SinSR: Diffusion-Based Image Super-Resolution in a Single Step

Yufei Wang[1,2†], Wenhan Yang[3], Xinyuan Chen[2*], Yaohui Wang[2], Lanqing Guo[1],
Lap-Pui Chau[4], Ziwei Liu[1], Yu Qiao[2], Alex C. Kot[1], Bihan Wen[1*]
[1]Nanyang Technological University [2]Shanghai Artificial Intelligence Laboratory
[3]PengCheng Laboratory [4]The Hong Kong Polytechnic University

Abstract

*While super-resolution (SR) methods based on diffusion models exhibit promising results, their practical application is hindered by the substantial number of required inference steps. Recent methods utilize the degraded images in the initial state, thereby shortening the Markov chain. Nevertheless, these solutions either rely on a precise formulation of the degradation process or still necessitate a relatively lengthy generation path (e.g., 15 iterations). To enhance inference speed, we propose a simple yet effective method for achieving single-step SR generation, named **SinSR**. Specifically, we first derive a deterministic sampling process from the most recent state-of-the-art (SOTA) method for accelerating diffusion-based SR. This allows the mapping between the input random noise and the generated high-resolution image to be obtained in a reduced and acceptable number of inference steps during training. We show that this deterministic mapping can be distilled into a student model that performs SR within only one inference step. Additionally, we propose a novel consistency-preserving loss to simultaneously leverage the ground-truth image during the distillation process, ensuring that the performance of the student model is not solely bound by the feature manifold of the teacher model, resulting in further performance improvement. Extensive experiments conducted on synthetic and real-world datasets demonstrate that the proposed method can achieve comparable or even superior performance compared to both previous SOTA methods and the teacher model, in just one sampling step, resulting in a remarkable up to* $\times 10$ *speedup for inference. Our code will be released at* https://github.com/wyf0912/SinSR/.

1. Introduction

Image super-resolution (SR) aims to reconstruct a high-resolution image from a given low-resolution (LR) counterpart [45]. Recently, diffusion models, known for their effectiveness in modeling complex distributions, have gained widespread adoption and demonstrated remarkable performance in SR tasks, particularly in terms of perceptual quality.

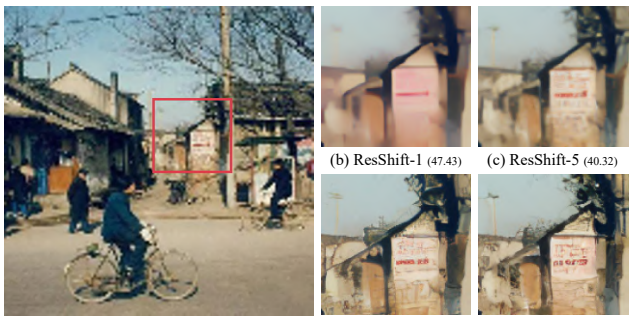
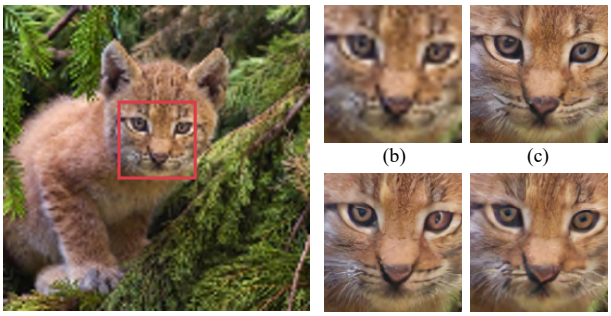


Figure 1. A comparison between the most recent SOTA method ResShift [46] for the acceleration of diffusion-based SR and the proposed method. We achieve on-par or even superior perceptual quality using only one inference step. ("-N" behind the method name represents the number of inference steps, and the value in the bracket is the quantitative result measured by MUSIQ↑ [15].)

Specifically, current strategies for employing diffusion models can be broadly categorized into two streams: concatenating the LR image to the input of the denoiser in the diffusion models [32, 33], and adjusting the inverse process of a pre-trained diffusion model [4, 5, 14]. Despite achieving promising results, both strategies encounter computational efficiency issues. Notably, the initial state of these conditional diffusion models is a pure Gaussian noise without using the prior knowledge from the LR image. Consequently, a substantial number of inference steps are required to achieve satisfactory performance, significantly hindering the practical applications of diffusion-based SR techniques.

Efforts have been made to enhance the sampling efficiency of diffusion models, leading to various techniques proposed [22, 28, 37]. However, in the realm of low-level vision where maintaining high fidelity is critical, these techniques often fall short as they achieve acceleration at the cost of performance. More recently, innovative techniques have emerged to reformulate the diffusion process in image restoration tasks, focusing on improving the signal-to-noise ratio of the initial diffusion state and thereby shorten the

† Work done as an intern at Shanghai AI Lab. * Corresponding authors.



(a) LR image (b) (c) (d) (e)

Figure 2. An illustration of the generative ability of the proposed method in only one step. Given the same LR image (Fig. (a) and (b)), by using different noise added to the input, HR images (Fig. (c)-(e)) with different details are generated, *e.g.*, eyes of different shapes and colors. Best zoom in for details.

Markov chain. For instance, [43] initiates the denoising diffusion process with the input noisy image, while in the SR task, [46] models the initial step as a combination of the LR image and random noise. Nonetheless, even in these most recent works [43, 46], limitations persist. For instance, while [43] shows promising results within just three inference steps, it requires a clear formulation of the image degradation process. Besides, [46] still necessitates 15 inference steps and exhibits degraded performance with noticeable artifacts if the number of inference steps is further reduced.

