SkipDiff: Adaptive Skip Diffusion Model for High-Fidelity Perceptual Image Super-resolution

Xiaotong Luo[1], **Yuan Xie**[2*], **Yanyun Qu**[1*], **Yun Fu**[3]

[1]School of Informatics, Xiamen University, Fujian, China
[2]School of Computer Science and Technology, East China Normal University, Shanghai, China
[3]Northeastern University
xiaotluo@stu.xmu.edu.cn, yyqu@xmu.edu.cn

Abstract

It is well-known that image quality assessment usually meets with the problem of perception-distortion (p-d) tradeoff. The existing deep image super-resolution (SR) methods either focus on high fidelity with pixel-level objectives or high perception with generative models. The emergence of diffusion model paves a fresh way for image restoration, which has the potential to offer a brand-new solution for p-d trade-off. We experimentally observed that the perceptual quality and distortion change in an opposite direction with the increase of sampling steps. In light of this property, we propose an adaptive skip diffusion model (SkipDiff), which aims to achieve high-fidelity perceptual image SR with fewer sampling steps. Specifically, it decouples the sampling procedure into coarse skip approximation and fine skip refinement stages. A coarse-grained skip diffusion is first performed as a high-fidelity prior to obtaining a latent approximation of the full diffusion. Then, a fine-grained skip diffusion is followed to further refine the latent sample for promoting perception, where the fine time steps are adaptively learned by deep reinforcement learning. Meanwhile, this approach also enables faster sampling of diffusion model through skipping the intermediate denoising process to shorten the effective steps of the computation. Extensive experimental results show that our SkipDiff achieves superior perceptual quality with plausible reconstruction accuracy and a faster sampling speed.

Introduction

Single image super-resolution (SISR) aims to reconstruct a high-resolution (HR) image from its degraded low-resolution (LR) version. Image quality is typically evaluated by several fidelity/distortion measures (e.g., PSNR, SSIM) or subjective metrics that quantify perceptual quality (e.g., NIQE, LPIPS), which are at odds with each other and lead to the problem of perception-distortion (p-d) tradeoff (Blau and Michaeli 2018). The existing deep SR methods include PSNR-oriented (Chen et al. 2023; Wang et al. 2023) and perception-oriented models (Zhang et al. 2022a; Park, Moon, and Cho 2023). The former aims to minimize pixel-level distortion (e.g., l_1 or l_2), which usually leads to over-smooth results and looks not realistic. The latter pursues to produce authentic texture details by modeling the target distribution with generative models, e.g., generative adversarial network (GAN) (Zhang et al. 2022a) and normalizing flow (Lugmayr et al. 2020). These SR methods focus on either high fidelity or high perception, but few discuss how to make a flexible solution for the p-d tradeoff.

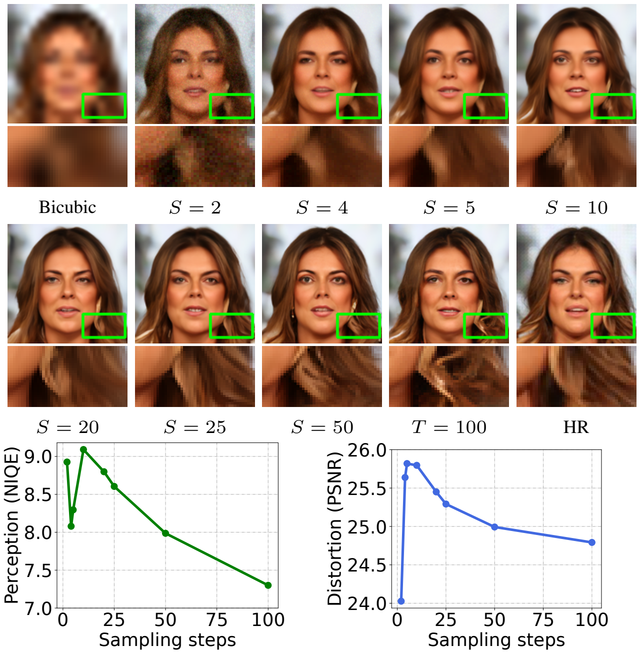

Figure 1: The observation that different sampling steps induce an approximate monotonic tendency to the image quality. *Top:* The visual examples of different sampling steps S (full steps $T = 100$) on CelebA-HQ for $8\times$ face SR. The SR result tends to transition from smooth to realistic with the increase of sampling steps. *Bottom:* The perception and distortion performance comparison of different sampling steps.

Recent advances (Nichol and Dhariwal 2021) in diffusion models (DDPM) shed light on a new pathway for image restoration (Wang, Yu, and Zhang 2023). They generate high-quality SR images by progressive sampling from a latent distribution (Ho, Jain, and Abbeel 2020). It is observed that different sampling steps induce an approximate

[*]Corresponding authors.

monotonic tendency to the image quality. Meanwhile, the SR result tends to transition from smooth to realistic with the sampling steps increased in Fig. 1. The observation indicates that diffusion models have the potential to balance perception and distortion. However, it is unexplored how to use DDPM to present a solution for the p-d tradeoff.

In this paper, we aim to explore a new way of using DDPM for the p-d tradeoff. Considering that different sampling steps essentially activate different denoising granularities, coarse time steps lead to blurry samples with low distortion, whereas fine time steps lead to crisp samples with high perceptual quality. It inspires us to combine different noise levels with adaptive skipping operations along the full diffusion to obtain a more accurate and realistic SR image.

