Optimized View and Geometry Distillation from Multi-view Diffuser

Youjia Zhang[1] , **Zikai Song**[1] , **Junqing Yu**[1] , **Yawei Luo**[2] and **Wei Yang**[1†]
[1]Huazhong University of Science and Technology
[2]Zhejiang University
{youjiazhang, weiyangcs}@hust.edu.cn

Abstract

Generating multi-view images from a single input view using image-conditioned diffusion models is a recent advancement and has shown considerable potential. However, issues such as the lack of consistency in synthesized views and over-smoothing in extracted geometry persist. Previous methods integrate multi-view consistency modules or impose additional supervisory to enhance view consistency while compromising on the flexibility of camera positioning and limiting the versatility of view synthesis. In this study, we consider the radiance field optimized during geometry extraction as a more rigid consistency prior, compared to volume and ray aggregation used in previous works. We further identify and rectify a critical bias in the traditional radiance field optimization process through score distillation from a multi-view diffuser. We introduce an **Unbiased Score Distillation (USD)** that utilizes unconditioned noises from a 2D diffusion model, greatly refining the radiance field fidelity. We leverage the rendered views from the optimized radiance field as the basis and develop a two-step specialization process of a 2D diffusion model, which is adept at conducting object-specific denoising and generating high-quality multi-view images. Finally, we recover faithful geometry and texture directly from the refined multi-view images. Empirical evaluations demonstrate that our optimized geometry and view distillation technique generates comparable results to the state-of-the-art models trained on extensive datasets, all while maintaining freedom in camera positioning. Source code of our work is publicly available at: https://youjiazhang.github.io/USD/.

1 Introduction

Traditionally, the process of generating a three-dimensional model from a singular image necessitates extensive and meticulous efforts by highly skilled artists. However, recent advancements in neural networks, particularly through the adaptation of 2D diffusion models for 3D synthesis, have rendered the conversion of a single image into a 3D object feasible. The early breakthrough comes from the text to 3D domain, where DreamFusion [Poole *et al.*, 2023] and Score Jacobian Chaining (SJC) [Wang *et al.*, 2023a] proposes a Score Distilling Sampling (SDS) strategy to distill the scores learned by 2D diffusion models from large-scale images to optimize a Neural Radiance Field (NeRF) [Mildenhall *et al.*, 2020], circumventing the need for 3D data. Successive approaches further improve the quality and diversity of generated geometries from textural prompts [Wang *et al.*, 2023b; Lin *et al.*, 2023; Chen *et al.*, 2023]. Particularly, RealFusion [Melas-Kyriazi *et al.*, 2023] migrates the scheme to generate plausible 3D reconstruction matches to a single input image via textual inversion adapted supervision.

More relevantly, 3DiM [Watson *et al.*, 2023] and MVDream [Shi *et al.*, 2023] develop a pose-conditional image-to-image diffusion model, which generates the novel view at a target pose from a source view. Zero-1-to-3 [Liu *et al.*, 2023b] adopts a similar framework and learns control of viewpoints through a synthetic dataset and demonstrates zero-shot generalization to in-the-wild images. Though Zero-1-to-3 demonstrates plausible novel views, they are not multi-view consistent and the geometry distilled from SDS tends to be oversmoothed. To enhance the multi-view consistency, SyncDreamer [Liu *et al.*, 2024] devises a volume-encoded multi-view noise predictor to share information across different views. Wonder3D [Long *et al.*, 2024] predicts the multi-view color images along with their normal maps from a cross-domain diffusion model. Though enhancing the multi-view consistency of image generation, these methods compromise the flexibility of camera positioning and only allow synthesis for a limited number of views.

In this study, we observe that the predicted unconditional noise from the multi-view diffuser, *i.e.*, the Zero-1-to-3 model, appears to be biased. That is, even if we only add very low-level noise to a normal image and use the unconditional noise predicted by a Zero-1-to-3 model for denoising, the result still tends to deviate greatly from the original image. We analyze and rectify the critical bias by using an unconditioned noise from a pre-trained 2D diffusion model and greatly refining the geometry fidelity. Moreover, previous approaches use either 3D volume [Liu *et al.*, 2024] or ray aggregation [Tseng *et al.*, 2023] to share information

[†] Corresponding author.

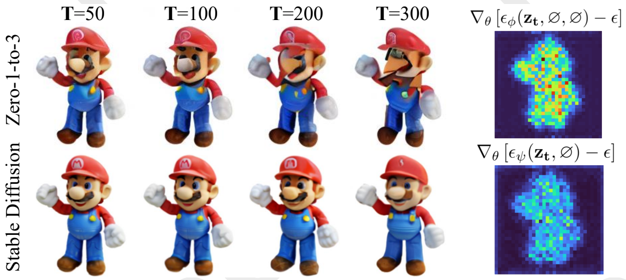

Figure 1: The unconditional noise predicted by Zero-1-to-3 model tends to be biased. As a demonstration, we use the '*Mario*' image as a toy example and add various levels of noise to the image (larger **T** means more noise has been added). We use the predicted unconditional noise to recover the original image from noisy input and find the results of Zero-1-to-3 deviate from the input image greatly even for very small amount of noise. The right sub-figure shows the averaged difference between the predicted noise and the added noise.

across views. We consider the radiance field as the consistency prior, and encourage the generated multi-view images to be consistent with the NeRF renderings. We develop a two-step specialization process of a 2D diffusion model, which is adept at conducting target-specific denoising and generating high-quality multi-view images from NeRF renderings. We then further use the refined views to generate the geometry and texture from NeuS [Wang *et al.*, 2021], and in the meanwhile enforce input view consistency using view score distillation. Our approach generates comparable-quality of multi-view images and geometry to the SOTA approaches, including SyncDreamer and Wonder3D, without enforcing any restriction on camera poses. Consequently, we posit that our approach offers superior adaptability and effectiveness in addressing the challenges associated with generating consistent and high-quality multi-view imagery and geometry, affirming its substantial potential for widespread application in relevant fields.

