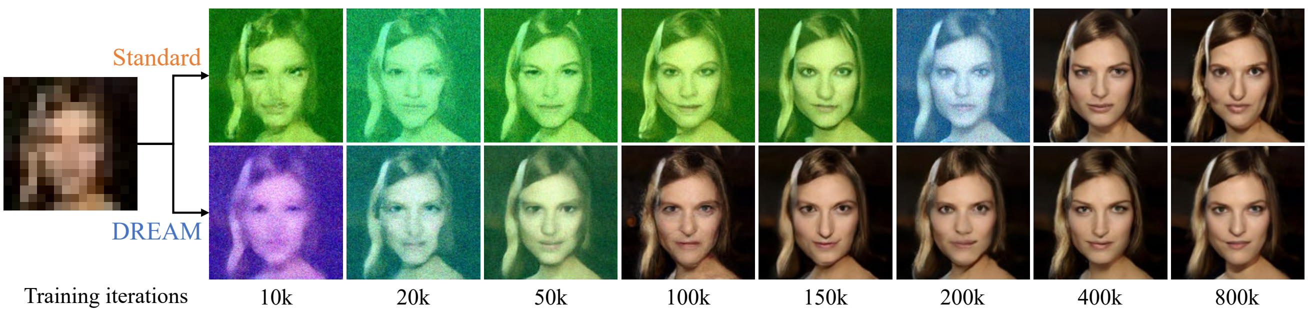
DREAM: Diffusion Rectification and Estimation-Adaptive Models

Jinxin Zhou[1*] Tianyu Ding[2*†] Tianyi Chen[2] Jiachen Jiang[1] Ilya Zharkov[2]
Zhihui Zhu[1] Luming Liang[2†]
[1]Ohio State University [2]Microsoft
{zhou.3820,jiang.2880,zhu.3440}@osu.edu, {tianyuding,tiachen,zharkov,lulian}@microsoft.com



Turning the top to the bottom by adding only ***three*** lines of code (line 4-6 in Algorithm 3).

Figure 1. Comparative training of conditional diffusion models for super-resolution. Top: standard conditional DDPM [44]. Bottom: enhancing the same model training with just *three* additional lines of code, leaving the sampling process unchanged. DREAM facilitates notably faster and more stable training convergence, significantly surpassing baseline models in key metrics of perception and distortion.

Abstract

*We present DREAM, a novel training framework representing **D**iffusion **R**ectification and **E**stimation-**A**daptive **M**odels, requiring minimal code changes (just three lines) yet significantly enhancing the alignment of training with sampling in diffusion models. DREAM features two components:* diffusion rectification*, which adjusts training to reflect the sampling process, and* estimation adaptation*, which balances perception against distortion. When applied to image super-resolution (SR), DREAM adeptly navigates the tradeoff between minimizing distortion and preserving high image quality. Experiments demonstrate DREAM's superiority over standard diffusion-based SR methods, showing a* 2 *to* $3\times$ *faster training convergence and a* 10 *to* $20\times$ *reduction in sampling steps to achieve comparable results. We hope DREAM will inspire a rethinking of diffusion model training paradigms. Our source code is available at link.*

1. Introduction

Single-image super-resolution (SISR) [3, 12, 50, 59] involves generating high-resolution (HR) images from low-resolution (LR) counterparts, a process crucial in various applications including video surveillance, medical diagnosis, and photography. SISR is challenging due to the diverse real-world degradation patterns and the inherent ill-posed nature of the task, where different HR images can correspond to the same LR image.

SISR methods are generally categorized into regression-based and generation-based approaches. Regression-based methods [7, 31, 34, 69] focus on minimizing pixel-level discrepancies, *i.e.*, distortion, between SR predictions and HR references. However, this approach often fails to capture the perceptual quality of images. To address this, generation-based methods employ deep generative models, including autoregressive models [40, 41], variational autoencoders (VAEs) [27, 53], normalizing flows (NFs) [11, 26], and generative adversarial networks (GANs) [16, 24, 33, 42], aiming to improve the perceptual aspects of SR images.

Recently, Diffusion Probabilistic Models (DPMs) [19, 48], a novel class of generative models, have attracted increased interest for their impressive generative abilities, especially in the SISR task [14, 20, 43, 44, 62]. Nonetheless, DPM-based methods face challenges due to their dependence on a long sampling chain, which can lead to error accumulation and reduce training and sampling efficiency. A further issue is the discrepancy between training and sampling [39, 61]: training typically involves denoising noisy images conditioned on ground truth samples, whereas test-

*Equal contribution. This work was done when Jinxin Zhou was an intern at Applied Sciences Group, Microsoft.
†Corresponding author.

ing (or sampling) conditions on previously self-generated results. This disparity, inherent in the multi-step sampling process, tends to magnify with each step, thereby constraining the full potential of DPMs in practice.

To bridge the gap between training and sampling in diffusion models, we introduce DREAM, an end-to-end training framework denoting Diffusion Rectification and Estimation-Adaptive Models. DREAM consists of two key elements: *diffusion rectification* and *estimation adaptation*. Diffusion rectification extends traditional diffusion training with an extra forward pass, enabling the model to utilize its own predictions. This approach accounts for the discrepancy between training (using ground-truth data) and sampling (using model-generated estimates). However, solely relying on this self-alignment can compromise perceptual quality for the sake of reducing distortion. To counter this, our estimation adaptation strategy balances standard diffusion and diffusion rectification by adaptively incorporating ground-truth information. This approach smoothly transitions focus between the two by adaptively injecting ground-truth information. This integration harmonizes the advantages of both approaches, effectively reducing the training-sampling discrepancy, as demonstrated in Figure 3.

