Diverse and Stable 2D Diffusion Guided Text to 3D Generation with Noise Recalibration

Xiaofeng Yang[1], Fayao Liu[2], Yi Xu[3], Hanjing Su[4], Qingyao Wu[5], Guosheng Lin[1*]

[1]Nanyang Technological University, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]OPPO US Research Center, USA
[4]Tencent, China
[5]South China University of Technology, China
{yang.xiaofeng, gslin}@ntu.edu.sg

Abstract

In recent years, following the success of text guided image generation, text guided 3D generation has gained increasing attention among researchers. Dreamfusion is a notable approach that enhances generation quality by utilizing 2D text guided diffusion models and introducing SDS loss, a technique for distilling 2D diffusion model information to train 3D models. However, the SDS loss has two major limitations that hinder its effectiveness. Firstly, when given a text prompt, the SDS loss struggles to produce diverse content. Secondly, during training, SDS loss may cause the generated content to overfit and collapse, limiting the model's ability to learn intricate texture details. To overcome these challenges, we propose a novel approach called Noise Recalibration algorithm. By incorporating this technique, we can generate 3D content with significantly greater diversity and stunning details. Our approach offers a promising solution to the limitations of SDS loss.

Introduction

Text guided 3D generation is a challenging task that aims to generate 3D content based on textual prompts. This approach has numerous applications in various fields such as gaming, virtual environments, automation, AI augmented design, and 3D data augmentations. However, the lack of annotated 3D data makes this task extremely difficult. Current 3D generation methods (Chan et al. 2022; Liao et al. 2020; Henzler, Mitra, and Ritschel 2019; Nguyen-Phuoc et al. 2019, 2020; Wu et al. 2016; Zhu et al. 2018; Zhou et al. 2021; Yu et al. 2021), which typically focus on generating categorical objects, often require pose supervision during training, resulting in a significant gap between 3D generation and text guided generation.

To address this challenge and achieve photo-realistic 3D object and scene generation, recent methods have utilized 2D models trained on 2D image data. For example, Dreamfield (Jain et al. 2022) leverages the contrastive image text model CLIP (Radford et al. 2021) to train Neural Radiance Field (NeRF) (Mildenhall et al. 2021) by measuring the similarity between rendered images and text. Meanwhile, Dreamfusion (Poole et al. 2022) represents a significant breakthrough in improving generation quality by using 2D language-guided diffusion models to train NeRF.

Specifically, Dreamfusion (Poole et al. 2022) starts from a random initialized NeRF and adjusts the NeRF weights by calculating the Score Distillation Sampling (SDS) loss between the NeRF rendered images and the text prompt based on the 2D diffusion model's output. A detailed illustration of the Dreamfusion method can be found in the third Section.

Despite the success of SDS loss, it faces two major issues. First, as also witnessed in Dreamfusion, the SDS loss can hardly generate diversified content given different random seeds during NeRF training. The authors attribute the reason to that the smoothed density may not contain many distinct modes at high noise levels (Poole et al. 2022). In experiments, we also witness the similar issue – given a fixed text prompt, the randomness in NeRF optimization does not guarantee sufficient generation variety. Second, the degeneration issue. The SDS loss can cause the learned NeRF to gradually collapse, preventing it from learning high-quality texture details. This issue occurs not only with the original SDS loss, but also with its successors like VSD loss (Wang et al. 2023). We show a visual illustration of the two problems in Fig. 2.

In this paper, we thoroughly investigate the two issues and propose a Noise-Recalibration SDS (NR-SDS) algorithm to overcome them. The NR-SDS algorithm contains two parts: the single noise training and the Noise Recalibration loss. First, we propose a single noise training scheme to address the diversity issue. We demonstrate that the original SDS loss is searching for the optimal mode using random noises from the entire Gaussian space. However, the generation process can be limited to a single noise sampled from the Gaussian space, leading to more diverse results. Second, we propose Noise-Recalibration loss to address the degeneration issue. We attribute the degeneration problem to the high guidance weight used in SDS loss. While a high guidance factor is essential for the NeRF model to learn text-specific content, it tends to cause the learned NeRF to degrade during training. To resolve this dilemma, we make the "single noise" in the single noise training method learnable and

*Corresponding Author

Figure 1: Examples of generation results with NR-SDS loss. Given a text prompt, we show our method can generate high-quality diversified 3D objects. Results are generated using NeRF only, without DMTet finetuning.

gradually adjust the learnable noise based on the Noise Recalibration loss, such that the noise could generate high quality contents even when operating at a high guidance weight. By using the NR-SDS algorithm, we achieve impressive results with improved texture details and diverse 3D content generation. Some generation results can be found in Fig. 1.

