Reducing Spatial Fitting Error in Distillation of Denoising Diffusion Models

Shengzhe Zhou[1], Zejian Li[1], Shengyuan Zhang[2], Lefan Hou[2], Changyuan Yang[3], Guang Yang[3], Zhiyuan Yang[3], Lingyun Sun[2]

[1]School of Software Technology, Zhejiang University
[2] College of Computer Science and Technology, Zhejiang University
[3] Alibaba Group
{zhoujj7248, zejianlee, zhangshengyuan, houlefan, sunly}@zju.edu.cn, {changyuan.yangcy, adam.yzy}@alibaba-inc.com, qingyun@taobao.com

Abstract

Denoising Diffusion models have exhibited remarkable capabilities in image generation. However, generating high-quality samples requires a large number of iterations. Knowledge distillation for diffusion models is an effective method to address this limitation with a shortened sampling process but causes degraded generative quality. Based on our analysis with bias-variance decomposition and experimental observations, we attribute the degradation to the spatial fitting error occurring in the training of both the teacher and student model. Accordingly, we propose **S**patial **F**itting-**E**rror **R**eduction **D**istillation model (**SFERD**). SFERD utilizes attention guidance from the teacher model and a designed semantic gradient predictor to reduce the student's fitting error. Empirically, our proposed model facilitates high-quality sample generation in a few function evaluations. We achieve an FID of 5.31 on CIFAR-10 and 9.39 on ImageNet 64×64 with only one step, outperforming existing diffusion methods. Our study provides a new perspective on diffusion distillation by highlighting the intrinsic denoising ability of models.

Introduction

Diffusion-based (DPMs) (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015; Song and Ermon 2019) and score-based (Song et al. 2021b) generative models have recently achieved outstanding performance in synthesizing high-quality images. They have already shown comparable or even superior results to GAN (Goodfellow et al. 2020) in multiple fields such as 3D generation (Luo and Hu 2021; Zhou, Du, and Wu 2021), text-to-image generation (Rombach et al. 2022; Ruiz et al. 2023), image restoration (Lugmayr et al. 2022; Wang et al. 2022), controllable image editing (Kawar et al. 2022; Couairon et al. 2022) and graph generation (Huang et al. 2023).

A major problem with diffusion models is the slow sampling speed, as it often requires hundreds of iterations to achieve satisfactory generative quality. To address this issue, there are two mainstream improvements: fast sampling schemes without extra training (Song, Meng, and Ermon 2020; Lu et al. 2022a; Liu et al. 2022; Kong and Ping 2021) and trained acceleration schemes (Salimans and Ho 2022; Dockhorn, Vahdat, and Kreis 2022; Zhang, Zhao, and Lin 2022). The former aims to reduce errors resulting from large-step sampling by utilizing better numerical methods for integration, whereas the latter enhances performance through additional model training and fine-tuning. With the advent of large-scale pre-training models, the latter approach has become more notable. Among trained schemes, diffusion models based on knowledge distillation have demonstrated considerable potential for fast sampling. An illustrative example is Progressive Distillation model (PD) (Salimans and Ho 2022), where the student model learns the two-step inference output of the teacher model in a single step without compressing the model size. However, these diffusion models based on distillation face degraded generation quality given limited distillation sampling step.

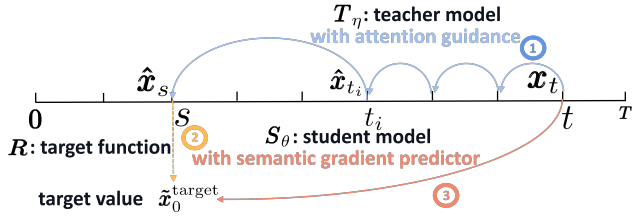


Figure 1: The process of Spatial Fitting-Error Reduction Distillation model. In the first step, $\hat{\boldsymbol{x}}_s$ is predicted by the teacher model $\boldsymbol{T}_\eta$ from $\boldsymbol{x}_t$ with attention guidance. The target function $\boldsymbol{R}$ further calculates the target value $\boldsymbol{x}_0^{\text{target}}$ based on $\hat{\boldsymbol{x}}_{\boldsymbol{s}}$ in the second step. The student model $\boldsymbol{S}_\theta$ tries to regress the target value with semantic gradient predictor in the final step.

In this paper, we investigate scalable enhancements for the diffusion distillation model to improve the quality of the images generated within few steps or even a single step. We begin by reframing the process of diffusion distillation models and designing a multi-step training framework for distillation. We then conduct analysis with bias-variance decomposition and preliminary experiments to identify the fitting errors of both the teacher and student models. Our results show that fitting errors have a broad impact on model performance within our framework. Especially, we observe a positive correlation between the self-attention maps of the diffusion model and the spatial fitting error in the predicted noise.

Based on our observations, we propose **S**patial **F**itting-

Error **R**eduction **D**istillation model (**SFERD**). SFERD utilizes internal and external representations to reduce the fitting error of the teacher and the student models, respectively. First, we design attention guidance to improve the denoised prediction using intrinsic information in the self-attention maps of the teacher model. We define the spatial regions having high self-attention scores as "Risky Regions". According to the observed correlation, these regions exhibits the high fitting error of the teacher model, which is further inherited by the student model. Therefore, by reducing the error in Risky Regions, we expect to enhance the quality of denoised prediction in the student model. This method does not require additional supervised information or extra auxiliary classifiers, and it can be combined with various diffusion models as well. Inspired by the classifier-based gradient guidance (Dhariwal and Nichol 2021), we also introduce a semantic gradient predictor and reformulate the training loss for the student model. The design is to reduce the student model's fitting error by providing additional information for image reconstruction from a learned latent space.

Empirically, SFERD efficiently reduces the fitting error of the student model, leading to superior performance as compared to other distillation models on CIFAR-10 (Krizhevsky 2009) and ImageNet 64×64 (Deng et al. 2009). Notably, it achieves single-step FID scores of 9.39 and 5.31 for ImageNet 64×64 and CIFAR-10 respectively. Furthermore, fine-tuning pre-trained diffusion model itself with our proposed method also results in improved performance. Project link: https://github.com/Sainzerjj/SFERD.