The DREAM framework excels in its simplicity, easily integrating into existing diffusion-based models with only three lines of code and requiring no alterations to the network architecture or sampling process. When applied to the SR task, DREAM has notably improved generation quality across various diffusion-based SR methods and datasets. For example, on the $8\times$ CeleA-HQ dataset, it boosts the SR3 [44] method's PSNR from 23.85 dB to 24.63 dB while reducing the FID score from 61.98 to 56.01. Additionally, DREAM accelerates training convergence by 2 to 3 times and improves sampling efficiency, requiring 10 to 20 times fewer steps for comparable or superior results. It also demonstrates enhanced out-of-distribution (OOD) SR results compared to baseline methods.

Our contributions are summarized as follows:

- We introduce DREAM, a simple yet effective framework to alleviate the training-sampling discrepancy in standard diffusion models, requiring minimal code modifications.
- We demonstrate the application of DREAM to various diffusion-based SR methods, resulting in significant improvements in distortion and perception metrics.
- The proposed DREAM also notably speeds up training convergence, enhances sampling efficiency, and delivers superior out-of-distribution (OOD) results.

2. Related work

Super-resolution. In single-image super-resolution, substantial efforts [2, 9, 10, 15, 22, 28, 33, 47, 63, 64, 68, 69] have been devoted to two primary categories: regression-based and generation-based. Regression-based methods, such as EDSR [34], RRDB [57], and SWinIR [31], focus on a direct mapping from LR to HR images, employing pixel-wise loss to minimize differences between SR images and their HR references. While effective in reducing distortion, these methods often yield overly smooth, blurry images. Generation-based methods, on the other hand, aim to produce more realistic SR images. GAN-based models, like SRGAN [28], combine adversarial and perceptual losses [65] to enhance visual quality. Methods of this line include SFTGAN [56] and GLEAN [5], which integrate semantic information to improve texture realism. ESRGAN [57] further refines SRGAN's architecture and loss function. However, GAN-based methods often face challenges like complex regularization and optimization to avoid instability. Autoregressive models (*e.g.*, Pixel-CNN [54], Pixel-RNN [41], VQVAE [55], and LAR-SR [17]) are computationally intensive and less practical for HR image generation. Normalizing Flows (NFs) [11, 26] and VAEs [27, 53] also contribute to the field, but these methods sometimes struggle to produce satisfactory results.

Diffusion model. Inspired by non-equilibrium statistical physics, [48] first proposes Diffusion Probabilistic Models (DPMs) to learn complex distributions. These models have since advanced significantly [8, 19, 37, 49], achieving state-of-the-art results in image synthesis. Beyond general image generation, diffusion models have shown remarkable utility in low-level vision tasks, particularly in SR. Notable examples include SR3 [44], which excels in image super-resolution through iterative refinement, and IDM [14], which blends DPMs with explicit image representations to enable flexible generation across various resolutions. SRDiff [29] uniquely focuses on learning the residual distribution between HR and LR images through diffusion processes. LDM [43] deviates from traditional pixel space approaches, employing cross-attention conditioning for diffusion in latent space. Building upon LDM, ResShift [62] employs a refined transition kernel for sequentially transitioning the residual from LR embeddings to their HR counterparts.

Training-sampling discrepancy. [39] first analyzes the training-sampling discrepancy in unconditional diffusion models, proposing to represent estimation errors with a Gaussian distribution for improved DPM training. This discrepancy was later attributed by [61] to a constant training weight strategy, suggesting a reweighted objective function based on the signal-to-noise ratio at different diffusion steps. In addition, [30] adjusts the distribution during the sampling process by choosing the optimal step within a predefined windows for denoising at each stage. [38] applies a predefined linear function to adjust noise variance during sampling, and [13] recommends starting the sampling from an approximate distribution that mirrors the training process in terms of frequency and pixel space.

Algorithm 1 Conditional DDPM Training

1: **repeat**
2: $(\boldsymbol{x}_0, \boldsymbol{y}_0) \sim p(\boldsymbol{x}_0, \boldsymbol{y}_0), t \sim \mathbb{U}(1, T), \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
3: Compute $\boldsymbol{y}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t$
4: Update θ with gradient $\nabla_\theta ||\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t, t)||_1$
5: **until** converged

Algorithm 2 Conditional DDPM Sampling

1: $\boldsymbol{y}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
2: **for** $t = T \cdots 1$ **do**
3: $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ if $t > 1$ else $\boldsymbol{z} = \mathbf{0}$
4: $\boldsymbol{y}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t, t)) + \sigma_t \boldsymbol{z}$
5: **end for**
6: return $\boldsymbol{y}_0$

Our approach, distinct from previous unconditional methods, addresses discrepancies based on predictions relative to the conditional input data, ensuring a tailored and accurate solution for complex visual prediction tasks like SISR. Our method also draws inspiration from step-unrolling techniques in depth estimation [21, 46] and text generation [45], leveraging the model's own predictions for error estimation. However, we uniquely integrate self-estimation with adaptive incorporation of ground-truth data. This integration, guided by the pattern of estimation errors, effectively balances perceptual quality and distortion, enhancing generated image qualities.

3. Method

3.1. Preliminaries

The goal of SR is to recover a high-resolution (HR) image from its low-resolution (LR) counterpart. This task is recognized as ill-posed due to its one-to-many nature [44, 62], and is further complicated by various degradation models in real-world scenarios. Notably, diffusion models [19, 48] have emerged as powerful generative models, showcasing strong capabilities in image generation tasks. Following [44], we address the SR challenge by adapting a *conditional* denoising diffusion probabilistic (DDPM) model. This adaptation, conditioned on the LR image, sets it apart from traditional, unconditional models which are primarily designed for unconstrained image generation.

We denote the LR and HR image pair as $(\boldsymbol{x}_0, \boldsymbol{y}_0)$. A conditional DDPM model involves a Markov chain, encompassing a *forward process* that traverses the chain, adding noise to $\boldsymbol{y}_0$, and a *reverse process*, which conducts reverse sampling from the chain for denosing from pure Gaussain noise to the HR image $\boldsymbol{y}_0$, conditioned on the LR image $\boldsymbol{x}_0$.