2 Related Work

2D Diffusion Models for 3D Generation. Recent advancements in 2D diffusion models [Rombach *et al.*, 2022; Croitoru *et al.*, 2023] and large-scale visual language models, notably the CLIP model [Radford *et al.*, 2021; Song *et al.*, 2024], have catalyzed new approaches for generating 3D assets. Pioneering efforts such as DreamFusion [Poole *et al.*, 2023] and SJC [Wang *et al.*, 2023a] have developed methods for transforming 2D text into images, subsequently facilitating the generation of 3D shapes from text. This approach has inspired a range of subsequent studies that adopt a shape-by-shape optimization scheme. Additionally, the integration of 2D diffusion models with robust vision language models, especially CLIP [Radford *et al.*, 2021], has emerged as a significant exploration in the generation of 3D assets [Xu *et al.*, 2023b]. The typical methodology involves optimizing a 3D representation, such as NeRF, mesh, or SDF, and then utilizing neural rendering to generate 2D images from various viewpoints. These images are processed through 2D diffusion models or the CLIP model to calculate SDS losses, which guide the optimization of the 3D shape. Building on the foundations laid by DreamFusion and SJC, numerous works have enhanced text-to-3D distillation methods in various aspects. Notably, Magic3D [Lin *et al.*, 2023] develops a two-stage coarse-to-fine optimization framework for high-resolution 3D content generation, and ProlificDreamer [Wang *et al.*, 2023b] proposes a Variational Score Distillation (VSD) for generating highly detailed geometry. However, challenges such as low efficiency and the multi-face Janus problem, where optimized geometry tends to produce multiple faces due to the lack of explicit 3D supervision, remain prevalent. Furthermore, some works [Radford *et al.*, 2021] have applied this distillation pipeline in single-view reconstruction tasks. While these methods have achieved impressive results, they often require extensive time for textual inversion and NeRF optimization, without always guaranteeing satisfactory outcomes. In contrast with the 2D diffusion to 3D extension, which is ignorant to multi-view consistency, our method focuses on the multi-view diffuser technique, which predicts the noises for novel views, and inherently avoids the Janus problem.

Multi-view Diffusion Models. In light of the complexities involved in ensuring the integrity of generated 3D content, recent efforts [Watson *et al.*, 2023; Gu *et al.*, 2023; Deng *et al.*, 2023; Tseng *et al.*, 2023; Chan *et al.*, 2023; Yu *et al.*, 2023; Tang *et al.*, 2023; Liu *et al.*, 2023c] have explored the feasibility of directly generating novel views from a single image input. Notably, the Scene Representation Transformer [Sajjadi *et al.*, 2022] extends the vision transformer to image sets, enabling global information integration for 3D reasoning. Similarly, 3DiM [Watson *et al.*, 2023] develops a pose-conditional image-to-image diffusion model, translating a single input view into consistent and sharp completions across multiple views. A seminal work in this area, Zero-1-to-3 [Liu *et al.*, 2023b], utilizes a similar network structure, trained on a large-scale synthetic 3D dataset, demonstrating notable generalizability. Recent advancements, such as One-2-3-45 [Xu *et al.*, 2023a], have leveraged the generalizable neural reconstruction method SparseNeuS [Long *et al.*, 2022] to directly produce 3D geometry from images generated by Zero-1-to-3. While this approach is more efficient and alleviates the Janus (multi-head) problem, it tends to produce lower-quality results with reduced geometric detail. In a different vein, SyncDreamer [Liu *et al.*, 2024] concentrates on object reconstruction, generating images in a single reverse process and utilizing attention to synchronize states among views. This contrasts with Viewset Diffusion [Szymanowicz *et al.*, 2023], which requires predicting a radiance field. SyncDreamer solely relies on attention for synchronization, fixing the viewpoints of generated views to enhance training convergence. A significant trend in recent research has been the design of various functional attention layers. Consistent123 [Weng *et al.*, 2023] employs a shared self-attention layer, where all views query the same key and value from the input view. ConsistNet [Yang *et al.*, 2024] introduces two sub-modules: a view aggregation module and a ray aggregation module, to extract features consistent across multiple views. MVDream [Shi *et al.*, 2023] utilizes 3D self-attention. Wonder3D [Long *et al.*,

