Reschedule Diffusion-based Bokeh Rendering

Shiyue Yan[1] , **Xiaoshi Qiu**[1] , **Qingmin Liao**[1] , **Jing-Hao Xue**[2] and **Shaojun Liu**[3]
[1]Shenzhen International Graduate School, Tsinghua University
[2]Department of Statistical Science, University College London
[3]College of Health Science and Environmental Engineering, Shenzhen Technology University
liusj14@tsinghua.org.cn

Abstract

Bokeh rendering for images shot with small apertures has drawn much attention in practice. Very recently people start to explore diffusion models for bokeh rendering, aiming to leverage the models' surging power of image generation. However, we can clearly observe two big issues with the images rendered by diffusion models: large fluctuation and severe color deviation. To address these issues, we propose in this paper a prior-aware sampling approach, which can adaptively control the noise scale through learned priors, and a prior-aware noise scheduling strategy, which can greatly reduce the number of inference steps without sacrificing performance. Extensive experiments show that our method can effectively alleviate the fluctuation problem of sampling results while ensuring similar color styles to the input image. In addition, our method outperforms state-of-the-art methods, sometimes even with only two steps of sampling. Our code is available at https://github.com/Loeiii/Reschedule-Diffusion-based-Bokeh-Rendering.

1 Introduction

Bokeh rendering refers to simulating the effect of images taken with a large aperture based on a given small aperture image. Compared with shooting with a large aperture camera, bokeh rendering is clearly much more convenient and thus has received widespread attention in practice.

Recently, diffusion models have achieved tremendous success on various generative tasks [Podell *et al.*, 2023; Voleti *et al.*, 2022; Kim *et al.*, 2022]. Very recently, people start to explore diffusion models for bokeh rendering [Luo *et al.*, 2023b]. However, for bokeh rendering, diffusion models still have two big issues that can be clearly observed.

Large Fluctuation. For bokeh rendering under the same camera parameter settings, the results are supposed to be unique. However, diffusion models are generative models that are inherently characterized by certain degrees of diversity in their outputs. Although specific conditions can guide the generation process, considerable variability persists in the outputs, resulting in relatively large variance of metrics when the same input is sampled multiple times.

Figure 1: Bokeh rendering results from diffusion models. From left to right: (a) ground truth; (b) results from starting from standard Gaussian noise, with severe color deviations being either too "cold" or too "hot"; (c) results from our prior-aware sampling method.

Severe Color Deviation. Diffusion models for bokeh rendering still introduce severe color deviation. As shown in Figure 1, the model tends to yield the bokeh rendering results that are either overly bright or too dark, along with a color shift to either too "cold" or too "hot".

To address these two issues, we propose in this paper a prior-aware sampling approach, which can adaptively control the noise scale through learned priors, and a prior-aware noise scheduling strategy, which can greatly reduce the number of inference steps without sacrificing performance. Our proposal is based on the following motivations.

Firstly, there should exist much similarity between a small aperture image and its corresponding large aperture image. Hence, we utilize the information from small aperture images as a useful prior to initialize the "noise" at the beginning of sampling. This can not only avoid sampling from a standard Gaussian noise, but also address both issue simultaneously.

Secondly, we observe from experiments that a moderate increase in the proportion of prior information within the initial state of diffusion models can lead to the generation of higher-quality sampling outcomes. However, the proportion of prior information progressively decreases with increasing noise. Hence, during sampling, we reschedule the noise based on priors to avoid excessively large noise and better leverage the priors.

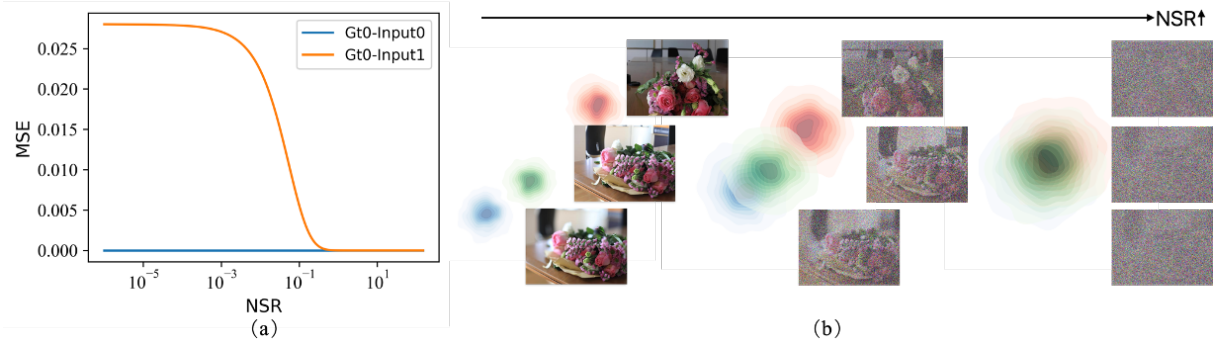

Figure 2: Progressive noise addition to images. (a) Relationship between MSE and NSR of noisy images, for the same image pair (blue curve) and different pairs (orange curve). (b) Data distribution at various NSR ($10^{-6}, 10^{0.4}, 10^{1.8}$), where blue and green contour plots are for the corresponding large-aperture and small-aperture images, respectively, and red contour plot is for another image. As the noise level increases, the distinctions between images (and data distributions) gradually diminish, eventually converging towards standard Gaussian noise. This suggests that initiating sampling with standard Gaussian noise fails to leverage any prior information.

