RichDreamer: A Generalizable Normal-Depth Diffusion Model for Detail Richness in Text-to-3D

Lingteng Qiu[1,3][*][†]   Guanying Chen[2,1][*]   Xiaodong Gu[3][*]

Qi Zuo[3]   Mutian Xu[1]   Yushuang Wu[2,1]   Weihao Yuan[3]   Zilong Dong[3]

Liefeng Bo[3],   Xiaoguang Han[1,2][‡]

[1]SSE, CUHKSZ   [2]FNii, CUHKSZ   [3]Alibaba Group

Figure 1. **3D Generation Results and Applications of _RichDreamer_**. RichDreamer can generate highly-detailed and diverse 3D content from free-form user prompts. Our method achieves this by first generating the object geometry based on a generalizable Normal-Depth diffusion model, followed by modeling the physically-based rendering (PBR) materials. Notably, the diverse crocodile-theme objects at the bottom highlights the generalization ability of our method. The abbreviation of text prompts are shown beside the corresponding objects (full prompts can be found in the supplementary materials).

Abstract

Lifting 2D diffusion for 3D generation is a challenging problem due to the lack of geometric prior and the complex entanglement of materials and lighting in natural images. Existing methods have shown promise by first creating the geometry through score-distillation sampling (SDS) applied to rendered surface normals, followed by appearance modeling. However, relying on a 2D RGB diffusion model to optimize surface normals is suboptimal due to the distribution discrepancy between natural images and normals maps, leading to instability in optimization. In this paper, recognizing that the normal and depth information effectively describe scene geometry and be automatically estimated from images, we propose to learn a generalizable Normal-Depth diffusion model for 3D generation. We achieve this by training on the large-scale LAION dataset together with the generalizable image-to-depth and normal prior models. In an attempt to alleviate the mixed illumination effects in the generated materials, we introduce an albedo diffusion model to impose data-driven constraints on the albedo component. Our experiments show that when integrated into existing text-to-3D pipelines, our models significantly enhance the detail richness, achieving state-of-the-art results. Our project page is at https://aigc3d.github.io/richdreamer/.

1. Introduction

Image generation models have witnessed notable advancements in controllable image synthesis [56, 61]. This remarkable progress can be attributed to the scalability of generative models [21, 68] and utilization of the large-scale training datasets consisting of billions of image-caption pairs scrapped from the internet [63]. Conversely, due to the limited scale of the publicly available 3D datasets, existing 3D generative models are primarily evaluated for category-

---
[*] Equal contribution.
[†] Work done during internship at Alibaba.
[‡] Corresponding author: hanxiaoguang@cuhk.edu.cn.

specific generation and face challenges when attempting to generate novel, unseen categories [15, 18, 93]. It remains an open problem to create a comprehensive 3D dataset to facilitate generalizable 3D generation.

Recently developments in the field of text-to-3D, such as DreamFusion [52], have demonstrated impressive capabilities in zero-shot generation. This is achieved by optimizing a neural radiance field [46] through score distillation sampling (SDS) [52, 76] using a 2D diffusion model [62]. Subsequent to this, several methods have been proposed to improve the quality of the generated objects [7, 35, 45, 78]. However, the approach of lifting from 2D to 3D has presented two primary challenges. Firstly, 2D diffusion models tend to lack multi-view constraints, often leading to the emergence of multi-face objects, a phenomenon referred to as the "Janus problem" [52, 76]. To address this issue, recent advancements in multi-view-based diffusion models have shown success in mitigating these multi-face artifacts [67, 95].

Secondly, given the inherent coupling of surface geometry, texture, and lighting in natural images, the direct use of 2D diffusion models for the simultaneous inference of geometry and texture is considered suboptimal [7]. This suggests a two-stage decoupled approach: first, the generation of geometry, followed by the generation of texture. The recent Fantasia3D [7] method has shown promise in this decoupling strategy, yielding notably improved geometric reconstructions. However, Fantasia3D relies on 2D RGB diffusion models to optimize normal maps, leading to data distribution discrepancies that compromise the quality of geometric generation and introduce instability in optimization. This limitation underscores the pressing need for a robust prior model to provide an effective geometric foundation.

In this work, we aim to develop a robust 3D prior model to push forward the decoupled text-to-3D generation approach. Creating a model that offers 3D geometric priors typically requires access to 3D data for training supervision. However, amassing a large-scale dataset containing high-quality 3D models across diverse scenes is a challenging endeavor due to the time-consuming and costly process of 3D object scanning and model design [11, 59, 83, 91]. The limited scale of the publicly available 3D datasets presents a critical challenge: *how to learn a generalizable 3D prior model with limited 3D data*?

