67, 68], SinSR [56], and OSEDiff [58]. For LDM, StableSR, DiffBIR, and SeeSR, we all use 50 sampling steps for fair comparison. In the case of ResShift, SinSR, and OSEDiff, we adhere to the number of sampling steps suggested by their official guidelines.



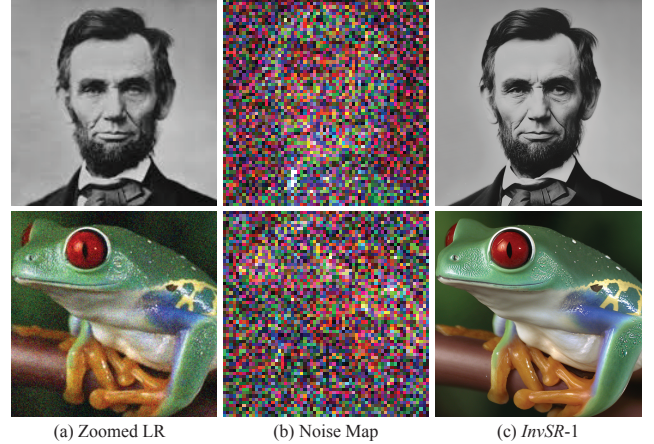(a) Zoomed LR     (b) Noise Map     (c) *InvSR*-1

Figure 3. From left to right: (a) zoomed LR image, (b) predicted noise map by our method for the initial timestep, (c) super-resolved results by our method with a single sampling step.

**Metrics**. The performance of various methods was assessed across seven metrics, including three reference metrics, namely PSNR, SSIM [57], LPIPS [73], as well as four non-reference metrics, namely NIQE [34], PI [2], MUSIQ [24], and CLIPIQA [51]. For evaluations on the datasets of *ImageNet-Test* and *RealSR*, all seven metrics were adopted to ensure a holistic assessment. For the dataset of *RealSet80*, however, only non-reference metrics were employed since the ground truth images are not accessible.

### 4.2. Model Analysis

**Arbitrary-steps Sampling**. Recent efficient diffusion-based SR approaches, such as ResShift [68], SinSR [56], and OSEDiff [58], constrain the sampling process to a pre-defined number of steps, consistent with their training configuration. In contrast, the proposed *InvSR* supports sampling with an arbitrary number of steps, significantly enhancing flexibility and adaptability to varying degradation types, as demonstrated in Fig. 1 and Fig. II of supp.

We further provide a comprehensive comparison of *InvSR* with one, three, and five sampling steps, as summarized in Table. 1. Three key observations can be ob-

Table 2. Quantitative comparisons of different methods on *ImageNet-Test* and *RealSR*. The number of sampling steps is marked in the format of "Method name-Steps" for diffusion-based methods. The best and second-best results are highlighted in **bold** and <u>underlined</u>.

