Self-Adaptive Reality-Guided Diffusion for Artifact-Free Super-Resolution

Qingping Zheng[1], Ling Zheng[3], Yuanfan Guo[2], Ying Li[1]*, Songcen Xu[2], Jiankang Deng[2], Hang Xu[2]

[1]Northwestern Polytechnical University [2]Huawei Noah's Ark Lab [3]Tsinghua-Fuzhou Institute for Data Technology

Abstract

*Artifact-free super-resolution (SR) aims to translate low-resolution images into their high-resolution counterparts with a strict integrity of the original content, eliminating any distortions or synthetic details. While traditional diffusion-based SR techniques have demonstrated remarkable abilities to enhance image detail, they are prone to artifact introduction during iterative procedures. Such artifacts, ranging from trivial noise to unauthentic textures, deviate from the true structure of the source image, thus challenging the integrity of the super-resolution process. In this work, we propose **S**elf-**A**daptive **R**eality-**G**uided **D**iffusion (SARGD), a training-free method that delves into the latent space to effectively identify and mitigate the propagation of artifacts. Our SARGD begins by using an artifact detector to identify implausible pixels, creating a binary mask that highlights artifacts. Following this, the Reality Guidance Refinement (RGR) process refines artifacts by integrating this mask with realistic latent representations, improving alignment with the original image. Nonetheless, initial realistic-latent representations from lower-quality images result in over-smoothing in the final output. To address this, we introduce a Self-Adaptive Guidance (SAG) mechanism. It dynamically computes a reality score, enhancing the sharpness of the realistic latent. These alternating mechanisms collectively achieve artifact-free super-resolution. Extensive experiments demonstrate the superiority of our method, delivering detailed artifact-free high-resolution images while reducing sampling steps by 2×. We release our code at* https://github.com/ProAirVerse/Self-Adaptive-Guidance-Diffusion.git.

1. Introduction

Super-resolution (SR) techniques strive to restore high-resolution images from their corresponding low-resolution counterparts [8, 29]. This process is inherently ill-posed since the same LR image can yield many plausible HR solutions. Traditional deep learning SR methods, leveraging techniques like squeeze layers [49], transposed convolutions [38], and sub-pixel convolutions [29], bind themselves to fixed upsampling scales, limiting output resolution adjustments without architectural changes. This causes inflexibility in real-world applications. To tackle the inherent one-to-many inverse challenge in SR, it is intuitive to envision the solution space as a distribution. Yet, this perspective often prompts generative models to introduce details that deviate from the authentic content of the original low-resolution image. As a result, a wide range of generative-based SR approaches have emerged to address this intrinsically ill-posed problem, particularly in diffusion models.

(a) Variation of Artifacts within StableSR

(b) Variation of Artifacts within Our SARGD

Figure 1. **Visual comparison of artifact variability between (a) StableSR and (b) Our SARGD method.** In the lower panel, the left image is the decoded output, while the right image is the artifact mask. Regions containing artifacts are highlighted in red, utilizing a binary map to serve as the artifact mask. Our SARGD method exhibits a superior capability in reducing artifacts.

Diffusion-based SR (Diffusion-SR for short) has exhibited significant potential in improving the resolution of images from their low-resolution counterparts. SRDiff [17] is the first diffusion-based model for image super-resolution, progressively transforming the Gaussian noise into a high-resolution (HR) image conditioned on its low-resolution (LR) input through a Markov chain. Building on the foundational concept of denoising score matching [31], Saharia *et al*. [28] propose SR3 model. This model advances the field of super-resolution by learning an empirical data distribution and leveraging U-Net architec-

*Corresponding author

ture [27] to progressively denoise and refine the image output. However, these methods fail to preserve accurate pixel-wise image structures, often necessitating further training or additional skipped connections for detail reproduction. While the pixel-aware stable diffusion network proposed by Yang *et al*. [37] significantly improves the robustness of the super-resolution process, it concurrently introduces the challenge of generating implausible content. In contrast, StableSR [32], the current state-of-the-art, leverages pre-trained diffusion priors to retain image details, avoiding assumptions about image degradation.

Nevertheless, as depicted in Figure 1, contemporary diffusion-based StableSR models invariably generate implausible content, frequently resulting in perceptual artifacts within specific areas of the image. Indeed, these artifacts frequently stem from the fundamental use of Denoising Diffusion Probabilistic Models (DDPMs) [13, 30] in modern diffusion models. Specifically, DDPMs, categorized as Partial Differential Equation (PDE)-based models [6], are designed to diffuse pixel intensities in order to reduce noise levels. They can enhance the preservation of edges and textural features by integrating non-linear terms, thus improving image details. However, despite these advancements, the introduction of non-linearity can occasionally yield unexpected outcomes, particularly in areas with intricate textures or noise patterns. This unintended consequence can result in a loss of fine details and a degradation of essential image elements such as edges and textures, especially when the diffusion process is overly aggressive or inadequately constrained. Additionally, any deficiencies in the model's architecture or parameter settings can be amplified during diffusion, aggravating artifact problems and producing results that deviate from the desired photorealistic quality [36]. These challenges pose a significant obstacle to the preservation of the authentic detail and clarity expected in high-resolution images.

In this work, we propose ***S**elf-**A**daptive **R**eality-**G**uided **D**iffusion* (SARGD), a pioneering training-free approach, aiming to produce "artifact-free" SR images. In particular, SARGD operates through two stages: Reality-Guided Refinement (RGR) to reduce artifacts and Self-Adaptive Guidance (SAG) to boost image fidelity. During the initial RGR phase, an artifact detector is employed to identify unrealistic pixels within the latent space at each inference step, leading to a binary artifact mask. This mask collaborates with a realistic latent guidance extracted from an upscaled LR image to refine artifacts and enhance intrinsic details. However, depending on upscaled LR inputs for guidance can potentially result in over-smoothed images, as it may overlook high-frequency details that are essential for SR. To alleviate this limitation, the subsequent SAG phase introduces a 'reality score to self-adjust the authenticity of the latent guidance, significantly boosting the accuracy of detail and texture representation. Two phases alternate to progressively achieve high-fidelity super-resolution, effectively eliminating artifacts in the diffusion process. To sum up, the contributions of this paper are as follows:

- We introduce a Reality-Guided Refinement (RGR) strategy, designed to identify and rectify artifacts during the diffusion process with the aid of a reality guidance mechanism, effectively eliminating such artifacts.
- We propose a Self-Adaptive Guidance (SAG) mechanism that iteratively enhances the realistic-latent guidance, addressing over-smoothing issues and improving the fidelity and authenticity of the reference latent guidance.
- We are the first to develop the ***S**elf-**A**daptive **R**eality-**G**uided **D**iffusion* (SARGD), a training-free approach that sets a new benchmark by effectively dealing with artifacts and over-smoothing issues. Additionally, it outperforms StableSR in image quality while reducing the inference time by 2$\times$ for super-resolution tasks.