Forward process. The forward process, also referred to as the diffusion process, takes a sample $\boldsymbol{y}_0$ and simulates the non-equilibrium thermodynamic diffusion process [48]. It gradually adds Gaussian noise to $\boldsymbol{y}_0$ via a fixed Markov chain of length T:

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

$$q(\boldsymbol{y}_{1:T}|\boldsymbol{y}_0) = \prod_{t=1}^{T} q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}), \tag{2}$$

where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ is the variance scheduler. As the step t increases, the signal $\boldsymbol{y}_0$ gradually loses its distinguishable features. Ultimately, as $t \to \infty$, $\boldsymbol{y}_t$ converges to an isotropic Gaussian distribution. Moreover, we can derive the distribution for sampling at arbitrary step t from $\boldsymbol{y}_0$:

$$q(\boldsymbol{y}_t|\boldsymbol{y}_0) = \mathcal{N}\left(\boldsymbol{y}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0, (1-\bar{\alpha}_t)\boldsymbol{I}\right). \tag{3}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\alpha_t = 1 - \beta_t$.

Reverse process. The reverse process, also referred to as the denosing process, learns the conditional distributions $p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x}_0)$ for denoising from Gaussian noise to $\boldsymbol{y}_0$ conditioned on $\boldsymbol{x}_0$, through a reverse Markovian process:

$$p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{y}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t, t), \sigma_t^2 \boldsymbol{I}), \tag{4}$$

$$p_\theta(\boldsymbol{y}_{0:T}|\boldsymbol{x}_0) = p(\boldsymbol{y}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x}_0), \tag{5}$$

where σ_t is a predetermined term related to β_t [19].

Training. We train a denoising network $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t, t)$ to predict the noise vector $\boldsymbol{\epsilon}_t$ added at step t. Following [19, 44], the training objective can be expressed as:

$$\mathcal{L}(\theta) = \mathbb{E}_{(\boldsymbol{x}_0, \boldsymbol{y}_0), \boldsymbol{\epsilon}_t, t} \left\| \boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t, t) \right\|_1. \tag{6}$$

With Eq. (3), we parameterize $\boldsymbol{y}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t$, and summarize the training process in Algorithm 1.

Sampling. In essence, the training minimizes the divergence between the forward posterior $q(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{y}_0)$ and $p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x}_0)$, and the mean $\boldsymbol{\mu}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t, t)$ in Eq. (4) is parameterized [44] to match the mean of $q(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{y}_0)$:

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t, t) = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t, t)). \tag{7}$$

To sample $\boldsymbol{y}_0 \sim p_\theta(\boldsymbol{y}_0|\boldsymbol{x}_0)$, starting from $\boldsymbol{y}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, we reverse the Markovian process by iteratively sampling $\boldsymbol{y}_{t-1} \sim p(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x}_0)$ based on Eqs. (4) and (7), which completes the sampling process, as shown in Algorithm 2.

3.2. Challenge: training-sampling discrepancy

Training diffusion models for SR presents a critical challenge, stemming from a discrepancy between the training and inference phases, which we term as *training-sampling discrepancy*. During the training phase, the model operates on actual data, wherein the noisy image $\boldsymbol{y}_t$ at diffusion step t is derived from the *ground-truth* HR image $\boldsymbol{y}_0$ as per line 3 in Algorithm 1. However, during the inference phase, the ground truth $\boldsymbol{y}_0$ is unavailable. As outlined in line 4 in Algorithm 2, the model now operates on predicted data, where $\boldsymbol{y}_t$ is obtained from the preceding sampling step

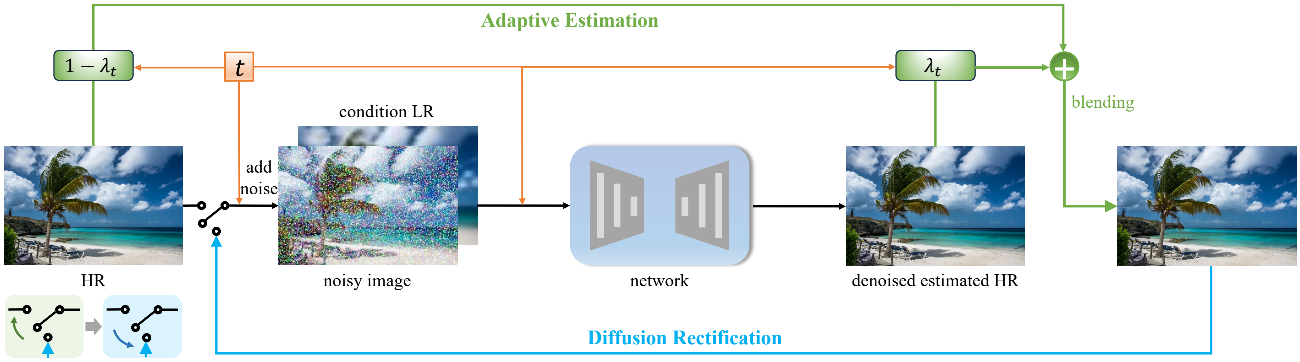


Figure 2. **Overview of the DREAM framework**. Starting with ground-truth HR images, a standard diffusion process with a frozen denoiser network generates denoised HR estimates. The Adaptive Estimation merges these estimated HR images with the original HR images, guided by the pattern of estimation errors. The Diffusion Rectification constructs the noisy images from this merged HR images, which are then fed into the denoiser network (now unfrozen). Similar to DDPM [19], the denoiser network is trained to eliminate both the introduced Gaussian noise and errors arising from the training-sampling discrepancy, as detailed in Eq. (14).

$t+1$. Due to the estimation error, the noisy image $\boldsymbol{y}_t$ constructed in these two processes usually differs, giving rise to the training-sampling discrepancy.