Preliminary

We briefly review backgrounds of diffusion distillation. Detailed content is deferred to Appendix F. Diffusion models typically include the forward process and the backward denoising process. Ho et al. (Ho, Jain, and Abbeel 2020) define the forward diffusion process as $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}\,,\, \beta_t\mathbf{I})$, where $\{\beta_t\}_{t=1}^T$ is a variance schedule used to control the noise intensity. Since the process satisfies Markov conditions, the forward process can be expressed as $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^T q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$.

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0\,,\, (1-\bar{\alpha}_t)\mathbf{I}) \tag{1}$$

where $\bar{\alpha}_t = 1-\beta_t$, $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. The sampling procedure becomes $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)\ , \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t))$. Its training objective is to minimize the upper bound of the variance of the negative log-likelihood, allowing for various predicted parameters, such as ϵ-prediction (Song et al. 2021a), v-prediction (Ho et al. 2022), x_0-prediction (Salimans and Ho 2022). To ensure a uniform representation, we define the training loss in distillation as:

$$\mathbb{E}_{t\sim U[0,1]\,,\,\boldsymbol{x}_0\sim p(\boldsymbol{x})\,,\,\boldsymbol{x}_t\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\omega(\lambda_t)\|\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t\,,\,t;\theta) - \boldsymbol{x}_0\|_2^2] \tag{2}$$

where $\hat{\boldsymbol{x}}_0$ denotes denoise prediction, $\lambda_t = \log[\bar{\alpha}_t/(1-\bar{\alpha}_t)]$ represents the signal-to-noise ratio (Kingma et al. 2021), and different choices of the weighting function $\omega(\cdot)$ correspond to different predicted variable θ.

DDIM (Song, Meng, and Ermon 2020) breaks the limitations of Markov chains and allows for more stable and fast reverse sampling, the deterministic process becomes:

$$\begin{aligned}\boldsymbol{x}_{t-1} =& \sqrt{\bar{\alpha}_{t-1}}\left(\frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\cdot\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)}{\sqrt{\bar{\alpha}_t}}\right) \\ &+ \sqrt{1-\bar{\alpha}_{t-1}}\cdot\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\end{aligned} \tag{3}$$

DDIM has been demonstrated to be a first-order discrete numerical solution (Lu et al. 2022a) of ordinary differential equation (ODE):

$$\begin{aligned}\boldsymbol{x}_t = &\underbrace{\sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_s}}\boldsymbol{x}_s - \frac{1}{2}\sqrt{\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(x_s, s)\int_s^t\left(\frac{d\lambda_\delta}{d\delta}\right)\sqrt{\frac{1-\bar{\alpha}_\delta}{\bar{\alpha}_\delta}}d\delta}_{A} \\ &+ \underbrace{\mathcal{O}\big((\lambda_t-\lambda_s)^2\big)}_{B}\end{aligned} \tag{4}$$

here $\lambda_\delta = \log[\bar{\alpha}_\delta/(1-\bar{\alpha}_\delta)]$. If $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_s, s)$ is assumed to be constant from s to t, the first term (A) of Eq.(4) is equivalent to Eq.(3) (Lu et al. 2022a). Such assumption brings high-order approximation error formalized in the second term (B), which leads to the discretization error in sampling process.

Generalizing Diffusion Distillation Model

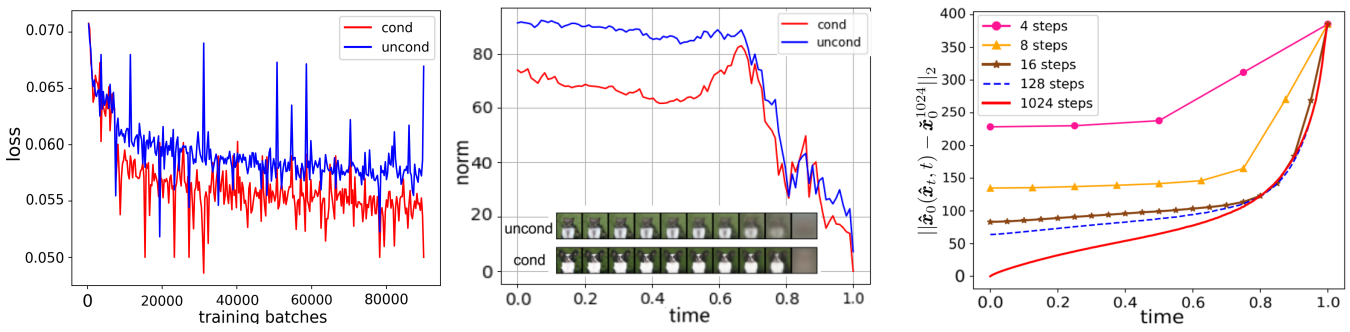
The Process of Diffusion Distillation Model

The distillation of diffusion models is the process that trains the student model $\boldsymbol{S}_\theta$ to approximate the corresponding target distribution in a single step, bypassing the costly multi-step sampling process. Based on this comprehension, we divide the training process of the diffusion distillation model into three steps (Figure 1), formulated in Eq.(5):

$$\hat{\boldsymbol{x}}_s = \boldsymbol{T}_\eta^{(t-s)}(\boldsymbol{x}_t, s, t) \coloneqq \boldsymbol{S}_\theta^{(1)}(\boldsymbol{x}_t, s, t) \tag{5}$$

$$\tilde{\boldsymbol{x}}_0^{\text{target}} = \boldsymbol{R}(\hat{\boldsymbol{x}}_{t_i}, s, t_i)\ , t_i \in [s\,, t-1] \tag{6}$$