2. Related Work

In this section, we summarize past super-resolution (SR) techniques, explore diffusion-based models in SR, and address artifact detection and mitigation in the SR process.

Image Super-Resolution. Single image super-resolution (SISR) is a process designed to construct high-resolution (HR) images from their low-resolution (LR) counterparts, which is inherently a complex task due to the multiple HR images that can correspond to a single LR image. The literature has seen a variety of solutions aiming to learn mappings from LR to HR images. These solutions fall into two primary groups: PSNR-oriented methods and generative model-based methods. PSNR-oriented methods [8, 9, 18, 20, 29, 48] employ L1 or L2 norms as objectives, achieving notable PSNR metrics. However, the reliance on L1/L2 losses tends to predict an average of all possible HR images [15, 16, 22], resulting in over-smoothed outputs. In contrast, generative model-based methods [2, 12, 16, 17, 19, 22, 23, 33, 34, 42, 45] seek to address the ill-posed nature of the SR problem by learning the distribution of potential HR images through techniques such as GAN-based, diffusion-based, flow-based, and AR-based super-resolution. These models aim to generate HR images that not only have high pixel accuracy but also possess the rich textural details necessary for visual realism.

Diffusion Models. Diffusion probabilistic models [13, 30] are a class of generative models that utilize a Markov chain to convert latent variables from simple distributions (*e.g*., Gaussian) into data with complex distributions. Recognizing the potential of these models to address "one-to-many" mapping problems, Li *et al*. [17] have applied them to the task of super-resolution, generating a variety of high-resolution images from a single low-resolution input. This

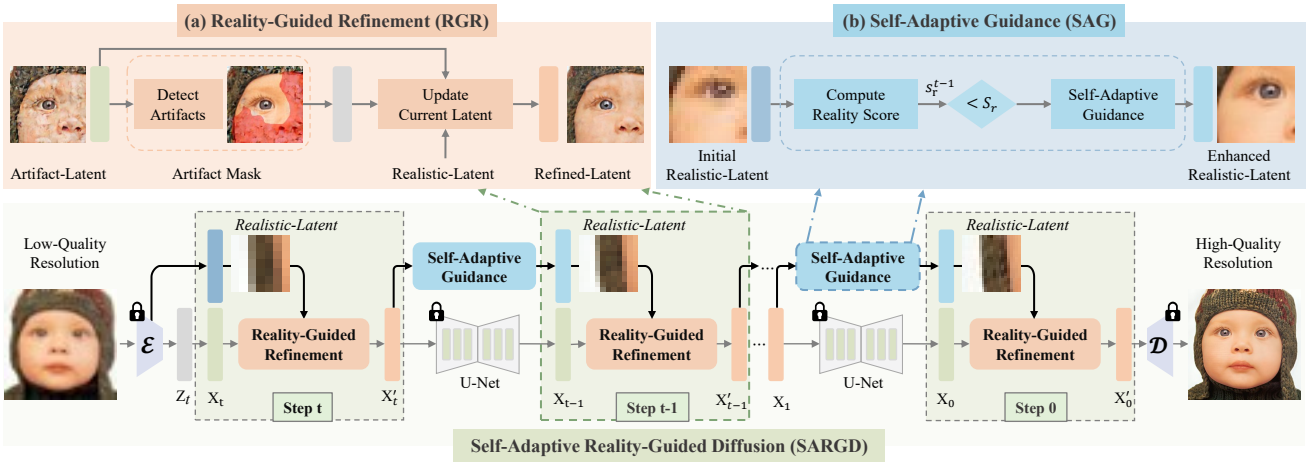


Figure 2. **Self-Adaptive Reality-Guided Diffusion (SARGD)** for artifact-free super-resolution. Our proposed SARGD is a training-free approach that consists of two principal components: 1) a Reality-Guided Refinement (**RGR**) that identifies and corrects artifacts within the latent representation by using a realistic latent as a guide to maintain the inherent details of the original image during the diffusion process, and 2) a Self-Adaptive Guidance (**SAG**) mechanism that enhances the fidelity of the initial realistic latent guidance, derived from upscaled low-resolution images, thereby effectively addressing the issue of over-smoothing in the final outputs.

application not only yields diverse outcomes but also simultaneously addresses challenges related to over smoothing, mode collapse, and the extensive computational footprint that are often encountered in super-resolution tasks. A prevalent method [26, 28] in diffusion-based SR involves integrating a low-resolution image directly into the input of existing diffusion models (*e.g.*, DDPM [13]), and retraining the model with super-resolution data from scratch. Another approach [4, 5, 40] leverages an unconditional pretrained diffusion model as a prior and adapts its reverse diffusion process to synthesize the desired high-resolution image. Nonetheless, both strategies are constrained by the intrinsic Markov chain sequence of DDPM, which can potentially lead to artifacts during inference, thus affecting the quality of the resulting high-resolution image.

Artifact Detection Methods. Despite impressive improvement, diffusion-based super-resolution methods tend to present undesired outcomes, known as artifacts. The artifact detection approach aims to probe the perceptually implausible contents rooted in the generated images. Existing deep-learning-based image artifact detection methods can be divided into spatial and frequency domain-based approaches. The former approaches [39] focus on capturing differences in texture between real and generated images. Liu *et al*. [21] leverage Gram matrices to capture the difference in texture between real and synthetic images. Dang *et al*. [7] adopt the attention mechanism to improve the accuracy of artifact detection by highlighting informative regions. Zhao *et al*. [50] utilize the attention mechanism to amplify subtle differences in the shallow layers, improving the detection performance. The latter approaches [46] study the artifacts in the frequency domain. Durall *et al*. [10] classify generated images according to the characteristics of the artifacts in high-frequency components. Dzanic *et al*. [11] demonstrate that image artifacts could be detected by leveraging the degree of partial decay at high frequencies. Frank *et al*. [11] find that the use of upsampling operations causes the artifacts in generated images, and employs the DCT transforms to detect them.