Therefore, we propose an adaptive skip diffusion model (SkipDiff) for high-fidelity perceptual image SR. Specifically, our SkipDiff consists of coarse skip approximation (CSA) and fine skip refinement (FSR) stages, where CSA is performed first to provide initialization for FSR. Based on the characteristic of typical diffusion models shown in Fig. 1, we design CSA as a fidelity-driven process, which conducts a coarse-grained diffusion with predefined steps to approximate the full diffusion for obtaining blurry samples with low distortion. Conversely, FSR is a perception-driven process, which performs a fine-grained diffusion with learned time steps to refine the output of CSA for generating sharp images with high perceptual quality. To determine the precise fine time steps adaptively, a lightweight policy network is introduced to maximize the reward function measured by perceptual metric. Note that we hold the training process of DDPM unchanged and only modify the inference process so as to sample from a conditional distribution for obtaining multiple outputs. This scheme essentially provides a new solution to p-d tradeoff through a prior-guided (Fig. 1) adaptive rearrangement on the noise level to be removed at each step of the full diffusion model.

The main contributions of this work are three-fold:

- We propose an adaptive skip diffusion model (SkipDiff) for image SR, which can lead to a perceptually better and more accurate prediction of the HR image.
- We perform the sampling with a coarse-grained skip diffusion and a fine-grained skip diffusion as a way to traverse through the p-d plane, where the fine time steps are adaptively learned by deep reinforcement learning.
- Extensive experiments show that our SkipDiff achieves the best perceptual metrics with higher reconstruction accuracy on the face and natural image SR.

Related Work

Deep SR Model

PSNR-oriented models. This kind of method mainly uses pixel-level loss as the optimization objective. As a preliminary, SRCNN (Dong et al. 2016) utilizes a three-layer convolutional neural network for SISR. Then, numerous excellent deep SR models have sprung up, where residual connections and recursive learning are widely adopted (Zhang et al. 2018; Xin et al. 2022). Besides, attention mechanism (Zhang et al. 2022b; Li et al. 2023; Gao et al. 2023a) has been explored for mining the underlying non-local self-similarity for image SR. Although these methods have achieved significant performance on reconstruction accuracy, they often result in over-smooth results and lack realism.

Perception-oriented models. This kind of method aims to generate realistic SR results by fitting the distribution of target data, where most works are based on generative adversarial network (GAN) (Zhang et al. 2022a). RankSRGAN (Zhang et al. 2022a) introduces Ranker to optimize the generator in the direction of perceptual metrics. FSRNet (Chen et al. 2018) utilizes the facial geometry prior and the adversarial training to reconstruct realistic faces. PULSE (Menon et al. 2020) adopts an alternative formulation with a downscaling loss to generate realistic faces in an entirely self-supervised fashion. DGP (Pan et al. 2022) exploits the image prior captured by GAN on large-scale datasets for image restoration. GCFSR (He et al. 2022) proposes a controllable SR framework to reconstruct faithful face identity information while not adding extra priors. Flow-based methods (Li et al. 2021; Yuan et al. 2023) aim to map the target data to a latent space, where the distribution is factorized by a sequence of learnable invertible functions. The perception-oriented methods can obtain satisfying visual results but sacrifice the reconstruction fidelity.

Denosing Diffusion Probability Model

Nowadays, DDPM has exposed great potential in image generation (Bansal et al. 2023b; Zhou et al. 2023). ILVR (Choi et al. 2021) induces the denoising process in DDPM to produce high-quality results according to a given reference image. SR3 (Saharia et al. 2022) and SRDiff (Li et al. 2022) apply DDPM into image SR while exhibiting strong performance. Whang et al. (Whang et al. 2022) proposes a conditional DDPM for perceptual blind image deblurring, achieving significantly improved perceptual quality and competitive distortion metrics. Cold diffusion (Bansal et al. 2023a) generalizes diffusion models with arbitrary image transformations rather than built on Gaussian noise. DDRM (Kawar et al. 2022) proposes a general linear inverse problem solver based on unconditional or class-conditional DDPM. DADA (Metzger, Daudt, and Schindler 2023) unites a convolutional network and guided anisotropic diffusion, which achieves superior performance for guided depth SR.

Typically, DDPM requires numerous diffusion steps to generate a high-quality sample, which leads to very slow inference. To address this, the existing works mainly focus on constructing deterministic sampling path (Song, Meng, and Ermon 2021; Nichol and Dhariwal 2021), estimating the noise level to adjust the noise schedule (San-Roman, Nachmani, and Wolf 2021), searching for the optimal discrete time schedules (Watson et al. 2021) or truncating the reverse diffusion early by relying on a pre-trained model to obtain an initialization (Chung, Sim, and Ye 2022; Lyu et al. 2022). All these methods rely on a critical decoupling property of DDPMs, i.e., the training schedule can differ from the inference schedule. Our work mainly explores a new way for the p-d tradeoff in image SR based on DDPM, which can improve SR performance while reducing diffusion steps.

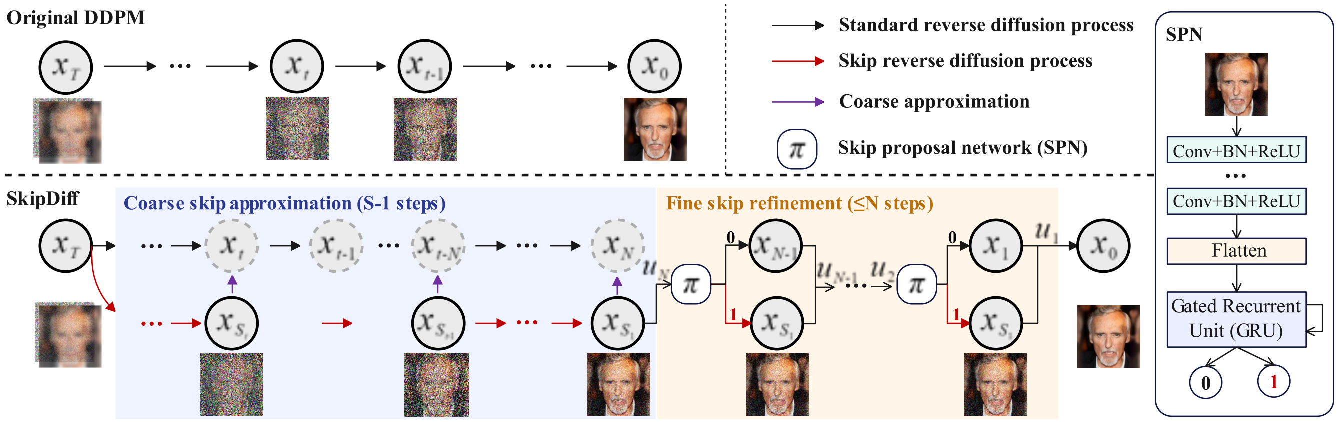

Figure 2: The overall framework of SkipDiff. Original DDPM generates SR images by sequential reverse diffusion. Instead, SkipDiff first performs coarse-grained skip diffusion to obtain a rough approximation of the full diffusion for low distortion, and then conducts fine-grained skip diffusion guided by deep reinforcement learning for promoting perception.