To address these challenges, we introduce a novel approach that can generate high-resolution (HR) images in only one sampling step, without compromising the diversity and perceptual quality of the diffusion model, as shown in Fig. 1 and Fig. 2. Specifically, we propose to directly learn a well-paired bi-directional deterministic mapping between the input random noise and the generated HR image from a teacher diffusion model. To accelerate the generation of well-matched training data, we first derive a deterministic sampling strategy from the most recent state-of-the-art work [46], designed for accelerating diffusion-based SR, from its original stochastic formulation. Additionally, we propose a novel consistency-preserving loss to leverage ground-truth images, further enhancing the perceptual quality of the generated HR images by minimizing the error between ground-truth (GT) images and those generated from the predicted initial state. Experimental results demonstrate that our method achieves comparable or even better performance compared to SOTA methods and the teacher diffusion model [46], while greatly reducing the number of inference steps from 15 to 1, resulting in up to a $\times 10$ speedup in inference.

Our main contributions are summarized as follows:

- We accelerate the diffusion-based SR model to a single inference step with comparable or even superior performance for the first time. Instead of shortening the Markov chain of the generation process, we propose a simple yet effective approach that directly distills a deterministic generation function into a student network.
- To further fasten training, we derive a deterministic sampling strategy from the recent SOTA method [46] on accelerating the SR task, enabling efficient generation of well-matched training pairs.
- We propose a novel consistency-preserving loss that can utilize the ground-truth images during training, preventing the student model from only focusing on fitting the deterministic mapping of the teacher diffusion model, therefore leading to better performance.
- Extensive experiments on both synthetic and real-world datasets show that our proposed method can achieve comparable or even superior performance compared to SOTA methods and the teacher diffusion model, while greatly reducing the number of inference steps from 15 to 1.

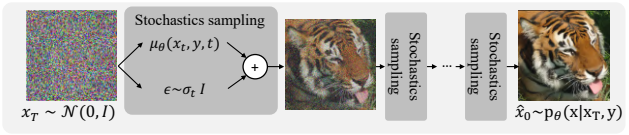
2. Related Work

2.1. Image Super-Resolution

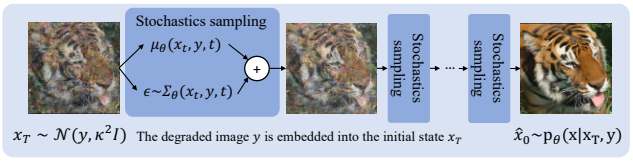
With the rise of deep learning, deep learning-based techniques gradually become the mainstream of the SR task [8, 45]. One prevalent approach of early works is to train a regression model using paired training data [1, 2, 16, 44]. While the expectation of the posterior distribution can be well modeled, they inevitably suffer from the over-smooth problem [17, 27, 34]. To improve the perceptual quality of the generated HR images, generative-based SR models attract increasing attention, *e.g.*, autoregressive-based models [6, 26, 29, 30]. While significant improvements are achieved, the computational cost of autoregressive models is usually large. Subsequently, normalizing flows [23, 42] are demonstrated to have good perceptual quality under an efficient inference process, while its network design is restricted by the requirements of the invertibility and ease of calculation. Besides, GAN-based methods also achieve great success in terms of perceptual quality [9, 13, 17, 27, 34]. However, the training of GAN-based methods is usually unstable. Recently, diffusion-based models have been widely investigated in SR [4, 5, 14, 32, 33]. The diffusion-based SR methods can be roughly summarized into two categories, concatenating the LR image to the input of the denoiser [32, 33], and modifying the backward process of a pre-trained diffusion model [4, 5, 10, 14]. While promising results are achieved, they rely on a large number of inference steps, which greatly hinders the application of diffusion-based models.

2.2. Acceleration of Diffusion Models

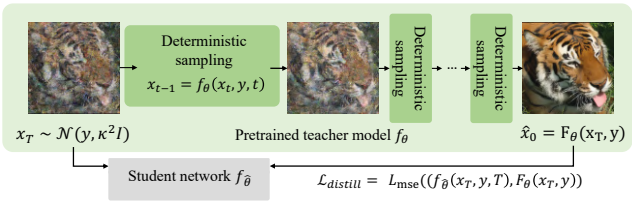
Recently, the acceleration of diffusion models has attracted more and more attention. Several algorithms are proposed for general diffusion models [22, 28, 37, 38] and proved quite effective for image generations. One intuitive strategy among them is to distill the diffusion models to a student model. However, the huge training overhead to solve



(a) The inference of SR3 [33] starts from a pure noise, which requires a large number of inference steps (T=100 after using DDIM [37]).



(b) The recent SOTA method ResShift [46] shortens the Markov chain to speed up the inference process by incorporating the information of the LR image y to the initial state x_T (T=15).



(c) A simplified pipeline of the proposed method *SinSR* (distill only). It directly learns the deterministic mapping between x_T and x_0, therefore the inference process can be further compressed into only one step (T=1).