3. Methodology

In this section, we propose a novel training-free method called "Self-Adaptive Reality-Guided Diffusion" (SARGD) to tackle the ongoing challenge of artifact suppression in diffusion-based super-resolution. This approach, depicted in Figure 2, introduces two essential mechanisms.

- **Reality-Guided Refinement (RGR)**, mitigates artifact emergence during inference by employing guidance from a realistic latent space, thereby preserving the original characteristics of the images.
- **Self-Adaptive Guidance (SAG)**, enhances the authenticity of the reference realistic latent representation, effectively addressing the over-smoothing tendencies associated with the initial realistic latent representation.

3.1. Preliminaries

Diffusion-based super-resolution (SR) techniques [17, 28, 32] have emerged as a compelling paradigm for the enhancement of image resolutions. Consider the diffusion process as a function that aims to translate a given low-resolution (LR) image into a high-resolution (HR) image, each successive HR image at time $t+1$ is obtained by refining the current estimate, factoring in the original LR image. This refinement is mathematically represented as:

$$\mathbf{I}_{HR}^{(t+1)} = \mathbf{I}_{HR}^{(t)} + f_\theta(\mathbf{I}_{HR}^{(t)}, \mathbf{I}_{LR}), \tag{1}$$

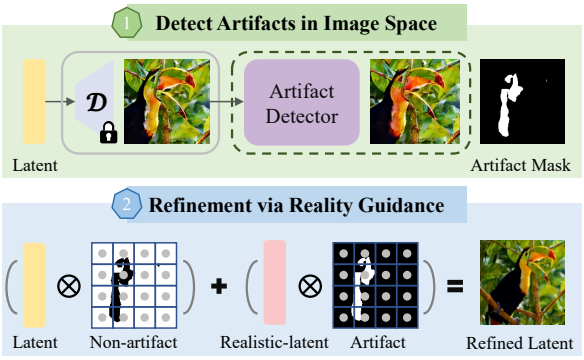


Figure 3. **Overview of the Reality-Guided Refinement (RGR) workflow**: 1) Decoding the current latent into an RGB image and using an artifact detector to create a binary mask identifying areas with artifacts; and 2) Utilizing realistic latent guidance to refine the masked regions, enhancing the image's fidelity and authenticity.

where $\mathbf{I}_{HR}^{(t)}$ is the current state of the high-resolution image at iteration t. f_θ is the diffusion function parameterized by θ, which updates the current high-resolution state based on both its current state and the given low-resolution image. t is the current iteration during the inference.

In the iterative diffusion process for super-resolution, the input LR-HR image pairs in the training set are used to train the diffusion models with the total diffusion step T. The diffusion process initially leverages a pre-trained visual encoder to convert the upsampled LR image into a latent $z_t = \mathcal{E}(up(\mathbf{I}_{LR}))$, and gradually adding the Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ into it, resulting the noised latent variable $\boldsymbol{x}$. The reverse diffusion process aims to guide the generation to the corresponding HR image through iterative refinement of the noised latent $\boldsymbol{x}$, taking the form of:

$$\boldsymbol{y}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta\left(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t\right)\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t, \tag{2}$$

where $\alpha_t \in (0,1)$ is the corresponding coefficient and γ_t is the variance of the noise at iteration t. Lastly, the decoder $\mathcal{D}$ is employed to project the resulting latent into the final high-resolution image.

3.2. Reality-Guided Refinement (RGR)

To overcome the prevalent dilemma of aggravated noise and the emergence of artifacts in images subjected to super-resolution (SR) via diffusion-based techniques, we propose a pioneering training-free approach: Reality-Guided Refinement (RGR). This method couples an artifact detector with a latent refinement to address these issues effectively. As illustrated in Figure 3, a pixel-level segmentation model is initially applied to locate any artifacts on the current latent variable. Once detected, these imperfections are adeptly rectified through a sophisticated procedure of realistic-latent guidance refinement. This diffusion process ensures that the resulting images retain the original, authentic textures and intricate details, effectively circumventing the usual distortions and artifacts associated with traditional diffusion-based upsampling techniques for SR tasks.

Artifact Detection in Diffusion Process. In the context of pixel-sensitive super-resolution, it is crucial to identify the fine-grained artifacts during the diffusion process. Hence, a robust artifact detection system should precisely pinpoint these implausible pixels within the latent space for effective identification. To achieve this, we utilize the established artifact detection method named PAL [43] to isolate these imperfections in the latent space. The process entails transforming the latent variable into the image domain at each iteration using the decoder, depicted by the following equation:

$$\mathbf{I}_{HR}^{(t-1)} = \mathcal{D}(\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\boldsymbol{x_t}, \boldsymbol{x}, t)\right) + \sigma_\theta(x_t, t)\boldsymbol{\epsilon})), \tag{3}$$

Subsequently, the resulting image is input into the PAL detector, yielding a binary mask E_A that denotes the presence of artifacts (referenced as line 12 in Algorithm 1). This process can be formulated as:

$$E_A^{(t-1)} = \mathcal{A}(\mathbf{I}_{HR}^{(t-1)}). \tag{4}$$

Finally, the resultant artifact mask is resized to match the scale of the latent variable, maintaining the proportionality between the two.

Refinement via Reality Guidance. For the removal of artifacts during sampling, we refine the latent variable via a proposed training-free RGR. Following Algorithm 1, the RGR inference, consisting of T steps, begins by taking an upscaled low-resolution image $up(\mathbf{I}_{LR})$ as the starting point. A noise vector $\boldsymbol{\epsilon}$ is then drawn from a standard Gaussian distribution. The process iterates backward from $t = T$, with each step generating a latent variable $\boldsymbol{x}_{t-1}$ in accordance with DDPM [13] framework. Once artifacts E_A are identified within the current latent $\boldsymbol{x}_{t-1}$, we proceed to refine these imperfections through reality guidance. This correction process unfolds as follows:

$$\boldsymbol{x}_{t-1} = \boldsymbol{x}_{t-1} \times (1 - E_A) + \boldsymbol{x}_r \times E_A. \tag{5}$$

This process allows us to effectively eliminate artifacts from the latent space. Notably, in this formula, we employ the initially encoded latent, which has not undergone the denoising process, as a source of realistic latent guidance, since it retains the intrinsic structural integrity of the image.