| Datasets | Methods | Metrics | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | PSNR↑ | SSIM↑ | LPIPS↓ | NIQE↓ | PI↓ | CLIPIQA↑ | MUSIQ↑ | #Params (M) |
| *ImageNet-Test* | BSRGAN [72] | 27.05 | 0.7453 | 0.2437 | 4.5345 | 3.7111 | 0.5703 | 67.7195 | 16.70 |
| | RealESRGAN [54] | 26.62 | <u>0.7523</u> | 0.2303 | 4.4909 | 3.7234 | 0.5090 | 64.8186 | 16.70 |
| | LDM-50 [42] | <u>27.19</u> | 0.7285 | 0.2286 | 5.2411 | 4.2554 | 0.5554 | 62.8776 | 113.60 |
| | StableSR-50 [52] | 24.77 | 0.6908 | 0.2591 | 4.5120 | <u>3.1473</u> | <u>0.7067</u> | <u>71.2811</u> | 152.70 |
| | DiffBIR-50 [30] | 25.72 | 0.6695 | 0.2795 | 4.5875 | 3.2260 | 0.6900 | 69.7089 | 385.43 |
| | SeeSR-50 [59] | 26.69 | 0.7422 | <u>0.2187</u> | <u>4.3825</u> | 3.4742 | 0.5868 | 71.2412 | 751.50 |
| | ResShift-4 [68] | **27.33** | **0.7530** | **0.1998** | 5.8700 | 4.3643 | 0.6147 | 65.5860 | 118.59 |
| | SinSR-1 [56] | 26.98 | 0.7304 | 0.2209 | 5.2623 | 3.8189 | 0.6618 | 67.7593 | 118.59 |
| | OSEDiff-1 [58] | 23.95 | 0.6756 | 0.2624 | 4.7157 | 3.3775 | 0.6818 | 70.3928 | 8.50 |
| | *InvSR*-1 (Ours) | 24.14 | 0.6789 | 0.2517 | **4.3815** | **3.0866** | **0.7093** | **72.2900** | 33.84 |
| *RealSR* | BSRGAN [72] | <u>26.51</u> | <u>0.7746</u> | <u>0.2685</u> | 4.6501 | 4.4644 | 0.5439 | 63.5869 | 16.70 |
| | RealESRGAN [54] | 25.85 | 0.7734 | 0.2728 | 4.6766 | 4.4881 | 0.4898 | 59.6803 | 16.70 |
| | LDM-50 [42] | **26.75** | 0.7711 | 0.2945 | 4.8712 | 5.0025 | 0.4907 | 54.3910 | 113.60 |
| | StableSR-50 [52] | 26.27 | **0.7755** | **0.2671** | 5.1745 | 4.8209 | 0.5209 | 60.1758 | 152.70 |
| | DiffBIR-50 [30] | 24.83 | 0.6642 | 0.3864 | **3.7366** | **3.3661** | 0.6857 | 65.3934 | 385.43 |
| | SeeSR-50 [59] | 26.20 | 0.7555 | 0.2806 | 4.5358 | 4.1464 | 0.6824 | <u>66.3757</u> | 751.50 |
| | ResShift-4 [68] | 25.77 | 0.7453 | 0.3395 | 6.9113 | 5.4013 | 0.5994 | 57.5536 | 118.59 |
| | SinSR-1 [56] | 26.02 | 0.7097 | 0.3993 | 6.2547 | 4.7183 | 0.6634 | 59.2981 | 118.59 |
| | OSEDiff-1 [58] | 23.89 | 0.7030 | 0.3288 | 5.3310 | 4.3584 | **0.7008** | 65.4806 | 8.50 |
| | *InvSR*-1 (Ours) | 24.50 | 0.7262 | 0.2872 | <u>4.2189</u> | <u>3.7779</u> | <u>0.6918</u> | **67.4586** | 33.84 |

tained from these results: i) With a fixed number of sampling steps, *e.g.*, one or three, varying the starting timestep enables a trade-off between fidelity (measured by reference metrics) and realism (measured by non-reference metrics). Specifically, using larger starting timesteps favors improved realism at the expense of fidelity. ii) As expected, reference metrics deteriorate with increased sampling steps due to the introduction of additional randomness. iii) Interestingly, non-reference metrics also exhibit a decline when using more sampling steps. This is mainly because most testing images contain some noise, which can lead to undesired artifacts if multiple sampling steps are used, thus degrading the overall image quality. However, using more sampling steps can effectively recover intricate fine-grained structures in cases involving substantial blur, as evidenced by the first examples in Fig. 1 and Fig. II of supp.

**Initial Noise Prediction** Figure 3 presents the noise map predicted by our method for the initial timestep, exhibiting a strong correlation with image content. This visualization aligns well with the theoretical analysis in Sec 3.2.3, empirically validating that our noise predictor can effectively find an LR-dependent noise map to facilitate the SR task.

### 4.3. Comparison with the State of the Art

Considering that recent studies [56, 58] mainly focus on developing one-step diffusion-based methods, we thus evaluate *InvSR* against these methods under a one-step configuration to ensure a fair comparison.

**Synthetic Dataset**. Table 2 reports a comprehensive evalu-

Table 3. Quantitative comparisons of various methods on *RealSet80*. The number of sampling steps is marked in the format of "Method name-Steps" for diffusion-based methods. The best and second-best results are highlighted in **bold** and <u>underlined</u>.

| Methods | Metrics | | | |
|---|---|---|---|---|
| | NIQE↓ | PI↓ | CLIPIQA↑ | MUSIQ↑ |
| BSRGAN [72] | 4.4408 | 4.0276 | 0.6263 | 66.6288 |
| RealESRGAN [54] | 4.1568 | 3.8852 | 0.6189 | 64.4957 |
| LDM-50 [42] | 4.3248 | 4.2545 | 0.5511 | 55.8246 |
| StableSR-50 [52] | 4.5593 | 4.0977 | 0.6214 | 62.7613 |
| DiffBIR-50 [30] | **3.8630** | **3.2117** | **0.7404** | 67.9806 |
| SeeSR-50 [59] | 4.3678 | 3.7429 | 0.7114 | <u>69.7658</u> |
| ResShift-4 [68] | 5.9866 | 4.8318 | 0.6515 | 61.7967 |
| SinSR-1 [56] | 5.6243 | 4.2830 | 0.7228 | 64.0573 |
| OSEDiff-1 [58] | 4.3457 | 3.8219 | 0.7093 | 68.8202 |
| *InvSR*-1 (Ours) | <u>4.0284</u> | <u>3.4666</u> | <u>0.7291</u> | **69.8055** |