Figure 3. A comparison between the vanilla diffusion-based SR method [33], a most recent method for acceleration of the diffusion-based SR [46], and the proposed one-step SR. Different from recent works that shorten the Markov chain to speed up the inference process [43, 46], the proposed method directly learns the deterministic generation process and the details can be found in Fig. 4.

the ordinary differential equation (ODE) of the inference process makes this scheme less attractive on a large-scale dataset [24]. To alleviate the training overhead, progressive distillation strategies are usually adopted [25, 35]. Meanwhile, instead of simply simulating the behavior of a teacher diffusion model through distillation, better inference paths are explored in an iterative manner [20, 21]. While progressive distillation effectively decreases the training overhead, the error accumulates at the same time, leading to an obvious performance loss in SR. Most recently, targeting the image restoration task, some works reformulate the diffusion process by either using the knowledge of degradation process [43] or a pre-defined distribution of the initial state [46], yielding a shortened Markov chain of the generation process and better performance than directly applying DDIM [37] in low-level tasks. However, they either require a clear formulation of the degradation or still require a relatively large number of inference steps.

3. Motivation

Preliminary. Given an LR image y and its corresponding HR image x_0, existing diffusion-based SR methods aim to model the conditional distribution $q(x_0|y)$ through a Markov chain where a forward process is usually defined as $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$ with an initial state $x_T \sim \mathcal{N}(0, I)$. The role of the diffusion model can be regarded as transferring the input domain (standard Gaussian noise) to the HR image domain conditioned on the LR image. Since the matching relationship between x_T and x_0 is unknown, usually a diffusion model [11, 21, 33] through an iterative manner is required to learn/infer from an unknown mapping between x_T and x_0. Our method is grounded in the idea that having an SR model that effectively captures the conditional distribution $q(x_0|y)$ and establishes a deterministic mapping between x_T and $\hat{x}0$ given an LR image y, we can streamline the inference process to a single step by employing another network, denoted as $f_{\hat{\theta}}$, to learn the correspondence between $\hat{x}_0$ and x_T, as illustrated in Fig. 3.

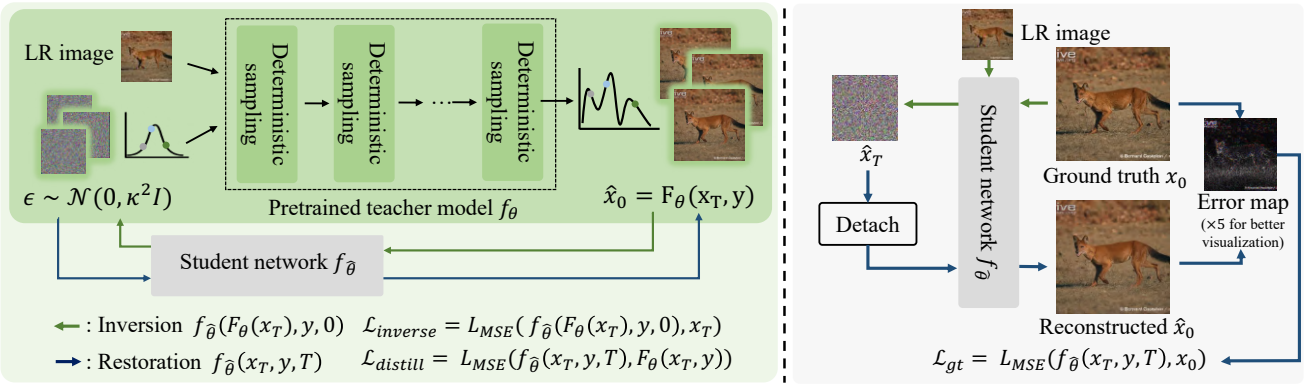
Distillation for diffusion SR models: less is more. While the concept of distilling the mapping between x_T and $\hat{x}_0$ to a student network has been previously explored [21], its application to SR introduces several challenges:

- The training overhead becomes substantial for one-step distillation due to a large number of inference steps of previous models, *e.g.*, LDM [32] still need 100 steps after using DDIM [37] for inference to generate high-quality pairs $(\hat{x}_0, x_T, y)$ as the training data of the student model.
- The performance degradation is attributed to the introduction of a more intricate distillation strategy involving iteration. For example, to reduce the training overhead, an iterative distillation strategy [35] is adopted which gradually decreases the number of inference steps during training. However, despite achieving satisfactory results in generation tasks, the cumulative error significantly impacts the fidelity of the SR results, as SR tasks are relatively more sensitive to image quality.

To address the aforementioned two challenges, we propose to distill the diffusion SR process into a single step in a simple but effective way based on the following observations. More details of the observations can be seen in Sec. 5.3

- We demonstrate that the most recent SOTA method for accelerating the diffusion-based SR [46], which achieves comparable performance in 15 steps as LDM [32] in 100 DDIM steps, has a deterministic mapping between x_T and x_0. Besides, the greatly reduced number of inference steps and the existence of the deterministic mapping make the training of a single-step distillation possible as shown in Fig. 6 and Table 4.
- Learning the mapping between x_T and $\hat{x}_0$ is found to be easier than denoising x_t under different noise levels as shown in Table 5. Therefore, it is feasible to directly learn the mapping between x_T and $\hat{x}_0$ so that the accumulated error by the iterative distillation can be avoided.
- Due to the accumulated error, a more sophisticated distillation strategy (iterative-based) does not contribute to the improvement in our setting as shown in Table 6.