3.3. Self-Adaptive Guidance (SAG)

However, relying on the initial latent space as a guide for artifact suppression during inference often results in over-smoothing results, primarily attributed to reliance on upscaled low-resolution images. To address this, we propose a

novel Self-Adaptive Guidance (SAG) renewal mechanism, an automated system that perpetually refines realistic latent representations, ensuring the guidance is continually calibrated to enhance realistic detail. This mechanism is indispensable for diffusion-based super-resolution, aiming to yield high-resolution and high-fidelity images that strike a balance between details and sharpness.

Reality Score Computation. To regulate the update mechanism of reality guidance in image super-resolution, we introduce an innovative metric known as the reality score. This metric quantifies the realism of latent guidance. As delineated in Algorithm 1, we begin with an artifact-free latent variable $\boldsymbol{x}_{t-1}$and utilize our decoding function to synthesize a realistic image $\mathbf{I}_r = \mathcal{D}(\boldsymbol{x}_{t-1})$. Subsequently, we deploy a realistic-region function $\mathcal{R}$ by inverting the output of PAL [43] on $\mathbf{I}_r$, resulting in a binary reality mask M_R with dimensions corresponding to the channels, height, and width of the latent $C \times H \times W$, respectively. The reality score $\boldsymbol{s}_r$ at the current $t-1$ step is calculated as the normalized sum of the reality mask across the spatial dimensions of the latent, indicating the distribution of realistic features. The formula is given by:

$$s_r^{t-1} = \mathcal{S}(M_R) = \frac{1}{H \times W} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} M_R^{(ij)}. \tag{6}$$

Herein, H and W are the height and width of the latent, and $M_R^{(ij)}$ indicates the reality measure at each pixel. This approach ensures a quantitative evaluation of the SR process, focusing on producing images that boast both high resolution and a high degree of fidelity to reality.

Self-Adaptive Reality Guidance. Self-adaptability in the realistic latent space refers to the ability of the system to adjust the reality guidance to correct the details and textures that are characteristic of high-quality images. This is especially important to mitigate the smoothing effect of the initial realistic-latent reliance on bicubic upsampling. As outlined in Algorithm 1, we begin by considering a realistic image $\mathbf{I}_r$ at the current step, decoded from an artifact-free latent variable $\boldsymbol{x}_{t-1}$. This artifact-free image is then encoded into a realistic latent $\boldsymbol{x}_r^{t-1} = \mathcal{E}(\mathbf{I}_r)$. Grounded on the computed reality score, self-adaptability is characterized as a guidance refinement process. During this process, the realistic latent $\boldsymbol{x}_r$ at the current $t-1$ step is iteratively refined to converge towards an ideal latent representation $\boldsymbol{x}r$, precisely aligning with the high-resolution output $\mathbf{I}HR$. The equation below illustrates this adaptive mechanism:

$$\mathcal{G}(\boldsymbol{x}_r, \boldsymbol{x}_r^{t-1}) = \boldsymbol{x}_r \times (1 - M_R) + \boldsymbol{x}_r^{t-1} \times M_R, \text{ if } \boldsymbol{s}_r^{t-1} > \boldsymbol{s}_r, \tag{7}$$

where $\boldsymbol{s}_r$ represents the reality score associated with $\boldsymbol{x}r$, and M_R is the binary reality mask corresponding to $\boldsymbol{x}r^{t-1}$. This self-adaptive mechanism ensures that the guidance is autonomously updated, leading to enhanced super-resolution outcomes.

Algorithm 1 Self-Adaptive Reality-Guided Diffusion

Input: LR image $\mathbf{I}_{LR}$, and total diffusion steps T
Load: Encoder $\mathcal{E}$, artifact detector $\mathcal{A}$ and LR decoder $\mathcal{D}$
► **Step 1: Initialization**
Upscale LR image as $up(\mathbf{I}_{LR})$
Encode the upsampled image as $\boldsymbol{x} = \mathcal{E}(up(\mathbf{I}_{LR}))$
Initialize the $\boldsymbol{x}$ as a realistic latent $\boldsymbol{x}_r$ and set it as guidance
Compute the realty score of the realistic latent $\boldsymbol{s}_r$
► **Step 2: Sampling**
for $t = T, \cdots, 1$ **do**
  Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\boldsymbol{\epsilon} = 0$
  Compute the latent variable at the current step $\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\boldsymbol{x_t}, \boldsymbol{x}, t) \right) + \sigma_\theta(\boldsymbol{x}_t, t)\boldsymbol{\epsilon}$
  ★ **Reality-Guided Refinement**
  Detect artifacts of the current latent $E_A = \mathcal{A}(\mathcal{D}(\boldsymbol{x}_{t-1}))$
  Refine the latent $\boldsymbol{x}_{t-1} = \boldsymbol{x}_{t-1} \times (1 - E_A) + \boldsymbol{x}_r \times E_A$
  ★ **Reality Score Computation**
  Decode the refined latent into an image $\mathbf{I}_r = \mathcal{D}(\boldsymbol{x}_{t-1})$
  Generate the current binary reality map $M_R = \mathcal{R}(\mathbf{I}_r)$
  Calculate the current reality score $\boldsymbol{s}_r^{t-1} = \mathcal{S}(M_R)$
  ★ **Self-Adaptive Guidance**
  Encode the current realistic latent $\boldsymbol{x}_r^{t-1} = \mathcal{E}(\mathbf{I}_r)$
  Update the guidance $\boldsymbol{x}_r = \mathcal{G}(\boldsymbol{x}_r, \boldsymbol{x}_r^{t-1})$ if $\boldsymbol{s}_r^{t-1} > \boldsymbol{s}_r$
  Update the reality score $\boldsymbol{s}_r = \boldsymbol{s}_r^{t-1}$ if $\boldsymbol{s}_r^{t-1} > \boldsymbol{s}_r$
end for
return the artifact-free SR $\mathbf{L}_{HR} = \mathcal{D}(\boldsymbol{x}_0)$

4. Experimental Results

In this section, we present our experimental results, discuss insights from ablation studies, and delve into the broader implications of our findings.