To better illustrate the discrepancy, we conduct an experiment utilizing a pre-trained SR3 model [44], denoted by $\boldsymbol{\epsilon}_\theta$, adhering to the standard diffusion training framework. The goal is to understand the implications for HR signal $\boldsymbol{y}_0$ reconstruction under two distinct scenarios:

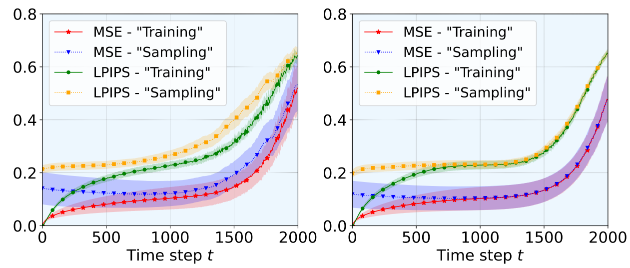
- "Training". Simulating the training process, we *assume access* to the ground-truth $\boldsymbol{y}_0$, and construct the noisy image at time step t as per line 3 in Algorithm 1, denoting the image as $\boldsymbol{y}_t^{\text{train}}$.
- "Sampling". Simulating the sampling process, we *assume no access* to $\boldsymbol{y}_0$ and iteratively construct the noisy image at each time step t by sampling from the previous step, as per line 4 in Algorithm 2. The noisy image thus obtained is denoted by $\boldsymbol{y}_t^{\text{sample}}$.

To retrieve the HR image $\boldsymbol{y}_0$ from the noisy image in both scenarios, we utilize Eq. (3) and the pre-trained network $\boldsymbol{\epsilon}_\theta$ to compute the predicted HR signal as follows:

$$\widetilde{\boldsymbol{y}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{y}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_0, \boldsymbol{y}_t, t\right)\right) =: h_\theta(\boldsymbol{y}_t). \tag{8}$$

Following this, we compute $\widetilde{\boldsymbol{y}}_0^{\text{train}} = h_\theta(\boldsymbol{y}_t^{\text{train}})$ and $\widetilde{\boldsymbol{y}}_0^{\text{sample}} = h_\theta(\boldsymbol{y}_t^{\text{sample}})$ as the predicted HR images in the "training" and "sampling" scenarios, respectively. For performance evaluation, we take 100 samples from FFHQ [25] and calculate the averaged MSE and LPIPS [65] metrics between the predicted HR images and the ground-truth $\boldsymbol{y}_0$ across various time step t under the defined settings.

We present the findings in Figure 3a, where both MSE and LPIPS exhibit a decline with a smaller t, as expected, since the network can reconstruct more accurate HR signal from less noisy input. Importantly, discernible disparities are observed between the curves representing the "training" and "sampling" settings—the "training" curves consistently exhibit lower error compared to the "sampling" ones, suggesting the advantage of having access to the ground-truth $\boldsymbol{y}_0$ for improved prediction accuracy. In contrast, Figure 3b illuminates a remarkable alleviation in this discrepancy when employing our DREAM framework to train the identical SR3 architecture: *the "sampling" curve closely aligns with the "training" curve, despite the lack of access to the ground-truth $\boldsymbol{y}_0$, across both MSE and LPIPS metrics.* This underscores the efficacy of our approach in bridging the training-sampling discrepancy and thereby facilitating more accurate predictions.



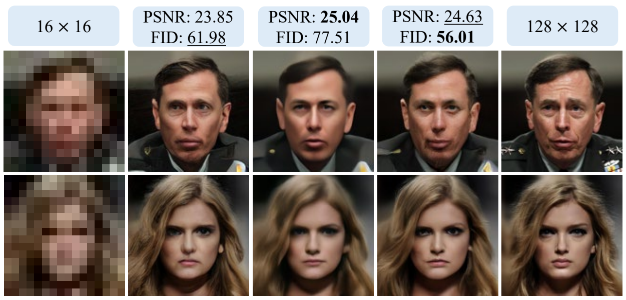
(a) Standard diffusion (b) DREAM

Figure 3. Evaluation of training-sampling discrepancy and its alleviation through our DREAM framework. The mean curve over 100 samples at each time step t is plotted, with the shaded area representing the standard deviation of each metric. Here, $T = 2000$.

3.3. The DREAM framework

We now present our DREAM framework (see Figure 2), an end-to-end training strategy designed to bridge the gap between training and sampling in diffusion models. It consists of two core components: *diffusion rectification* and *estimation adaptation*, which we elaborate as follows.

Diffusion rectification. The goal of diffusion rectification is to modify the behavior of the diffusion training to account for the training-sampling discrepancy, which arises



(a) LR (b) Standard (c) DRM (d) DREAM (e) HR

Figure 4. 8× SR on the CelebA-HQ dataset [23].

from the manner in which we construct the intermediate signals—either from the ground-truth or from the model's own estimation. Hence, we extend the diffusion training framework to align more closely with the sampling process, enabling the model to utilize its own output for prediction.

Specifically, during training, upon acquiring $\boldsymbol{y}_t^{\text{train}}$ as per line 3 in Algorithm 1, we refrain from directly minimizing $\mathcal{L}(\theta)$. Instead, we construct our own prediction of the HR image as $\widetilde{\boldsymbol{y}}_0^{\text{train}}$ according to Eq. (8), formulated as:

$$\begin{aligned}\widetilde{\boldsymbol{y}}_0^{\text{train}} &= \frac{1}{\sqrt{\bar{\alpha}_t}}\big(\boldsymbol{y}_t^{\text{train}} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t^{\text{train}}, t)\big) \\ &= \frac{1}{\sqrt{\bar{\alpha}_t}}\big(\sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t \qquad \triangleright \text{ line } 3 \\ &\qquad - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t^{\text{train}}, t)\big) \\ &= \boldsymbol{y}_0 + \sqrt{(1-\bar{\alpha}_t)/\bar{\alpha}_t}\Delta\boldsymbol{\epsilon}_{t,\theta}\end{aligned} \tag{9}$$