The organization of the following sections is as follows:



(a) One-step bi-directional distillation

(b) Consistency preserving loss

Figure 4. The overall framework of the proposed method. By minimizing $\mathcal{L}_{distill}$ and $\mathcal{L}_{inverse}$, the student network $f_{\hat{\theta}}$ learns the deterministic bi-directional mapping between x_T and $\hat{x}_0$ obtained from a pre-trained teacher diffusion model in one step. Meanwhile, the proposed consistency preserving loss $\mathcal{L}_{gt}$ is optimized during training to utilize the information from the GT images to pursue better perceptual quality instead of simply fitting the deterministic mappings from the teacher model. Specifically, the GT image is first converted to its latent code $\hat{x}_T = f_{\hat{\theta}}(x_0, y, 0)$, and then converted back to calculate its reconstruction loss $L_{MSE}(f_{\hat{\theta}}(\hat{x}_T, y, T), x_0)$.

we first demonstrate that ResShift [46], in which the inference process is originally stochastic, can be converted to a deterministic model without retraining in Sec 4.1, and then the proposed consistency preserving distillation in Sec 4.2.

4. Methodology

4.1. Deterministic Sampling

A core difference between ResShift [46] and LDM [32] is the formulation of the initial state x_T. Specifically, in ResShift [46], the information from the LR image y is integrated into the diffusion step x_t as follows

$$q(x_t|x_0, y) = \mathcal{N}(x_t; x_0 + \eta_t(y - x_0), \kappa^2\eta_t\mathbf{I}), \quad (1)$$

where η_t is a serial of hyper-parameters that monotonically increases with the timestep t and obeys $\eta_T \to 1$ and $\eta_0 \to 0$. As such, the inverse of the diffusion process starts from an initial state with rich information from the LR image y as follows $x_T = y + \kappa\sqrt{\eta_T}\epsilon$ where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. To generate a HR image x from a given image y, the original inverse process of [46] is as follows

$$p_\theta(x_{t-1}|x_t, y) = \mathcal{N}(x_{t-1}|\mu_\theta(x_t, y, t), \kappa^2\frac{\eta_{t-1}}{\eta_t}\alpha_t\mathbf{I}), \quad (2)$$

where $\mu_\theta(x_t, y, t)$ is reparameterized by a deep network. As shown in Eq. 2, given an initial state $x_T = y + \kappa\sqrt{\eta_T}\epsilon$, the generated image is stochastic due to the existence of the random noise during the sampling from $p_\theta(x_{t-1}|x_t, y)$. Inspired by DDIM sampling [37], we find that a non-Markovian reverse process $q(x_{t-1}|x_t, x_0, y)$ exists which keeps the marginal distribution $q(x_t|x_0, y)$ unchanged so that it can be directly adopted to a pre-trained model. The reformulated deterministic reverse process is as follows

$$q(x_{t-1}|x_t, x_0, y) = \delta(k_t x_0 + m_t x_t + j_t y), \quad (3)$$

where δ is the unit impulse, and k_t, m_t, j_t are as follows

$$\begin{cases} m_t = \sqrt{\frac{\eta_{t-1}}{\eta_t}} \\ j_t = \eta_{t-1} - \sqrt{\eta_{t-1}\eta_t} \\ k_t = 1 - \eta_{t-1} + \sqrt{\eta_{t-1}\eta_t} - \sqrt{\frac{\eta_{t-1}}{\eta_t}}. \end{cases} \quad (4)$$

The details of the derivation can be found in the supplementary material. As a consequence, for inference, the reverse process conditioned on y is reformulated as follows

$$\begin{aligned} x_{t-1} &= k_t\hat{x}_0 + m_t x_t + j_t y \\ &= k_t f_\theta(x_t, y, t) + m_t x_t + j_t y, \end{aligned} \quad (5)$$

where $f_\theta(x_t, y, t)$ is the predicted HR image from a pre-trained ResShift [46] model. By sampling from the reformulated process in Eq. 5, a deterministic mapping between x_T (or ϵ) and $\hat{x}_0$ can be obtained and is denoted as $F_\theta(x_T, y)$.

4.2. Consistency Preserving Distillation

Vanilla distillation. We propose utilizing a student network $f_{\hat{\theta}}$ to learn the deterministic mapping F_θ between the random initialized state x_T and its deterministic output $F_\theta(x_T, y)$ from a teacher diffusion model. The vanilla distillation loss is defined as follows

$$\mathcal{L}_{distill} = L_{MSE}(f_{\hat{\theta}}(x_T, y, T), F_\theta(x_T, y)), \quad (6)$$

where $f_{\hat{\theta}}(x_T, y, T)$ is the student network that directly predicts the HR image in only one step, and F_θ represents the proposed deterministic inference process of ResShift [46] in Sec. 4.1 through an iterative manner using a pre-trained

For ease of presentation, the LR image is y is pre-upsampled to the same spatial resolution with the HR image x. Besides, similar to [32, 46], the diffusion is conducted in the latent space.

network parameterized by θ. We observe that the student model trained solely with the distillation loss in Eq. 6 already achieves promising results in just one inference step, as indicated by "(distill only)" in the result tables.