ation of various methods on *ImageNet-Test* dataset, encompassing seven quantitative metrics, with additional qualitative comparisons included in Fig. IV of supp. Notably, compared to the recent state-of-the-art (SotA) one-step method OSEDiff [58], *InvSR* demonstrates evident superiority across all the seven metrics. Moreover, even compared to multi-step methods with 50 sampling steps, such as StableSR and DiffBIR, *InvSR* still achieves comparable performance in distortion-oriented metrics, including PSNR and SSIM, while outperforming these methods in perception-oriented metrics, such as LPIPS, NIQE, PI, and MUSIQ. These results indicate that *InvSR* effectively balances both performance and efficiency, advancing the
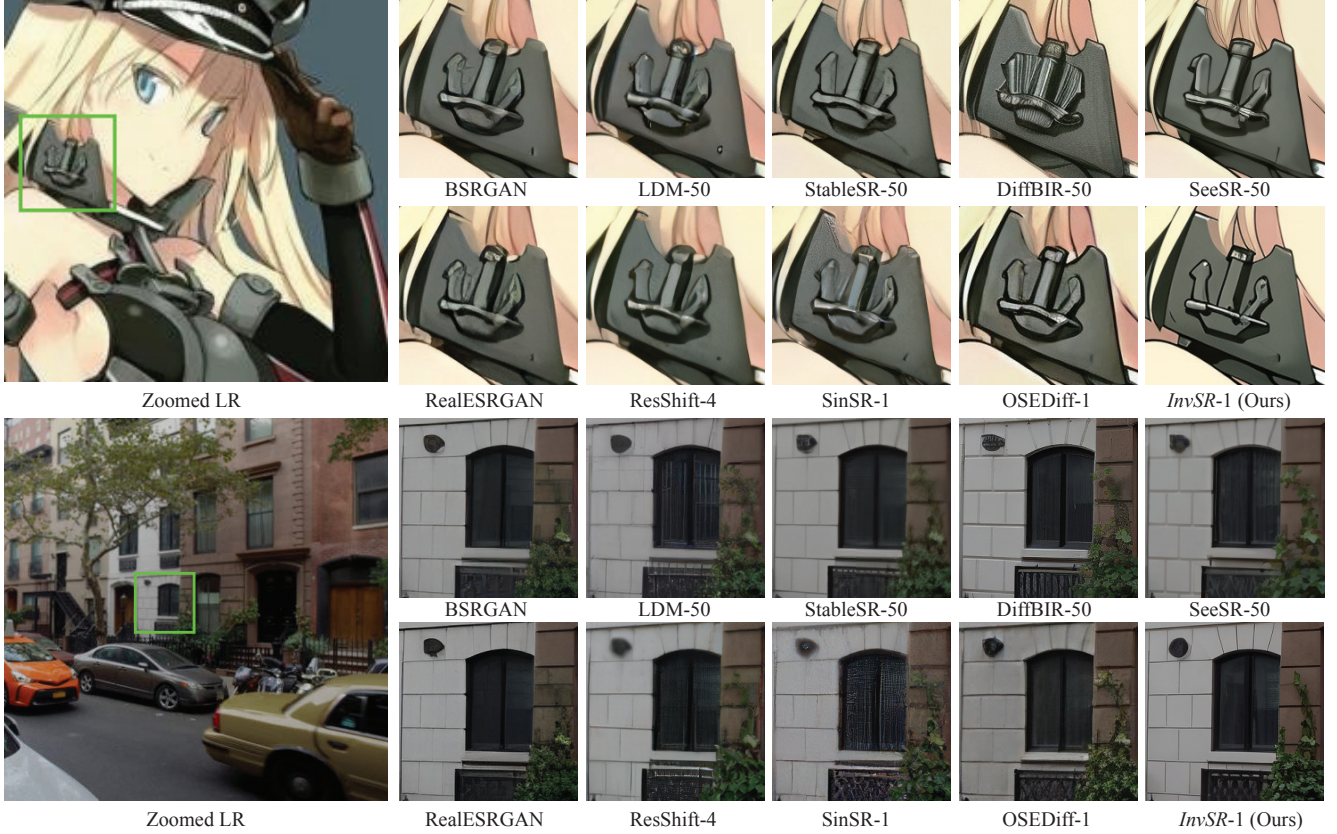
Figure 4. Visual results of different methods on two typical real-world examples from *RealSet80* dataset. For clear comparisons, the number of sampling steps is annotated in the format "Method name-Steps" for diffusion-based approaches. *(Zoom-in for best view)*

field of diffusion-based SR approaches. Additionally, *InvSR* maintains a moderate model size with about 34 million learnable parameters, further enhancing its practicality for real-world applications.