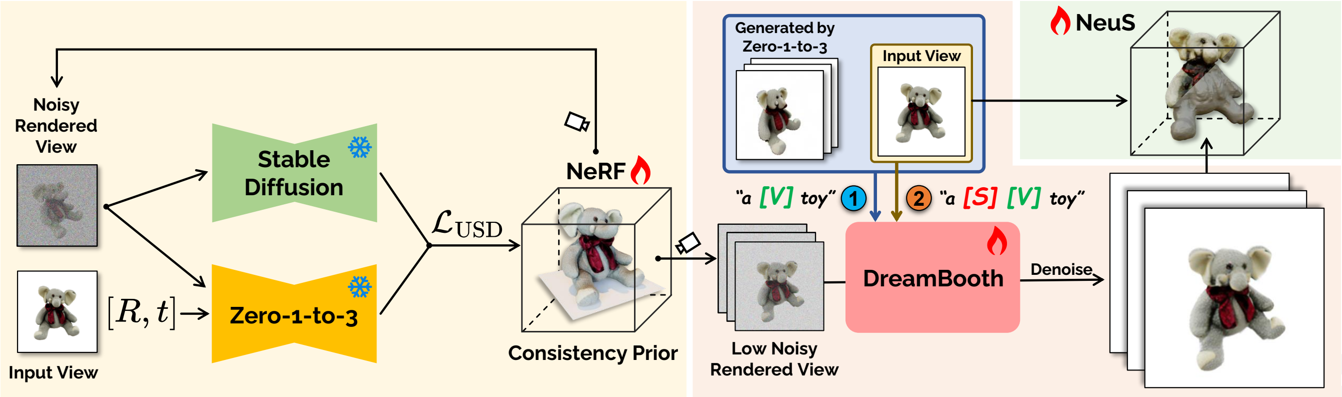

Figure 2: The overall pipeline of our approach. We first use our Unbiased Score Distillation to extract an optimized underlying radiance field. And then we use the NeRF as our consistency prior, *i.e.*, the generated views should be consistent with the NeRF renderings. We propose a two-stage specialization scheme to obtain a specified DreamBooth specifically for the target. We then denoise the NeRF renderings to obtain high-quality views and subsequently use NeuS technique to recover the geometry. Our optimized scheme generates comparable, sometimes better particularly for irregular camera poses, results to the SOTA works without training on large-scale data.

2024] goes a step further by not only outputting multi-view images but also outputting the normal map for each perspective, using cross-domain attention to maintain the consistency of the 3D structure. Both SyncDreamer and Wonder3D, resulting from fixed view outputs, exhibit sensitivity to the camera angle of the input image. In contrast, our method can reconstruct accurate 3D structures from inputs captured at varying camera angles. While existing approaches have focused on developing novel modules for enforcing multi-view consistency, our work demonstrates that an optimized distillation strategy can yield views and geometries comparable to models trained on large-scale datasets. This insight may inspire further exploration in improved strategies for geometry and view extraction.

3 Method

Our objective is to synthesize consistent multi-view images and high-quality geometric representations from a single input image. Notable prior work, such as the Zero-1-to-3, has demonstrated impressive results by utilizing an image and camera pose-conditioned diffusion model. However, this approach encounters limitations, particularly in terms of generating inconsistent multi-view images and a tendency for over-smoothing in the geometric output. The state-of-the-art research, including SyncDreamer [Liu *et al.*, 2024] and Wonder3D [Long *et al.*, 2024], addresses these challenges by incorporating additional modules for consistency or employing normal supervision. However, this often compromises the flexibility of positioning the target camera at will. Contrastingly, our work adopts a distinct methodology, showcasing that comparable quality in views and geometry can be achieved through a meticulously crafted distillation strategy. Central to our approach is the insight that the unconditional noise predictions from Zero-1-to-3 are inherently biased. We propose the utilization of unconditional noise from the Stable Diffusion [Takagi and Nishimoto, 2023] model to rectify this issue, as elaborated in Sec. 3.1. Our method, termed Unbiased Score Distillation (USD), significantly enhances the quality of the radiance field relative to previously used SDS/SJC methods. Furthermore, we employ the optimized NeRF [Mildenhall *et al.*, 2020] as a consistency prior, in contrast to previous implicit constraints such as 3D volume or ray aggregation. We ensure that the generated views and geometry align coherently with the distilled NeRF to achieve consistency. We posit that specializing a diffusion model to denoise the target object is crucial. To this end, we implement the DreamBooth [Ruiz *et al.*, 2023] technique and engage in a two-stage fine-tuning process, detailed in Sec. 3.2. The first stage involves using multi-view images from Zero-1-to-3 as positive samples, contrasting them with text-prompt generated images as negative samples for image style learning. In the subsequent stage, the input image serves as the positive sample, with all Zero-1-to-3 generated images as negatives, focusing on learning finer details. Subsequently, we introduce a low level of noise into the NeRF renderings and employ the fine-tuned diffusion model for denoising. The final step involves applying the NeuS technique for mesh reconstruction, while in the meanwhile using View Score Distillation to ensure input view consistency. Fig. 2 illustrates the complete pipeline of our methodology.

3.1 Unbiased Score Distillation

In this section, we highlight the significant bias in Zero-1-to-3's unconditional noise due to insufficient unconditional training and object-level dataset bias, affecting geometry quality in SDS-based distillation, both theoretically and empirically. We then propose a rectification method.

Bias in Unconditional Noise. As a multi-view diffuser, Zero-1-to-3 predicts noises of a target image latent $\mathbf{z}$ given two conditions: the input image c_I and *relative* camera pose c_P, *i.e.*, it learns the probability distribution $P(\mathbf{z}|c_I, c_P)$ for image latent $\mathbf{z}$. We apply *Bayes' Rule* to decompose the conditional probability:

$$P(\mathbf{z}|c_I, c_P) = \frac{P(\mathbf{z}, c_I, c_P)}{P(c_I, c_P)} = \frac{P(c_I|c_P, \mathbf{z})P(c_P|\mathbf{z})P(\mathbf{z})}{P(c_I, c_P)} \tag{1}$$