Here $\boldsymbol{T}_\eta$ and $\boldsymbol{S}_\theta$ denote the teacher model and the student model, respectively, whose superscripts represent the corresponding number of sampling steps. Note that our framework can be generalized to different diffusion models. The sampling in Eq.(5) may use different numerical solvers g for the teacher or the student, and we will introduce the choice of g in implementation details. In the first step, we sample $\hat{\boldsymbol{x}}_{t_i}$ with the teacher model $\boldsymbol{T}_\eta$ in the same ODE generation path as $\boldsymbol{x}_t$. The process ensures the consistency of the target distribution (Song et al. 2023), which is crucial for effective distillation (Hinton, Vinyals, and Dean 2015). In the second step, a target value $\tilde{\boldsymbol{x}}_0^{\text{target}}$ is obtained for the student's learning with the target function $\boldsymbol{R}$ as Eq. 6. The target value is typically the denoised prediction of $\hat{\boldsymbol{x}}_{\boldsymbol{s}}$. $\boldsymbol{R}$ is defined differently given various distillation principles. It can be the previous teacher from the first step (Salimans and Ho 2022; Meng et al. 2022; Luhman and Luhman 2021), another newly proposed teacher (Dockhorn, Vahdat, and Kreis 2022) or a competent student (Song et al. 2023). $\boldsymbol{T}_\eta$ is a pre-trained diffusion generative model, whose network parameters η are often fixed during the distillation process. Finally, in the third



(a) (b) (c)

Figure 2: (a) shows the training MSE loss of conditional and unconditional diffusion models based on ϵ-prediction. (b) shows the ℓ_2 distances of the predicted real samples $\hat{\boldsymbol{x}}_0^t$ between the baseline teacher model and the student models in each time step ($1 \rightarrow 0$). Both unconditional and conditional student models are compared here. (c) shows the ℓ_2 distances from $\hat{\boldsymbol{x}}_0^t$ given by different student models in each time step ($1 \rightarrow 0$) to the final generated sample $\check{x}_0^{1000}$ given by the baseline teacher model with the same initial noise. Examples in these figures are based on models trained on CIFAR-10 (Krizhevsky 2009).

step, the student model $\boldsymbol{S}_\theta$ is trained to fit the target value $\tilde{\boldsymbol{x}}_0^{\text{target}}$ as Eq.(7) choosing appropriate $\omega(\lambda_t)$.

$$\mathcal{L}_S = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_t}[\omega(\lambda_t)\|\hat{\boldsymbol{x}}_0^{\text{student}}(\boldsymbol{x}_t\,,\,t) - \tilde{\boldsymbol{x}}_0^{\text{target}}\|_2^2] \qquad (7)$$

Such three-step division of the distillation process provides a new insight into identifying the errors that may occur during the process and how to mitigate them. We argue that the common errors arising in the training of diffusion distillation models mainly originate from the sampling error of the teacher model in the first step, as well as the fitting error of the student model in the third step.

Sampling error in the teacher model The sampling error of the teacher model can be divided into the fitting error in training and the discretization error in sampling. Without loss of generality, we consider teacher model of $\boldsymbol{\epsilon}$-prediction, whose fitting error is reflected by the mismatch between the predicted noise $\boldsymbol{\epsilon}_\eta$ and the real noise $\boldsymbol{\epsilon}$. This error is caused by the model's limited capacity or the divergence during training. On the other hand, the discretization error is largely introduced in the sampling when the step size is large and the high-order variation of $\boldsymbol{\epsilon}_\eta$ is omitted, as in Eq.(4).

Fitting error in the student model The error is generally caused by the student model's inability to regress the denoised prediction $\tilde{\boldsymbol{x}}_0^{\text{target}}$ when the loss in Eq.(7) fails to converge to zero. To further investigate the main cause, we examine denoised prediction given student models with different total steps and visualize errors in Figure 2c. The model with 1024 total steps is the teacher model, while others are students obtained through Progressive Distillation method. The curves depict the mean square error between the intermediate denoised samples and the final sample generated by the teacher. Empirical evidence shows that the error increases as the number of sampling steps decreases. The gap widens substantially when the student models sample only 4 or 8 steps. Furthermore, this error typically occurs during the middle or final stages of sampling.

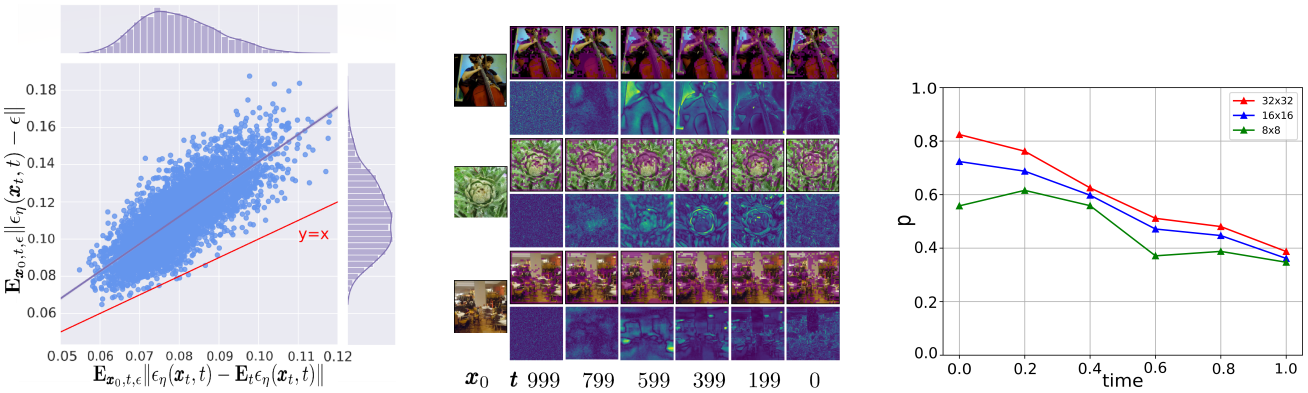
Previous works have mainly reduced the discretization error of the teacher model efficiently (Lu et al. 2022a,b). In this study, we focus on reducing the fitting errors of both the teacher and student models to pave a new way for the distillation of diffusion models.

The Exploration of Reducing Fitting Error

Reducing fitting error of the student Figure 2a shows the training loss of the conditional diffusion model is lower than that of the unconditional diffusion model. We also find that the conditional student model gives a denoised prediction $\hat{\boldsymbol{x}}_0$ closer to the teacher's prediction with better quality than the unconditional student during the sampling process (Figure 2b). These results suggest that semantic information (like labels) reduces the fitting error in the diffusion model. We will embed semantic information in a learned latent space for error reduction.