4.1. Experimental Setups

Testing Datasets. Our proposed SARGD method is rigorously assessed for arbitrary-scale single image super-resolution (SR) across a comprehensive set of benchmark datasets, including Set5 [1], Set14 [41], B100 [24], Urban100 [14], and Manga109 [25]. The Set5 and Set14 datasets serve as standard benchmarks within the super-resolution domain, comprising five and fourteen high-resolution images, respectively. The B100 dataset is a collection of 100 natural images used for testing the versatility of super-resolution algorithms across various real-world scenarios. Urban100 comprises 100 detailed urban images, challenging super-resolution algorithms to accurately enhance intricate structures. Manga109 contains 109 high-resolution manga images, testing the edge preservation capabilities of super-resolution models.

Evaluation Metrics. We utilize established perceptual metrics including PSNR, SSIM [35], LPIPS [44], and DISTS [47] to conduct a comprehensive comparison with existing diffusion-based SR approaches. PSNR and SSIM, particularly when assessed on the Y channel of the YCbCr

Table 1. **Quantitative comparison with the state-of-the-art Diffusion-SR methods across benchmark datasets for super-resolution at $\times 2$, $\times 3$, and $\times 4$ scales.** The dagger symbol (†) signifies LDM trained on the identical dataset as used for StableSR. **Bold** highlights the best performance. All evaluations are conducted on a 32G GPU. Our training-free SARGD attains the most favorable results.

Benchmarks		Set5 [1]			Set14 [41]			B100 [24]			Urban100 [14]			Manga109 [25]		
Metrics	Methods	$\times 2$	$\times 3$	$\times 4$	$\times 2$	$\times 3$	$\times 4$	$\times 2$	$\times 3$	$\times 4$	$\times 2$	$\times 3$	$\times 4$	$\times 2$	$\times 3$	$\times 4$
PSNR ↑	LDM†	27.86	27.94	28.30	26.42	26.81	27.21	26.91	27.54	27.67	25.18	25.72	25.91	26.49	26.96	27.04
	StableSR	28.78	29.09	29.36	27.26	27.69	28.04	27.61	28.28	28.40	26.00	26.44	26.55	27.58	28.01	27.94
	Ours	**32.64**	**33.06**	**32.27**	**31.40**	**30.94**	**30.01**	**30.92**	**30.84**	**30.23**	**29.34**	**28.84**	**27.93**	**31.21**	**31.13**	**30.23**
SSIM ↑	LDM†	0.765	0.777	0.769	0.673	0.688	0.696	0.651	0.672	0.679	0.712	0.719	0.719	0.814	0.821	0.815
	StableSR	0.805	0.814	0.812	0.716	0.731	0.733	0.694	0.714	0.717	0.745	0.748	0.744	0.843	0.847	0.838
	Ours	**0.899**	**0.897**	**0.871**	**0.842**	**0.823**	**0.778**	**0.817**	**0.800**	**0.763**	**0.849**	**0.817**	**0.771**	**0.909**	**0.897**	**0.851**
LPIPS ↓	LDM†	0.176	0.172	0.163	0.217	0.214	0.236	0.222	0.238	0.252	0.167	0.176	0.186	0.132	0.128	0.138
	StableSR	0.165	0.153	0.150	0.197	0.194	0.207	0.200	0.214	0.226	0.150	0.159	**0.166**	0.118	0.115	**0.123**
	Ours	**0.118**	**0.126**	**0.131**	**0.139**	**0.149**	**0.204**	**0.164**	**0.118**	**0.224**	**0.117**	**0.138**	0.177	**0.083**	**0.085**	0.128
DISTS ↓	LDM†	0.172	0.173	0.161	0.157	0.158	0.162	0.167	0.171	0.176	0.129	0.134	0.141	0.102	0.098	0.103
	StableSR	0.167	0.165	0.157	0.152	0.151	**0.150**	0.160	0.160	0.164	0.127	0.128	**0.131**	0.102	0.097	0.099
	Ours	**0.136**	**0.134**	**0.142**	**0.126**	**0.121**	0.156	**0.132**	**0.140**	**0.156**	**0.111**	**0.119**	0.139	**0.086**	**0.083**	**0.096**

Image (500×362) (a) Bicubic (b) StableSR (c) Ours Image (228×344) (a) Bicubic (b) StableSR (c) Ours

Figure 4. **Visual comparison with diffusion-SR methods** for $\times 2$, $\times 3$, and $\times 4$ super-resolution, including (a) **Bicubic upsampling**, (b) **StableSR**, and (c) our ***SARGD***. The red solid-lined boxes represent the ground truth (GT), focusing on regions zoomed for detailed inspection. The green boxes illustrate how our SARGD method preserves significantly more detail and clarity compared to the alternatives.

color space, are fidelity-focused measures traditionally favored in super-resolution tasks. Typically, higher values of PSNR and SSIM correspond to better performance. Nevertheless, these metrics do not adequately assess the quality of generative models. Therefore, we also incorporate LPIPS and DISTS in our assessment suite, as they have been acknowledged for their effectiveness in evaluating the perceptual quality of images produced by generative models. Generally, lower scores in LPIPS and DISTS are indicative of superior perceptual image quality.

Implementation Details. We adhere to the experimental protocol established by previous research [32]. All of our experiments are conducted using PyTorch and tested on a consistent computing platform. The uniform model trained by Zhang *et al*. [43] is employed as an artifact detector to generate the artifact mask and compute the reality score. For the inference phase, we set the number of denoising steps to 200. Following established protocols [3], we derive low-resolution test samples using the imresize function in Matlab, employing bicubic interpolation with downsampling scales of $\times 2$, $\times 3$, and $\times 4$, in line with the precedents set by prior research. In diffusion-based super-resolution methods, it is standard practice to upsample the inputs to the same size as the outputs prior to the inference stage.

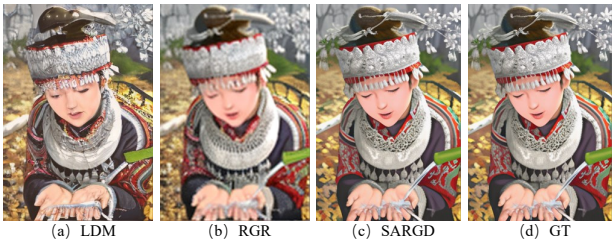
4.2. Comparison with State-of-the-Art

To verify the effectiveness of our training-free SARGD, we conducted comparative analyses with various state-of-the-art methods, such as LDM [26] and StableSR [32]. To ensure a fair comparison, LDM is finetuned using the same training configurations and dataset as those employed for StableSR. For other methods, we directly use the official codes and pre-trained models for testing. Comprehensive results, along with additional comparative analyses, are available in the supplementary material.