In short, our main contributions are three-fold:

- We propose a prior-aware diffusion sampling approach that suppresses color deviations during model sampling while greatly reducing fluctuation in sampling results.
- We propose a prior-aware noise scheduling strategy that can greatly reduce the number of sampling steps, allowing the model to render desired bokeh effects within only 10 steps, substantially (100 times in principle) improving inference speed.
- Our method achieves the state-of-the-art (SOTA) performance on real-world bokeh rendering tasks using diffusion models.

2 Related Work

2.1 Diffusion Models

Diffusion models [Ho *et al.*, 2020] can generate high-quality images through a large number of iterative denoising operations, and have achieved tremendous success on various image restoration tasks.

[Song *et al.*, 2020b] proposes a unified diffusion model framework from the perspective of stochastic differential equations, which inspires a large number of training-free samplers [Song *et al.*, 2020a; Lu *et al.*, 2022; Zhang and Chen, 2022], greatly accelerating the model inference process. In addition, many subsequent works investigate model training [Karras *et al.*, 2022; Choi *et al.*, 2022], latent space encoding [Rombach *et al.*, 2022], and variance estimation during inference [Bao *et al.*, 2021; Bao *et al.*, 2022], greatly improved the model utility.

In conditional image restoration tasks, unlike most works that combine the condition with the noisy image [Saharia *et al.*, 2022] to guide the inference process, [Yue *et al.*, 2023; Luo *et al.*, 2023a] reformulate the diffusion models by introducing conditional information into the forward process.

2.2 Bokeh Rendering

Using post-processing rendering to achieve bokeh effects is a very popular method. Traditional single image bokeh rendering methods estimate the defocus map [Bae and Durand, 2007; Zhuo and Sim, 2011] or foreground regions [Xue *et al.*, 2013] and manually render bokeh effects. Although easy to implement, the rendering effects often lack realism.

Recently, learning-based methods have quickly become the mainstream in this field due to their ability to render more realistic bokeh effects.

[Wadhwa *et al.*, 2018; Purohit *et al.*, 2019; Ignatov *et al.*, 2020; Yang *et al.*, 2023] rely on obtaining depth information, such as masks and saliency maps, in advance to render bokeh effects. These method are highly dependent on the accuracy of depth map or saliency map, and thus the performance degrades dramatically when the depth information is not accurate enough.

[Dutta *et al.*, 2021; Qian *et al.*, 2020] directly use small aperture images to achieve end-to-end bokeh rendering, eliminating the dependency of depth information. With feature pyramids [Dutta *et al.*, 2021; Liu *et al.*, 2022] or Laplacian pyramids [Georgiadis *et al.*, 2022], the model parameters can be greatly reduced, making the networks lightweight. [Nagasubramaniam and Younes, 2022; Yang *et al.*, 2023] utilize Vision Transformers for bokeh rendering to better preserve image details. By incorporating aperture information in networks, [Seizinger *et al.*, 2023; Yang *et al.*, 2023] achieve controllable multi-aperture bokeh rendering to some extent. [Qian *et al.*, 2020; Choi *et al.*, 2020] explore the use of generative adversarial networks for bokeh rendering, providing a new idea for this task.

Very recently, people start to explore diffusion models for bokeh rendering tasks. [Luo *et al.*, 2023b] attempt the task on a synthetic dataset by leveraging the diffusion model framework for image restoration proposed by [Luo *et al.*, 2023a]. Although this method can better preserve the input image style, it introduces priors by modifying the forward process, making it incompatible with mainstream approaches based on Denosing Diffusion Probabilistic Model (DDPM), thereby hindering effective utilization of existing achievements such as [Zhang *et al.*, 2023].

Therefore, unlike [Luo *et al.*, 2023b], we utilize the existing DDPM framework [Ho *et al.*, 2020] and introduce priors during sampling, which not only can better guide model

	ΔPSNR	ΔSSIM	ΔLPIPS
	mean ± std	mean ± std	mean ± std
DDPM	1.17 ± 0.96	0.0130 ± 0.0108	0.0158 ± 0.0131
DDIM	0.65 ± 0.56	0.0863 ± 0.0660	0.0739 ± 0.0546
Same	0.99 ± 0.77	0.0106 ± 0.0080	0.0111 ± 0.0095

Table 1: Comparison of variability between DDPM, DDIM, and using the same initial state. The term "Same" refers to using the same initial state, and the metrics are computed over a set of 61 images.

generation without changing the model training, but also can better leverages existing achievements.

3 Preliminary: Conditional Diffusion Model

The forward process of the diffusion model is a process of progressively adding Gaussian noise to perturb the image through a Markov chain, which can be expressed as

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$ determines the variance of the noise and controls the noise-to-signal ratio (NSR) of the noisy image. $\beta_t \in (0,1)$ is a monotonically increasing parameter over time defined in advance. When t is very large, the noisy image in the forward process tends towards standard Gaussian noise.