Reducing fitting error of the teacher In this part, we present our findings that the fitting error is dominated by the prediction variation, and the variation is correlated to the self-attention map spatially. Firstly, we perform a bias-variance decomposition of the ϵ-prediction loss, which measures the fitting error (Eq.(8)). The first term (A) represents the prediction variance across the sampling procedure, while the second component (B) represents bias. With the training data, we estimate the fitting error and prediction variance values. As visualized in Figure 3a, the fitting error is mainly determined by the variance spanning the entire global range, and the bias plays a minor role. Besides, the error is also positively correlated with the variance (see Appendix C.1 for more details). Therefore, the fitting error can be largely reduced by restricting the prediction variance. Such insight conceptually coincides with the consistency assumption in Consistency Models (Song et al. 2023).

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}_0,t,\epsilon}\|\epsilon_\eta(\boldsymbol{x}_t,t) - \epsilon\| = &\underbrace{\mathbb{E}_{\boldsymbol{x}_0,t,\epsilon}\|\epsilon_\eta(\boldsymbol{x}_t,t) - \mathbb{E}_t\epsilon_\eta(\boldsymbol{x}_t,t)\|}_{A} \\ &+ \underbrace{\mathbb{E}_{\boldsymbol{x}_0,\epsilon}\|\mathbb{E}_t\epsilon_\eta(\boldsymbol{x}_t,t) - \epsilon\|}_{B}\end{aligned} \qquad (8)$$



(a) (b) (c)

Figure 3: (a) The correlation trend between the values of $\|\epsilon_\eta(\boldsymbol{x}_t,t)-\epsilon\|$ (fitting error) and $\|\epsilon_\eta(\boldsymbol{x}_t,t)-\mathbb{E}_t\epsilon_\eta(\boldsymbol{x}_t,t)\|$ (variance) at $t=399$ using ϵ-prediction pre-trained diffusion model on ImageNet, given different $\boldsymbol{x}_0$'s and $\boldsymbol{\epsilon}$'s. (b) Visualization of attention maps and predicted noise variance on ImageNet diffusion. The left image in each group is the original image. The first row is the annotated attention maps while the second the noise variance at different t. (c) The Pearson correlation between the attention maps from different resolutions (8, 16, 32) and the predicted noise variance during sampling ($1\rightarrow 0$). Using ϵ-prediction pre-trained diffusion model on ImageNet.

By analyzing the spatial configuration of the prediction variance, we find high variance tends to occur in regions with high self-attention scores. Inspired by (Baranchuk et al. 2021; Tumanyan et al. 2022; Kwon, Jeong, and Uh 2022), we opt to utilize attention modules from the decoder part of U-Net (Ronneberger, Fischer, and Brox 2015) in the diffusion model to extract self-attention maps A_t^l, which have been experimentally demonstrated to contain more representations. Specially, we perform global average pooling and upsampling on A_t^l to match the resolution of $\hat{\boldsymbol{x}}_t$.

As illustrated in Figure 3b, the predicted noise variance and the regions with high self-attention scores exhibit similar spatial distributions throughout the sampling process. By investigating the attention maps, we suggest that main subjects are mostly generated during the middle stages of sampling, while visual details are added during the final stages. This finding is supported by the experimental view of (Baranchuk et al. 2021; Wu et al. 2022). Consequently, a well-trained method should prioritize enhancing features in different granularity that require emphasis at various timesteps. We conduct Pearson correlation analysis between the attention maps and the spatial prediction variance during the sampling process (Figure 3c). As time t goes from 1 to 0, the correlation between the variance and all attention maps at different resolutions gradually increases. Notably, the attention score maps resolved at 32×32 display the strongest positive correlation. This finding implies the potential use of attention maps in error reduction.

Method

Based on the observations in the previous section, we propose Spatial Fitting-Error Reduction Distillation of denoising diffusion models (SFERD). SFERD uses attention guidance and an extrinsic semantic gradient predictor to reduce the fitting error of the teacher and the student models, respectively. To better illustrate the improvements, we use DDPM process, DDIM sampler g and $\boldsymbol{\epsilon}$-prediction pre-trained teacher model $\boldsymbol{T}_\eta$ by default in this section. Moreover, our methods can be extended to other samplers and predicted parameters including v-prediction or x_0-prediction (Salimans and Ho 2022; Ho et al. 2022) easily.

Teacher Model with Attention Guidance

Our approach focuses on identifying and optimizing high attention score regions ("Risky Regions"), which are strongly correlated with the teacher's fitting error in generated images. The process visualized in Figure 4 involves three key operations, including Gaussian blurring, attention injection and attention guidance sampling.

Gaussian blurring The real image $\boldsymbol{x}_0$ is first sampled by a forward process to obtain $\boldsymbol{x}_t$ (Step 1 in Figure 4). Then, given $\boldsymbol{x}_t$ and with reparameterization, the denoised prediction $\hat{\boldsymbol{x}}_{\boldsymbol{0}}^{\boldsymbol{t}}$ is predicted by the teacher model $\boldsymbol{T}_\eta$ at time t (Step 2 in Figure 4). Next, we utilize Gaussian blur to introduce interference for the construction of unbalanced information. Specially, we deliberately destroy the Risky Regions from $\hat{\boldsymbol{x}}_{\boldsymbol{0}}^{\boldsymbol{t}}$ with Gaussian blur. This allows us to extract and optimize them later. Finally, we employ inverse DDIM to generate $\tilde{\boldsymbol{x}}_t=\sqrt{\bar{\alpha}_t}B(\hat{\boldsymbol{x}}_0^t)+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{\boldsymbol{\eta}}^{\boldsymbol{t}}$ (Step 3 in Figure 4). $B(\cdot)$ denotes the Gaussian blur operation.

The main reason for using Gaussian blur is twofold. Conceptually, Gaussian blur can reduce trivial details in images while preserving the original manifold of the generated image. Empirically, further experiments demonstrate in most time during sampling, the blurred denoised prediction is closer to the final generated sample than the original denoised prediction (Appendix E).