Quantitative Results. Table 1 demonstrates the outstanding capability of our proposed Self-Adaptive Reality-Guided Diffusion (SARGD) method in producing artifact-free super-resolution images. Notably, our method significantly surpasses the previous StableSR model in terms of PSNR, with improvements of 3.86, 3.97, and 2.91 on the Set5 dataset at upscaling factors of $\times 2$, $\times 3$, and $\times 4$, respec-

Table 2. **Analysis of SARGD inference strategies.** The 'Baseline' model represents the absence of Reality-Guided Refinement (RGR) and Self-Adaptive Guidance (SAG). 'RGR' utilizes initial realistic latent guidance, while 'SAG' employs the self-adaptive guidance mechanism. All experiments are conducted under identical settings. The best scores are highlighted in **bold**.

Metric	RGR	SAG	Set5 [1]			Set14 [41]			B100 [24]			Manga109 [25]		
			×2	×3	×4	×2	×3	×4	×2	×3	×4	×2	×3	×4
PSNR ↑	✗	✗	27.86	27.94	28.30	26.42	26.81	27.21	26.91	27.54	27.67	26.49	26.96	27.04
	✓	✗	29.82	30.39	30.27 (+ 6.96)	28.16	27.99	27.79 (+ 2.13)	28.08	28.42	28.30 (+ 2.28)	25.10	26.41	27.88 (+ 3.11)
	✓	✓	**32.42**	**32.93**	**31.79** (+ 12.3)	**30.16**	**30.33**	**29.28** (+ 7.61)	**30.29**	**30.38**	**29.92** (+ 8.13)	**30.40**	**30.58**	**29.23** (+ 8.10)
SSIM ↑	✗	✗	0.765	0.777	0.769	0.673	0.688	0.696	0.651	0.672	0.679	0.814	0.821	0.815
	✓	✗	0.823	0.843	0.824 (+ 7.15)	0.728	0.732	0.730 (+ 4.89)	0.702	0.711	0.704 (+ 3.68)	0.775	0.808	0.822 (+ 0.86)
	✓	✓	**0.897**	**0.895**	**0.870** (+ 13.1)	**0.821**	**0.815**	**0.770** (+ 10.6)	**0.807**	**0.792**	**0.760** (+ 11.9)	**0.900**	**0.893**	**0.851** (+ 4.41)
LPIPS ↓	✗	✗	0.176	0.172	0.163	0.217	0.214	0.236	0.222	0.238	0.252	0.132	0.128	0.138
	✓	✗	0.144	0.212	0.161 (- 1.23)	0.227	0.270	0.242 (+ 2.54)	0.201	0.248	0.292 (+ 15.8)	0.128	0.139	0.164 (+ 18.9)
	✓	✓	**0.122**	**0.141**	**0.131** (- 19.6)	**0.144**	**0.152**	**0.208** (- 11.9)	**0.172**	**0.201**	**0.228** (- 9.52)	**0.087**	**0.087**	**0.132** (- 4.35)
DISTS ↓	✗	✗	0.172	0.173	0.161	0.157	0.158	0.162	0.167	0.171	0.176	0.102	0.098	0.103
	✓	✗	0.145	0.173	0.192 (+ 19.3)	0.160	0.179	0.214 (+ 32.1)	0.191	0.213	0.238 (+ 35.2)	0.094	0.098	0.111 (+ 7.77)
	✓	✓	**0.137**	**0.141**	**0.145** (- 9.94)	**0.127**	**0.129**	**0.158** (- 2.47)	**0.138**	**0.141**	**0.158** (- 10.2)	**0.091**	**0.084**	**0.099** (- 3.88)



(a) LDM (b) RGR (c) SARGD (d) GT

Figure 5. **Visual comparison of SARGD components.** Our SARGD exhibits the best outcomes for ×3 super-resolution.

tively. Within the SSIM metric, our SARGD method realizes significant improvements (*e.g.* the increase from 0.716 to 0.842 on Set14 and from 0.694 to 0.817 on B100), illustrating a substantial enhancement in structural fidelity. Additionally, in terms of the DISTS metric, SARGD significantly reduces the score from 0.102 to 0.086 in Manga109, reflecting an improved alignment with human perceptual quality. These improvements establish our SARGD as a leading approach in the field of super-resolution, setting a new standard for image clarity and detail preservation.

Visualization. Figure 4 offers a visual comparison between our SARGD and StableSR [32], further highlighting the superior performance of our method. Compared to columns (a) and (c), bicubic upscaling maintains satisfactory detail but often leads to an over-smoothing issue. While StableSR succeeds in generating high-resolution images, its dependence on conventional DDPM methods makes it susceptible to introducing additional artifacts. When compared to columns (b) and (c), our approach successfully produces images with high authenticity, thereby avoiding the over-smoothing problem commonly seen in other methods.

4.3. Ablation Study

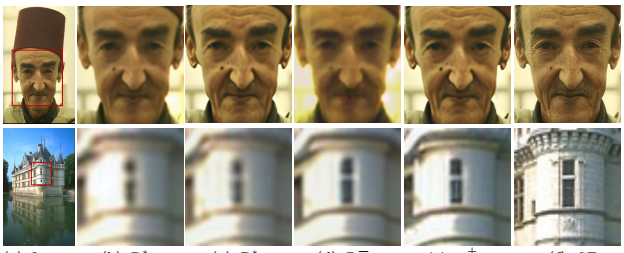
In this section, we thoroughly evaluate and analyze the performance of the proposed training-free SARGD method, primarily focusing on its effectiveness in denoising.

Table 3. **Comparison of varied denoising approaches within RGR** across Set14, B100, and Urban100 datasets for ×2, ×3, ×4 upscaling. '$\mathbf{D}^*_{1s}$' refers to artifact refinement in the latent space at each step, '$\mathbf{D}^*_{10s}$' at every ten steps, while '$\mathbf{D}^-_{100s}$' and '$\mathbf{D}^+_{100s}$' indicate refinement before and after 100 steps, respectively.