where $\Delta\boldsymbol{\epsilon}_{t,\theta} = \boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t^{\text{train}}, t)$. Utilizing this self-estimated HR image $\widetilde{\boldsymbol{y}}_0^{\text{train}}$, we generate the noisy image $\widetilde{\boldsymbol{y}}_t^{\text{train}}$ to serve as input[1] to the network $\boldsymbol{\epsilon}_\theta$ once more:

$$\begin{aligned}\widetilde{\boldsymbol{y}}_t^{\text{train}} &= \sqrt{\bar{\alpha}_t}\widetilde{\boldsymbol{y}}_0^{\text{train}} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t' \\ &= \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_t}(\boldsymbol{\epsilon}_t' + \Delta\boldsymbol{\epsilon}_{t,\theta}),\end{aligned} \tag{10}$$

where $\boldsymbol{\epsilon}_t' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Then, the training objective for this diffusion rectification model (DRM) can be expressed as:

$$\mathcal{L}^{\text{DRM}}(\theta) = \mathbb{E}_{(\boldsymbol{x}_0,\boldsymbol{y}_0),\boldsymbol{\epsilon}_t,\boldsymbol{\epsilon}_t',t}\left\|\left(\boldsymbol{\epsilon}_t' + \Delta\boldsymbol{\epsilon}_{t,\theta}\right) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \widetilde{\boldsymbol{y}}_t^{\text{train}}, t)\right\|_1. \tag{11}$$

Essentially, Eq. (11) suggests that this DRM approach strives not only to eliminate the sampled noise $\boldsymbol{\epsilon}_t'$ but also to address the error term $\Delta\boldsymbol{\epsilon}_{t,\theta}$ arising from the discrepancy between the imperfect estimation $\widetilde{\boldsymbol{y}}_0^{\text{train}}$ and the ground-truth $\boldsymbol{y}_0$, as seen in Eq. (9); hence the term "rectification". Notably, leveraging the model's own prediction during training as in Eq. (10) mirrors the sampling process of DDIM [49] with a particular choice of σ_t, thereby imposing enhanced supervision. We remark that DRM is closely related to the approaches in [21, 45, 46] where they perform similar step-unrolling techniques for perceptual vision tasks or text generation tasks. However, we are the first to tailor it to low-level vision tasks and provide a clear analysis.

[1] To match the actual sampling process, there might be a desire to reconstruct $\widetilde{\boldsymbol{y}}_{t-1}^{\text{train}}$, yet this could notably complicate the entire procedure. Nonetheless, we have observed similar performance by simply using $\widetilde{\boldsymbol{y}}_t^{\text{train}}$.

Algorithm 3 Conditional DREAM Training

repeat
  $(\boldsymbol{x}_0, \boldsymbol{y}_0) \sim p(\boldsymbol{x}_0, \boldsymbol{y}_0), t \sim \mathbb{U}(1, T), \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  Compute $\boldsymbol{y}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t$
  Compute $\Delta\boldsymbol{\epsilon}_{t,\theta} = \boldsymbol{\epsilon}_t - \texttt{StopGradient}(\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \boldsymbol{y}_t, t))$
  Compute $\widehat{\boldsymbol{y}}_t = \boldsymbol{y}_t + \sqrt{1-\bar{\alpha}_t}\lambda_t\Delta\boldsymbol{\epsilon}_{t,\theta}$
  Update θ with gradient $\nabla_\theta\|\boldsymbol{\epsilon}_t + \lambda_t\Delta\boldsymbol{\epsilon}_{t,\theta} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \widehat{\boldsymbol{y}}_t, t)\|_1$
until converged

Estimation adaptation. While DRM incorporates additional rectification supervision to account for the sampling process, its naive application to the SR task might not deliver satisfactory results. As shown in Figure 4, a distortion-perception tradeoff [4] is observed in the generated SR images. Despite achieving a state-of-the-art PSNR (less distortion), the images produced by DRM tend to be smoother and lack fine details, reflecting a high FID score (poor perception). This is particularly evident when compared to the standard conditional diffusion model, namely SR3 [44]. This limitation could be traced back to DRM's static self-alignment mechanism, which may inappropriately guide the generated images to regress towards the mean.

To address the issue, and inspired by the powerful generative capability of the standard diffusion model, we propose an estimation adaptation strategy. This aims to harness both the superior quality of standard diffusion and the reduced distortion offered by diffusion rectification. Specifically, rather than naively using our own prediction $\widetilde{\boldsymbol{y}}_0^{\text{train}}$ computed in Eq. (9), we adaptively inject ground-truth information $\boldsymbol{y}_0$ by blending it with $\widetilde{\boldsymbol{y}}_0^{\text{train}}$ as follows:

$$\widehat{\boldsymbol{y}}_0 = \lambda_t\widetilde{\boldsymbol{y}}_0^{\text{train}} + (1-\lambda_t)\boldsymbol{y}_0, \tag{12}$$

where $\lambda_t \in (0, 1)$ is an increasing function such that $\widehat{\boldsymbol{y}}_0$ emphasizes more on $\boldsymbol{y}_0$ at smaller t, aligning with the network's tendency to achieve more accurate predictions, as observed in Figure 3. Intuitively, as t decreases, $\widehat{\boldsymbol{y}}_0$ closely approximates the ground-truth, making it more beneficial to resemble the standard diffusion, yielding images with realistic details. Conversely, as t increases and the prediction leans towards random noise, it is advantageous to focus more on the estimation itself, effectively aligning the training and sampling processes through the rectification.