Attention injection This operation is to further highlight the Risky Regions since $\tilde{\boldsymbol{x}}_t$ after Gaussian blurring is globally

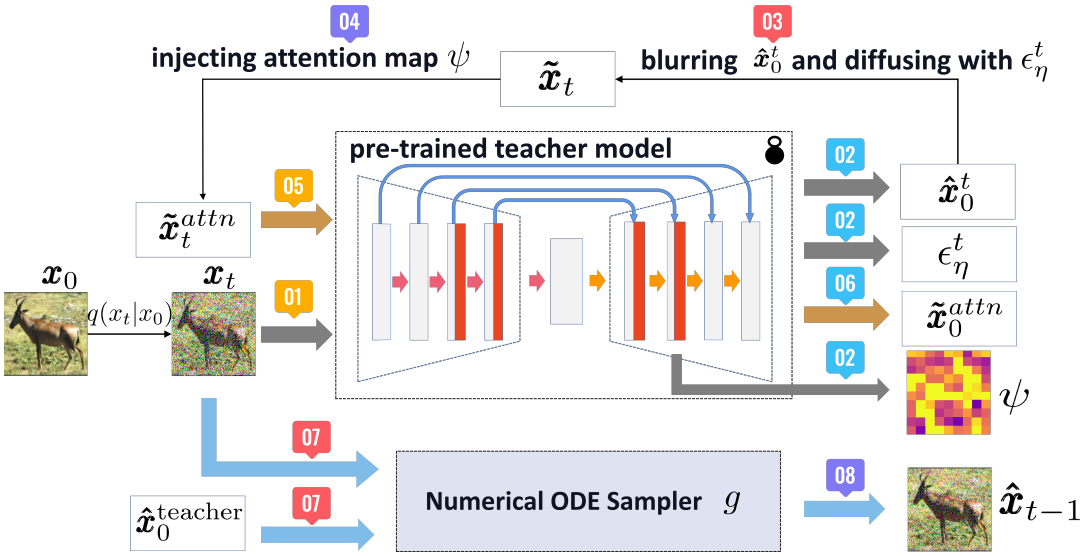


Figure 4: The illustration of attention guidance in the teacher model.

blurred instead of being applied to the region only. Let f_t^l denote the hidden feature fed to the attention blocks in layer l at time t. After N heads of self-attention blocks, the output self-attention map A_t^l can be expressed as:

$$\begin{aligned} A_t^l &= \text{softmax}(Q_t^{l,a}(K_t^{l,a})^T/\sqrt{d}) \\ \text{where} \quad Q_t^{l,a} &= f_t^l W_Q^{l,a}, K_t^{l,a} = f_t^l W_K^{l,a} \end{aligned} \tag{9}$$

Here $a \in [0, N-1]$ denotes an attention head, and d is the output dimension of queries $Q_t^{l,a}$ and keys $K_t^{l,a}$. After that, we upsample A_t^l to the image size and extract the regions with high attention scores with Eq.(10).

$$\psi = \mathbb{I}(A_t^l > k)\ , \quad \tilde{\boldsymbol{x}}_t^{attn} = (1-\psi)\odot \boldsymbol{x}_t + \psi \odot \tilde{\boldsymbol{x}}_t \tag{10}$$

Here ψ denotes a Boolean matrix where a pixel is 1 if its attention value is over a given threshold k and 0 otherwise. $\odot$ denotes the Hadamard product. $\tilde{\boldsymbol{x}}_t^{attn}$ is identical to $\boldsymbol{x}_t$ in regions with low self-attention scores but becomes blurred in regions with high scores (Step 4 in Figure 4).

Attention guidance sampling In this operation, the denoised prediction $\tilde{\boldsymbol{x}}_0^{attn}$ is calculated from the teacher model with $\tilde{\boldsymbol{x}}_t^{attn}$ (Steps 5, 6 in Figure 4). Together with $\hat{\boldsymbol{x}}_0^t$, the improved denoised prediction $\tilde{\boldsymbol{x}}_0^{\text{teacher}}$ can be obtained. Both calculations are in Eq.(11). Subsequently, the DDIM sampler conditioned on $\boldsymbol{x}_t$ and $\tilde{\boldsymbol{x}}_0^{\text{teacher}}$ is applied to get $\hat{\boldsymbol{x}}_{t-1}$ (Step 7, 8 in Figure 4). Repeating the above three operations until $\hat{\boldsymbol{x}}_{t_i}$ is obtained. Finally, we get $\tilde{\boldsymbol{x}}_0^{\text{target}}$ by Eq.(6).

$$\begin{aligned} \tilde{\boldsymbol{x}}_0^{attn} &= \frac{\tilde{\boldsymbol{x}}_t^{attn} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{\boldsymbol{\eta}}(\tilde{\boldsymbol{x}}_t^{attn}, t)}{\sqrt{\bar{\alpha}_t}} \\ \tilde{\boldsymbol{x}}_0^{\text{teacher}} &= \hat{\boldsymbol{x}}_0^t + w \times (\hat{\boldsymbol{x}}_0^t - \hat{\boldsymbol{x}}_0^{attn}) \end{aligned} \tag{11}$$

Here w denotes the attention guidance strength. $(\hat{\boldsymbol{x}}_0^t - \hat{\boldsymbol{x}}_0^{attn})$ contains the semantic information differences for guidance highlighted by the high attention scores, helping the teacher to improve the quality of denoised prediction. The approach

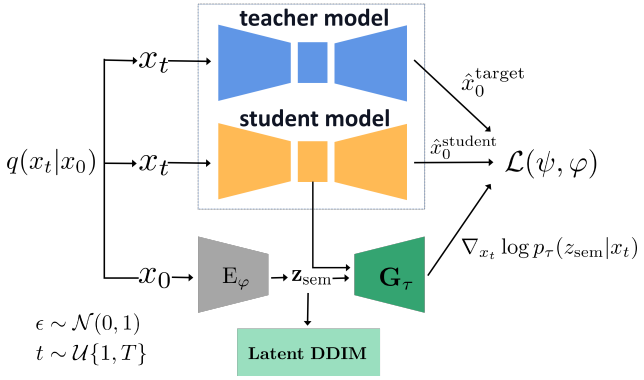


Figure 5: The training process of distillation with semantic gradient predictor. Note that the original student model is trained.

relies entirely on the diffusion model's intrinsic representations and provides a new perspective of guidance. It is unsupervised and still supports the extension of incorporating external conditions for better performance.