The inference (reverse) process starts from a standard Gaussian noise, and gradually transfers it to the target data distribution according to the condition through Gaussian transition $q_\theta\left(x_{t-1} \mid x_t, y\right) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t; y), \sigma_t)$. [Song *et al.*, 2020b] claims that this process can be non-Markovian, and gives the sampling formula in the reverse process as

$$\begin{aligned} x_{t-1} =& \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t; y)\right) + \\ &+\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_\theta(x_t, t; y) + \sigma_t\epsilon, \end{aligned} \tag{2}$$

where y represents the input image, $\sigma_t = \eta\sqrt{\frac{(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}}\sqrt{\beta_t}$; $\eta \in [0,1]$; when $\eta = 0$, the sampling process is deterministic, known as Denosing Diffusion Implicit Model (DDIM), whereas at $\eta = 1$, the sampling process aligns with that of DDPM [Ho *et al.*, 2020]; $\epsilon_\theta(x_t, t; y)$ is the noise estimated by using x_t, t, y through the denoising network, and its training loss function is $\mathcal{L}_1$ loss between the estimated noise and the real noise, shown as

$$\mathcal{L} = \mathbf{E}_{t,\epsilon}[\|\epsilon - \epsilon_\theta(x_t, t; y)\|^2]. \tag{3}$$

4 Methodology

4.1 Technical Motivations

Stochasticity in Sampling. From Equation (2), we can see that the randomness of the sampling results is affected by the initial sampling state x_T and the random noise ϵ introduced during inference. To evaluate the fluctuation, we employ a method comparing the variations between twice independent sampling of the same input image in terms of Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). Specifically, we calculate the mean and standard deviation of the absolute difference values of these metrics. When the number of images is relatively large, this method effectively reflects the stability of the sampling results under different metrics. As shown in Table 1, we compare DDPM, DDIM and the method sampling from the same Gaussian noise, which eliminates the randomness in the initial state. Although DDIM yields a more stable PSNR by avoiding random noise introduction during sampling, its structural metric fluctuation is significantly higher than even that of DDPM. Conversely, using a fixed initial state is capable of mitigating fluctuations in all measured metrics. Thus, employing a suitable method to generate the initial state can result in more stable sampling outcomes with only slightly higher PSNR fluctuation than that of DDIM.

Data Distribution. Figure 2(b) illustrates that a pair of images sharing a substantial amount of information have quite close data distributions. Conversely, different pairs of images exhibit significant differences. As the noise magnitude increases, the data distribution of the noisy images progressively deviates from the original and approaches standard Gaussian distribution, diminishing differences between different pairs. Figure 2(a) illustrates that the Mean Squared Error (MSE) between the different image pair gradually decreases with increase of the NSR. When the noise scale approaches infinity, the noisy image transforms into standard Gaussian noise, losing all prior information. Consequently, sampling directly from Gaussian noise introduces a big obstacle in transitioning to the target data distribution.

Noise Schedule. Due to the similarity between small and large aperture images, only an intermediate noise magnitude is needed to bring their data distributions very close. Typically, using smaller noise means fewer inference steps. Therefore, avoiding excessively large noise during sampling can not only improve model performance but also accelerate inference speed.

Based on the aforementioned technical motivations, we propose a prior-aware sampling method in section 4.2 and a prior-aware noise scheduling strategy in section 4.3.

4.2 Prior-aware Sampling

We set $T = 1000$ and use the small aperture input y to replace x_0. Then, using Equation (4) as the initial state for sampling, we generate the noisy image:

$$x_t = \sqrt{\bar{\alpha}_t}y + \sqrt{1-\bar{\alpha}_t}\epsilon. \tag{4}$$

Compared with standard Gaussian noise, such a generated noisy image contains a certain level of prior information, thus greatly reducing the distance between the initial and target data distributions.

As can be seen from Figure 1(b), directly using the diffusion model for sampling leads to severe color deviations. By introducing priors, we greatly alleviate the color problem during sampling and achieve considerable performance gains. In Figure 3, we use the isolated point on the far right to denote sampling from standard Gaussian noise, as its NSR is infinity. As shown in Figure 3, the end of the blue curve on the

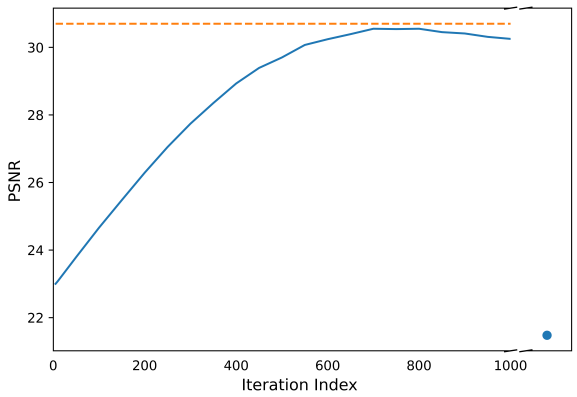

Figure 3: Results of sampling under different prior scales. The orange dashed line represents the results from Algorithm 1, while the blue curve and the blue dot represent sampling results obtained from starting from x_t (with a prior scale of $\bar{\alpha}_t$) and standard Gaussian noise (without prior), respectively. Given that $\bar{\alpha}_t$ decreases over t, this plot suggests that appropriately scaled prior information (e.g. by Alg. 1) is more conducive to the model's inference.

Algorithm 1 Prior-aware Sampling

Input: noise schedule $\bar{\alpha}$, denoising network $\epsilon_\theta(\cdot)$, small aperture image y

Sample: $\epsilon \sim \mathcal{N}(0, \mathbf{I})$

$A = \frac{1}{1+\left[\frac{1}{\sqrt{6}}\left(3n \max_{c\in[R,G,B]} \sigma_{y;c}\right)\right]^2}$

Find an index i such that $\bar{\alpha}_{i-1} > A$ and $\bar{\alpha}_i < A$.