Scale	Type	Set14		B100		Urban100	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
×2	$\mathbf{D}^*_{1s}$	28.16	0.728	28.08	0.702	25.28	0.708
	$\mathbf{D}^*_{10s}$	27.99	0.725	28.05	0.692	25.11	0.703
	$\mathbf{D}^-_{100s}$	26.79	0.685	27.32	0.672	24.42	0.676
	$\mathbf{D}^+_{100s}$	**28.42**	**0.739**	**28.11**	**0.715**	**25.61**	**0.716**
×3	$\mathbf{D}^*_{1s}$	27.99	0.732	28.42	0.711	26.25	0.751
	$\mathbf{D}^*_{10s}$	27.83	0.726	28.47	0.705	26.14	0.747
	$\mathbf{D}^-_{100s}$	26.64	0.686	27.66	0.688	25.78	0.731
	$\mathbf{D}^+_{100s}$	**28.02**	**0.730**	**28.54**	**0.718**	**26.59**	**0.759**
×4	$\mathbf{D}^*_{1s}$	27.79	0.730	28.30	0.704	25.92	0.711
	$\mathbf{D}^*_{10s}$	27.83	0.729	28.48	0.702	25.88	0.708
	$\mathbf{D}^-_{100s}$	26.94	0.708	27.83	0.689	25.47	0.702
	$\mathbf{D}^+_{100s}$	**28.33**	**0.734**	**28.51**	**0.714**	**25.96**	**0.713**

Analysis of Improvement. Table 2 presents the effects of our proposed modules, Reality-Guided Refinement (RGR) and Self-Adaptive Guidance (SAG). The baseline corresponds to the LDM configured and trained similarly to StableSR [32]. Analysis of the initial two columns reveals that the integration of RGR notably enhances performance, with gains of 6.96%, 2.13%, 2.28%, and 3.11% observed in ×4 super-resolution on the Set5, Set14, B100, and Manga109 datasets, respectively. The integration of RGR into the DDPM architecture leads to an increase in SSIM scores, validating the effectiveness of RGR in enhancing image quality. The visual results in Figure 5 show that while RGR manages to maintain the essential attributes of the original images, it also leads to over-smoothing in the super-resolved outputs. By contrasting figures (b) and (c), it is evident that incorporating our SAG during inference enhances image quality and successfully addresses the issue of over-smoothing, thus confirming the effectiveness of SAG.

Effect of Varied Denoising Strategies within RGR. Table 3 demonstrates that the artifact refinement in the last 100 steps achieves the best results. Analyzing the first two



(a) Image (b) $\mathbf{D}^{*}_{1s}$ (c) $\mathbf{D}^{*}_{10s}$ (d) $\mathbf{D}^{-}_{100s}$ (e) $\mathbf{D}^{+}_{100s}$ (f) GT

Figure 6. **Visual comparison of RGR denoising strategies impacting over-smoothing.** They are represented as '$\mathbf{D}^{*}_{1s}$' for continuous artifact refinement, '$\mathbf{D}^{*}_{10s}$' for refinement every ten steps, and '$\mathbf{D}^{-}_{100s}$' vs. '$\mathbf{D}^{+}_{100s}$' for pre- and post-100-step refinement.

Table 4. **Comparison of different realistic latent udpate strategies.** '$\mathbf{R}^{*}_{1s}$' refers to realistic-latent enhancement at each step and '$\mathbf{R}^{+}_{100s}$' indicates enhancement after 100 steps, respectively.

Scale	Type	Set14			B100		
		PSNR ↑	SSIM ↑	DISTS ↓	PSNR ↑	SSIM ↑	DISTS ↓
×2	$\mathbf{R}^{*}_{1s}$	30.16	0.821	0.127	30.29	0.807	0.138
	$\mathbf{R}^{+}_{100s}$	**31.40**	**0.842**	**0.126**	**30.92**	**0.817**	**0.132**
×3	$\mathbf{R}^{*}_{1s}$	30.33	0.815	0.129	30.38	0.792	0.141
	$\mathbf{R}^{+}_{100s}$	**30.94**	**0.823**	**0.121**	**30.84**	**0.800**	**0.140**
×4	$\mathbf{R}^{*}_{1s}$	29.28	0.770	0.158	29.92	0.760	0.158
	$\mathbf{R}^{+}_{100s}$	**30.01**	**0.778**	**0.156**	**30.23**	**0.763**	**0.156**

lines per scale shows that refining artifacts at each step improves outcomes over ten-step intervals, as seen with higher PSNRs (*28.16 vs. 27.99* on Set4 at ×2), due to variable artifact patterns in the latent space. Comparing artifact refinement strategies indicates that post-100-step processing yields greater enhancements, as demonstrated by a PSNR increase from 26.94 to 28.33 at ×4 super-resolution. Observations from panels (d) and (c) in Figure 6 suggest that artifact refinement early in the noisy latent phase causes more over-smoothing than later-stage refinement.

Approaches to Updating the Realistic Latent in SAG. Table 4 presents a comparative analysis of different strategies for updating the reference realistic-latent within SAG. Notably, updating the realistic latent after 100 steps yields superior results across all metrics when compared to enhancements at every step. For instance, this strategy leads to an increase in PSNR from 30.16 to 31.40 at ×2, from 30.33 to 30.94 at ×3, and from 29.28 to 30.1 at ×4 on Set14. These findings suggest that enhancing the reference latent later in the process correlates with improved performance.

Influence of Artifact Detection on SARGD. Table 5 contrasts SARGD's performance with and without the implementation of artifact detection in the latent space during inference, providing a line-by-line comparative analysis. It is observed that the elimination of perceptual artifacts in the latent space significantly enhances super-resolution quality, as evidenced by improved metrics such as PSNR (*e.g.*, *30.84 vs. 30.06* on B100 at ×3 scaling) and DISTS (*e.g.*, *0.219 vs. 0.140* on B100 at ×3 scaling), highlighting the benefits of integrating artifact detection within SARGD.

Table 5. **Performance of SARGD with artifact detection ('$\mathcal{A}$') versus without** in latent space across ×2, ×3, and ×4 scales. Here, '✗' signifies the substitution of RGR with a direct sum of the latent and the initial realistic reference latent.