The paper is organized as follows: in the related works section, we first briefly review the 3D generation methods and text guided 3D generation methods. After that, we introduce the preliminaries of diffusion models, the Dreamfusion algorithm, the SDS loss and the problems of SDS loss. Consequently, we present our proposed NR-SDS algorithm. In the experiment section, we show the ablation experiments and more experimental results.

Unless otherwise stated, our analysis and experiments are primarily based on the latent diffusion models (Rombach et al. 2022). However, we show in the experiment section that the identified issues are not unique to the latent diffusion models. Our proposed method could improve the Pixel-Pixel diffusion models (Saharia et al. 2022) as well.

To summarize our contributions:

- We identify and systematically study the diversity and degeneration issues of SDS loss.
- We propose the NR-SDS algorithm. The NR-SDS algorithm consists of two key components: single noise training to solve the diversity issue and the Noise Recalibration loss to solve the degeneration issue.
- With the proposed NR-SDS algorithm, we are able to generate high-quality, multi-view consistent 3D objects using 2D diffusion models.

Related Works

Our work belongs to the field of 3D generation, specifically in text guided 3D generation. Previous research (Chan et al.

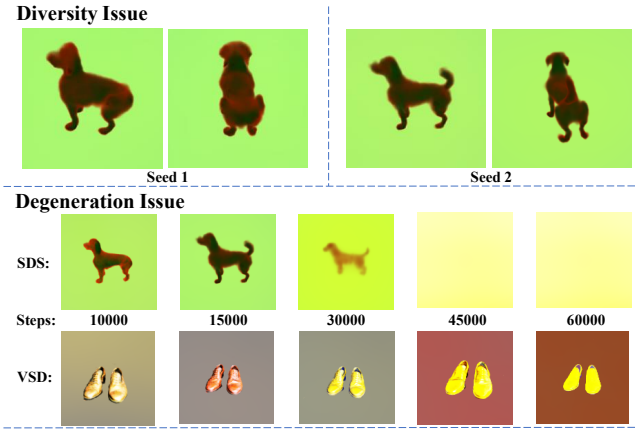

Figure 2: An illustration of the two identified issues. For the diversity issue, we run SDS loss using text prompt "a dog" with different random seeds. It can be observed that the generated content barely changes with different seeds. For the degeneration issue, we observe that it happens with both SDS loss and VSD loss.

2022; Liao et al. 2020; Henzler, Mitra, and Ritschel 2019; Nguyen-Phuoc et al. 2019, 2020; Wu et al. 2016; Zhu et al. 2018; Zhou et al. 2021; Yu et al. 2021) has focused on generating synthetic data or data of a single category, requiring the network to be trained on multi-view images of the same scene or images with pose annotations. These supervised learning methods have limitations in training scale and generalizing ability. Moreover, the lack of annotated 3D data with language constraints supervised language-guided 3D generation (Liu et al. 2022; Canfes et al. 2023) to simple shapes or avatars.

Compared to language-guided image generation models, which are usually trained on billions of images (Rombach et al. 2022; Ramesh et al. 2022; Saharia et al. 2022; Nichol et al. 2022), it seems impossible to achieve the same training scale for 3D data by supervised learning. As a result, researchers have turned to using large-scale trained multi-modal 2D models to improve text-guided 3D generation. Previous works (Jain et al. 2022; Wang et al. 2022) have used CLIP (Radford et al. 2021) to guide NeRF training or editing, but CLIP as a contrastive model struggles to recover high-frequency surface details and accurate object shapes. Large-scale text-guided diffusion models (Rombach et al. 2022; Ramesh et al. 2022; Saharia et al. 2022; Nichol et al. 2022) provide a more tractable way to distill 2D generative priors. Dreamfusion (Poole et al. 2022), a pioneering work, proposes a score distillation sampling (SDS) method to distill the prior of a 2D diffusion model for training Neural Radiance Field (NeRF) since diffusion models are a type of score function.