**Real-world Dataset**. To evaluate real-world datasets, we mainly focus on the non-reference metrics. Table 2 and 3 provide a detailed comparison of *InvSR* against recent SotA methods on the datasets of *RealSR* and *RealSet80*, respectively. It can be easily observed that *InvSR* achieves superior performance across most non-reference metrics compared to recent one-step methods under fair comparison and second-best results compared to existing multi-step methods. To further substantiate these conclusions, we present visual comparisons of two real-world examples in Fig. 4, and more examples can be found in Fig. V of supp. In the first example, where the LR image contains evident compression noise, *InvSR* successfully removes these artifacts and generates clear results, while other methods struggle with remaining artifacts. In the second example, which is degraded by noticeable blurriness, *InvSR* produces sharper image structures, such as the tile edges on the wall. These quantitative and qualitative evaluations highlight the great potential of *InvSR* to solve the real-world SR task.

## 5. Conclusion

We proposed *InvSR*, a new SR method based on diffusion inversion. Our method introduces a noise prediction network designed to estimate an optimal noise map, enabling the construction of an intermediate state of a pre-trained diffusion model as the starting sampling point. This design is appealing in two aspects: first, *InvSR* can sufficiently harness the prior knowledge encapsulated in the pre-trained diffusion model, thereby facilitating SR performance. Second, *InvSR* offers a flexible sampling strategy capable of starting from various intermediate states of the diffusion model by incorporating a time-dependent architecture of the noise predictor. This flexibility allows users to freely adjust the sampling steps according to the degradation type or their specific requirements. Even after reducing the sampling steps to just one, *InvSR* still exhibits significant superiority beyond recent one-step diffusion-based methods, suggesting its effectiveness and efficiency.

# References

[1] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 6

[2] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Eur. Conf. Comput. Vis. Worksh.*, pages 0–0, 2018. 6

[3] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Int. Conf. Comput. Vis.*, pages 3086–3095, 2019. 6

[4] Kelvin CK Chan, Xiangyu Xu, Xintao Wang, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for image super-resolution and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3154–3168, 2022. 2, 5

[5] Hamadi Chihaoui, Abdelhak Lemkhenter, and Paolo Favaro. Blind image restoration via fast diffusion inversion. In *Adv. Neural Inform. Process. Syst.*, 2024. 3

[6] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Adv. Neural Inform. Process. Syst.*, pages 25683–25696, 2022. 3

[7] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12413–12422, 2022. 2, 3

[8] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *Int. Conf. Learn. Represent.*, 2023. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009. 6

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12873–12883, 2021. 5, 6

[11] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9935–9946, 2023. 3

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[13] Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. ReNoise: Real image inversion through iterative noising. In *Eur. Conf. Comput. Vis.*, 2024. 2, 3

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Adv. Neural Inform. Process. Syst.*, 27, 2014. 2, 5

[15] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3012–3021, 2020. 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst.*, pages 6840–6851, 2020. 2, 3, 5

[17] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5197–5206, 2015. 6

[18] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Int. Conf. Comput. Vis.*, 2017. 6

[19] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *Int. Conf. Learn. Represent.*, 2024. 3

[20] Wonjun Kang, Kevin Galim, and Hyung Il Koo. Eta inversion: Designing an optimal eta function for diffusion-based real image editing. In *Eur. Conf. Comput. Vis.*, 2024. 3

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4401–4410, 2019. 5

[22] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Adv. Neural Inform. Process. Syst.*, pages 26565–26577, 2022. 2, 4

[23] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Adv. Neural Inform. Process. Syst.*, pages 23593–23606, 2022. 2, 3

[24] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Int. Conf. Comput. Vis.*, pages 5148–5157, 2021. 6

[25] Diederik P Kingma. Auto-encoding variational bayes. In *Int. Conf. Learn. Represent.*, 2014. 3

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent.*, 2015. 5

[27] Jianze Li, Jiezhang Cao, Zichen Zou, Xiongfei Su, Xin Yuan, Yulun Zhang, Yong Guo, and Xiaokang Yang. Distillation-free one-step diffusion for real-world image super-resolution. *arXiv preprint arXiv:2410.04224*, 2024. 3

[28] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Lsdir: A large scale dataset for image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 1775–1787, 2023. 5

[29] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conf. on Applications of Comput. Vision*, pages 5404–5411, 2024. 4

[30] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. DiffBIR:

Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 2, 3, 6, 7

[31] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Int. Conf. Comput. Vis.*, pages 416–423, 2001.