Scale	$\mathcal{A}$	Set14			B100		
		PSNR ↑	SSIM ↑	DISTS ↓	PSNR ↑	SSIM ↑	DISTS ↓
×2	✗	31.24	0.825	0.138	30.56	0.801	0.158
	✓	**31.40**	**0.842**	**0.126**	**30.92**	**0.817**	**0.132**
×3	✗	30.47	0.796	0.193	30.06	0.783	0.219
	✓	**30.94**	**0.823**	**0.121**	**30.84**	**0.800**	**0.140**
×4	✗	29.98	0.773	0.234	30.15	0.753	0.260
	✓	**30.01**	**0.778**	**0.156**	**30.23**	**0.763**	**0.156**

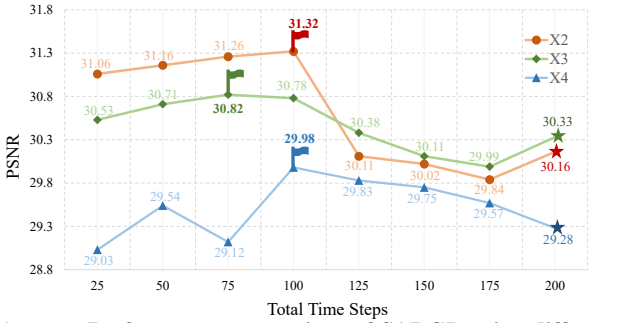


Figure 7. **Performance comparison of SARGD using different time steps** on Set14 for ×2, ×3, ×4 upscaling using PSNR. Our SARGD reaches peak score between 75 and 100 steps, effectively reducing the total inference steps by half, from 200 to 100.

Impact on Inference Steps for SARGD. Figure 7 illustrates that our proposed SARGD not only enhances image quality but also decreases the number of total inference steps required. Specifically, the performance of SARGD in super-resolution steadily improves from the start and achieves optimal PSNR scores between 75 and 100 steps, such as a PSNR of 31.32 at 100 steps compared to 30.16 at 200 steps, which effectively cuts the inference time in half. This suggests that the removal of artifacts during the diffusion phase serves to improve image quality in SR while simultaneously boosting the speed of the inference process.

5. Conclusions

In this study, we tackle the issue of artifacts that arise in diffusion-based super-resolution. Our SARGD incorporates two key mechanisms: Reality-Guided Refinement (RGR) and Self-Adaptive Guidance (SAG). RGR is designed to identify and correct artifacts within the latent space by utilizing a realistic latent as a reference, thereby improving the image's fidelity. To overcome the over-smoothing effect that comes from employing an initial realistic latent for guidance, SAG is introduced to refine the reference latent, improving the quality of the super-resolved images.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (62271400), and the Shaanxi Provincial Key R&D Program, China (2023-GHZD-02).

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. 5, 6, 7

[2] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *ECCV*, 2018. 2

[3] 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 (CVPR)*, pages 8628–8638, 2021. 6

[4] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. in 2021 ieee. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14347–14356, 2021. 3

[5] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12413–12422, 2022. 3

[6] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[7] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5781–5790, 2020. 3

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(2):295–307, 2015. 1, 2

[9] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 391–407. Springer, 2016. 2

[10] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7890–7899, 2020. 3

[11] Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. *Advances in Neural Information Processing Systems (NIPS)*, 33:3022–3032, 2020. 3

[12] 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 (CVPR)*, pages 1909–1918, 2022. 2

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

[14] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015. 5, 6

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 694–711, 2016. 2

[16] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017. 2

[17] 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. 1, 2, 3

[18] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using wwin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1833–1844, 2021. 2

[19] 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 (ICCV)*, pages 4076–4085, 2021. 2

[20] 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/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 136–144, 2017. 2

[21] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8060–8069, 2020. 3

[22] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 715–732, 2020. 2

[23] Andreas Lugmayr, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Normalizing flow as a flexible fidelity objective for photo-realistic super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1756–1765, 2022. 2

[24] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 416–423, 2001. 5, 6, 7

[25] 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. 5, 6, 7

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3, 6

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015. 2

[28] 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 (TPAMI)*, 45(4): 4713–4726, 2022. 1, 3

[29] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016. 1, 2

[30] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICLR)*, pages 2256–2265, 2015. 2

[31] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems (NIPS)*, 32, 2019. 1

[32] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 2, 3, 6, 7

[33] Tao Wang, Kaihao Zhang, Xuanxi Chen, Wenhan Luo, Jiankang Deng, Tong Lu, Xiaochun Cao, Wei Liu, Hongdong Li, and Stefanos Zafeiriou. A survey of deep face restoration: Denoise, super-resolution, deblur, artifact removal. *arXiv:2211.02831*, 2022. 2

[34] 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

[35] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4):600–612, 2004. 5

[36] Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1704–1713, 2023. 2

[37] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023. 2

[38] Xin Yang, Haiyang Mei, Jiqing Zhang, Ke Xu, Baocai Yin, Qiang Zhang, and Xiaopeng Wei. Drfn: Deep recurrent fusion network for single-image super-resolution with large factors. *IEEE Transactions on Multimedia (TMM)*, 21(2): 328–337, 2018. 1

[39] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7556–7566, 2019. 3

[40] Zongsheng Yue and Chen Change Loy. Diffface: Blind face restoration with diffused error contraction. *arXiv preprint arXiv:2212.06512*, 2022. 3

[41] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, pages 711–730. Springer, 2012. 5, 6, 7

[42] Kaihao Zhang, Dongxu Li, Wenhan Luo, Jingyu Liu, Jiankang Deng, Wei Liu, and Stefanos Zafeiriou. Edface-celeb-1 m: Benchmarking face hallucination with a million-scale dataset. *TPAMI*, 2022. 2

[43] Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual artifacts localization for image synthesis tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7579–7590, 2023. 4, 5, 6

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 5

[45] 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 (ICCV)*, pages 3096–3105, 2019. 2

[46] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *IEEE international Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2019. 3

[47] 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. 5

[48] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2481, 2018. 2

[49] Yulun Zhang, Kai Li, Kunpeng Li, and Yun Fu. Mr image super-resolution with squeeze and excitation reasoning attention network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13425–13434, 2021. 1

[50] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2185–2194, 2021. 3