In the following section, we will introduce the preliminaries of Dreamfusion, the SDS loss, and the problems of Dreamfusion.

Dreamfusion and SDS Loss Revisit

Diffusion Model

The diffusion models (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015; Song, Meng, and Ermon 2020; Song et al. 2020) as a new family of state-of-the-art generative models treat the image generation process as a noise removing process. Starting from a randomly sampled noise from the Gaussian space, diffusion process gradually removes a small portion of Gaussian noise step by step. Next, we discuss the training and testing phases of the diffusion model following DDPM (Ho, Jain, and Abbeel 2020).

Training of Diffusion Model. To generate training data, given a sample from the real data distribution $x_0 \sim q(x)$, the diffusion process adds random Gaussian noise by following:

$$q(\mathbf{x_t} \mid \mathbf{x_{t-1}}) = \mathcal{N}(\mathbf{x_t}; \sqrt{1-\beta_t}\mathbf{x_{t-1}}, \beta_t \mathbf{I}), \quad (1)$$

where β is called a variance schedule with value $\beta_t \in (0, 1)$. Since the process is defined as a Markov process, we can also get:

$$q(\mathbf{x_{1:T}} \mid \mathbf{x_0}) = \prod_{t=1}^{T} q(\mathbf{x_t} \mid \mathbf{x_{t-1}}). \quad (2)$$

Considering Eq 2, Eq 1 can be further simplified as:

$$q(\mathbf{x_t} \mid \mathbf{x_0}) = \mathcal{N}(\mathbf{x_t}; \sqrt{\bar{\alpha}_t}\mathbf{x_0}, (1-\bar{\alpha}_t)\mathbf{I}), \quad (3)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{t=1}^{T} \alpha_t$. In other words, the distribution at time step t can be directly calculated from x_0 without considering intermediate steps. Once training data is generated, the diffusion model is trained by optimizing the MSE loss between the added random noise and the model prediction:

$$\mathcal{L}(\phi) = \mathbb{E}[\| \epsilon - \epsilon_\phi(\sqrt{\bar{\alpha}_t}\mathbf{x_0} + \sqrt{1-\bar{\alpha}_t}\epsilon, t) \|^2], \quad (4)$$

where ϵ is the noise added to the image and ϵ_ϕ is the noise predicted by the diffusion model.

The generation process is indeed the reverse process of diffusion process: to find $p(x_{t-1})$ given $p(x_t)$. Formally, the reverse diffusion process is:

$$p_\phi(\mathbf{x_{t-1}} \mid \mathbf{x_t}) = \mathcal{N}(\mathbf{x_{t-1}}; \mu_\phi(\mathbf{x_t}, \mathbf{t}), \Sigma_\phi(\mathbf{x_t}, \mathbf{t})), \quad (5)$$

where μ_ϕ and Σ_ϕ can be calculated from the trained diffusion model ϵ_ϕ given the output of the previous step and current step t.

Dreamfusion and SDS Loss

In this section, we provide a description of the Dreamfusion algorithm and the SDS loss. Methodologically, Dreamfusion shares the same underlying principles as other gradient inversion techniques, such as Deepdream (Mordvintsev, Olah, and Tyka 2015), Dreaming to distill (Yin et al. 2020), and Gradient Inversion (Yin et al. 2021). These methods seek to optimize the input of a trained model, rather than the model parameters. However, previous methods mainly employ pre-trained image classification networks like ResNet, whereas Dreamfusion uses a diffusion model for distillation

Figure 3: A high guidance is necessary to learn good shapes. With the guidance weight reduced from 100, the learned shapes get weaker.

purposes. The significant differences between a classification model and a diffusion model are twofold. Firstly, a diffusion model is a generative model that can potentially perform better in generation tasks when compared with a classification model. Secondly, a diffusion model is a score function that directly generates an update gradient.