Student Model with Semantic Gradient Predictor

The gradients with semantic information calculated by the classifier have been validated for their ability to compensate for information bias and loss that arise during sampling (Dhariwal and Nichol 2021). To reduce the fitting error of the trained distillation student model, we introduce a learned semantic encoder in the student model, which provides a latent vector containing more intact reconstruction information. In detail, we integrate a semantic encoder $\boldsymbol{z}_{\text{sem}} = \mathrm{E}_\varphi(\boldsymbol{x}_0)$ and a predictor $G_\tau(\boldsymbol{x}_t, \boldsymbol{z}_{\text{sem}}, t)$ for the student to learn representations from the real image x_0. This would help in fitting to $\tilde{\boldsymbol{x}}_0^{\text{target}}$. Previous work (Dhariwal and Nichol 2021) introduce an extra label y in the conditional

Model	NFE	CIFAR-10 32×32 FID (↓)	CIFAR-10 32×32 IS (↑)	ImageNet 64×64 FID (↓)	ImageNet 64×64 IS (↑)
SFERD-PD (ours)	1	7.54	8.61	14.85	36.55
PD (Salimans and Ho 2022)	1	14.85	7.96	19.23	22.73
SFERD-CD (ours)	1	**5.31**	**9.24**	**9.39**	**47.19**
CD (Song et al. 2023)	1	8.21	8.41	13.87	29.98
DFNO* (Zheng et al. 2023)	1	4.12	/	8.35	/
ViTGAN (Lee et al. 2021)	1	6.66	9.30	/	/
DiffAugment-BigGAN (Zhao et al. 2020)	1	5.61	9.16	/	/
TransGAN (Jiang, Chang, and Wang 2021)	1	9.26	9.02	/	/
SFERD-PD (ours)	2	6.37	8.92	7.53	45.72
PD (Salimans and Ho 2022)	2	7.64	8.85	9.71	25.37
SFERD-CD (ours)	2	**4.19**	**9.45**	**6.08**	**54.05**
CD (Song et al. 2023)	2	6.26	9.17	8.24	32.60
SFERD-PD (ours)	4	3.44	9.32	5.93	55.19
PD (Salimans and Ho 2022)	4	4.28	9.25	7.22	30.72
SFERD-CD (ours)	4	**2.68**	**9.79**	**4.41**	**57.98**
CD (Song et al. 2023)	4	3.39	9.71	5.81	35.41
SFERD-EDM (ours)	35	**2.12**	**9.88**	/	/
EDM (Karras et al. 2022)	35	2.41	9.83	/	/
SFERD-EDM (ours)	79	/	/	**2.43**	**74.73**
EDM (Karras et al. 2022)	79	/	/	2.99	39.09
SFERD-DDIM (ours)	1024	2.27	9.80	2.82	69.17
DDIM (Song, Meng, and Ermon 2020)	1024	2.58	9.76	3.34	37.55

Table 1: Sample quality on CIFAR-10 and ImageNet 64×64. SFERD-* represents the implementation of the corresponding model within SFERD framework. For example, SFERD-PD and SFERD-RD represent the models that refer to ideas of Progressive Distillation (PD) and Consistency Distillation (CD) and imply within SFERD, respectively. Both attention guidance method and semantic encoding-based gradient predictor are introduced.

diffusion model. When we replace class label conditions $\boldsymbol{y}$ with the learned latent equation $\boldsymbol{z}_{\text{sem}}$:

$$p_{\theta,\varphi}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{z}_{\text{sem}}) \approx \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) + \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t) \cdot \nabla_{\boldsymbol{x}_t} \log p_\tau(\boldsymbol{z}_{\text{sem}}|\boldsymbol{x}_t)\,,\, \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)) \tag{12}$$

Based on Eq.(12), the training objective of the distillation model can be reformulated as Eq.(14).

$$\begin{aligned}
&\mathcal{L}(\theta, \varphi, \tau) \\
=&\mathbb{E}_{\boldsymbol{x}_0,t,\epsilon}\Big[\omega(\lambda_t)\big\|\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t) \cdot \nabla_{\boldsymbol{x}_t} \log p_\tau(\boldsymbol{z}_{\text{sem}}|\boldsymbol{x}_t) \\
&- \big(\widetilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_0^{\text{target}}) - \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)\big)\big\|^2\Big] \qquad (13) \\
=&\mathbb{E}_{\boldsymbol{x}_0,t,\epsilon}\Big[\omega(\lambda_t)\big\|\frac{(1-\bar{\alpha}_t)\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)}{\sqrt{\bar{\alpha}_{t-1}}\beta_t} \cdot \nabla_{\boldsymbol{x}_t} \log p_\tau(\boldsymbol{z}_{\text{sem}}|\boldsymbol{x}_t) \\
&- \big(\tilde{\boldsymbol{x}}_0^{\text{target}} - \hat{\boldsymbol{x}}_0^{\text{student}}(\boldsymbol{x}_t, t)\big)\big\|^2\Big] \qquad (14)
\end{aligned}$$

where $\boldsymbol{\Sigma}_\theta = \sigma_t^2\mathbf{I} = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\mathbf{I}$. Eq.(14) uses the gradient $\nabla_{\boldsymbol{x}_t} \log p_\tau(\boldsymbol{z}_{\text{sem}}|\boldsymbol{x}_t)$ with semantic information to compensate for the fitting error with $\tilde{\boldsymbol{x}}_0^{\text{target}}$. We design a predictor $G_\tau(\boldsymbol{x}_t, \boldsymbol{z}_{\text{sem}}, t)$ to approximate $\nabla_{\boldsymbol{x}_t} \log p_\tau(\boldsymbol{z}_{\text{sem}}|\boldsymbol{x}_t)$. Figure 5 shows the training network and data flow. By default, G_τ is not trained directly to fit $\nabla_{\boldsymbol{x}_t} \log p_\tau(\boldsymbol{z}_{\text{sem}}|\boldsymbol{x}_t)$ but combined with the distillation model. The trained student model is frozen during the first half of the training epochs and jointly optimized with G_τ and E_φ using a low learning rate in the latter half. We find in this case the optimized G_τ produces gradients close to $\nabla_{\boldsymbol{x}_t} \log p_\tau(\boldsymbol{z}_{\text{sem}}|\boldsymbol{x}_t)$. We validate the crucial role of incorporating $\boldsymbol{z}_{\text{sem}}$ in ensuring the effectiveness of the improved method through ablation experiments. The student model can be combined with G_τ to perform a few-step sampling using the formula similar to the classifier guidance. In addition, we also train another DDIM sampler $p_\omega(\boldsymbol{z}_{\text{sem}}^{t-1}|\boldsymbol{z}_{\text{sem}}^t)$ to sample $\boldsymbol{z}_{\text{sem}}$ in the latent space, following the approach as (Preechakul et al. 2022). Both E_φ and G_τ are independent of the original distillation model and can be integrated into the trained distillation process, achieving less training time.