Let $T^{'} = i$ and $x_{T'} = \sqrt{\bar{\alpha}_{T'}}y + \sqrt{1-\bar{\alpha}_{T'}}\epsilon$

for $t = T^{'}, \ldots, 1$ **do**

  Sample $\epsilon \sim \mathcal{N}(0, I)$ if $t > 1$, else $\epsilon = 0$

  $x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t; y)) + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}\epsilon$

end for

return x_0

left, where the scale of prior information is $\bar{\alpha}_{1000}$, exhibits an improvement of over 9 dB compared with starting sampling with standard Gaussian noise.

Since $\bar{\alpha}_{1000}$ is very close to 0, the proportion of priors in the initial noisy image obtained through Equation 4 is very small. By enumerating the starting point over t, we find that appropriately increasing the proportion of priors in the initial image can further improve performance, as shown in Figure 3. Therefore, we propose to determine the initial state's proportion of priors, i.e. the starting point of model inference, based on an adaptively learned prior magnitude from the conditional information.

Similarly to [Ye *et al.*, 2023], by dividing both sides of Equation (4) by $\sqrt{\bar{\alpha}_t}$, we obtain

$$\frac{1}{\sqrt{\bar{\alpha}_t}}x_t = y + \sqrt{NSR} \cdot \epsilon, \tag{5}$$

where $NSR = (1-\bar{\alpha}_t)/\bar{\alpha}_t$ is the NSR.

As shown in Figure 2, as the NSR increases, the data distributions of the noisy large and small aperture images gradually

Algorithm 2 Refined Prior-aware Sampling

Input: denoising network $\epsilon_\theta(\cdot)$, inference stps N, small aperture image y

Sample: $\epsilon \sim \mathcal{N}(0, \mathbf{I})$

$A = \frac{1}{1+\left[\frac{1}{\sqrt{6}}\left(3n \max_{c\in[R,G,B]} \sigma_{y;c}\right)\right]^2}$

Calculate β_{end} from $\prod_{i=0}^{N-1}(1-\beta_{i;end}+\frac{\gamma}{N-1}) = \xi(N)A$ and set new noise schedule according to β_{end} ▷ use Newton's Iteration Method

$x_N = \sqrt{\bar{\alpha}_N}y + \sqrt{1-\bar{\alpha}_N}\epsilon$

for $t = N, \ldots, 1$ **do**

  Sample $\epsilon \sim \mathcal{N}(0, I)$ if $t > 1$, else $\epsilon = 0$

  $x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t; y)) + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}\epsilon$

end for

return x_0

become closer. Considering the 3σ rule, for condition y_i, we could comfortably argue that, when the NSR is greater than the level shown in Equation (6), the noisy images are almost identical:

$$NSR_{i;min} = \left[\frac{1}{\sqrt{6}}\left(3n \max_{c\in[R,G,B]} \sigma_{i;c}\right)\right]^2, \tag{6}$$

where n indicates the aperture ratio, since the blur disc diameter is proportional to the f-number [Liu *et al.*, 2016], and $\sigma_{i;c}$ is the standard deviation of each color channel of y_i.

By solving Equation (6), for condition y_i, there exists an A_i such that

$$A_i = \frac{1}{1+NSR_{i;min}}, \tag{7}$$

and then we simply use A_i to determine the initial noisy image and time t during sampling, as shown in Algorithm 1.

4.3 Prior-aware Noise Scheduling

As with [Saharia *et al.*, 2022], we use a piece-wise distribution method to sample the noise level during training, where $p(\bar{\alpha}_t) = U(\bar{\alpha}_{t-1}, \bar{\alpha}_t)$. This ensures greater flexibility in noise scheduling during inference, aligning with the concept proposed in [Chen *et al.*, 2020].

According to Equation (7), NSR above a certain level will not appear during sampling. Therefore, to ensure each input gets the same number of iterations, we reschedule the noise level during sampling based on the following strategy:

$$\prod_{n=0}^{N-1}\left(1-\beta_{i;end}+\frac{\gamma}{N-1}n\right) = A_i, \tag{8}$$

where N indicates the number of sampling steps, and γ is a hyperparameter to adjust the initial noise magnitude. Unlike [Chen *et al.*, 2020], we impose fewer constraints on β. Under the condition of Equation (8), we combine DDPM with priors for sampling. Remarkably, even when the sampling steps are reduced to only 10, our model maintains its performance without any degradation. In some cases, such a steps reduction would produce even better results.