Formally, suppose $g(\theta)$ is the NeRF model to learn, Dreamfusion optimizes the parameter θ by the following SDS loss:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x} = g(\theta)) \triangleq \mathbb{E}_{t,\epsilon}\left[(\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) - \epsilon)\frac{\partial \mathbf{x}}{\partial \theta}\right], \tag{6}$$

where ϵ is a randomly sampled noise and $\hat{\epsilon}_\phi$ is calculated from the trained diffusion model. Intuitively, the Dreamfusion algorithm adds a randomly sampled noise to the NeRF rendered image. The combined image is then fed into a trained diffusion model to predict the added noise. Ideally, if the rendered image is realistic, the diffusion model should predict the noise accurately. The difference between the predicted noise and the added noise is known as the SDS loss. It is witnessed in SDS loss that ignoring the UNet Jacobian will improve the generation quality.

In practice, the 2D diffusion model is a conditional diffusion model and the calculation of $\hat{\epsilon}_\phi$ is based on the classifier-free guidance (Ho and Salimans 2022):

$$\hat{\epsilon}_\phi = \epsilon_\phi(\mathbf{z}_t) + w(\epsilon_\phi(\mathbf{z}_t; y) - \epsilon_\phi(\mathbf{z}_t)), \tag{7}$$

where $\epsilon_\phi(\mathbf{z}_t; y)$ and $\epsilon_\phi(\mathbf{z}_t)$ represent the diffusion process using the given text prompt and the null embedding. w is the guidance weight. If the w is set too low, the generated image will be less related to the condition y. If the w is set too high, the generation quality will be reduced. In image generation tasks, a typical w choice is 5 to 20. However, SDS loss (Poole et al. 2022) requires training the 3D model with a guidance weight w as high as 100. Without such a high guidance value, the 3D models will not learn good shapes as described in Dreamfusion (Poole et al. 2022) (see Sec. 3.2 and Appendix Figure 9). In our experiments, we witness similar behaviours. We show a brief example in Fig. 3 using text prompt "a cat".

Understanding the Issues of SDS Loss

In this section, we will discuss the two major issues of SDS loss: the diversity issue and the degeneration issue. These issues are summarized in Fig. 2.

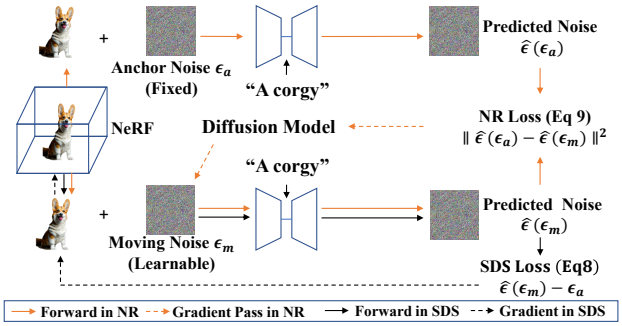

Figure 4: The forward and backward path of our NR-SDS algorithm. The SDS loss directly applies gradient on the NeRF rendered image, while the NR loss requires back-propagation on the diffusion model parameters. For the case of latent diffusion models, the rendered image will first pass an encoder and the noises are added to the latent features generated by the encoder.

The diversity issue is a problem that has been previously observed in the original Dreamfusion work (see appendix Sec. A.5) and also in our own experiments. As shown in Fig. 2, even when adjusting the random seeds, the SDS loss struggles to generate diverse content. For example, in the case of the dog images, the shape and texture of the dog remain relatively unchanged despite varying the random seeds.

The degeneration issue, as shown in Fig. 2 at the bottom, is another problem with SDS loss and its variances. Although the SDS loss can help the model learn good shapes in the early stages of training, the training process eventually collapses after longer iterations, resulting in a degeneration issue. Even for the VSD loss in prolific dreamer (Wang et al. 2023), the texture of the generated object could sometimes gradually disappear. This issue restricts the NeRF model from learning object details such as dog fur. Furthermore, we observe that the degeneration issue occurs unpredictably and not on a fixed training step. This makes it challenging to address this issue through methods like early stopping.

Our Method

In this section, we present our method – the Noise Recalibration SDS algorithm.