Following the adaptive estimation $\widehat{\boldsymbol{y}}_0$ in Eq. (12), we construct the new noisy image $\widehat{\boldsymbol{y}}_t$ similarly as before:

$$\begin{aligned}\widehat{\boldsymbol{y}}_t &= \sqrt{\bar{\alpha}_t}\widehat{\boldsymbol{y}}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t' \\ &= \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_t}(\boldsymbol{\epsilon}_t' + \lambda_t\Delta\boldsymbol{\epsilon}_{t,\theta}).\end{aligned} \tag{13}$$

Finally, the training objective for our full *Diffusion Rectification and Estimation-Adaptive Model (DREAM)* can be expressed as:

$$\mathcal{L}^{\text{DREAM}}(\theta) = \mathbb{E}_{(\boldsymbol{x}_0,\boldsymbol{y}_0),\boldsymbol{\epsilon}_t,\boldsymbol{\epsilon}_t',t}\left\|\left(\boldsymbol{\epsilon}_t' + \lambda_t\Delta\boldsymbol{\epsilon}_{t,\theta}\right) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \widehat{\boldsymbol{y}}_t, t)\right\|_1. \tag{14}$$

Table 1. Comparison on face and general scene datasets against three baselines for various p values, with best and second-best colorized.

p	CelebA-HQ [23] SR3 [44] PSNR↑	SSIM↑	LPIPS↓	FID↓	CelebA-HQ [23] IDM [14] PSNR↑	SSIM↑	LPIPS↓	FID↓	DIV2K [1] SR3 [44] PSNR↑	SSIM↑	LPIPS↓	FID↓	DIV2K [1] ResShift [62] PSNR↑	SSIM↑	LPIPS↓	FID↓
0 (DRM)	25.04	0.76	0.204	77.51	25.06	0.76	0.188	67.46	28.67	0.81	0.189	16.72	29.98	0.83	0.233	17.76
1 (DREAM)	24.63	0.74	0.177	56.01	24.50	0.73	0.167	53.22	28.10	0.79	0.121	14.32	29.24	0.80	0.158	16.23
2 (DREAM)	24.62	0.74	0.180	61.72	24.32	0.72	0.169	55.38	28.06	0.79	0.140	15.54	28.77	0.79	0.134	15.72
3 (DREAM)	24.15	0.71	0.182	58.89	24.09	0.72	0.172	54.04	27.88	0.79	0.123	14.83	28.44	0.79	0.124	15.67
∞ (standard)	23.85	0.71	0.184	61.98	24.01	0.71	0.172	56.01	27.02	0.76	0.121	16.72	25.30	0.68	0.211	25.91

Choice of λ_t. Comparing Eq. (14) with Eq. (11), the key difference lies in the introduction of λ_t for adaptively modulating the intensity of the rectification term $\Delta\boldsymbol{\epsilon}_{t,\theta}$. Note that we only need $\lambda_t \in (0, 1)$ to be increasing to leverage the benefits of both standard diffusion and rectification. In practice, we set $\lambda_t = (\sqrt{1-\bar{\alpha}_t})^p$, where p adds an extra layer of flexibility: at $p = 0$, λ_t remains at 1, reverting the method to DRM with consistent static rectification; as $p \to \infty$, $\lambda_t \to 0$, transitioning our approach towards the standard diffusion model. As shown in Figure 4, the images produced by DERAM with $p = 1$ achieve a superior balance between perception and distortion, significantly outperforming the standard SR3 [44] across both metrics.

Training details. It's important to highlight that while the same network $\boldsymbol{\epsilon}_\theta$ is utilized for calculating both the rectification term $\Delta\boldsymbol{\epsilon}_{t,\theta}$ and the predicted noise $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_0, \widehat{\boldsymbol{y}}_t, t)$ in Eq. (14), a key distinction exists: we refrain from propagating the gradient when computing $\Delta\boldsymbol{\epsilon}_{t,\theta}$, and thus, it is derived from the frozen network. The actual supervision is imposed following its adaptive adjustment. Moreover, we empirically observe that using the same Gaussian noise (*i.e.*, $\boldsymbol{\epsilon}_t \equiv \boldsymbol{\epsilon}_t'$) in DREAM yields superior performance, further simplifying Eq. (13) to:

$$\widehat{\boldsymbol{y}}_t = \boldsymbol{y}_t^{\text{train}} + \sqrt{1-\bar{\alpha}_t}\lambda_t \Delta\boldsymbol{\epsilon}_{t,\theta}. \tag{15}$$

We summarize our DREAM framework in Algorithm 3, tailored for enhanced diffusion training, while Algorithm 2 remains applicable for sampling purposes.

4. Experiments

4.1. Implementation details

Baselines and datasets. Our experiments involve three diffusion-based SR methods as baselines, spanning datasets for faces, general scenes, and natural images. For face image datasets, we adopt SR3[2] [44] and IDM [14] as baselines, with training conducted on FFHQ [25] and evaluations on CelebA-HQ [23]. For general scenes, we use the DIV2K dataset [1], employing SR3 [44] and ResShift[3] [62] as baseline models. Notably, SR3 and IDM operate in pixel space, whereas ResShift conducts diffusion process in latent space. In addition, to assess out-of-distribution (OOD) performance, we train SR3 as baseline on the DIV2K dataset and evaluate on CAT [66] and LSUN datasets [60].

[2]Due to the unavailability of official code, we use a widely-recognized implementation [link].

[3]To ensure consistency across baselines, we standardize the transition kernel to align with DDPM's approach for noise prediction.

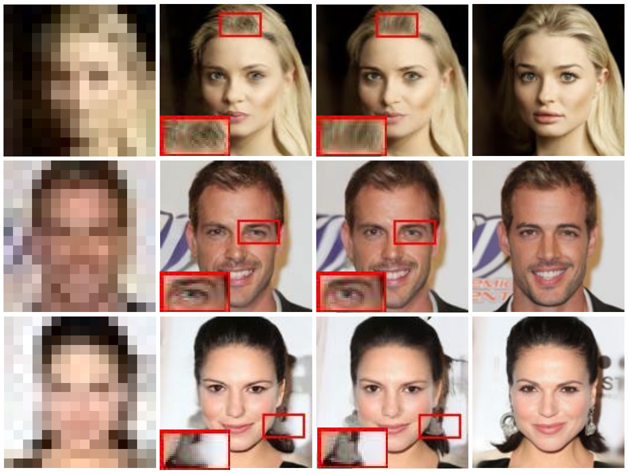
Table 2. Quantitative comparison for 16×16 to 128×128 face super-resolution on CelebA-HQ [23]. Consistency measures the MSE ($\times 10^{-5}$) between LR and downsampled SR images.