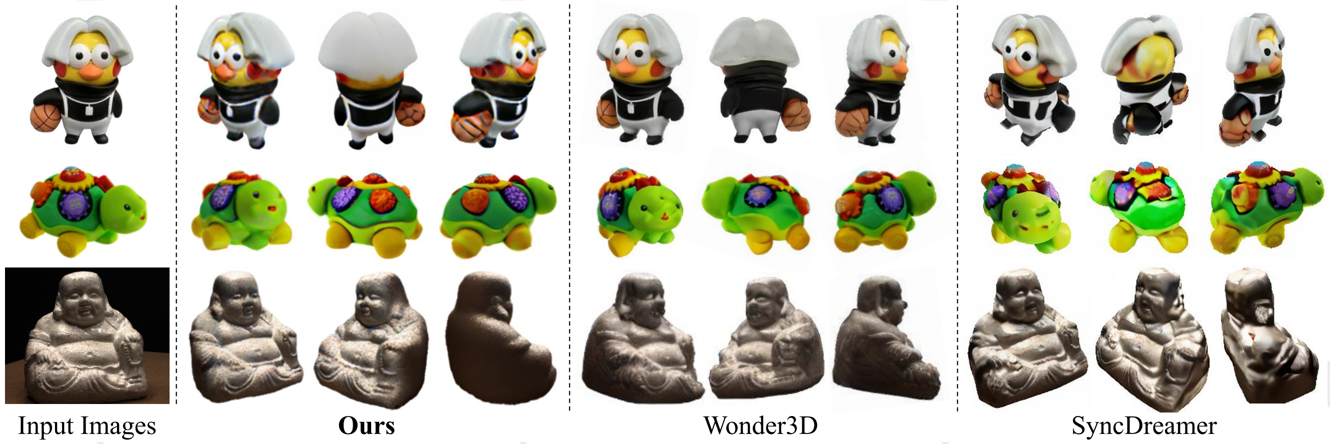

Figure 3: The qualitative comparisons with baseline models on multi-view color images. Our approach generates consistent multi-view images while preserving the image details.

Diffusion models estimate the score [Hyvärinen and Dayan, 2005] of the data distribution, *i.e.*, the derivative of the log probability, giving us the following expression:

$$\begin{aligned}\nabla_{\mathbf{z}}\log(P(\mathbf{z}|c_I,c_P)) = &\nabla_{\mathbf{z}}\log(P(c_I|c_P,\mathbf{z}))+\\ &\nabla_{\mathbf{z}}\log(P(c_P|\mathbf{z}))+\\ &\nabla_{\mathbf{z}}\log(P(\mathbf{z}))\end{aligned} \tag{2}$$

This leads to the score estimate with classifier-free guidance (CFG [Ho and Salimans, 2022]):

$$\begin{aligned}\epsilon_\phi^{\textbf{CFG}}(\mathbf{z_t},c_I,c_P) = &\alpha_1[\epsilon_\phi(\mathbf{z_t},c_I,c_P)-\epsilon_\phi(\mathbf{z_t},c_P,\varnothing)]+\\ &\alpha_2[\epsilon_\phi(\mathbf{z_t},c_P,\varnothing)-\epsilon_\phi(\mathbf{z_t},\varnothing,\varnothing)]+\\ &\epsilon_\phi(\mathbf{z_t},\varnothing,\varnothing)\end{aligned} \tag{3}$$

where ϵ_ϕ is a neural network for predicting noises given conditions, α_1 and α_2 are guidance scales that enable separately trading off the strength of conditions c_I and c_P separately. To ensure accurate prediction of $\epsilon_\phi(\mathbf{z_t},c_P,\varnothing)$ and $\epsilon_\phi(\mathbf{z_t},\varnothing,\varnothing)$, one needs to randomly drop c_I and c_P during training ϵ_ϕ. However, we observed that Zero-1-to-3 did not truly follow this procedure to supervise unconditional noise. As shown in Fig. 4, Zero-1-to-3 only randomly dropped image conditions c_I while keeping c_P untouched. During inference, it replaces the tensor after fully connected (FC) layer $f(\cdot)$ with zeros (not exactly equal to setting both c_I and c_P to zeros as $f(0)\neq 0$), leads to a bias. Moreover, the dataset used for fine-tuning is majorly object-centric, which may also introduce additional domain bias. Fully addressing this bias problem requires re-training the multi-view diffuser on a broader and more balanced dataset, which requires tremendous effort. Here we propose an effortless fix in the following.

Rectification. To alleviate the inaccurate unconditional noise problem, we first set $\alpha_1 = \alpha_2 = \omega$ to eliminate $\epsilon_\phi(\mathbf{z_t},c_P,\varnothing)$. We then and yield a special case of $\epsilon_\phi^{\textbf{CFG}}(\mathbf{z_t},c_I,c_P)$ as:

$$\begin{aligned}\epsilon_\phi^{\textbf{CFG}}(\mathbf{z_t},c_I,c_P) = \omega[\epsilon_\phi(\mathbf{z_t},c_I,c_P)-\epsilon_\phi(\mathbf{z_t},\varnothing,\varnothing)]+\\ \epsilon_\phi(\mathbf{z_t},\varnothing,\varnothing)\end{aligned} \tag{4}$$

To demonstrate our setting of α_1 and α_2 is valid, we provide an empirical validation can be found in the Appendix E. Further, we consider $\epsilon_\phi(\mathbf{z_t},\varnothing,\varnothing)$ predicts noises from only the noisy latent and is equivalent to the unconditional noise $\epsilon_\psi(\mathbf{z_t},\varnothing)$ of Stable Diffusion (SD) as it uses the same variational autoencoders (VAE [Kingma and Welling, 2013]) as Zero-1-to-3. As such, the bias has been well rectified. To verify the effect of our rectification, Tab. 1 shows the denoising effect using various unconditional noise settings: (1) use SD unconditional noise $\epsilon_\psi(\mathbf{z_t},\varnothing)$. (2) use Zero-1-to-3 noise $\epsilon_\phi(\mathbf{z_t},f(0,c_P))$ with $f(0,c_P)$ as a condition (same to the training process as in Fig 4). (3) use Zero-1-to-3 noise $\epsilon_\phi(\mathbf{z_t},0)$ with 0 as a condition (referring to the inference process, as shown in Fig 4). The SD noise generates the best result, and setting (2) is slightly better than (3) as it follows the training set up. More details can be found in the Appendix A.