Regularization by the ground-truth image. A limitation of the aforementioned vanilla distillation strategy is that the GT image is not utilized during training, thereby restricting the upper performance bound of the student model. To further enhance the student's performance, we propose a novel strategy that incorporates a learned inversion of the HR image to provide additional regularization from the ground-truth images. In addition to the vanilla distillation loss, the student network concurrently learns the inverse mapping during training by minimizing the following loss,

$$\mathcal{L}_{inverse} = L_{MSE}(f_{\hat{\theta}}(F_{\theta}(x_T, y), y, 0), x_T), \quad (7)$$

where the last parameter of $f_{\hat{\theta}}$ is set from T in Eq. 6 to 0, indicating that the model is predicting the inversion instead of the $\hat{x}_0$. Then the GT image x_0 can be employed to regularize the output SR image given its predicted inversion $\hat{x}_T$ as follows

$$\begin{aligned} \hat{x}_T &= detach(f_{\hat{\theta}}(x_0, y, 0)) \\ \mathcal{L}_{gt} &= L_{MSE}(f_{\hat{\theta}}(\hat{x}_T, y, T), x_0), \end{aligned} \quad (8)$$

where $\mathcal{L}_{gt}$ is the proposed consistency preserving loss. By reusing $f_{\hat{\theta}}$ to learn both $f_{\hat{\theta}}(\cdot, \cdot, T)$ and $f_{\hat{\theta}}(\cdot, \cdot, 0)$ simultaneously, we can initialize the parameter $\hat{\theta}$ of the student model from the teacher one θ to speed up the training.

The overall training objective. The student network is trained to minimize the aforementioned three losses at the same time as follows

$$\hat{\theta} = \arg\min_{\hat{\theta}} \mathbb{E}_{y, x_0, x_T}[\mathcal{L}_{distill} + \mathcal{L}_{reverse} + \mathcal{L}_{gt}], \quad (9)$$

where the losses are defined in Eq. 6, 7, and 8 respectively. We assign equal weight to each loss term, and ablation studies are in the supplementary material. The overall of the proposed method is summarized in Algorithm 1 and Fig. 4.

Algorithm 1 Training

Require: Pre-trained teacher diffusion model f_θ
Require: Paired training set (X, Y)
Init $f_{\hat{\theta}}$ from the pre-trained model, *i.e.*, $\hat{\theta} \leftarrow \theta$.
while not converged **do**
  sample $x_0, y \sim (X, Y)$
  sample $\epsilon \sim \mathcal{N}(\mathbf{0}, \kappa^2 \eta_T \mathbf{I})$
  $x_T = y + \epsilon$
  for $t = T, T-1, ..., 1$ **do**
    if $t = 1$ **then**
      $\hat{x}_0 = f_\theta(x_1, y, 1)$
    else
      $x_{t-1} = k_t f_\theta(x_t, y, t) + m_t x_t + j_t y$
    end if
  end for
  $\mathcal{L}_{distill} = L_{MSE}(f_{\hat{\theta}}(x_T, y, T), \hat{x}_0)$
  $\mathcal{L}_{inverse} = L_{MSE}(f_{\hat{\theta}}(\hat{x}_0, y, 0), x_T)$
  $\hat{x}_T = f_{\hat{\theta}}(x_0, y, 0)$,
  $\mathcal{L}_{gt} = L_{MSE}(f_{\hat{\theta}}(detach(\hat{x}_T), y, T), x_0)$
  $\mathcal{L} = \mathcal{L}_{distill} + \mathcal{L}_{inverse} + \mathcal{L}_{gt}$
  Perform a gradient descent step on $\nabla_{\hat{\theta}} \mathcal{L}$
end while
return The student model $f_{\hat{\theta}}$.

5. Experiment

5.1. Experimental setup

Training Details. For a fair comparison, we follow the same experimental setup and backbone design as that in [46]. Specifically, the main difference is that we finetuned the model for 30K iterations instead of training from scratch for 500K in [46]. We find that the student model can converge quickly so that even if for each iteration we need extra time to solve the ODE to get paired training data, the overall training time is still much shorter than retraining a model from scratch following [46]. We train the models on the training set of ImageNet [7] following the same pipeline with ResShift [46] where the degradation model is adopted from RealESRGAN [41].

Compared methods. We compare our method with several representative SR models, including RealSR-JPEG [12], ESRGAN [40], BSRGAN [47], SwinIR [18], RealESRGAN [41], DASR [19], LDM [32], and ResShift [46]. For a comprehensive comparison, we further evaluate the performance of diffusion-based models LDM [32] and ResShift [46] with a reduced number of sampling steps. Besides, we compare the proposed method with Rectified-Flow [21], a SOTA method that can compress the generation process into a single step, in Table 6.

Metrics. For the evaluation of the proposed method on the synthetic testing dataset with reference images, we utilize PSNR, SSIM, and LPIPS [48] to measure the fidelity performance. Besides, two recent SOTA non-reference metrics are used to justify the realism of all the images, *i.e.*, CLIP-IQA [39] which leverages a CLIP model [31] pre-trained on a large-scale dataset (Laion400M [36]) and MUSIQ [15].