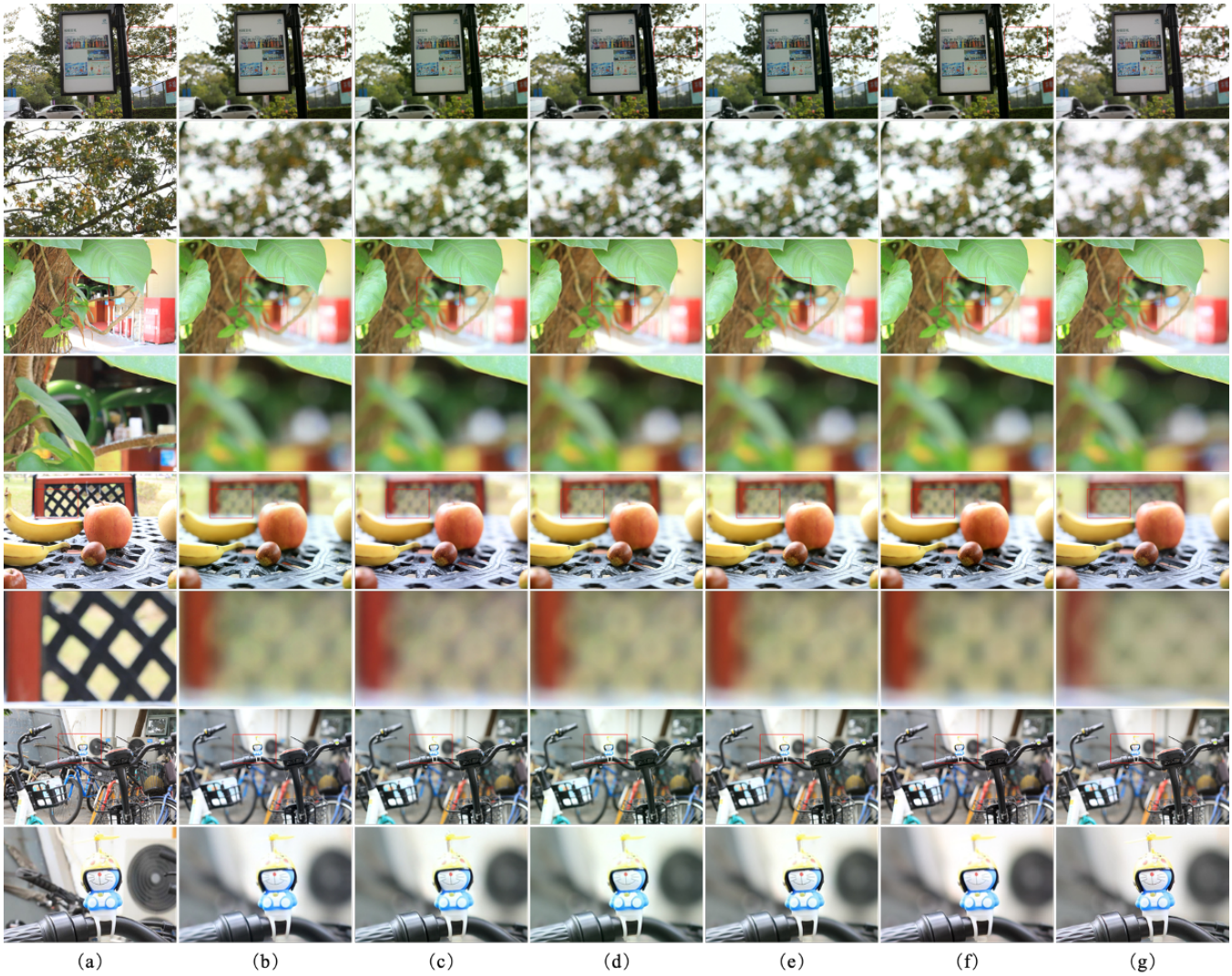

Figure 4: Qualitative comparison. (a) Input image; (b) BRViT; (c) FPBNet; (d) SDMSHN; (e) Refusion; (f) Ours; (g) ground truth. It can be observed that our method ensures the clarity of the foreground while rendering a more natural bokeh effect.

As the number of inference steps is reduced, the relative proportion of information provided by the condition in the denoising process increases. This shift results in a tendency for the model's inference outcomes to converge to the given conditions. Consequently, continuing to use priors of the same scale fails to achieve the desired performance. Therefore, when the number of inference steps is small, we propose to appropriately reduce the scale of priors based on the number of inference steps by weighting A_i. Our final formulation is

$$\prod_{n=0}^{N-1}\left(1-\beta_{i;end}+\frac{\gamma}{N-1}n\right)=(1-e^{-\frac{N}{20}})A_i. \quad (9)$$

After implementing the refined strategy following Equation (9), the model is capable of maintaining robust performance even when constrained to a mere two inference steps. In Algorithm 2, we present our refined sampling algorithm, which incorporates the refined noise scheduling strategy.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
BRViT	30.02	0.9387	0.0837
SDMSHN	30.13	0.9242	0.0945
FPBNet	29.98	0.9358	0.0834
Refusion	<u>30.82</u>	0.9255	0.0953
Baseline	30.25	0.9421	<u>0.0682</u>
Ours ($N=2$)	30.54	<u>0.9428</u>	0.0691
Ours ($N=5$)	**30.99**	**0.9485**	**0.0678**

Table 2: Quantitative comparison with SOTA bokeh rendering methods on the FPBNet dataset. The best is in bold; the second best is underlined.

	BRViT	SDMSHN	FPBNet	Refusion	Ours
P(M) ↓	123.14	<u>10.84</u>	**9.38**	126.22	97.81

Table 3: Amounts of parameters of different methods.

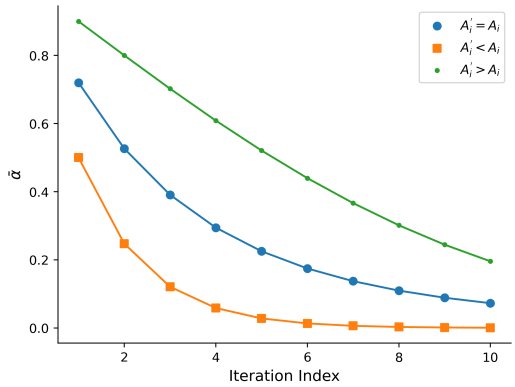

Figure 5: Noise schedule with different scaling levels for A_i. According to Equations. (4) and (8), adjusting A_i in noise schedule is equivalent to altering the proportion of prior information in the initial state.