Unbiased Score Distillation. One major application of a multi-view diffuser is to distill 3D content, represented as NeRF paramater θ, using the SDS loss for optimization:

$$\nabla_\theta\mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,c_P,\epsilon}\left[w(t)\left(\epsilon_\phi^{\textbf{CFG}}(\mathbf{z_t},c_I,c_P)-\epsilon\right)\frac{\partial\mathbf{z_t}}{\partial\theta}\right] \tag{5}$$

where $w(t)$ is a weighting function, ϵ is standard Gaussian noise, and $\mathbf{z_t}$ refers to the noisy latent as $\mathbf{z_t} = \sqrt{\alpha_t}\mathbf{z} + \sqrt{1-\alpha_t}\epsilon$, with α_t being the noise scheduler. We rewrite the noise difference in Formula 5 by adding an additional weighting factor $\boldsymbol{\lambda}$ to $[\epsilon_\psi(\mathbf{z_t},\varnothing)-\epsilon]$:

$$\begin{aligned}\epsilon_\phi^{\textbf{CFG}}(\mathbf{z_t},c_I,c_P)-\epsilon = \omega\left[\epsilon_\phi(\mathbf{z_t},c_I,c_P)-\epsilon_\psi(\mathbf{z_t},\varnothing)\right]+\\ \boldsymbol{\lambda}\left[\epsilon_\psi(\mathbf{z_t},\varnothing)-\epsilon\right]\end{aligned} \tag{6}$$

where setting $\boldsymbol{\lambda} = 1$ we get Formula 5. Inspired by DDS [Hertz *et al.*, 2023] and CSD [Yu *et al.*, 2024], we observed that setting $\boldsymbol{\lambda} = 0$ can significantly improve the details of the 3D details generated using SDS. Further details and an in-depth analysis are provided in the Appendix D. We obtain

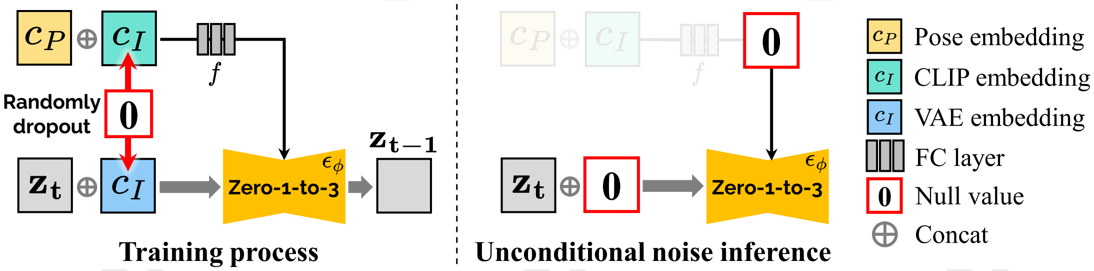

Figure 4: Training and inference process of Zero-1-to-3. Training for predicting unconditional noise involves setting the c_I conditions to 0 at regular intervals.

Noise level	T=50			T=100			T=200			T=300		
Metrics	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
$\epsilon_\psi(\mathbf{z_t}, \varnothing)$	40.26	0.992	0.008	36.94	0.988	0.014	33.99	0.981	0.023	31.98	0.974	0.027
$\epsilon_\phi(\mathbf{z_t}, f(0, c_P))$	36.17	0.981	0.015	34.85	0.979	0.024	33.07	0.975	0.031	30.44	0.969	0.046
$\epsilon_\phi(\mathbf{z_t}, 0)$	36.13	0.979	0.015	34.70	0.970	0.026	32.76	0.973	0.033	29.94	0.963	0.055

Table 1: Quantitative evaluation of different unconditional noises. We randomly samples 1,000 different images from the GSO dataset, RTMV, and Objaverse, add a certain level of noise w.r.t. **T**, where larger **T** means more noise is added, and compare the denoising effect under different conditions of Stable Diffusion and Zero-1-to-3. The top three for each metric are highlighted in red , orange, and yellow respectively.

our Unbiased Score Distillation (USD) as:

$$\nabla_\theta \mathcal{L}_{\text{USD}} = \mathbb{E}_{t,c_P,\epsilon}\left[w(t)\left[\omega\left(\epsilon_\phi\left(\mathbf{z_t}, c_I, c_P\right) - \epsilon_\psi\left(\mathbf{z_t}, \varnothing\right)\right)\right]\frac{\partial \mathbf{z_t}}{\partial \theta}\right] \tag{7}$$

Our USD generates much better and consistent 3D than the SDS/SJC method used in Zero-1-to-3, the results can be found in the Appendix E.

3.2 Consistent View and Geometry Distillation

Although USD can be used to optimize NeRF for view synthesis, the resulting images still tend to be blurry, and directly extracting geometry from the NeRF density field introduces noise into the mesh. Since our ultimate goal is to extract high-quality geometry and consistent multi-views, we propose to utilize the generated NeRF as the view consistency prior, *i.e.*, the final high-quality view images should be consistent with the NeRF rendering. Thus, our problem transformed into a denoising problem.

Two-Stage Specified Diffusion. We leverage the recent advance in diffusion model specialization, *i.e.*, the DreamBooth [Ruiz *et al.*, 2023], to fine-tune a 2D diffusion model for the specific target in the input image. We observe that the novel view images generated by Zero-1-to-3, though not multi-view consistent, tend to have the same style as the input image. We design a two-stage tuning method to gradually let the diffusion model learn the object details. In the first stage, we use the multi-view images generated by Zero-1-to-3 as positive samples, contrasting them with text-prompt generated images as negative samples for learning the visual style of the target. In the subsequent stage, the input image serves as the positive sample, with all Zero-1-to-3 generated images as negatives, focusing on learning finer details. During optimization, we use a unique identifier [V] to capture the visual style of the target. In the second stage, we set only the input image as the positive sample and the images generated by Zero-1-to-3 as negative ones, and use the additional identifier [S] to capture the identity of the specific target.