Deep Reinforcement Learning

Deep reinforcement learning (DRL) (Yu et al. 2018a) has been widely applied in robotic control (Rana et al. 2023), game player (Stephens and Exton 2022) and computer vision (Le et al. 2022). It is used to solve the problems of decision optimization. The basic idea of DRL is that agents constantly adjust the strategies based on the rewards received from the interaction with the environment to achieve the optimal decision. EAST (Huang, Lucey, and Ramanan 2017) learns an agent to make a decision for adaptive object tracking. Attention-FH (Cao et al. 2017) utilizes RL to explore the rich correlation cues among different facial parts for face hallucination. RL-Restore (Yu et al. 2018b) investigates restoration tool selection in an RL framework for complex degradation. GFNet (Wang et al. 2020) proposes a dynamic decision framework for efficient image classification. In this paper, we use DRL to adaptively learn the precise fine timesteps to obtain better perceptual quality.

Proposed Method

Overview

As illustrated in Fig. 2, original DDPM generates high-quality samples by starting from a pure Gaussian noise conditioned on the bicubic interpolated LR image to perform a series of reverse diffusion. It is a standard Markov chain, i.e., the distribution of the current state only depends on the previous state, which is time-consuming. Based on the observation that coarse time steps lead to blurry samples with low distortion while fine time steps lead to sharp images with high perceptual quality, we propose an adaptive skip diffusion method (SkipDiff) to achieve high-fidelity perceptual image SR, which can be viewed as a Markov jump process. SkipDiff consists of coarse skip approximation (CSA) and fine skip refinement (FSR) stages. To reduce distortion, CSA performs a coarse-grained diffusion to approximate the latent sample of full diffusion. To promote perception, the output of CSA is input to FSR for fine-grained diffusion. Due to the *unmatched noise level* between the latent sample of CSA (x_{S_1}) and the approximated one of full diffusion (x_N), the actual refined steps for FSR are hard to predict and not the more the better. Therefore, we introduce deep reinforcement learning to learn the specific fine time steps by maximizing the reward function measured by perceptual metrics.

Notation and Formulation

We first provide a basic formulation review of DDPM (Ho, Jain, and Abbeel 2020). It defines two Markov chains: a parameter-free forward diffusion process q and a learned reverse denoising process p. The former perturbs the target data distribution x_0 to the latent variable x_T by gradually adding Gaussian noise, which can be formulated as:

$$q\left(x_1, \cdots, x_T \mid x_0\right) = \prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right), \tag{1}$$

$$q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t | \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}\right), \tag{2}$$

where $\mathbf{I}$ is an identity matrix. T is the total number of diffusion steps. $x_1, \cdots, x_T$ are the latent samples with the same dimensionality as x_0. $\beta_1, \cdots, \beta_T \in (0, 1)$ are the monotonically increasing noise variance schedule. $\mathcal{N}(x_t|\mu, \sigma)$ denotes the Gaussian distribution with mean μ and variance σ for the variable x_t. Note that the forward diffusion can also be conducted through a single-step diffusion, i.e.,

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \tag{3}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{0}$ is an all-zero matrix. Besides, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ for $\alpha_t = 1 - \beta_t$.

The reverse denoising diffusion process recovers $q(x_0)$ by gradually predicting the added noise at each step from a pure Gaussian noise x_T, which can be represented as:

$$p(x_T) = \mathcal{N}(x_T | \mathbf{0}, \mathbf{I}), \tag{4}$$

$$p_\theta(x_0, \cdots, x_T) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{5}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\mu_\theta(x_t, t), \delta_t^2 \mathbf{I}), \tag{6}$$

where $\mu_\theta(x_t, t)$ is parameterized by a neural network. δ_t is a timestep dependent constant. We set $\delta_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t$ as defined in (Ho, Jain, and Abbeel 2020). This reverse process generates high-quality samples iteratively with the neural network repeated for T times.

Specifically, the objective function minimizes the variant of the variational lower bound with x_0 and t as inputs:

$$\min_{\theta} L_{t-1}(\theta) = \mathbb{E}_{x_0,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2\right], \tag{7}$$

where ϵ_θ is a noise predictor. More details about DDPM can refer to (Ho, Jain, and Abbeel 2020).

Coarse Skip Approximation

CSA performs uniform skip operations on the full diffusion model to keep high fidelity. It is a Markov jump process used as a rough approximation for the full Markov chain (Opper and Sanguinetti 2007). Next, we present the diffusion process and the diffusion steps of CSA.