Resolving Diversity Issue with Single Noise Training

We hypothesize that the lack of diversity in SDS loss and Dreamfusion is primarily due to the large noise sampling space. To generate the 3D representation of a scene, the SDS algorithm samples random noise from the entire Gaussian space. This enforces the generated NeRF to satisfy all noises. In this case, the trained NeRF will finally become the average model. In fact, the generation of one data instance, whether it is an image or a 3D scene, is simply a data point sample from the data space. In the case of 2D generation based on Eq. 5, only T random noises are sam-

Algorithm 1: The NR-SDS Algorithm.

Input: Diffusion Model ϵ_ϕ, Language Prompt y, Hyper-parameters: Total Step N.
Output: NeRF Model $g(\theta)$

Initialize NeRF Model $g(\theta)$
Initialize fixed anchor noise ϵ_a
Set initial learnable moving noise $\epsilon_m[t] = \epsilon_a$ for all timestep t;
for *n = 1 to N* **do**
  SAMPLE diffusion time step t
  UPDATE θ Based on Eq 9 and $\epsilon_m[t]$
  UPDATE $\epsilon_m[t]$ Based on Eq 8
RETURN $g(\theta)$;

pled to generate a single image, with T values ranging from 25 to 1000, depending on the generation steps. **Therefore, we assume in 3D generation, the generated scene does not need to satisfy all the random noises sampled from the entire Gaussian space as well**. In its extreme case, a 3D scene only needs to satisfy one single noise when using SDS loss. With that in mind, we propose a single noise training scheme to restrict the noise sampling process to a single random noise sampled from the Gaussian space, i.e., training one scene with one random noise. In the experiment section and Fig. 6, we show that the single noise training method helps to generate more varied content without reducing the generation quality.

Resolving Degeneration Issue with Noise Recalibration Loss

To understand the degeneration issue, we first consider the SDS loss given different input images. Ideally, the SDS should generate a high value for the unreal images and a zero value for the real images. However, this can not be achieved with the current SDS loss due to the high guidance weight w (Eq. 6, Eq. 7). The reason for this is as follows: Given the training and generation formula of the classifier-free guidance diffusion model (Eq. 4 and Eq. 7), the training of the original diffusion model is carried out by randomly choosing the correct language embedding and the null embedding. In other words, the training is carried out with a guidance weight $w = 1$. Therefore, if the diffusion model is well trained, it could only guarantee a zero SDS loss value with guidance weight $w = 1$. Under the setting of guidance weight $w = 100$, the difference between the language embedding output and the null embedding output will be greatly amplified. We show a detailed visual proof in **Supplementary Material**. Finally, the SDS loss will still generate a relatively high response to the real images, causing the 3D model to get worse and be unable to learn good details.

Solving this problem is non-trivial because reducing the guidance weight in the SDS loss is not a viable option. As stated in the previous section and shown in Fig. 3, the NeRF model requires a large guidance weight to learn correct shapes. To address this issue, we propose the Noise-Recalibration (NR) loss.

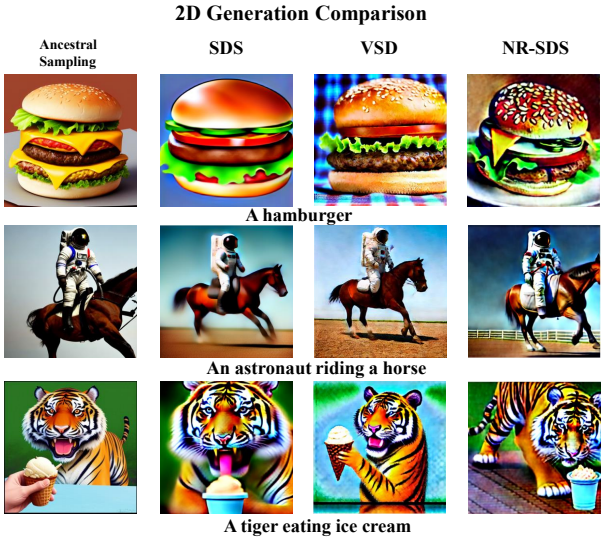

Figure 5: Generation comparison in 2D space.