Related Work

The goal of our methods is to accelerate sampling while maintaining high image quality, in line with many existing works. For instance, Watson et al. (2021, 2022) exploit using traditional dynamic programming methods to accelerate sampling, Zhang et al. (2022) proposes the use of an exponential integrator (DEIS) to speed up sampling. The use of knowledge

distillation in diffusion models can be traced back as far as the DDIM-based one-step denoising model (DS) implemented in (Luhman and Luhman 2021), whose main drawback is that the full steps of the original teacher model have to be fully applied for each training. The Classifier-Guided Distillation (CFD) proposed in (Sun et al. 2022) uses a classifier to extract the sharpened feature distribution of the teacher into the students and uses the KL divergence as training loss, allowing it to focus on the focal image features. Higher-Order Denoising Diffusion Solvers (GENIE) (Dockhorn, Vahdat, and Kreis 2022), on the other hand, achieve accelerated sampling by adding a prediction module capable of receiving information from distillation Higher Order Solvers to the network backbone. Progressive distillation (PD) (Salimans and Ho 2022) and guided distillation (GD) (Meng et al. 2022) implement unconditional and conditional distillation diffusion models, respectively, at the cost of halving the number of sampling steps per iteration, in addition to GD which extends the training to the latent space and implements random sampling. Consistency Distillation (CD) (Song et al. 2023) exploits the self-consistency of the ODE generation process by minimizing the difference between two noisy data points on the same ODE path to achieve few-step distillation. Our two improvement methods can be applied to the majority of diffusion distillation models above.

Experiments

Experimental Setting

We mainly examine SFERD on two standard image generation benchmarks, including CIFAR-10 and class-conditional ImageNet 64×64. We measure the performance using Fréchet Inception Distance (FID) (Heusel et al. 2017) and Inception Score (IS) (Salimans et al. 2016). All results are computed from 50,000 sampled generated images and averaged over 3 random seeds. All students are uniformly initialized by the corresponding teachers.

Few-steps Image Generation

Our distillation framework allows different configurations of diffusion process (like Variance Preserving or Exploding (Song et al. 2021b)), numerical solver, total diffusion timesteps and so on. For fair comparisons, we choose to implement SFERD by referring to the ideas of Progressive Distillation models (PD) (Salimans and Ho 2022) and Consistency Distillation models (CD) (Song et al. 2023). Specifically, PD can be interpreted as setting the target function $\boldsymbol{R}$ in SFERD to $\boldsymbol{T}_\eta$, while $\boldsymbol{R}$ is set to the student model $\boldsymbol{S}_{\theta^-}$ updated by the exponential moving average (EMA) in CD. In our experiments, PD is applied to the teacher network from unconditional ADM (Dhariwal and Nichol 2021) using DDPM noise schedule and DDIM (Euler) sampler, and CD is applied to the teacher from EDM (Karras et al. 2022) and 2^{nd} Heun sampler. We unify the metric function to ℓ_2 distance. We pretrain all teachers using the configurations specified in the original papers. The initial teacher model of PD is set to 1024 sampling steps, whereas for CD, it is set to 18 or 40 steps. By default, the students of the previous distillation stage serve as the teachers for the next stage in SFERD-PDs.

We compare our SFERD with PD and CD on the CIFAR-10 and ImageNet 64×64. Results in Table 1 demonstrate that the performance of all models improved as the sampling steps increased. Notably, SFERD-PD and SFERD-CD achieved better performance than their baseline models, and SFERD-CD displays superior performance across all datasets and sampling steps. Our improved methods can be applied not only to mainstream diffusion distillation models but also to enhance the performance of pre-trained models directly through fine-tuning. Specifically, we apply our methods on pre-trained unconditional ADM and EDM directly, aligning the distillation steps of the student with that of the teacher, which can be easily achieved by setting $s = t - 1$. The results shown in Table 1 indicate superior performances of ADM and EDM, which are improved through the integration of attention guidance and semantic gradient predictor.

Ablation Studies

We conduct ablation experiments on the design of critical hyperparameters in the training of SFERD. All ablations are performed on the conditional ImageNet 64×64 using SFERD-PD (no semantic gradient predictor in the student) with 4 sampling steps unless otherwise stated.

Attention threshold. In order to determine the attention threshold, we compute the scales of 0.8, 0.9, 1.0, 1.1 and 1.2. The best metrics are obtained when the ψ is 1.0. A threshold that is too high or too low can deteriorate the performance.

Attention guidance strength. We evaluate the effect of attention guidance strength, calculating the scales from 0 to 0.5. The best FID was achieved when $w = 0.3$.

Gaussian blur strength. We evaluate the effect of Gaussian blur strength σ on performance. We test the strength values of 1, 3, 5, and obtain the best FID at σ = 3.

Denoising ability. Specially, we randomly select 500 real images $\boldsymbol{x}_0$ from ImageNet 64×64 and perform forward diffusion on them to obtain $\boldsymbol{x}_t$. We then compute the ℓ_2 distances between $\boldsymbol{x}_0$ and $\hat{\boldsymbol{x}}_0^t$ generated by pre-trained 4-sampling-step CD, CD with attention guidance and SFERD-CD, averaging over 500 random images. The results indicate that both enhancements of SFERD obviously reduce the ℓ_2 distances of the denoised prediction at each timestep. Moreover, the FID of CD, CD with attention guidance, and SFERD-CD are presented in order: 5.81, 5.27, and 4.41, respectively.