Recognizing that the normal and depth information can effectively describe scene geometry and can be automatically estimated from images [57], we propose to learn a generalizable *Normal-Depth diffusion model* for 3D generation (see Fig. 1). This is achieved by training on the large-scale LAION dataset [63] to learn diverse distributions of normal and depth of real-world scenes with the help of the generalizable image-to-depth and normal prior models [3, 58]. To improve the capability in modeling a

more accurate and sharp distribution of normal and depth, we fine-tune the proposed diffusion model on the synthetic Objaverse dataset [11]. Our results demonstrate that by pre-training on a large-scale real world dataset, the proposed Normal-Depth diffusion model can retain its generalization ability after fine-tuning on the synthetic dataset, indicating that our model learns a good distribution of diverse normal and depth in real-world scenes.

Given the inherent ambiguity in the decomposition of surface reflectance and illumination effects, textures generated by existing methods often retain shadows and specular highlights [60]. In an attempt to address this problem, we introduce an *albedo diffusion model* to provide data-driven constraints on the albedo component, enhancing the disentanglement of reflectance and illumination effects.

In summary, the key contributions of this paper are as follows:

- We propose a novel Normal-Depth diffusion model to provide strong geometric prior for high-fidelity text-to-3D geometry generation. By training on the extensive LAION dataset, our method exhibits remarkable generalization abilities.
- We introduce an albedo diffusion model that acts as a data-driven regularization for albedo, resulting in a more accurate separation of reflectance and illumination effects.
- Experiments demonstrate that integrating our models into existing text-to-3D pipelines yields state-of-the-art results in both geometry and appearance modeling.

2. Related Work

3D Generative Model The creation of high-quality 3D content is gaining increasing importance in various applications. Generative models directly learn the data distributions of the 3D data, enabling data sampling. Existing methods have yielded promising results by representing scenes using various 3D representations, including voxels [20, 82], point clouds [42, 51, 92], meshes [15, 39], and implicit fields [1, 6, 10, 13, 18, 28, 29, 48, 77, 90, 93]. However, these methods primarily demonstrate their generative capabilities within limited categories of objects due to the restricted scale of 3D training datasets. In contrast, our approach addresses the text-to-3D problem by extending 2D diffusion to the 3D domain, allowing for better generalization across diverse scenes specified by user prompts.

2D Diffusion for 3D Generation Recent research has demonstrated the generation of 3D objects from user prompts, leveraging pre-trained models like CLIP model [27, 47, 54, 85] or 2D diffusion models [61, 62]. Notably, DreamFusion [52] achieves zero-shot text-to-3D generation by optimizing a neural radiance field (NeRF) [46] through score distillation sampling (SDS) with a 2D diffu-
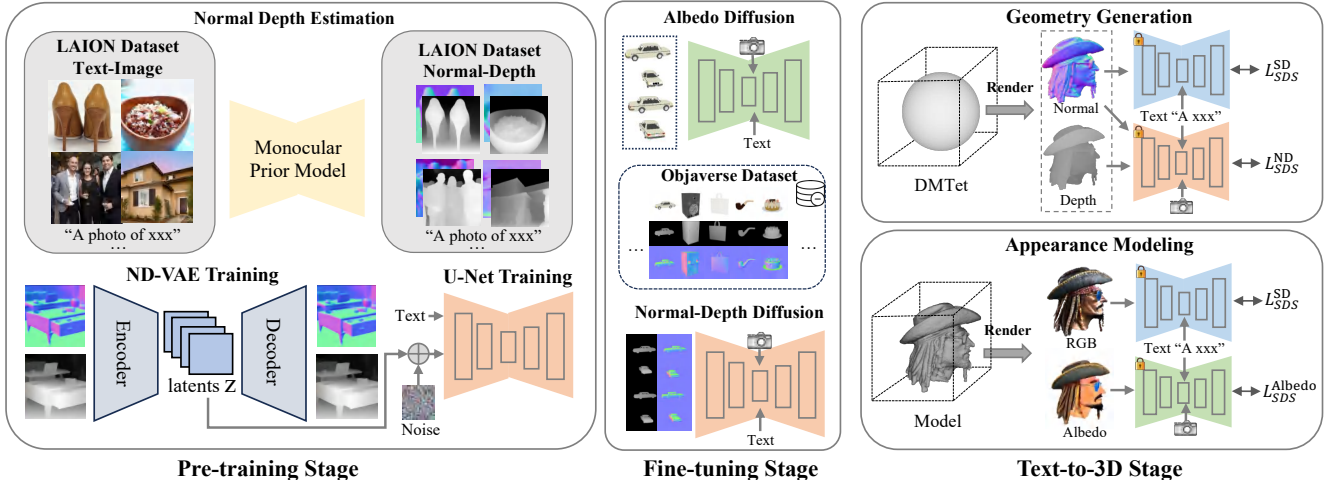
Figure 2. **Overview of the proposed *RichDreamer*.** We introduce a generalizable Normal-Depth diffusion model that is trained on the LAION-2B dataset with normal and depth predicted by Noraml-Bae [3] and Midas [58], followed by fine-tuning on the synthetic dataset. Our model can be incorporated with the DMTet and NeRF representations to enhance the geometry generation. To alleviate the ambiguity in appearance modeling, we propose an albedo diffusion model to impose data-drive prior on the albedo component.