Geometry and Texture Distillation with Input View Supervision. With the specialization diffusion model, we add a small noise, Stable Diffusion scheduler $t = 200$, to the NeRF render images and conduct the denoising process. Then, we use the NeuS technique to reconstruct the geometry from the high-quality and clear images (*i.e.*, 100 input images).

We observe that the input view is rarely used for optimization of the geometry and texture in multi-view diffusers, despite the fact that the input image is most faithful to the target object. To exploit the input view information, we further develop a reference view score distillation during the NeuS reconstruction process. Specifically, we consider the image render from NeuS at the input image viewpoint as $\mathbf{z_t}^* = \mathcal{R}(\Theta, p^*)$, where $\mathcal{R}$ is the rendering function from NeuS model defined by Θ, p^* is the camera pose of the input image, can be set to a particular relative pose. We define the following reference view distillation loss as:

$$\mathcal{L}_{RV} = \mathbb{E}_{t,\epsilon}\left[w(t)||\epsilon_\psi\left(\mathbf{z_t}^*, \varnothing\right) - \epsilon_\psi\left(\mathbf{y_t}, \varnothing\right)||_2^2\right] \tag{8}$$

where $\mathbf{y_t}$ is input view image. We add this reference view distillation loss $\mathcal{L}_{RV}$ to original photometric loss with optimized images for optimizing the NeuS parameter Θ. This scheme is a better supervision strategy than directly applying MSE loss on the input view. The reason for this is that, compared to computing MSE loss directly at the pixel-level, our patch-aware noise(latent)-level approach places greater emphasis on the perceptual quality of the image.

4 Experiments

We conduct extensive experiments, both qualitatively and quantitatively, to demonstrate the effectiveness of our method.

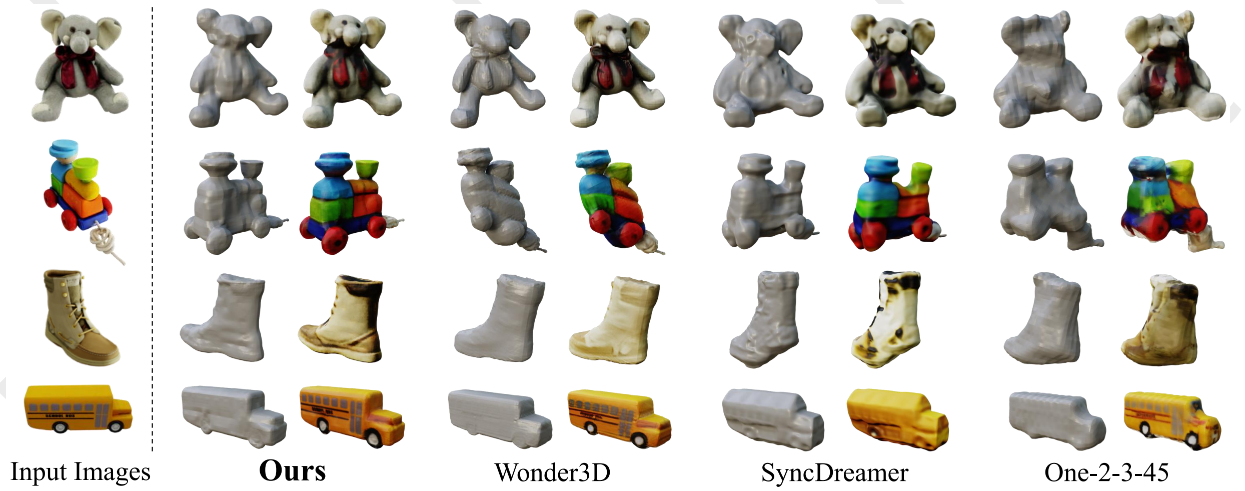

Figure 5: Qualitative comparisons of our method with baseline approaches, namely Wonder3D, SyncDreamer, and One-2-3-45, on the GSO dataset, focusing on the quality of the reconstructed textured meshes.

4.1 Implementation Details

Optimizing the Pipeline. We use exactly the same set of hyperparameters for all experiments and do not perform any per-object hyperparameter optimization. We use the implicit-volume implementation in threestudio [Guo *et al.*, 2023] as our 3D representation (NeRF), which includes a multi-resolution hash-grid and a MLP to predict density and RGB. The NeRF is optimized for 10,000 steps with an Adam optimizer at a learning rate of 0.01, weight decay of 0.05, and betas of (0.9, 0.95). For USD, the maximum and minimum time steps are decreased from 0.98 to 0.5 and 0.02, respectively, over the first 5,000 steps. We adopt the Stable Diffusion [Takagi and Nishimoto, 2023] model of V2.1. The classifier-free guidance (CFG) scale of the USD is set to 7.5 following [Wang *et al.*, 2023b]. The DreamBooth backbone is implemented using Stable Diffusion V2.1. In the first stage, we use Stable Diffusion to generate 200 images as negative samples. Additionally, we utilize 6 positive sample images with 360° surrounding camera poses (at 60° intervals) for training. The USD (NeRF) process takes about 1.5 hours on a NVIDIA Tesla V100 (32GB) GPU. To achieve reduced running time, we provide additional discussions and experimental results in the Appendix C. For DreamBooth fine-tuning, we train the model around for 600 steps with a learning-rate as 2e-6, weight decay as 0.01 and a batch size of 2. More details on the experimental setup are provided in the Appendix B.

Camera Setting. Since the reference image is unposed, we assume its camera parameters [Liu, 2023] are as follows. We set the field of view (FOV) of the camera is 40°, and the radial distance is 1.5 meters. Note this camera setting works for images subject to the front-view assumption. For images taken deviating from the front view, a manual change of polar angle or a camera estimation is required.