5.2. Experimental Results

Evaluation on real-world datasets. RealSR [3] and RealSet65 [46] are adopted to evaluate the generalization ability of the model on unseen real-world data. Specifically, in RealSR [3], there are 100 real images captured by two different cameras in different scenarios. Besides, RealSet65 [46] includes 65 LR images in total, collected from widely used datasets and the internet. The results on these two datasets

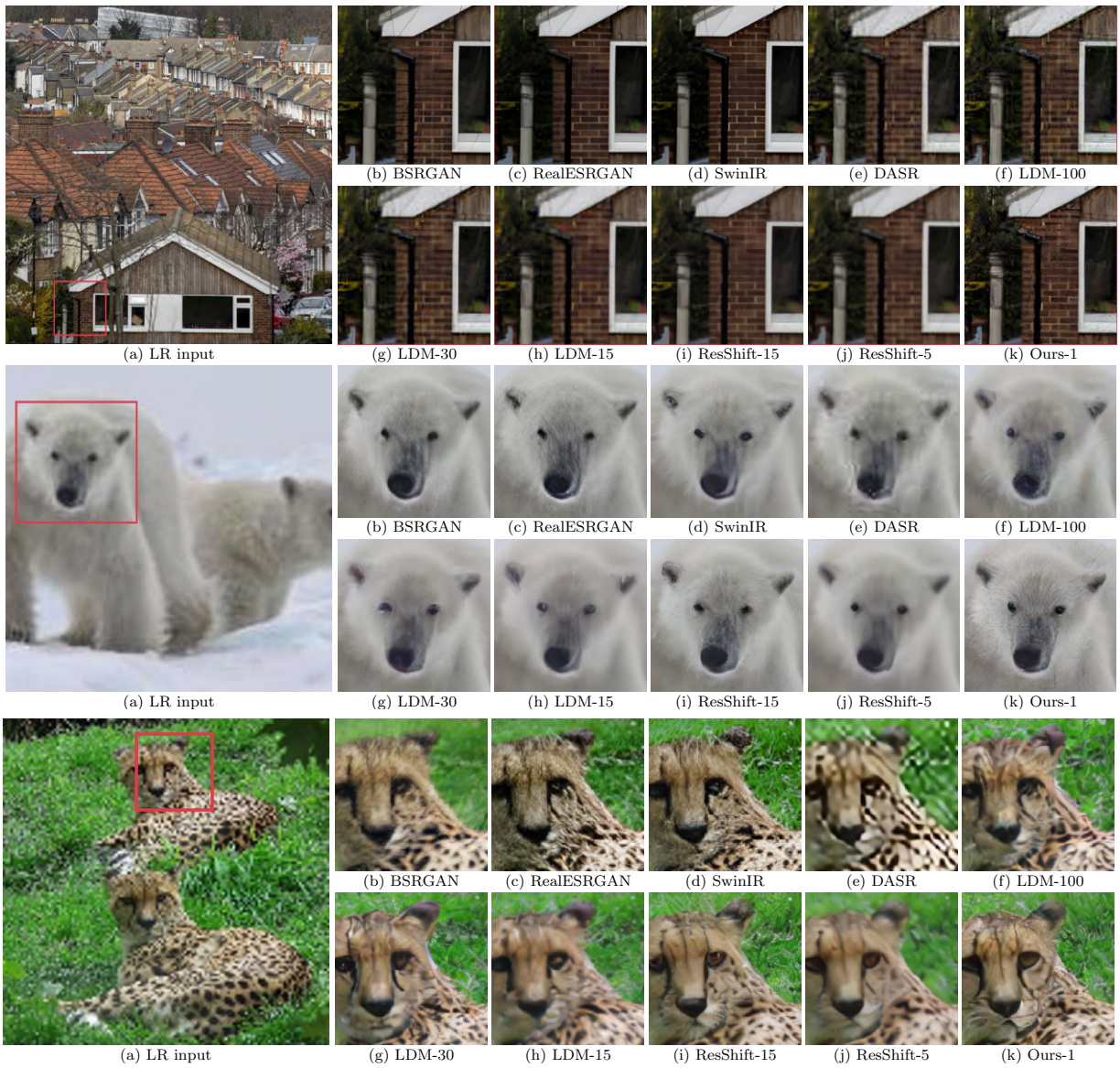


Figure 5. Visual comparison on real-world samples. Please zoom in for more details.

Methods	Datasets			
	RealSR		*RealSet65*	
	CLIPIQA↑	MUSIQ↑	CLIPIQA↑	MUSIQ↑
ESRGAN [40]	0.2362	29.048	0.3739	42.369
RealSR-JPEG [12]	0.3615	36.076	0.5282	50.539
BSRGAN [47]	0.5439	**63.586**	0.6163	**65.582**
SwinIR [18]	0.4654	59.636	0.5782	<u>63.822</u>
RealESRGAN [41]	0.4898	59.678	0.5995	63.220
DASR [19]	0.3629	45.825	0.4965	55.708
LDM-15 [32]	0.3836	49.317	0.4274	47.488
ResShift-15 [46]	0.5958	59.873	0.6537	61.330
SinSR-1 (distill only)	<u>0.6119</u>	57.118	<u>0.6822</u>	61.267
SinSR-1	**0.6887**	<u>61.582</u>	**0.7150**	62.169

Table 1. Quantitative results of models on two real-world datasets. The best and second best results are highlighted in **bold** and <u>underline</u>.

are reported in Table 1. As shown in the table, the proposed method with only one inference step can outperform the teacher model that we used by a large margin. Besides, for the latest metric CLIPIQA, the proposed method archives the best performance among all the competitors. Some visual comparisons are shown in Fig. 5, in which the proposed method achieves promising results using only one step.