Method	PSNR↑	SSIM↑	Consistency↓
PULSE [36]	16.88	0.44	161.1
FSRGAN [6]	23.85	0.71	33.8
Regression [44]	23.96	0.69	2.71
SR3 [44]	23.85	0.71	2.33
IDM [14]	24.01	0.71	2.14
SR3 [44]+DREAM	24.63	0.74	2.12
IDM [14]+DREAM	24.50	0.73	1.26

4.2. Results and analysis

Effect of p in λ_t. In DREAM implementation, we set $\lambda_t = (\sqrt{1-\bar{\alpha}_t})^p$, where p manages the balance between ground-truth and self-estimation data as in Eq. (12). We conduct experiments with three baselines (SR3, IDM and ResShift) for 8× face SR on CelebA-HQ and 4× general scene SR on DIV2K at various p settings, as shown in Table 1. Baselines use the standard diffusion process ($p \to \infty$). For $p = 0$ ($\lambda_t \equiv 1$), corresponding to the DRM model in Eq. (11), there is a notable reduction in distortion (higher PSNR and SSIM), but at the cost of perceptual quality (lower LPIPS and FID), confirming our findings in Figure 4. Increasing p to 1 (our full DREAM approach) leads to a slight decrease in distortion but significantly improves the balance between distortion and perception. Further increase in p shows continual distortion degradation, while perceptual quality initially improves then declines. *DREAM demonstrates clear advantages over baseline models across all metrics.* We found $p = 1$ yields the *best overall performance* compared to other p values and baselines, making it our choice for subsequent experiments.

Face super-resolution. Figures 4 and 5 show qualitative comparisons for face super-resolution from 16 × 16 to 128 × 128, applying our DREAM approach to state-of-the-art diffusion-based methods, SR3 and IDM. While SR3

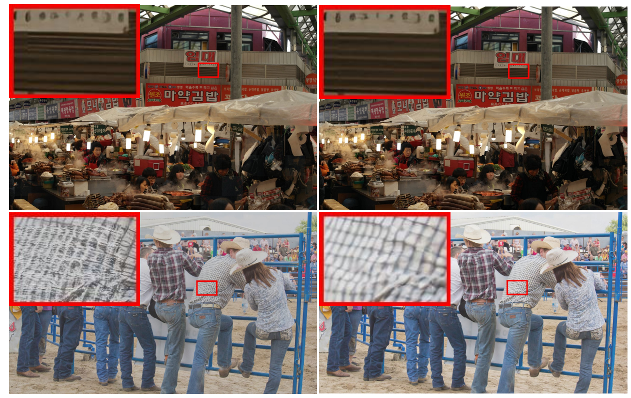


(a) LR (b) Standard (c) DREAM (d) HR

Figure 5. Qualitative comparison for $8\times$ SR using IDM [14] on the CelebA-HQ dataset [23]. Results highlight DREAM's superior fidelity and enhanced identity preservation, leading to more realistic detail generation in features like hair, eyes, and rings.

and IDM generally have decent image qualities, they often miss intricate facial details like hair and eyes, resulting in somewhat unrealistic appearance, and even omit accessories like rings. In contrast, our DREAM approach operated on the these baseline more faithfully preserves facial identity and details. Table 2 shows a quantitative comparison of our DREAM approach applied to SR3 and IDM against other methods, using metrics such as PSNR, SSIM, and consistency [44]. While GAN-based models are known for their fidelity to human perception at higher SR scales, their lower consistency scores suggest a notable deviation from the original LR images. Applying DREAM to SR3 and IDM, we observe considerable enhancements across all metrics. Notably, the simpler SR3, a pure conditional DDPM, when augmented with DREAM, outperforms the more complex IDM, underscoring DREAM's effectiveness.

General scene super-resolution. Figure 6 shows a visual comparison of $4\times$ SR results on the DIV2K dataset [1], using our DREAM approach against standard diffusion methods, with SR3 and ResShift as baselines. Standard training tends to produce images with blurred details and compromised realism, evident in unclear window outlines and distorted shirt textures. In contrast, DREAM maintains structural integrity and delivers more realistic textures. Following [17], we conduct a comprehensive comparison with various regression-based and generative methods on the DIV2K dataset. The results, detailed in Table 3 and benchmarked against models from [32], demonstrate DREAM's effectiveness. Notably, DREAM has led to an increase of 1.08dB and 3.14dB in PSNR, and improvements of 0.03 and 0.11 in SSIM for SR3 and ResShift, respectively, outperforming other generative methods. Moreover, these methods demonstrate comparable or superior performance in perceptual quality metrics, marked by a 0.087 reduction in LPIPS for ResShift. Although LPIPS scores are not as favorable as those obtained by HCFlow++, even with DREAM applied, further improvements in image quality could be achieved through advanced network designs and incorporating GAN loss, as in HCFlow++. However, such approaches are orthogonal to DREAM, and we leave these explorations for future work.



(a) Standard (b) DREAM

Figure 6. Qualitative comparison for $4\times$ SR on DIV2K [1]. Top with SR3 [44] the baseline; bottom with ResShift [62] the baseline.

Table 3. Quantitative comparison for $4\times$ SR on DIV2K. All models are trained on DIV2K plus Flickr2K [52]. The best and second-best results among generative models are colorized.