4.2 Evaluation Protocol

Evaluation Datasets. Following prior research [Liu *et al.*, 2023b; Liu *et al.*, 2024; Long *et al.*, 2024], we adopt the Google Scanned Object dataset [Downs *et al.*, 2022] for our evaluation, which includes a wide variety of common everyday objects. Our evaluation dataset matches that of SyncDreamer [Liu *et al.*, 2024], consisting of 30 objects that span from everyday items to animals. For each object in the evaluation set, we render an image with a size of 256 × 256 and use it as the input. Additionally, to assess the generalization ability of our model, we include images with diverse styles collected from the website in Zero-1-to-3, SyncDreamer and Wonder3D.

Method	Chamfer Dist.↓	Volume IoU↑
Realfusion [Melas-Kyriazi *et al.*, 2023]	0.0819	0.2741
Magic123 [Qian *et al.*, 2024]	0.0516	0.4528
One-2-3-45 [Liu *et al.*, 2023a]	0.0629	0.4086
Point-E [Nichol *et al.*, 2022]	0.0426	0.2875
Shap-E [Jun and Nichol, 2023]	0.0436	0.3584
Zero-1-to-3 [Liu *et al.*, 2023b]	0.0339	0.5035
SyncDreamer [Liu *et al.*, 2024] (NeuS)	0.0261	0.5421
Wonder3D [Long *et al.*, 2024] (iNGP+NeuS)	0.0199	0.6244
Ours (NeuS)	0.0240	0.5688
Ours (iNGP+NeuS)	0.0177	0.6330

Table 2: Quantitative comparison with baseline methods. We report Chamfer Distance and Volume IoU on the GSO dataset. The original implementation of SyncDreamer uses vanilla NeuS for extracting 3D meshes, while Wonder3D uses Instant NGP (iNGP) adapted NeuS. We report results using both techniques for better demonstration.

Method	PSNR↑	SSIM↑	LPIPS↓
Realfusion [Melas-Kyriazi *et al.*, 2023]	15.26	0.722	0.283
Zero-1-to-3 [Liu *et al.*, 2023b]	18.93	0.779	0.166
SyncDreamer [Liu *et al.*, 2024]	20.05	0.798	0.146
Wonder3D [Long *et al.*, 2024]	26.07	0.924	0.065
Ours	25.38	0.927	0.049

Table 3: The quantitative comparison in novel view synthesis. We report PSNR, SSIM, LPIPS on the GSO dataset.

Metrics. To evaluate the quality of single view reconstruc-

Method	Chamfer Dist.↓	Volume IoU↑
w/o USD	0.0253	0.5515
w/o DB[1st+2nd]	0.0217	0.6023
w/o DB[2nd]	0.0190	0.6216
w/o $\mathcal{L}_{RV}$	0.0185	0.6229
Ours	**0.0177**	**0.6330**

Table 4: Quantitative results of ablation studies. We report Chamfer Distance and Volume IoU on the GSO dataset.

tion, we used two commonly used metrics: the chamfer distance (CD) between the ground truth shape and the reconstructed shape, and the volume IoU. Due to different methods using different normative systems, before calculating these two metrics, we first align the generated shapes with the basic fact shapes. Moreover, we adopt the metrics PSNR, SSIM and LPIPS for evaluating the generated color images.

4.3 Ablation Study

We validate our design choices by ablating 4 major model variants, that are without Unbiased Score Distillation, without reference view distillation loss, without the two-stage DreamBooth (DB). As shown in Fig. 6, the 3D models generated without USD exhibit biased texture colors, and their shapes are not smooth. Not using reference view supervision will result in the inability to recover the same texture details as the input image, and not using DB will result in blurring of texture details. We also conducted a quantitative analysis on the GSO dataset, presented in Tab. 4. The results demonstrate that USD is crucial for geometric accuracy, and all our submodules collectively enhance overall performance.

4.4 Different Viewing Angle Comparisons

We found that SyncDreamer [Liu *et al.*, 2024] and Wonder3D [Long *et al.*, 2024] are very sensitive to viewing angles. If a relatively high viewing angle is input, SyncDreamer and Wonder3D will predict incorrect multi-view images. The results can be found in the Appendix E.

4.5 Single View Reconstruction

We evaluate the quality of the reconstructed geometry of different methods. The quantitative results are summarized in Tab. 2, and the qualitative comparisons are presented in Fig. 5. The quality of Wonder3D shape reconstruction depends on the perspective of the input view, such as '*Train*' shown in Fig. 5, where Wonder3D generated incorrect prediction results. The shape generated by SyncDreamer undergoes deformation due to the camera pose in the input view. One-2-3-45 [Liu *et al.*, 2023a] attempts to reconstruct meshes from the multiview-inconsistent outputs of Zero-1-to-3 [Liu *et al.*, 2023b]. While it can capture coarse geometries, it loses important details in the process. In contrast, our method can achieve good reconstruction quality and texture in terms of geometric structure and texture. In the paper, all the surface extraction demonstrated by our method is built on the Instant NGP (iNGP) [Müller *et al.*, 2022] based SDF reconstruction method [Guo, 2022], and we use Blender Cycles [Community, 2018] to render the results.

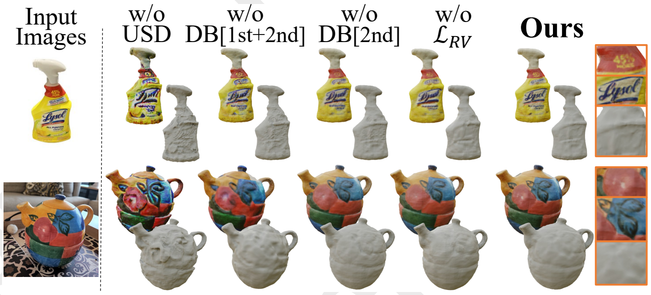

Figure 6: Ablation study on the effect of each components, including the USD (Sec. 3.1), the reference view score distillation (Sec. 3.2), the DreamBooth (Sec. 3.2) refinement of our method.