Coarse diffusion. Inspired by (Nichol and Dhariwal 2021; Song, Meng, and Ermon 2021), the noise variances $\delta^2_{S_t}$ of coarse timesteps (S) is formulated as:

$$\delta^2_{S_t} = \frac{1-\bar{a}_{S_{t-1}}}{1-\bar{a}_{S_t}}\beta_{S_t}, \tag{8}$$

where $\beta_{S_t} = 1 - \frac{\bar{a}_{S_t}}{\bar{a}_{S_{t-1}}}$ and $S_t \in \{1, 2, \cdots, S\}$. We adopt the cosine noise variance schedule, since it can be automatically rescaled for the coarse diffusion. It is defined as:

$$\bar{a}_t = \frac{f(t)}{f(0)}, f(t) = \cos\left(\frac{t/T+s}{1+s}\cdot\frac{\pi}{2}\right)^2, \tag{9}$$

where T is the diffusion steps, s is a hyper-parameter and is usually set to 0.008. Therefore, the reverse diffusion process of the coarse timesteps can be formulated as:

$$p_\theta(x_{S_{t-1}}|x_{S_t}) = \mathcal{N}(x_{S_{t-1}}|\mu_\theta(x_{S_t}, S_t), \delta^2_{S_t}\mathbf{I}). \tag{10}$$

Coarse steps. As shown in Fig. 2, CSA first performs $S-1$ diffusion steps at a uniform skip interval of N to obtain the x_{S_1}, which aims to approximate the corresponding x_N of the full diffusion. Then, the last step of CSA is replaced with fine-grained diffusion of FSR for further refinement.

In fact, the latent diffusion results depend on the parameterized μ_θ and the untrained δ_t in Eq. (6). We make the skip interval N divisible by T, which can satisfy:

$$S_t/S = t_N/T, \tag{11}$$

where t_N is the timestep with an interval of N. Furthermore, we can get $\bar{\alpha}_{S_t} = \bar{\alpha}_{t_N}$ and $\beta_{S_t} = \beta_{t_N}$ by Eq. (9). Then, the noise variance δ^2 between the coarse diffusion and the corresponding full diffusion can be matched, i.e.,

$$\delta^2_{S_t} = \delta^2_{t_N}. \tag{12}$$

Therefore, the coarse steps are divisible by full steps to obtain a coarse approximation. Note that we simply provide one way to achieve coarse-grained sampling, and other sampling ways are also encouraged.

Fine Skip Refinement

As mentioned above, the last latent sample from CSA is input to FSR to promote perception by fine-grained diffusion. While the perceptual quality is monotonically reduced with the increasing sampling steps in Fig. 1, using more sampling steps for refinement may not achieve optimality. The reasons lie in the unmatched noise level between x_N and x_{S_1}, since we only match the noise variance in Eq. (12). Therefore, it is necessary to adaptively learn the fine time steps rather than artificially setting a simple condition judgment.

Considering that FSR is a perception-driven process, it aims to obtain SR output with high perceptual quality. However, it is not easy to map the timesteps to perceptual metrics (e.g., NIQE, LPIPS). There is no way to implement the backpropagation to learn the timesteps by directly optimizing NIQE with gumbel softmax or straight through estimator. Therefore, we introduce deep reinforcement learning (DRL) to adaptively determine a proper timestep based on different latent observations to start fine-grained reverse diffusion. Specifically, we treat the problem of fine time steps determination as a decision-making procedure, where the agent can adaptively choose whether to skip the current diffusion step as so to obtain the optimal perceptual quality. In the following part, we will give the definition of action, state, reward, and the network structure of the agent.

Action. The action space A represents all possible actions for the agent to take. To determine whether to skip the current diffusion step, A is defined as a binary set like:

$$A = \{0, 1\}, \tag{13}$$

where 0 is to perform the reverse diffusion process (D) with the current state, while 1 is to skip the diffusion process.

State. The state refers to the current observation information, which is input to the agent for making the next decision. Here, we define the state set $U = \{u_N, \cdots, u_1\}$ with

$$u_t = \begin{cases} x_{S_1}, & t = N \\ D(u_{t+1}), & if \quad a_t = 0 \\ x_{S_1}, & if \quad a_t = 1 \end{cases} \tag{14}$$

where a_t is the action taken at timestep t. The coarse approximation x_{S_1} is used as the initial state. Note that the decision process is performed for $N-1$ times, and the state u_1 is input to ϵ_θ for generating the final SR result.

Reward. The reward is crucial in DRL, which acts as the objective function to guide the training of the policy network. The agent learns to make different decision trajectories for maximizing the accumulated reward. Here, the agent is expected to learn an optimal policy for generating samples with high perceptual quality. Thus, the reward function R is well-designed as:

$$R = \begin{cases} 0, & t = 1, \cdots, N \\ 1/M, & t = 0 \end{cases} \tag{15}$$

where M denotes the perceptual metric like NIQE or LPIPS, which is usually the lower the better. Besides, it is an alternative to utilize other perceptual measurements as the reward, e.g., perceptual loss. Since we only concern about the effect of each decision on the perceptual quality of the final output, the reward of the intermediate process is all set to 0.

Skip proposal network (SPN). The agent is modeled as a policy network π to output the discrete action proposal about whether to skip the current diffusion step. As shown in Fig. 2, SPN is a lightweight network with almost negligible computation burden, which is composed of several convolutional layers and each layer follows a batch normalization and ReLU function, and a gate recurrent unit (GRU) with flattened vectors as input for storing the historical hidden state information. Then, we can get the probability distribu-

Method	Bicubic	FSRNet	PULSE	GCFSR	SR3	IDM	SkipDiff (ours)
PSNR	23.49	24.73	21.37	25.06	24.79	24.08	**25.60**
SSIM	0.6003	**0.7086**	0.4839	0.6774	0.6766	0.6798	0.6838
NIQE	13.45	9.55	8.57	6.73	7.30	7.13	**6.47**
LPIPS	0.5374	0.2179	0.2026	0.1725	0.0992	0.1359	**0.0967**

Table 1: Quantitative results on CelebA-HQ for 8× face SR. The best and second best are highlighted in bold and underline.

Method	Bicubic	DGP (1000)	DDRM (20)	SR3 (100)	SkipDiff (20) (ours)
PSNR	25.60	23.01	**26.68**	26.38	26.65
SSIM	0.6594	0.5223	**0.7089**	0.6871	0.6884
NIQE	8.68	5.31	9.48	5.00	**4.95**
LPIPS	0.4716	0.2531	0.3499	0.1909	**0.1886**

Table 2: Quantitative results on ImageNet for 4× natural SR. The best and second best are highlighted in bold and underline.

tion p_t of the output action vector by:

$$p_t = \pi(u_t, \theta_s), \tag{16}$$

where θ_s denotes the network parameters of SPN. At test time, the action a_t with the maximum probability value, i.e., $a_t = \text{argmax}(p_t)$, is used for the current decision.