Conclusion

In conclusion, we propose the Spatial Fitting-Error Reduction Distillation model (SFERD) for Denoising Diffusion Models. SFERD effectively enhances performance using intrinsic and extrinsic representations, generating high-quality samples in only few steps. The core idea behind SFERD is to reduce the fitting error of the student and the teacher in distillation. It is achieved through the independent use of attention guidance we proposed for the teacher and an external semantic gradient predictor in the student model.

Acknowledgements

This paper is funded by National Key R&D Program of China (2022YFB3303301), National Natural Science Foundation of China (Grant No. 62006208), and Youth Program of Humanities and Social Sciences of the Ministry of Education (No.23YJCZH338). This paper is also supported by Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies. The corresponding author of the paper is Zejian Li. We are grateful to Jionghao Bai and Jingran Luo for supplementing the appendix. We also thank to Chenye Meng and Haoran Xu for their helpful comments and discussion, and to Jiahui Zhang, Qi Liu, Ying Zhang, and Yibo Zhao for proofreading the draft.

References

Baranchuk, D.; Voynov, A.; Rubachev, I.; Khrulkov, V.; and Babenko, A. 2021. Label-Efficient Semantic Segmentation with Diffusion Models. In *International Conference on Learning Representations*.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Dockhorn, T.; Vahdat, A.; and Kreis, K. 2022. Genie: Higher-order denoising diffusion solvers. *Advances in Neural Information Processing Systems*, 35: 30150–30166.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative Adversarial Networks. *Communications of the ACM*, 63(11): 139–144.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two-time-scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems*, 30: 6626–6637.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Huang, H.; Sun, L.; Du, B.; and Lv, W. 2023. Conditional Diffusion Based on Discrete Graph Structures for Molecular Graph Generation. In *AAAI Conference on Artificial Intelligence*.

Jiang, Y.; Chang, S.; and Wang, Z. 2021. TransGAN: Two pure transformers can make one strong GAN, and that can scale up. *Advances in Neural Information Processing Systems*, 34: 14745–14758.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2022. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*.

Kingma, D.; Salimans, T.; Poole, B.; and Ho, J. 2021. Variational diffusion models. *Advances in Neural Information Processing Systems*, 34: 21696–21707.

Kong, Z.; and Ping, W. 2021. On Fast Sampling of Diffusion Probabilistic Models. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Kwon, M.; Jeong, J.; and Uh, Y. 2022. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*.

Lee, K.; Chang, H.; Jiang, L.; Zhang, H.; Tu, Z.; and Liu, C. 2021. ViTGAN: Training GANs with Vision Transformers. In *International Conference on Learning Representations*.

Liu, L.; Ren, Y.; Lin, Z.; and Zhao, Z. 2022. Pseudo Numerical Methods for Diffusion Models on Manifolds. In *International Conference on Learning Representations*.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022a. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In *Advances in Neural Information Processing Systems*.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022b. DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. *arXiv preprint arXiv:2211.01095*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Gool, L. V. 2022. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11451–11461.

Luhman, E.; and Luhman, T. 2021. Knowledge Distillation in Iterative Generative Models for Improved Sampling Speed. *arXiv preprint arXiv:2101.02388*.

Luo, S.; and Hu, W. 2021. Diffusion probabilistic models for 3d point cloud generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2837–2845.

Meng, C.; Gao, R.; Kingma, D. P.; Ermon, S.; Ho, J.; and Salimans, T. 2022. On Distillation of Guided Diffusion Models. In *NeurIPS 2022 Workshop on Score-Based Methods*.

Preechakul, K.; Chatthee, N.; Wizadwongsa, S.; and Suwajanakorn, S. 2022. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10619–10629.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Part III 18*, volume 9351, 234–241. Springer.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training GANs. *Advances in Neural Information Processing Systems*, 29: 2226–2234.

Salimans, T.; and Ho, J. 2022. Progressive Distillation for Fast Sampling of Diffusion Models. In *International Conference on Learning Representations*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, Y.; Dhariwal, P.; Chen, M.; and Sutskever, I. 2023. Consistency Models. *arXiv preprint arXiv:2303.01469*.

Song, Y.; Durkan, C.; Murray, I.; and Ermon, S. 2021a. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34: 1415–1428.

Song, Y.; and Ermon, S. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. *Advances in Neural Information Processing Systems*, 32: 11895–11907.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021b. Score-based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Sun, W.; Chen, D.; Wang, C.; Ye, D.; Feng, Y.; and Chen, C. 2022. Accelerating Diffusion Sampling with Classifier-based Feature Distillation. *arXiv preprint arXiv:2211.12039*.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2022. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. *arXiv preprint arXiv:2211.12572*.

Wang, S.; Saharia, C.; Montgomery, C.; Pont-Tuset, J.; Noy, S.; Pellegrini, S.; Onoe, Y.; Laszlo, S.; Fleet, D. J.; Soricut, R.; et al. 2022. Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting. *arXiv preprint arXiv:2212.06909*.

Wu, Q.; Liu, Y.; Zhao, H.; Kale, A.; Bui, T. M.; Yu, T.; Lin, Z.; Zhang, Y.; and Chang, S. 2022. Uncovering the Disentanglement Capability in Text-to-Image Diffusion Models. *ArXiv*, abs/2212.08698.

Zhang, Z.; Zhao, Z.; and Lin, Z. 2022. Unsupervised representation learning from pre-trained diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 35: 22117–22130.

Zhao, S.; Liu, Z.; Lin, J.; Zhu, J.-Y.; and Han, S. 2020. Differentiable augmentation for data-efficient GAN training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 7559–7570.

Zheng, H.; Nie, W.; Vahdat, A.; Azizzadenesheli, K.; and Anandkumar, A. 2023. Fast sampling of diffusion models via operator learning. In *International Conference on Machine Learning*, 42390–42402. PMLR.

Zhou, L.; Du, Y.; and Wu, J. 2021. 3d shape generation and completion through point-voxel diffusion. In *IEEE/CVF International Conference on Computer Vision*, 5826–5835.