[32] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76:21811–21838, 2017. 6

[33] Barak Meiri, Dvir Samuel, Nir Darshan, Gal Chechik, Shai Avidan, and Rami Ben-Ari. Fixed-point inversion for text-to-image diffusion models. *arXiv preprint arXiv:2312.12540*, 2023. 3

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

[35] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023. 3

[36] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6038–6047, 2023. 2, 3

[37] Naoki Murata, Koichi Saito, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. Gibbsddrm: A partially collapsed gibbs sampler for solving blind inverse problems with denoising diffusion restoration. In *Int. Conf. Mach. Learn.*, pages 25501–25522. PMLR, 2023. 3

[38] Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via image prompting. In *Adv. Neural Inform. Process. Syst.*, 2024. 3

[39] Mehdi Noroozi, Isma Hadji, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. You only need one step: Fast super-resolution with stable diffusion via scale distillation. In *Eur. Conf. Comput. Vis.*, 2024. 3

[40] 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 Trans. Pattern Anal. Mach. Intell.*, 44(11):7474–7489, 2021. 2

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, 2019. 6

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 2, 4, 6, 7

[43] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Eur. Conf. Comput. Vis.*, pages 87–103. Springer, 2024. 5

[44] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Int. Conf. Mach. Learn.*, pages 2256–2265. PMLR, 2015. 2, 3

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Int. Conf. Learn. Represent.*, 2021. 2, 3, 4

[46] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *Int. Conf. Learn. Represent.*, 2023. 3

[47] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Int. Conf. Learn. Represent.*, 2021. 2, 3

[48] Haoze Sun, Wenbo Li, Jianzhuang Liu, Haoyu Chen, Renjing Pei, Xueyi Zou, Youliang Yan, and Yujiu Yang. Coser: Bridging image and language for cognitive super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 25868–25878, 2024. 3

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, 2017. 6

[50] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22532–22541, 2023. 3

[51] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 6

[52] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *Int. J. Comput. Vis.*, pages 1–21, 2024. 3, 6, 7

[53] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3

[54] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Int. Conf. Comput. Vis. Worksh.*, pages 1905–1914, 2021. 5, 6, 7

[55] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *Int. Conf. Learn. Represent.*, 2023. 2, 3

[56] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 25796–25805, 2024. 6, 7

[57] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 6

[58] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Adv. Neural Inform. Process. Syst.*, 2024. 3, 5, 6, 7

[59] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 25456–25467, 2024. 2, 3, 5, 6, 7

[60] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xing-long Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Int. Conf. Comput. Vis.*, pages 13095–13105, 2023. 3

[61] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3121–3138, 2022. 2

[62] Jie Xiao, Ruili Feng, Han Zhang, Zhiheng Liu, Zhantao Yang, Yurui Zhu, Xueyang Fu, Kai Zhu, Yu Liu, and Zheng-Jun Zha. DreamClean: Restoring clean image using deep diffusion prior. In *Int. Conf. Learn. Represent.*, 2024. 3

[63] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *Eur. Conf. Comput. Vis.*, 2024. 2, 3

[64] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *Adv. Neural Inform. Process. Syst.*, 2024. 5

[65] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 25669–25680, 2024. 3

[66] Zongsheng Yue and Chen Change Loy. DifFace: Blind face restoration with diffused error contraction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024. 3

[67] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *Adv. Neural Inform. Process. Syst.*, pages 13294–13307, 2023. 6

[68] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Efficient diffusion model for image restoration by residual shifting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024. 6, 7

[69] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Int. Conf. on Curves and Surfaces*, pages 711–730. Springer, 2012. 6

[70] Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058*, 2024. 3

[71] Guoqiang Zhang, Jonathan P Lewis, and W Bastiaan Kleijn. Exact diffusion inversion via bidirectional integration approximation. In *Eur. Conf. Comput. Vis.*, pages 19–36, 2024. 3

[72] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Int. Conf. Comput. Vis.*, pages 4791–4800, 2021. 5, 6, 7

[73] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018. 5, 6

[74] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *Eur. Conf. Comput. Vis.*, pages 592–608. Springer, 2020. 2