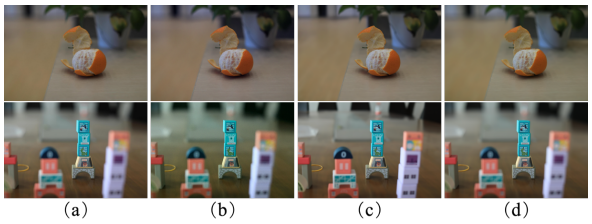

Figure 6: Results of using different prior scales: (a) ground truth; (b) $A_i^{'} < A_i$, results in color deviation; (c) $A_i^{'} > A_i$, leads to insufficient bokeh effect; (d) $A_i^{'} = A_i$. This suggests that adjustments should be made based on prior information when scheduling noise.

5 Experiments

5.1 Settings

Dataset. Currently, high-quality real-world bokeh rendering datasets are relatively scarce. One such dataset is the FPBNet dataset [Liu *et al.*, 2022], comprising 941 sets of highly aligned data with a resolution of 2232×1488. To mitigate the effects of data misalignment on the model training and maximize its potential in generating more natural results, we train and test our approach on the FPBNet dataset.

Implementation Details. During the training phase, we set T to 1000 and linearly increase the value of β from 0.00001 to 0.01. The Adam optimizer is employed, with an initial learning rate set at 1×10^{-4} without decay and the optimizer's weight decay is set at 0.01. Training is conducted over 1.2M iterations on an A100 GPU.

During the inference process, we employed sampling with the prior scale of $\alpha_{\bar{1000}}$ as our 'Baseline'. Setting $\gamma = 0.1$; sampling is then performed separately using $N = 2$ and $N = 5$. Additional sampling strategies will be showed in the ablation study.

To comprehensively evaluate the performance of our model, we conducted comparisons with several SOTA methods in the field of bokeh rendering, including BRVit [Nagasubramaniam and Younes, 2022], SDMSHN [Dutta *et al.*, 2021], FPBNet [Liu *et al.*, 2022], and Refusion [Luo *et al.*, 2023b] which is also based on diffusion models.

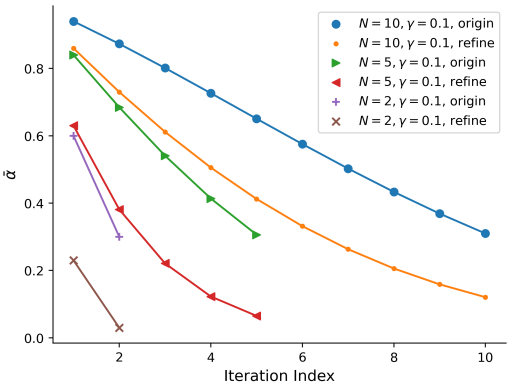

Figure 7: Comparison of the original and refined noise schedules.

The results are evaluated with mainstream metrics such as PSNR, SSIM, and LPIPS, with the calculations for PSNR and SSIM being conducted in the RGB space.

5.2 Results

Quantitative Evaluation. Tables 2 and 3 present the quantitative results of various methods. By integrating DDIM with a setting of $\eta = 1$, our approach achieves SOTA performance with just 5 steps.

As indicated in Table 2, even our 'Baseline' model demonstrates a considerable advantage in image perceptual quality under the LPIPS metric, exceeding a 15% improvement over existing models. Moreover, it also produces comparable PSNR and SSIM.

Furthermore, by strategically scheduling noise based on prior information, the model not only reduces inference time but also achieves a better performance. Employing a 5-step sampling strategy, our model achieves optimal performance across all evaluation metrics. Notably, even with a 2-step sampling, the model's performance remains essentially unaltered, surpassing the 'Baseline' in PSNR and SSIM metrics.

Qualitative Results. Figure 4 shows the visual comparison. Upon examining the visual quality of the experimental results, it is evident that our model produces images with superior clarity and noise suppression compared with competing methods. This is attributed to the incorporation of prior information into the initial state, resulting in a data distribution more closely aligning with the target distribution, thereby easing the diffusion process.

In terms of image content, our method accurately distinguishes the foreground from the background, rendering a more realistic bokeh effect. As demonstrated in the first and third rows of Figure 4, our model can simulate an effect closer to natural bokeh effect rather than a simple blur. Furthermore, as evident in the final row, our method better preserves the details in the foreground.

5.3 Ablation Studies

In this section, we conduct ablation studies to demonstrate the effectiveness of our proposed prior-aware sampling method and noise scheduling approach.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Baseline	30.25	0.9421	0.0682
Alg.1	30.60	0.9421	**0.0648**
$N = 700, \gamma = 0.005$	30.70	0.9461	0.0671
$N = 100, \gamma = 0.01$, w/o	16.67	0.7877	0.2210
$N = 100, \gamma = 0.01$	**30.83**	**0.9482**	0.0678
$N = 50, \gamma = 0.05$	<u>30.78</u>	<u>0.9481</u>	<u>0.0676</u>
$N = 20, \gamma = 0.1$	30.65	0.9477	0.0680
$N = 10, \gamma = 0.1$	30.52	0.9469	0.0691
$N = 5, \gamma = 0.1$	30.44	0.9466	0.0686
$N = 2, \gamma = 0.1$	29.95	0.9437	0.0708

Table 4: Ablation study on inference steps. When N is large, we appropriately reduce the value of γ to ensure the feasibility of β. The notation "w/o" means sampling from standard Gaussian noise without prior; 'Alg.1' means sampling using Algorithm 1.