4.6 Novel View Synthesis

We evaluate the quality of novel view synthesis for different methods. The quantitative results are presented in Tab. 3, and the qualitative results can be found in Fig. 3. Zero-1-to-3 produces visually reasonable images, but they lack multi-view consistency since it operates on each view independently (we don't show the results of Zero-1-to-3). Although SyncDreamer introduces a volume attention scheme to enhance the consistency of multi-view images, their model is sensitive to the elevation degrees of the input images and tends to produce unreasonable results. Wonder3D ensures 3D consistency by generating normal maps, but may result in incorrect results for some input viewpoints.

4.7 Text-to-Image-to-3D

As a case study, we combine text-to-image models, *i.e.*, the Stable Diffusion or Imagen [Saharia *et al.*, 2022] to generate 3D models from text. We show some examples in the Appendix E. Compared to DreamFusion [Poole *et al.*, 2023], ProlificDreamer [Wang *et al.*, 2023b] and MVDream [Shi *et al.*, 2023], our method shows no multi-face Janus problem and conforms to the text faithfully.

5 Conclusions

In this work, we introduce an optimized approach for distilling geometry and views from a multi-view diffuser, with a specific focus on the Zero-1-to-3 model. We observed that the direct application of the SDS/SJC technique to Zero-1-to-3 is often suboptimal, primarily due to bias issues inherent in unconditional noise. To address this, we propose an Unbiased Score Distillation (USD) strategy by leveraging unconditioned noises from a 2D diffusion model to effectively enhance the optimized radiance field. Moreover, we developed a two-stage DreamBooth refinement process to improve the rendering of views. This process ensures consistency across multiple perspectives while simultaneously enhancing image quality. While we have identified and addressed the bias issue in the Zero-1-to-3 model, the underlying causes remain to be fully understood. Future research will delve into the theoretical aspects of this bias problem. Additionally, we aim to explore the potential applications of USD in other fields, such as image translation and view synthesis.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC No. 62272184 and No. 62402189), the China Postdoctoral Science Foundation under Grant Number GZC20230894, the China Postdoctoral Science Foundation (Certificate Number: 2024M751012). The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

References

[Chan *et al.*, 2023] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *ICCV*, 2023.

[Chen *et al.*, 2023] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, 2023.

[Community, 2018] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[Croitoru *et al.*, 2023] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *T-PAMI*, 2023.

[Deng *et al.*, 2023] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *CVPR*, 2023.

[Downs *et al.*, 2022] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *ICRA*, 2022.

[Gu *et al.*, 2023] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *ICML*, 2023.

[Guo *et al.*, 2023] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. Accessed: 2024-10-18.

[Guo, 2022] Yuan-Chen Guo. Instant neural surface reconstruction. https://github.com/bennyguo/instant-nsr-pl, 2022. Accessed: 2024-11-03.

[Hertz *et al.*, 2023] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2328–2337, 2023.

[Ho and Salimans, 2022] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[Hyvärinen and Dayan, 2005] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[Jun and Nichol, 2023] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Lin *et al.*, 2023] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023.

[Liu *et al.*, 2023a] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *NeurIPS*, 2023.

[Liu *et al.*, 2023b] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023.

[Liu *et al.*, 2023c] Xinhang Liu, Shiu-hong Kao, Jiaben Chen, Yu-Wing Tai, and Chi-Keung Tang. Deceptive-nerf: Enhancing nerf reconstruction using pseudo-observations from diffusion models. *arXiv preprint arXiv:2305.15171*, 2023.

[Liu *et al.*, 2024] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *ICLR*, 2024.

[Liu, 2023] Minghua Liu. One-2-3-45. https://huggingface.co/spaces/One-2-3-45/One-2-3-45, 2023. Accessed: 2024-10-21.

[Long *et al.*, 2022] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *ECCV*, pages 210–227. Springer, 2022.

[Long *et al.*, 2024] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *CVPR*, 2024.

[Melas-Kyriazi *et al.*, 2023] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *CVPR*, 2023.

[Mildenhall *et al.*, 2020] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[Müller *et al.*, 2022] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.

[Nichol *et al.*, 2022] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

[Poole *et al.*, 2023] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023.

[Qian *et al.*, 2024] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *ICLR*, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[Ruiz *et al.*, 2023] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023.

[Saharia *et al.*, 2022] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022.

[Sajjadi *et al.*, 2022] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *CVPR*, pages 6229–6238, 2022.

[Shi *et al.*, 2023] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *ICLR*, 2023.

[Song *et al.*, 2024] Zikai Song, Ying Tang, Run Luo, Lintao Ma, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Autogenic language embedding for coherent point tracking. In *ACM International Conference on Multimedia*, 2024.

[Szymanowicz *et al.*, 2023] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. In *ICCV*, pages 8863–8873, 2023.

[Takagi and Nishimoto, 2023] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, pages 14453–14463, 2023.

[Tang *et al.*, 2023] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *NeurIPS*, 2023.

[Tseng *et al.*, 2023] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, 2023.

[Wang *et al.*, 2021] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021.

[Wang *et al.*, 2023a] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023.

[Wang *et al.*, 2023b] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023.

[Watson *et al.*, 2023] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *ICLR*, 2023.

[Weng *et al.*, 2023] Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, CL Chen, and Lei Zhang. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092*, 2023.

[Xu *et al.*, 2023a] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to A 3d object with 360° views. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 4479–4489. IEEE, 2023.

[Xu *et al.*, 2023b] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *CVPR*, 2023.

[Yang *et al.*, 2024] Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. In *CVPR*, 2024.

[Yu *et al.*, 2023] Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *ICCV*, 2023.

[Yu *et al.*, 2024] Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3d with classifier score distillation. In *ICLR*, 2024.