Evaluation on synthetic datasets. We further evaluate the performance of different methods on the synthetic dataset *ImageNet-Test* following the setting in [46]. Specifically,

Methods	Metrics				
	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MUSIQ↑
ESRGAN [40]	20.67	0.448	0.485	0.451	43.615
RealSR-JPEG [12]	23.11	0.591	0.326	0.537	46.981
BSRGAN [47]	24.42	0.659	0.259	0.581	**54.697**
SwinIR [18]	23.99	0.667	0.238	0.564	53.790
RealESRGAN [41]	24.04	0.665	0.254	0.523	52.538
DASR [19]	24.75	**0.675**	0.250	0.536	48.337
LDM-30 [32]	24.49	0.651	0.248	0.572	50.895
LDM-15 [32]	<u>24.89</u>	0.670	0.269	0.512	46.419
ResShift-15 [46]	**24.90**	<u>0.673</u>	0.228	0.603	<u>53.897</u>
SinSR-1 (distill only)	24.69	0.664	<u>0.222</u>	<u>0.607</u>	53.316
SinSR-1	24.56	0.657	**0.221**	**0.611**	53.357

Table 2. Quantitative results of models on *ImageNet-Test*. The best and second best results are highlighted in **bold** and <u>underline</u>.

Metrics	Methods							
	LDM-15	LDM-30	LDM-100	ResShift-1	ResShift-5	ResShift-10	ResShift-15	*SinSR-1*
LPIPS↓	0.269	0.248	0.244	0.383	0.345	0.274	0.228	**0.221**
CLIPIQA↑	0.512	0.572	**0.620**	0.340	0.417	0.512	0.603	<u>0.611</u>
Runtime (bs=64)	0.046s	0.080s	0.249s	**0.012s**	0.021s	0.033s	0.047s	**0.012s**
Runtime (bs=1)	0.408s	1.192s	3.902s	**0.058s**	0.218s	0.425s	0.633s	**0.058s**
# Parameters (M)	113.60			118.59				118.59

Table 3. Efficiency and performance comparisons with SOTA methods on *ImageNet-Test*. "-N" represents the number of sampling steps the model used. The running time per image is tested on a Tesla A100 GPU on the x4 (64→ 256) task averaged over the batch size (bs).

3000 high-resolution images are first randomly selected from the validation set of ImageNet [7]. The corresponding LR images are obtained by using the provided script in [46]. As shown in Table 2, while reducing the inference step from 15 to only 1 slightly decreases PSNR and SSIM, the proposed method achieves the best perceptual quality measured by LPIPS, a more recent full-reference image quality assessment (IQA) metric than SSIM. Besides, the proposed method also achieves the best performance among all the methods measured on the most recent SOTA metric CLIP-IQA [39], demonstrating that the proposed 1-step model is on par with or even slightly better than the teacher model with 15 inference steps in terms of perceptual qualities.

Evaluation of the efficiency. We assess the computational efficiency of the proposed method in comparison to SOTA approaches. As shown in Table 3, the proposed method demonstrates superior performance with only one inference step, outperforming ResShift [46]—the adopted teacher model, which had already significantly reduced the inference time compared to LDM [32]. It is worth noting that all methods presented in Table 3 run in latent space, and the computational cost of VQ-VAE is counted.

5.3. Analysis

How important is the deterministic sampling? We evaluate the performance of the model trained on generated paired samples from the proposed deterministic sampling and the default stochastic sampling strategy $(x_T, \hat{F}_\theta(x_T, y))$ in [46]. Due to the randomness of the generated samples $x \sim \hat{F}_\theta(x_T, y)$, given a random noise ϵ, the prediction is an expectation of its conditional distribution. The comparison in Fig. 6 further verifies that the results trained w/o deterministic teacher model exhibit blurred details. Besides, as shown in Table 4, there is a significant performance degradation when we replace the proposed deterministic sampling with the default one in [46], demonstrating the effectiveness and necessity of involving the proposed deterministic sampling.

Methods	CLIPIQA ↑	MUSIQ ↑
w/ default sampling in [46]	0.4166	51.53
SinSR (distill only)	**0.6822**	**61.27**

Table 4. A comparison between the model trained with the default stochastic sampling process in ResShift [46] and the proposed deterministic sampling in Eq. 5 using only distillation loss. We evaluate their performance on the RealSet65 testing set.

(a) input (b) w/ sampling in [46] (c) Ours (deterministic)

Figure 6. A comparison between the model trained with the default stochastic sampling process in ResShift [46] and the proposed deterministic sampling in Eq. 5. Best zoom in for more details.

Why does a single-step distillation work? Previous studies suggest that directly learning the mapping between x_T and x_0 is typically challenging due to the non-causal properties of the generation process [20]. However, our empirical find-

Methods	CLIPIQA ↑	MUSIQ ↑
ResShift [46] (24.32M)	0.5365	52.71
ResShift [46] (118.59M)	**0.6537**	**61.33**
SinSR (distill only) (24.32M)	0.6499	58.71

Table 5. A comparison of the models trained with different strategies on RealSet65. The model trained with the diffusion loss, *i.e.*, ResShift, is more sensitive to the model size than directly learning the deterministic mapping between x_T and $\hat{x}_0$, indicating that the deterministic mapping is relatively easier to learn.