Effectiveness of Sampling Starting Point. Figure 3 illustrates the sampling outcomes of the model under different prior scales. The horizontal dashed line in the figure represents the results from Algorithm 1, which shows considerable improvement over fixed prior scales. The results indicate that dynamically selecting an appropriate prior scale for each image according to Equation (7) could enhance the model's rendering performance.

Impact of Prior Scale. As shown in Figure 5, we manually adjust the value of A_i in Equation (8) to be larger or smaller, thereby exploring the impact of the prior weight on the sampling outcome. Observations from Figure 6 reveal that decreasing the value of A_i, thereby increasing the initial noise level, leads to noticeable color deviations in the inference results due to reduced prior information. This partially explains why starting sampling with Gaussian noise can result in color biases. Conversely, increasing the value of A_i, and thereby the weight of the prior in the initial noisy image, makes the model's inference results more dependent on the given conditions, leading to insufficient bokeh rendering effects. This also suggests that we should leveraging the shared information between the input and target images to identify the optimal prior scale in other image restoration tasks where the prior information is equally valuable.

Influence of Different Noise Schedules. Table 4 presents the sampling outcomes under different (N, γ) combinations. It is observed that when N is relatively large, the prior-aware sampling method does not compromise the model's performance; in fact, it may even enhance it to some extent. However, as N decreases, especially when $N < 20$, the quality of the sampling results considerably deteriorates. The deficiency in the bokeh rendering effect in these samples is evident. We hypothesize that the amplification of the condition's influence when the number of inferences is small leads to a significant bias in the sampling results towards the conditions. Based on these experiments, we adjust to Equation (9), especially for lower values of N, reducing the weight of the prior in the initial sampling image. This refinement makes the equation more suitable for our model's needs, with the comparison of the refined $\bar{\alpha}'$ and the original $\bar{\alpha}$ shown in Figure 7.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
$N = 20, \gamma = 0.1$	30.88	0.9483	0.0681
$N = 10, \gamma = 0.1$	**31.01**	**0.9494**	**0.0671**
$N = 5, \gamma = 0.1$	30.99	0.9485	0.0678
$N = 2, \gamma = 0.1$	30.54	0.9428	0.0691

Table 5: Ablation study on the refined noise schedule. Compared with Table 4, the model's performance has shown overall improvement after such a noise-schedule refinement.

Method	ΔPSNR	ΔSSIM	ΔLPIPS
	mean ± std	mean ± std	mean ± std
Baseline	1.07 ± 0.98	0.0037 ± 0.0111	0.0055 ± 0.0064
Alg.1	$\mathbf{0.35 \pm 0.33}$	0.0020 ± 0.0039	$\underline{0.0024} \pm 0.0034$
$N = 200$	$\underline{0.44} \pm \underline{0.36}$	0.0018 ± 0.0036	$\mathbf{0.0021} \pm 0.0029$
$N = 100$	$\underline{0.44} \pm 0.39$	0.0019 ± 0.0046	$\underline{0.0024} \pm 0.0042$
$N = 10$	0.52 ± 0.42	$\underline{0.0018} \pm \underline{0.0019}$	$\underline{0.0024} \pm \underline{0.0024}$
$N = 10^{\dagger}$	0.48 ± 0.49	$\mathbf{0.0013 \pm 0.0015}$	$\underline{0.0024} \pm \mathbf{0.0023}$
$N = 5$	0.71 ± 0.78	0.0022 ± 0.0047	0.0030 ± 0.0032
$N = 5^{\dagger}$	0.68 ± 0.71	0.0022 ± 0.0040	0.0034 ± 0.0041
$N = 2$	0.69 ± 0.47	0.0026 ± 0.0041	0.0042 ± 0.0039

Table 6: Ablation study of the fluctuations between employing priors of varying scales and using fixed initial states. † indicates sampling from a fixed initial state; 'Alg.1' means sampling using Algorithm 1; and for $N = 100, 200$, we set $\gamma = 0.01$. It can be observed that our method exhibits a level of stability comparable to that achieved by using a fixed initial state.

Furthermore, compared with Table 4, Table 5 demonstrates that this refinement enhances the overall performance of the model.

Comparison of Variability. As described in Section 4.1, we assess the variability of the model's sampling outcomes by analyzing the mean and standard deviation of the absolute differences in metrics (PSNR, SSIM, LPIPS) between twice independent sampling. To better evaluate the impact of prior information on the stability of sampling results, we fix the initial state as a benchmark, where the fixed initial state is generated using a constant standard Gaussian noise. According to Table 6, the introduction of prior information significantly enhances the stability of the model's sampling outcomes. Moreover, this improvement in stability is comparable with that achieved by using a fixed initial state.

6 Conclusion

In this paper, we propose a prior-aware sampling method and a prior-aware noise scheduling strategy, for diffusion model-based bokeh rendering. Our method adeptly mitigates significant color deviations commonly encountered during the model's inference by integrating prior information. The integration of priors not only significantly elevates the performance of the model, but also contributes to a substantial reduction in the number of required inference steps via prior-aware noise scheduling. In the future, we plan to explore and substantiate the versatility of our prior-aware sampling method in a wider array of image restoration challenges.