Training Strategy and Complexity Analysis

Training. We adopt a two-step strategy to train SkipDiff. In Step 1, we train the DDPM conditioned with bicubic interpolated LR images with Eq. (7). In Step 2, we perform the coarse inverse diffusion with predefined steps based on the pre-trained DDPM, and then train SPN with the classical proximal policy optimization (PPO) algorithm (Schulman et al. 2017), which has been proven to be effective and easy to converge. Note that we train and test SPN only on the final N diffusion steps for FSR. The whole optimization is easy to implement, since we only change the sampling way and introduce a lightweight policy network to indicate which skips should be skipped to maximize the perceptual reward with the high-fidelity prior of CSA. The detailed procedure is provided in supplementary materials.

Computational complexity. Here, we mainly analyze the computational complexity of our SkipDiff. The different p-d tradeoff results can be obtained by setting different (CSA_steps, FSR_steps) pairs, i.e., $(S-1, T/S)$. Actually, much of the sampling efficiency comes from CSA, which takes $S-1$ steps, while FSR takes at most T/S or N steps. Therefore, the diffusion steps can be reduced from T to at most $(S-1+T/S)$ theoretically.

Experiments

Experimental Settings

Datasets. Following (Kawar et al. 2022; Pan et al. 2022; Saharia et al. 2022), we train and evaluate our SkipDiff on faces (8×) and natural images (4×). For face SR, we train the models at $16 \times 16 \rightarrow 128 \times 128$ on Flickr-Faces-HQ (FFHQ) dataset, which includes 70k images in total, and we sample 1k images from CelebA-HQ dataset for evaluation. For natural SR, we train at $64 \times 64 \rightarrow 256 \times 256$ on the high-diversity ImageNet 1K (Russakovsky et al. 2015) dataset and evaluate on 1k images from its dev split.

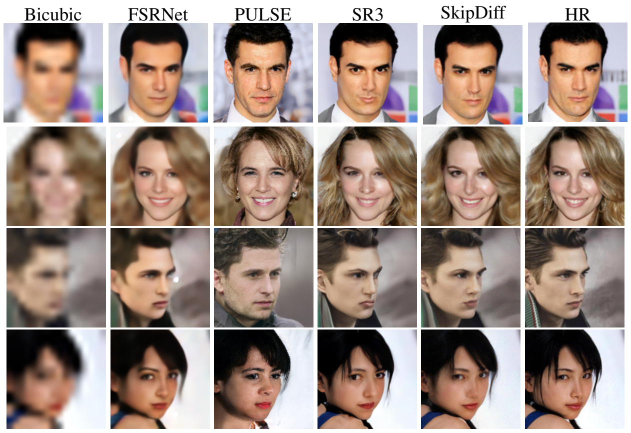

Figure 3: Visual comparisons on CelebA-HQ dataset for 8× face SR. Zoom in for a better view.

Implementation details. Following (Li et al. 2022), we set the total diffusion steps T as 100 and adopt a cosine noise schedule (Nichol and Dhariwal 2021) for β_t, which has been proven beneficial for training. Actually, β_t is tailored to be smaller than 0.999 for preventing singularities at the end of the diffusion, i.e, near $t = T$. Other experimental configurations are the same with (Saharia et al. 2022) to train the DDPM. Besides, the amount of convolutional layers in SPN is set to 4 and 5 for FFHQ and ImageNet datasets for the different image resolutions. We train SPN for 50k iterations with a learning rate of $3e^{-4}$. The hyperparameters for optimizing SPN are set like (Schulman et al. 2017; Wang et al. 2020): the clipping parameter $\epsilon = 0.2$, $\gamma = 0.7$, $c_1 = 0.5$, $c_2 = 0.01$ and $\lambda = 1$. All these hyperparameters are fixed across different datasets. The entire SkipDiff is trained and evaluated on 1 NVIDIA Tesla V100 cards.

Evaluation criterion. We evaluate the distortion measurement with PSNR and SSIM, which are computed on the Y channel (the luminance channel) of the YCbCr space. In addition, we evaluate the perceptual quality with NIQE (a non-reference metric) and LPIPS (a reference-based metric that calculates the perceptual similarity between the HR and the SR images).

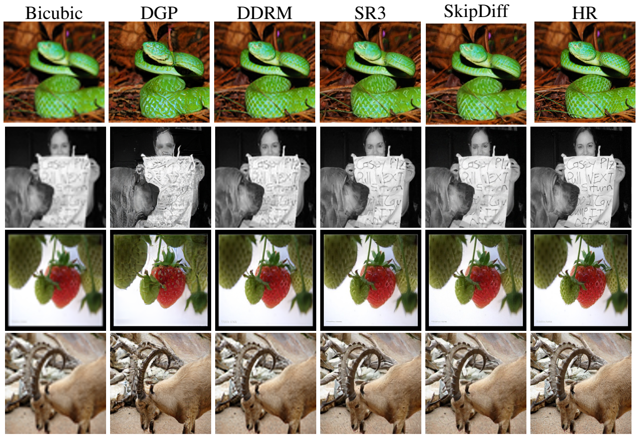

Figure 4: Visual comparisons on ImageNet validation dataset for 4× image SR. Zoom in for a better view.

Method	DGP	DDRM	SR3	SkipDiff
Time (s)	170.04	2.23	4.17	**0.69**
FLOP (G)	72,197	22,275	17,799	**2,848**

Table 3: Runtime and FLOPs comparisons on ImageNet.