The intuition behind NR loss is that we would like to make the "single noise" in the single noise training method learnable and the learnable noise could operate well even under guidance weight $w = 100$. We call this noise the learnable moving noise. To achieve this, we first sample a fixed anchor noise, such that the learnable moving noise running at $w = 100$ will finally converge to the fixed anchor noise running at $w = 1$.

Specifically, we define a fixed anchor noise ϵ_a and a learnable moving noise ϵ_m, where the moving noise ϵ_m is used to train the NeRF model and the anchor noise acts as an anchor to recalibrate the learnable moving noise. The Noise Recalibration loss is defined as:

$$\begin{aligned} &\|\hat{\epsilon}_\phi(\mathbf{z}_t; y, t, w = 1, \epsilon_a) \\ &\quad - \hat{\epsilon}_\phi(\mathbf{z}_t; y, t, w = 100, \epsilon_m[t])\ \|_2^2 \ . \end{aligned} \tag{8}$$

The NR loss gradually optimizes the moving noise to the direction of ϵ_a at $w = 1$. If fully optimized, the learnable moving noise operating at $w = 100$ will have the same behavior as the anchor noise operating at $w = 1$ and the degeneration issue can be resolved.

The NR-SDS Algorithm

To summarize the NR-SDS algorithm, our algorithm applies the following SDS loss and NR loss (Eq. 8):

$$\begin{aligned} \nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x} = g(\theta)) \triangleq \mathbb{E}_{t,\epsilon}[(\hat{\epsilon}_\phi(\mathbf{z}_t; y, t, w = 100, \epsilon_m[t]) \\ - \epsilon_a)\frac{\partial \mathbf{x}}{\partial \theta}], \end{aligned} \tag{9}$$

Concretely, Eq. 9 applies the SDS loss to update the NeRF parameters. Different from the original SDS loss, instead of randomly sampling the added noise from Gaussian space, we fix the noise to the learnable moving noise. Eq. 8 is the Noise Recalibration loss. We further summarize the NR-SDS algorithm in Algorithm 1. An illustration of the for-

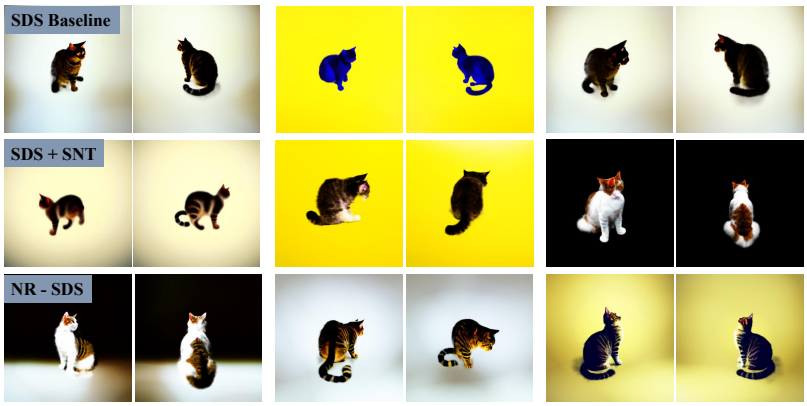

Figure 6: Ablation Experiments on Diversity. We run experiments with SDS baseline, SDS + Single Noise Training, and our final NR-SDS algorithm. For SDS baseline and SDS + SNT, we manually select the best time step before degeneration. The baseline method only generates cats with very similar gestures and textures. Our methods can better generate diversified content.