Acknowledgments

This work was supported in part by the National Science Foundation of China (grants no. 62301332 and no. U23B2030) and in part by the Natural Science Foundation of Top Talent of SZTU (grant no. GDRC202117).

Contribution Statement

Shiyue Yan and Xiaoshi Qiu contributed equally to this work. The corresponding author is Shaojun Liu (*liusj14@tsinghua.org.cn*).

References

[Bae and Durand, 2007] Soonmin Bae and Frédo Durand. Defocus magnification. In *Computer graphics forum*, volume 26, pages 571–579. Wiley Online Library, 2007.

[Bao *et al.*, 2021] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *International Conference on Learning Representations*, 2021.

[Bao *et al.*, 2022] Fan Bao, Chongxuan Li, Jiacheng Sun, Jun Zhu, and Bo Zhang. Estimating the optimal covariance with imperfect mean in diffusion probabilistic models. In *International Conference on Machine Learning*, pages 1555–1584. PMLR, 2022.

[Chen *et al.*, 2020] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020.

[Choi *et al.*, 2020] Min-Su Choi, Jun-Hyuk Kim, Jun-Ho Choi, and Jong-Seok Lee. Efficient bokeh effect rendering using generative adversarial network. In *2020 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pages 1–5. IEEE, 2020.

[Choi *et al.*, 2022] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022.

[Dutta *et al.*, 2021] Saikat Dutta, Sourya Dipta Das, Nisarg A Shah, and Anil Kumar Tiwari. Stacked deep multi-scale hierarchical network for fast bokeh effect rendering from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2398–2407, 2021.

[Georgiadis *et al.*, 2022] Konstantinos Georgiadis, Albert Saà-Garriga, Mehmet Kerim Yucel, Anastasios Drosou, and Bruno Manganelli. Adaptive mask-based pyramid network for realistic bokeh rendering. In *European Conference on Computer Vision*, pages 429–444. Springer, 2022.

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

[Ignatov *et al.*, 2020] Andrey Ignatov, Jagruti Patel, and Radu Timofte. Rendering natural camera bokeh effect with deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 418–419, 2020.

[Karras *et al.*, 2022] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

[Kim *et al.*, 2022] Heeseung Kim, Sungwon Kim, and Sungroh Yoon. Guided-tts: A diffusion model for text-to-speech via classifier guidance. In *International Conference on Machine Learning*, pages 11119–11133. PMLR, 2022.

[Liu *et al.*, 2016] Shaojun Liu, Fei Zhou, and Qingmin Liao. Defocus map estimation from a single image based on two-parameter defocus model. *IEEE Transactions on Image Processing*, 25(12):5943–5956, 2016.

[Liu *et al.*, 2022] Yi Liu, Juncheng Zhang, Qingmin Liao, Haoyu Ma, and Shaojun Liu. Feature pyramid boosting network for rendering natural bokeh. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.

[Lu *et al.*, 2022] 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.

[Luo *et al.*, 2023a] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. *arXiv preprint arXiv:2301.11699*, 2023.

[Luo *et al.*, 2023b] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1680–1691, 2023.

[Nagasubramaniam and Younes, 2022] Hariharan Nagasubramaniam and Rabih Younes. Bokeh effect rendering with vision transformers. 2022.

[Podell *et al.*, 2023] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[Purohit *et al.*, 2019] Kuldeep Purohit, Maitreya Suin, Praveen Kandula, and Rajagopalan Ambasamudram. Depth-guided dense dynamic filtering network for bokeh effect rendering. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3417–3426. IEEE, 2019.

[Qian *et al.*, 2020] Ming Qian, Congyu Qiao, Jiamin Lin, Zhenyu Guo, Chenghua Li, Cong Leng, and Jian Cheng. Bggan: Bokeh-glass generative adversarial network for

rendering realistic bokeh. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 229–244. Springer, 2020.

[Rombach *et al.*, 2022] 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.

[Saharia *et al.*, 2022] 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.

[Seizinger *et al.*, 2023] Tim Seizinger, Marcos V Conde, Manuel Kolmet, Tom E Bishop, and Radu Timofte. Efficient multi-lens bokeh effect rendering and transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1633–1642, 2023.

[Song *et al.*, 2020a] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.

[Song *et al.*, 2020b] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[Voleti *et al.*, 2022] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems*, 35:23371–23385, 2022.

[Wadhwa *et al.*, 2018] Neal Wadhwa, Rahul Garg, David E Jacobs, Bryan E Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018.

[Xue *et al.*, 2013] Weichen Xue, Xiangze Zhang, Bin Sheng, and Lizhuang Ma. Image-based depth-of-field rendering with non-local means filtering. In *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2013.

[Yang *et al.*, 2023] Zhihao Yang, Wenyi Lian, and Siyuan Lai. Bokehornot: Transforming bokeh effect with image transformer and lens metadata embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1542–1550, 2023.

[Ye *et al.*, 2023] Jiasheng Ye, Zaixiang Zheng, Yu Bao, Lihua Qian, and Mingxuan Wang. Dinoiser: Diffused conditional sequence learning by manipulating noises. *arXiv preprint arXiv:2302.10025*, 2023.

[Yue *et al.*, 2023] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *arXiv preprint arXiv:2307.12348*, 2023.

[Zhang and Chen, 2022] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[Zhuo and Sim, 2011] Shaojie Zhuo and Terence Sim. Defocus map estimation from a single image. *Pattern Recognition*, 44(9):1852–1858, 2011.