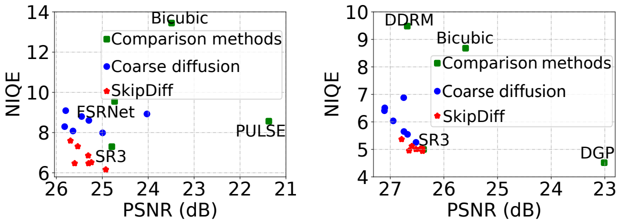

Figure 5: The perception-distortion plane on CelebA-HQ (left) and ImageNet (right). Zoom in for better view.

Case index	1	2	3	4	5
CSA	×	✓	×	✓	✓
FSR	×	×	✓ (100)	✓ (100)	✓ (10)
PSNR	24.79	25.80	24.81	25.59	25.60
SSIM	0.6766	0.7204	0.6761	0.6835	0.6838
NIQE	7.30	9.09	7.31	6.45	6.47
LPIPS	0.0992	0.1160	0.0987	0.0968	0.0967
Steps	100	10	100	16 (9+7)	16 (9+7)

Table 4: Ablation studies about SkipDiff: coarse skip approximation (CSA) and fine skip refinement (FSR) on CelebA-HQ dataset for 8× SR.

Comparison with State-of-the-arts

Face SR. We first evaluate the effectiveness of SkipDiff on face images and compare it with several face SR methods: Bicubic, FSRNet (Chen et al. 2018), PULSE (Menon et al. 2020), GCFSR (He et al. 2022), SR3 (Saharia et al. 2022) and IDM (Gao et al. 2023b) in Table 1. It shows that SkipDiff achieves superior results in a comprehensive performance under the coarse-to-fine step pair of (10,10). Therefore, our proposal achieves high-fidelity perceptual results on face SR. Besides, we present the visual comparisons in Fig. 3, which shows that SkipDiff recovers more realistic and high-fidelity face details than other methods.

Natural SR. We also evaluate our SkipDiff on natural images and compare it with Bicubic, DGP (Pan et al. 2022), DDRM (Kawar et al. 2022), and SR3 (Saharia et al. 2022) in Table 2. Note that the number in the bracket is the average model iteration steps. SkipDiff (20) obtains the best NIQE and LPIPS with comparable PSNR and SSIM against DDRM under the step pair of (5,20). Similarly, our method achieves excellent results on the p-d tradeoff for the natural image. We also give several visual comparisons in Fig. 4. It shows that DGP generates oversharp results with more artifacts, and DDRM produces smooth results. SR3 gets realistic details and texture, while SkipDiff obtains more satisfying results with less distortion. Furthermore, we provide the runtime and computation cost (FLOPs, Floating Point of Operations) comparison in Table 3, which shows that SkipDiff runs quite faster than other methods.

Perception-distortion plane. To illustrate various p-d results obtained by SkipDiff, we depict the p-d plane on CelebA-HQ and ImageNet in Fig. 5, which are obtained by setting different (CSA_steps, FSR_steps) pairs. We also add the results of other competitive methods given in Table 1 and Table 2, as well as the coarse diffusion with different sampling steps (2, 4, 5, 10, 20, 25, 50). It shows that our SkipDiff has an obvious advantage in comprehensive performance, which achieves a better perceptual quality with high fidelity than other methods.

Ablation Study

To demonstrate the effectiveness of SkipDiff, we conduct an ablation study on CelebA-HQ for 8× face SR to analyze different elements, including CSA, FSR, and reward function. Our baseline is the DDPM with $T = 100$ and adopts the same UNet with (Saharia et al. 2022) as the backbone. For the convenience of analysis, we mainly take the coarse-to-fine step pair $(10, 10)$ as an example.

Effect of coarse skip approximation (CSA). To evaluate the effect of CSA, we perform the coarse diffusion with the sampling steps $S = 10$. As Case 2 shows in Table 4, it obtains better PSNR and SSIM but poor NIQE and LPIPS than the baseline (Case 1). Besides, we give the visual comparison of the latent samples for the coarse diffusion and the corresponding full diffusion in Fig. 6. It shows that the latent results of $S = 10$ look similar to the ones of $T = 100$. Therefore, CSA helps to obtain SR results with low distortion as a rough approximation of the full diffusion.

Effect of fine skip refinement (FSR). To evaluate the effect of FSR, we conduct the following experiments:

1) We remove CSA and only perform FSR, i.e., learn to optimize which steps to denoise over the full diffusion process. As Case 3 shows in Table 4, it obtains a similar result as the baseline (Case 1), and no diffusion steps are skipped. Due to the lack of fidelity prior of CSA, the large decision space on the whole process (2^{T-1}) makes it hard to find the optimal solution. Our SkipDiff runs CSA first for high fidelity, and then runs FSR for promoting perception, which

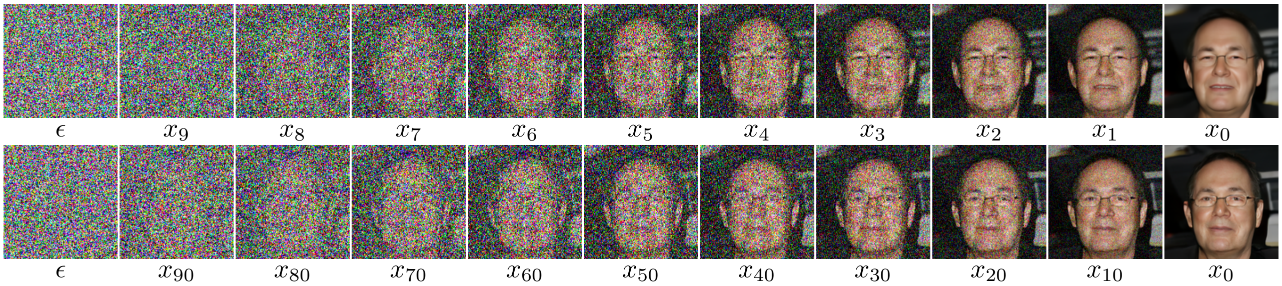

Figure 6: The visual comparisons for the latent reverse diffusion results of the coarse diffusion (top) and full diffusion (bottom).