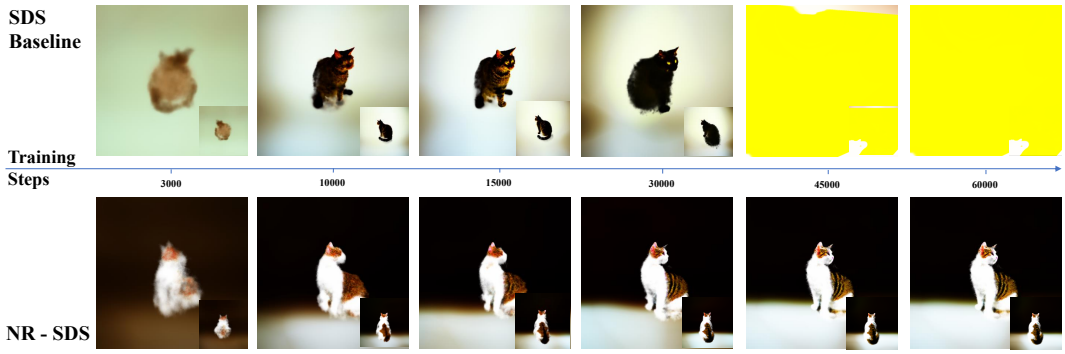

Figure 7: Ablation Experiments on Degeneration Issue. We observe that our method effectively resolves the degeneration issue. Our NeRF model learns object details when trained for long epochs, while the baseline method degenerates.

ward propagation and backward propagation can be found in Fig. 4.

In a perfectly trained case, the Noise-Recalibration SDS algorithm ensures the NeRF model converges to a stable state. That is, given a real image, both Eq. 9 and Eq. 8 will generate values close to zero.

Experiments

Implementation Details

The experiments are carried out on NVIDIA-A6000 GPUs with 48GB memory. We use the code base from (Tang 2022) and (Guo et al. 2023) for the implementation of SDS loss and VSD loss. We use Instant-NGP (Müller et al. 2022) as the backbone NeRF and stable diffusion (Rombach et al. 2022) as the 2D diffusion model.

The NeRF model is trained for 25000 iterations with 2048 sampled rays and a batch size of 1. We use a learning rate of 1x10-3 without learning rate decay. To ensure an equitable comparison and mitigate the potential influence of extraneous factors on the generated outcomes, we deliberately refrain from utilizing mesh finetuning to bolster the quality of our generation results. It's worth highlighting that our proposed approach seamlessly integrates with DMTet finetuning, as introduced in Magic3D (Lin et al. 2023) and Fantasia3D (Chen et al. 2023).

Qualitative Experiments

Comparison in 2D space. We first show the generation results in 2D space in Fig. 5 as a proof of concept. In this experiment, we start from a randomly initialized noise latent and optimize the latent with various optimization methods. After optimization, we use the decoder of latent diffusion model to convert the latents to images. Compared with SDS loss, our proposed method can generate much better results with stunning details. Compared with concurrently proposed VSD loss (Wang et al. 2023), our method generates comparable high-quality images. In the meantime, our method only optimizes the noise space, while VSD loss requires to finetune a Lora model. Therefore, our method only requires around 60% as much running memory as VSD loss.

3D Ablations on the identified issues. We demonstrate the ablation experiments on the identified issues in Fig. 6 and Fig. 7. For the diversity issue, we run the experiments with baseline SDS loss, SDS + single noise training, and our

Figure 8: Our Results vs. Other Methods. We compare our method against Dreamfusion and concurrently proposed VSD loss. We observe that our proposed method can generate high-fidelity 3D results.

	NR-SDS vs. SDS (Quality)	NR-SDS vs. SDS (Diversity)	NR-SDS vs. Dreamfusion	NR-SDS vs. VSD
Preference Score	80%	75%	67%	50%

Table 1: User Study Results

proposed NR-SDS method using the same text prompt "a cat". We observe that the single noise training method can effectively improve the diversity of generations. In Fig. 7, we observe that our method resolves the degeneration issue.

3D comparison with other methods. We also directly compare our method with Dreamfusion (Poole et al. 2022) by using the released images and VSD loss by running the method with the same setting. Results are shown in Fig. 8. We observe that our method can generate high-resolution results with better texture details compared with Dreamfusion and comparable results compared with VSD loss. More results of generation quality and diversity can be found in **Supplementary Materials**.

Quantitative Results

We conduct user studies to quantitatively evaluate our method. Participants are given NeRF rendered videos or multi-view images to evaluate in four different experiments. Some generation videos can be found in **Supplementary Materials**. First, a diversity comparison is conducted between the NR-SDS and SDS baseline. We generate 300 NeRFs using 100 text prompts, with each prompt training 3 NeRFs with different seeds. Participants are asked to select which set of three is more diverse. Second, we conduct quality experiments, with 100 NeRFs trained using the NR-SDS and SDS baseline (both use latent diffusion). Participants are asked to select which one is of higher quality. Thirdly, we compare our results to those Dreamfusion released generation examples and training NeRFs using the same prompts. The Dreamfusion released examples are generated with Imagen (Saharia et al. 2022) and the weights are not publicly available. Finally, we also compare ours against VSD loss. The user study preference scores are listed in Table. 1. The results of our user study indicate that our method attains higher or comparable user preference scores in comparison to both baseline and contemporary methods.