	LPIPS↓	MUSIQ↑	CLIPIQA↑
ResShift [46]	0.2275	**53.90**	0.6029
w/ Rectified Flow [21]	0.2322	51.05	0.5753
SinSR (distill only)	**0.2221**	53.32	**0.6072**

Table 6. A comparison between models accelerated by the proposed method and [46], which includes a reflow and a distillation operation. The models are evaluated on ImageNet-Test [46].

ings indicate that the matching between x_T and x_0 in the SR task is relatively easier to learn than denoising under different noise levels, as diffusion models do. Specifically, the capacity of the student network $f_{\hat{\theta}}$ is sufficient to effectively capture the ODE process F_θ using only one step. To verify our assumption, we evaluate the performance of smaller models trained under different strategies. Specifically, one model is trained following the experimental settings of [46] while the number of parameters decreases from 118.6M to 24.3M. Another model uses the same backbone as the aforementioned small model while directly learning the mapping relationship between x_T and $\hat{x}_0$ from the standard-size teacher diffusion model. A comparison between these two small models is reported in Table 5. As demonstrated by the results, the model trained for denoising under different noise levels suffers from a serious performance drop compared with the model that directly learns the deterministic mapping between. This strongly supports our assumption that directly learning the deterministic mapping is relatively easier.

Is a more sophisticated distillation strategy necessary? To explore the necessity of more advanced techniques that learn the mapping between x_T and x_0, we evaluate the performance of Rectified Flow [21], a recent method that learns the mapping to a single step through an iterative manner. Specifically, Reflow operations are conducted to avoid crossing the generation paths, and then followed by distilling the rectified generation process into a single step. However, as shown in Table 6, the involved iterative distillation degrades the performance of the final model due to the accumulated error as discussed by the author [21]. Besides, as verified by the previous section that the deterministic mapping between x_T and x_0 is easy to learn in the SR task, the benefit of a more sophisticated distillation strategy is not obvious.

Learned inversion. As the core of the consistency preserving loss, a comparison with the DDIM inversion [37] is shown in Fig. 7, where the proposed method achieves better fidelity performance. It indicates that the proposed method can obtain a more accurate estimation of x_T. Besides, more analyses regarding the consistency preserving loss are in the supplementary material.

LR image DDIM inversion Learned inversion GT image

Figure 7. A comparison between HR images generated from DDIM inversion and the proposed learned inversion. Zoom in for details.

Training overhead. While the proposed method involves solving ODEs during training, benefiting from a shortened inference process and initializing the student model from the pre-trained teacher model, the training cost of finetuning using the proposed training paradigm is still lower than that of retraining the diffusion model from scratch. Specifically, the training cost is shown in Table 7.

	Num of Iters	s/Iter	Training Time
ResShift [46]	500k	1.32s	∼7.64 days
SinSR (Ours)	30k	7.41s	∼2.57 days

Table 7. A comparison of the training cost on an NVIDIA A100.

6. Conclusion

In this work, we propose a novel strategy to accelerate the diffusion-based SR models into a single inference step. Specifically, a one-step bi-directional distillation is proposed to learn the deterministic mapping between the input noise and the generated high-resolution image and versa vice from a teacher diffusion model with our derived deterministic sampling. Meanwhile, a novel consistency preserving loss is optimized at the same time during the distillation so that the student model not only uses the information from the pre-trained teacher diffusion model but also directly learns from ground-truth images. Experimental results demonstrate that the proposed method can achieve on-par or even better performance than the teacher model in only one step.

Acknowledgements

The research is supported in part by the NTU-PKU Joint Research Institute (a collaboration between the Nanyang Technological University and Peking University that is sponsored by a donation from the Ng Teng Fong Charitable Foundation), National Key R&D Program of China under Grand NO.2022ZD0160100, the National Natural Science Foundation of China under Grant No. 62102150, the Science and Technology Commission of Shanghai Municipality under Grant No. 23QD1400800, No. 22511105800, and the Basic and Frontier Research Project of PCL and the Major Key Project of PCL.

References

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Image super-resolution via progressive cascading residual network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 791–799, 2018. 2

[2] FirstName Alpher. Frobnication. *IEEE TPAMI*, 12(1):234–778, 2002. 2

[3] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2019. 5

[4] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[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*, pages 12413–12422, 2022. 1, 2

[6] Ryan Dahl, Mohammad Norouzi, and Jonathon Shlens. Pixel recursive super resolution. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5, 7

[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*, 38(2):295–307, 2015. 2

[9] 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

[10] Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, et al. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. *arXiv preprint arXiv:2402.10491*, 2024. 2

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 3

[12] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *CVPR*, pages 466–467, 2020. 5, 6, 7

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 2

[14] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 1, 2

[15] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 1, 5

[16] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[17] 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

[18] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, pages 1833–1844, 2021. 5, 6, 7

[19] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *European Conference on Computer Vision*, pages 574–591. Springer, 2022. 5, 6, 7

[20] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 7

[21] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2022. 3, 5, 8

[22] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 1, 2

[23] 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. 2

[24] Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021. 3

[25] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023. 3

[26] Jacob Menick and Nal Kalchbrenner. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. *International Conference on Learning Representations,International Conference on Learning Representations*, 2018. 2

[27] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[28] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1, 2

[29] Aaronvanden Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, 2016. 2

[30] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, 2018. 2

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[32] 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, 3, 4, 5, 6, 7

[33] 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

[34] Mehdi S. M. Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[35] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2021. 3

[36] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 4, 8

[38] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023. 2

[39] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 5, 7

[40] 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. 5, 6, 7

[41] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 5, 6, 7

[42] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2604–2612, 2022. 2

[43] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C Kot, and Bihan Wen. Exposurediffusion: Learning to expose for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12438–12448, 2023. 2, 3

[44] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and ThomasS. Huang. Deep networks for image super-resolution with sparse prior. *Cornell University - arXiv,Cornell University - arXiv*, 2015. 2

[45] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3365–3387, 2020. 1, 2

[46] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[47] 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. 5, 6, 7

[48] 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. 5