Fine timestep	9	8	7	6	5	4	3	2	1	SkipDiff (16)
PSNR	**25.65**	25.61	25.56	25.50	25.31	25.03	24.64	24.15	23.60	25.60
SSIM	**0.7066**	0.7005	0.6836	0.6480	0.5925	0.5270	0.4615	0.4017	0.3512	0.6838
NIQE	7.13	6.53	**6.47**	7.22	8.26	9.27	10.34	11.39	12.63	**6.47**
LPIPS	0.0984	0.0965	**0.0964**	0.1022	0.1200	0.1526	0.1974	0.2518	0.3104	0.0967

Table 5: The quantitative results of manual fine refinement on CelebA-HQ dataset.

M	PSNR	SSIM	NIQE	LPIPS
NIQE	25.60	0.6838	6.47	0.0967
LPIPS	25.60	0.6839	6.47	0.0961

Table 6: The comparisons of different perceptual metrics M as the reward function on CelebA-HQ dataset.

provides an effective constraint to shrink the solution space. In our method, the decision space is reduced to 2^{N-1}, where $2^9 \ll 2^{99}$. Therefore, CSA and FSR are integral to achieving the high-fidelity perceptual image SR. Note that it is not reasonable to exchange the coarse and fine skip operations, since the last few diffusion steps contribute a lot to perceptual quality, which should be executed later.

2) We combine CSA with FSR with the coarse-to-fine step pair of (10, 100), i.e., x_{S_1} is used as the initial state at the timestep $t = 100$ and then perform fine-grained diffusion guided by DRL. It (Case 4) still obtains a similar performance as the step pair (10, 10) (Case 5) benefiting from the high fidelity prior of coarse diffusion. However, it took nearly five times as long to train due to the largely increased search space. Thus, the setting of initial steps for FSR is also important to reduce training expenses.

3) We combine CSA with FSR (SkipDiff), i.e., perform the fine-grained diffusion guided by DRL to refine the final latent diffusion results (x_{S_1}) of CSA. As Case 5 shows in Table 4, the coarse-to-fine process achieves the best NIQE and LPIPS with an acceptable degradation on PSNR and SSIM.

4) To verify the decision behaviors of the agent, we remove DRL and artificially set the fine-grained diffusion steps, i.e., refine x_{S_1} with manual-defined timesteps within the last N steps of full diffusion. Table 5 shows that the results at timestep $t = 9, 8, 7$ exceed the results of the baseline model (Case 1 in Table 4) on all metrics. Our SkipDiff (Case 5) obtains similar SR performance with the optimal manual timestep setting (7), which is in accord with the steps learned by the agent. Besides, it reflects that the monotonicity in Fig. 1 holds only for conventional DDPM, but not for our fine-grained diffusion, since the max timestep (9) fails to obtain the best perceptual metrics. Due to the unmatched noise level between the coarse diffusion and the full diffusion, more fine time steps may not be optimal for generating high-quality samples. Especially, when the initial fine time steps for FSR (N) are getting larger, it is pretty necessary to introduce DRL for adaptively determining the fine-grained diffusion steps instead of artificial trials.

Reward function. An appropriate reward function can help to guide the model to obtain optimal performance. To evaluate the effect of perceptual metric M as reward measurement, we adopt NIQE and LPIPS as M for comparison in Table 6. It shows that the two metrics obtain similar results. Since they both contribute to learning the optimal timesteps (7) as shown in Table 5 for fine refinement, the differences are insignificant. Note that we adopt the non-reference NIQE as the measurement in our methods. Nevertheless, we do not declare that NIQE is the best choice for the reward measurement, and other metrics are also encouraged.

Conclusion

In this paper, we propose SkipDiff, a novel adaptive skip diffusion scheme for high-fidelity perceptual image SR. The proposed technique is on account of the observation that coarse time steps lead to blurry samples with low distortion, whereas fine time steps lead to crisp samples with high perceptual quality. The whole framework decouples the sampling procedure into two stages: coarse-grained diffusion and fine-grained diffusion. Coarse diffusion is performed to provide a high-fidelity latent initialization for fine diffusion. Then, a lightweight policy network is trained via PPO to determine the precise diffusion steps for enhancing perception in the fine stage. Experiments demonstrate that our SkipDiff achieves superior results against other mainstream methods on the face and natural image SR tasks.

Acknowledgments

The authors Xiaotong Luo, Yuan Xie, Yanyun Qu were supported by National Natural Science Foundation of China under Grant No.62176224, No.62222602, No.62176092; Natural Science Foundation of Chongqing under No.CSTB2023NSCOJOX0007, CCF-Lenovo Blue Ocean Research Fund.

References

Bansal, A.; Borgnia, E.; Chu, H.; Li, J. S.; Kazemi, H.; Huang, F.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023a. Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise. In *ICLR*.

Bansal, A.; Chu, H.-M.; Schwarzschild, A.; Sengupta, S.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023b. Universal guidance for diffusion models. In *CVPR*.

Blau, Y.; and Michaeli, T. 2018. The Perception-Distortion Tradeoff. In *CVPR*.

Cao, Q.; Lin, L.; Shi, Y.; Liang, X.; and Li, G. 2017. Attention-Aware Face Hallucination via Deep Reinforcement Learning. In *CVPR*.

Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; and Dong, C. 2023. Activating More Pixels in Image Super-Resolution Transformer. In *CVPR*.

Chen, Y.; Tai, Y.; Liu, X.; Shen, C.; and Yang, J. 2018. FSRNet: End-to-End Learning Face Super-Resolution With Facial Priors. In *CVPR*.

Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In *ICCV*.

Chung, H.; Sim, B.; and Ye, J. C. 2022. Come-Closer-Diffuse-Faster: Accelerating Conditional Diffusion Models for Inverse Problems through Stochastic Contraction. In *CVPR*.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2016. Image Super-Resolution Using Deep Convolutional Networks. *TPAMI*.