NR-SDS with Even Larger Guidance Weight

In our experiments, we observe that our method remains effective with guidance weights exceeding 100, consistently producing accurate colors. We include these results in **Supplementary Materials**.

Conclusions, Limitation and Future Works

In this work, we identify and study the two commonly seen problems of SDS loss, the diversity issue and the degeneration issue. We propose NR-SDS algorithm to tackle the two problems. With NR-SDS loss, we could greatly improve the generation diversity and quality.

Future work can be done by improving the shapes of generated objects. One limitation of current method is that 2D diffusion guided 3D generation often falls short in learning object shapes. For example, they will suffer from the multi-head Janus problem. One potential solution could be to add additional 3D priors to the generation process.

Acknowledgements

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021). This research is also supported by an OPPO research grant.

References

Canfes, Z.; Atasoy, M. F.; Dirik, A.; and Yanardag, P. 2023. Text and image guided 3d avatar generation and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4421–4431.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16123–16133.

Chen, R.; Chen, Y.; Jiao, N.; and Jia, K. 2023. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. *arXiv preprint arXiv:2303.13873*.

Guo, Y.-C.; Liu, Y.-T.; Shao, R.; Laforte, C.; Voleti, V.; Luo, G.; Chen, C.-H.; Zou, Z.-X.; Wang, C.; Cao, Y.-P.; and Zhang, S.-H. 2023. threestudio: A unified framework for 3D content generation. https://github.com/threestudio-project/threestudio. Accessed: 2023-05-01.

Henzler, P.; Mitra, N. J.; and Ritschel, T. 2019. Escaping plato's cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9984–9993.

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

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Jain, A.; Mildenhall, B.; Barron, J. T.; Abbeel, P.; and Poole, B. 2022. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 867–876.

Liao, Y.; Schwarz, K.; Mescheder, L.; and Geiger, A. 2020. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5871–5880.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 300–309.

Liu, Z.; Wang, Y.; Qi, X.; and Fu, C.-W. 2022. Towards implicit text-guided 3d shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17896–17906.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Mordvintsev, A.; Olah, C.; and Tyka, M. 2015. Inceptionism: Going deeper into neural networks. *Blog*.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4): 1–15.

Nguyen-Phuoc, T.; Li, C.; Theis, L.; Richardt, C.; and Yang, Y.-L. 2019. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7588–7597.

Nguyen-Phuoc, T. H.; Richardt, C.; Mai, L.; Yang, Y.; and Mitra, N. 2020. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in neural information processing systems*, 33: 6767–6778.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, 16784–16804. PMLR.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Gontijo-Lopes, R.; Ayan, B. K.; Salimans, T.; et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*.

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. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Tang, J. 2022. Stable-dreamfusion: Text-to-3D with Stable-diffusion. Https://github.com/ashawkey/stable-dreamfusion.

Wang, C.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2022. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3835–3844.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *arXiv preprint arXiv:2305.16213*.

Wu, J.; Zhang, C.; Xue, T.; Freeman, B.; and Tenenbaum, J. 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29.

Yin, H.; Mallya, A.; Vahdat, A.; Alvarez, J. M.; Kautz, J.; and Molchanov, P. 2021. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16337–16346.

Yin, H.; Molchanov, P.; Alvarez, J. M.; Li, Z.; Mallya, A.; Hoiem, D.; Jha, N. K.; and Kautz, J. 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8715–8724.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4578–4587.

Zhou, P.; Xie, L.; Ni, B.; and Tian, Q. 2021. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*.

Zhu, J.-Y.; Zhang, Z.; Zhang, C.; Wu, J.; Torralba, A.; Tenenbaum, J.; and Freeman, B. 2018. Visual object networks: Image generation with disentangled 3D representations. *Advances in neural information processing systems*, 31.