	Method	PSNR↑	SSIM↑	LPIPS↓
	Bicubic	26.7	0.77	0.409
Reg.-based	EDSR [34]	28.98	0.83	0.270
	RRDB [57]	29.44	0.84	0.253
GAN-based	ESRGAN [57]	26.22	0.75	0.124
	RankSRGAN [67]	26.55	0.75	0.128
Flow-based	SRFlow [35]	27.09	0.76	0.121
	HCFlow [32]	27.02	0.76	0.124
Flow+GAN	HCFlow++ [32]	26.61	0.74	0.110
Diffusion	SR3 [44]	27.02	0.76	0.121
	SR3 [44]+DREAM	28.10	0.79	0.121
	ResShift [62]	25.30	0.68	0.211
	ResShift [62]+DREAM	28.44	0.79	0.124

4.3. Training and sampling acceleration

The DREAM strategy not only improves SR image quality but also accelerates the training. As shown in Figure 1, DREAM reaches convergence at around 100k to 150k iterations, a significant improvement over the standard diffusion-based SR3's 400k iterations. Moreover, Figure 7 illustrates the evolution of training in terms of distortion metrics (PSNR and SSIM) and perception metrics (LPIPS and FID) using SR3 as the baseline on the DIV2K dataset. DREAM not only converges faster but also surpasses SR3's final results before its own convergence. For example, DREAM achieves a PSNR of 28.07 and FID of

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 6, 7, 1, 2, 3, 4, 5

[2] Saeed Anwar and Nick Barnes. Densely residual laplacian super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1192–1204, 2020. 2

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

[4] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 5

[5] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14245–14254, 2021. 2

[6] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2492–2501, 2018. 6

[7] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 1

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[9] Tianyu Ding, Luming Liang, Zhihui Zhu, and Ilya Zharkov. Cdfi: Compression-driven network design for frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8001–8011, 2021. 2

[10] Tianyu Ding, Luming Liang, Zhihui Zhu, Tianyi Chen, and Ilya Zharkov. Sparsity-guided network design for frame interpolation. *arXiv preprint arXiv:2209.04551*, 2022. 2

[11] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. *arXiv:1605.08803*, 2016. 1, 2

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014. 1

[13] Martin Nicolas Everaert, Athanasios Fitsios, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. Exploiting the signal-leak bias in diffusion models. *arXiv preprint arXiv:2309.15842*, 2023. 2

[14] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10021–10030, 2023. 1, 2, 6, 7, 8, 3

[15] Zhicheng Geng, Luming Liang, Tianyu Ding, and Ilya Zharkov. Rstt: Real-time spatial temporal transformer for space-time video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17441–17451, 2022. 2

[16] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *NIPS*, 2014. 1

[17] Baisong Guo, Xiaoyun Zhang, Haoning Wu, Yu Wang, Ya Zhang, and Yan-Feng Wang. Lar-sr: A local autoregressive model for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1909–1918, 2022. 2, 7

[18] 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. *Advances in neural information processing systems*, 30, 2017. 1

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2, 3, 4

[20] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 1

[21] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. *arXiv preprint arXiv:2303.17559*, 2023. 3, 5

[22] Younghyun Jo, Seoung Wug Oh, Peter Vajda, and Seon Joo Kim. Tackling the ill-posedness of super-resolution through adaptive target generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16236–16245, 2021. 2

[23] 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. 5, 6, 7, 1, 3

[24] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 1

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 4, 6

[26] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. In *NIPS*, 2018. 1, 2

[27] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2013. 1, 2

[28] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken,

Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[29] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 2, 1

[30] Mingxiao Li, Tingyu Qu, Wei Sun, and Marie-Francine Moens. Alleviating exposure bias in diffusion models through sampling with shifted time steps. *arXiv preprint arXiv:2305.15583*, 2023. 2

[31] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1, 2

[32] Jingyun Liang, Andreas Lugmayr, Kai Zhang, Martin Danelljan, Luc Van Gool, and Radu Timofte. Hierarchical conditional flow: A unified framework for image super-resolution and image rescaling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4076–4085, 2021. 7

[33] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5657–5666, 2022. 1, 2

[34] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 1, 2, 7

[35] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 715–732. Springer, 2020. 7

[36] 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*, pages 2437–2445, 2020. 6, 1

[37] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[38] Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. *arXiv preprint arXiv:2308.15321*, 2023. 2

[39] Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. *arXiv preprint arXiv:2301.11706*, 2023. 1, 2

[40] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv preprint arXiv:1609.03499*, 2016. 1

[41] Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional Image Generation with PixelCNN Decoders. In *NIPS*, 2016. 1, 2

[42] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1

[43] 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*, pages 10684–10695, 2022. 1, 2

[44] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 1, 2, 3, 4, 5, 6, 7

[45] Nikolay Savinov, Junyoung Chung, Mikolaj Binkowski, Erich Elsen, and Aaron van den Oord. Step-unrolled denoising autoencoders for text generation. In *International Conference on Learning Representations*, 2022. 3, 5

[46] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023. 3, 5

[47] Jae Woong Soh, Gu Yong Park, Junho Jo, and Nam Ik Cho. Natural and realistic single image super-resolution with explicit natural manifold discrimination. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8122–8131, 2019. 2

[48] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1, 2, 3

[49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2, 5

[50] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Gradient profile prior and its applications in image super-resolution and enhancement. *IEEE Transactions on Image Processing*, 20(6):1529–1542, 2010. 1

[51] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 1

[52] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017. 7, 1, 2

[53] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *NeurIPS*, 2020. 1, 2

[54] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016. 2

[55] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[56] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. 2

[57] 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. 2, 7

[58] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[59] Qing Yan, Yi Xu, Xiaokang Yang, and Truong Q Nguyen. Single image superresolution based on gradient profile sharpness. *IEEE Transactions on Image Processing*, 24(10): 3187–3202, 2015. 1

[60] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 6, 8

[61] Hu Yu, Li Shen, Jie Huang, Man Zhou, Hongsheng Li, and Feng Zhao. Debias the training of diffusion models. *arXiv preprint arXiv:2310.08442*, 2023. 1, 2

[62] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 3, 6, 7, 5

[63] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3217–3226, 2020. 2

[64] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 2

[65] 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 conference on computer vision and pattern recognition*, pages 586–595, 2018. 2, 4

[66] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection-how to effectively exploit shape and texture features. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV 10*, pages 802–816. Springer, 2008. 6, 8

[67] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 7

[68] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 2

[69] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 1, 2