Gao, G.; Xu, Z.; Li, J.; Yang, J.; Zeng, T.; and Qi, G. 2023a. CTCNet: A CNN-Transformer Cooperation Network for Face Image Super-Resolution. *TIP*.

Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; and Zhang, B. 2023b. Implicit Diffusion Models for Continuous Super-Resolution. In *CVPR*.

He, J.; Shi, W.; Chen, K.; Fu, L.; and Dong, C. 2022. GCFSR: a Generative and Controllable Face Super Resolution Method Without Facial and GAN Priors. In *CVPR*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*.

Huang, C.; Lucey, S.; and Ramanan, D. 2017. Learning Policies for Adaptive Tracking with Deep Feature Cascades. In *ICCV*.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising Diffusion Restoration Models. In *NeurIPS*.

Le, N.; Rathour, V. S.; Yamazaki, K.; Luu, K.; and Savvides, M. 2022. Deep reinforcement learning in computer vision: a comprehensive survey. *Artif. Intell. Rev.*

Li, H.; Li, J.; Zhao, D.; and Xu, L. 2021. DehazeFlow: Multi-Scale Conditional Flow Network for Single Image Dehazing. In *NeurIPS*.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022. SRDiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*.

Li, Y.; Fan, Y.; Xiang, X.; Demandolx, D.; Ranjan, R.; Timofte, R.; and Van Gool, L. 2023. Efficient and explicit modelling of image hierarchies for image restoration. In *CVPR*.

Lugmayr, A.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2020. SRFlow: Learning the Super-Resolution Space with Normalizing Flow. In *ECCV*.

Lyu, Z.; Xu, X.; Yang, C.; Lin, D.; and Dai, B. 2022. Accelerating Diffusion Models via Early Stop of the Diffusion Process. arXiv preprint arXiv: 2205.12524.

Menon, S.; Damian, A.; Hu, S.; Ravi, N.; and Rudin, C. 2020. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. In *CVPR*.

Metzger, N.; Daudt, R. C.; and Schindler, K. 2023. Guided Depth Super-Resolution by Deep Anisotropic Diffusion. In *CVPR*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved Denoising Diffusion Probabilistic Models. In *ICML*.

Opper, M.; and Sanguinetti, G. 2007. Variational inference for Markov jump processes. In *NeurlPS*.

Pan, X.; Zhan, X.; Dai, B.; Lin, D.; Loy, C. C.; and Luo, P. 2022. Exploiting Deep Generative Prior for Versatile Image Restoration and Manipulation. *TPAMI*.

Park, S. H.; Moon, Y. S.; and Cho, N. I. 2023. Perception-Oriented Single Image Super-Resolution Using Optimal Objective Estimation. In *CVPR*.

Rana, K.; Dasagi, V.; Haviland, J.; Talbot, B.; Milford, M.; and Sünderhauf, N. 2023. Bayesian controller fusion: Leveraging control priors in deep reinforcement learning for robotics. *Int. J. Robotics Res.*

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV*.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image Super-Resolution via Iterative Refinement. *TPAMI*.

San-Roman, R.; Nachmani, E.; and Wolf, L. 2021. Noise Estimation for Generative Diffusion Models. arXiv preprint arXiv: 2104.02600.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. arXiv preprint arXiv: 1707.06347.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *ICLR*.

Stephens, C.; and Exton, C. 2022. Balancing Multiplayer Games across Player Skill Levels using Deep Reinforcement Learning. In *ICAART*.

Wang, H.; Chen, X.; Ni, B.; Liu, Y.; and Liu, J. 2023. Omni Aggregation Networks for Lightweight Image Super-Resolution. In *CVPR*.

Wang, Y.; Lv, K.; Huang, R.; Song, S.; Yang, L.; and Huang, G. 2020. Glance and Focus: a Dynamic Approach to Reducing Spatial Redundancy in Image Classification. In *NeurIPS*.

Wang, Y.; Yu, J.; and Zhang, J. 2023. Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model. In *ICLR*.

Watson, D.; Ho, J.; Norouzi, M.; and Chan, W. 2021. Learning to Efficiently Sample from Diffusion Probabilistic Models. arXiv preprint arXiv: 2106.03802.

Whang, J.; Delbracio, M.; Talebi, H.; Saharia, C.; Dimakis, A. G.; and Milanfar, P. 2022. Deblurring via Stochastic Refinement. In *CVPR*.

Xin, J.; Li, J.; Jiang, X.; Wang, N.; Huang, H.; and Gao, X. 2022. Wavelet-Based Dual Recursive Network for Image Super-Resolution. *TNNLS*.

Yu, K.; Dong, C.; Lin, L.; and Loy, C. C. 2018a. Crafting a Toolchain for Image Restoration by Deep Reinforcement Learning. In *CVPR*.

Yu, K.; Dong, C.; Lin, L.; and Loy, C. C. 2018b. Crafting a Toolchain for Image Restoration by Deep Reinforcement Learning. In *CVPR*.

Yuan, J.; Jiang, H.; Li, X.; Qian, J.; Li, J.; and Yang, J. 2023. Structure Flow-Guided Network for Real Depth Super-resolution. In Williams, B.; Chen, Y.; and Neville, J., eds., *AAAI*.

Zhang, W.; Liu, Y.; Dong, C.; and Qiao, Y. 2022a. RankSRGAN: Super Resolution Generative Adversarial Networks With Learning to Rank. *TPAMI*.

Zhang, X.; Zeng, H.; Guo, S.; and Zhang, L. 2022b. Efficient Long-Range Attention Network for Image Super-Resolution. In *ECCV*.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *ECCV*.

Zhou, Y.; Liu, B.; Zhu, Y.; Yang, X.; Chen, C.; and Xu, J. 2023. Shifted Diffusion for Text-to-Image Generation. In *CVPR*.