sion model. Concurrently, SJC [76] employs score Jacobian chaining for the 3D generation.

Encouraged by these promising results, numerous works have focused on improving the quality of generated objects through approaches such as coarse-to-fine optimization [9, 35], decoupled generation [7], new score distillation [78], improved optimization strategies [2, 2, 8, 22, 25, 45, 64, 73, 81, 86], and incorporating parametric shape model [19, 26, 34, 94] etc. As generating a 3D model typically involves hours of optimization, some methods explore efficient 3D representations (*e.g.*, hashgrid [49] and 3D Gaussian splatting [30]) or improved training strategy to accelerate the optimization [17, 41, 71, 88]. Alongside the rapid development of the text-to-3D field, several methods have also adopted 2D diffusion models for the problem of 3D generation from image condition [1, 12, 16, 36, 44, 53, 55, 72, 84, 96, 97]. However, as 2D diffusion models tend to lack multi-view constraints, these methods often suffer from the multi-view inconsistency issue, resulting in less desirable 3D generation outcomes.

Geometry Prior for Diffusion Models To enhance diffusion models for generative novel-view synthesis [5, 37, 79], Zero-1-to-3 [37] fine-tunes the Stable Diffusion model [61] to synthesize novel views conditioned on relative poses. Recent approaches have significantly improved the consistency of the generated multi-view images by performing multi-view diffusion [38, 66, 67, 70, 74, 80, 87, 95]. For example, MVDream [67] fine-tunes a pre-trained diffusion model on the synthetic Objaverse dataset [11] to simultaneously generate a set of 4-view images of the same object, conditioned on the camera poses. While these methods effectively address the multi-view inconsistency problem,

they perform diffusion in RGB image space, making them less suitable for the decoupled generation approach where the geometry is generated before appearance.

There are methods incorporating more explicit geometric constraints into the diffusion models. LDM3D [69] introduces an RGB-D diffusion model on the LAION-400M dataset. However, this model is not validated for text-to-3D generation. Concurrent to our work, SweetDreamer [33] proposes to align the geometric prior in 2D diffusion using a canonical coordinate map (CCM) representation. However, CCM implicitly requires the training objects to be aligned and can only be obtained from synthetic 3D datasets, potentially limiting its generalization and scalability. Wonder3D [40] introduces an RGB-Normal diffusion model on the Objaverse dataset. HumanNorm [24] proposes two disjoint diffusion models, one for normal and the other for depth, on a 3D Human dataset of 2952 body models. In contrast, our model jointly learns the distribution of normal and depth, and was pre-trained on the extensive LAION-2B dataset to improve the generalization ability. In addition, we introduce an albedo diffusion model to better model appearance.

3. Method

In this section, we introduce a Normal-Depth diffusion model and an albedo diffusion model to push forward the decoupled 3D generation pipeline, where the geometry is first generated followed by the appearance modeling (see Fig. 2).

Overview The existing approach for decoupled generation [7] adopts a text-to-image diffusion model (*i.e.*, Stable-Diffusion [61]) to optimize the rendered surface normals of the object for geometry generation. However, this direct

application is suboptimal due to the discrepancy in data distribution between natural images and normal maps, often leading to unstable optimization and compromised geometric fidelity [7]. As a result, appropriate geometry initialization (*e.g.*, 3D ellipsoids with different shapes and orientations, or coarse 3D models) is often needed to achieve good results for different prompts. In response, we propose to learn a 3D-aware diffusion model tailored for 3D geometry generation. Specifically, we introduce a Normal-Depth diffusion model pre-trained on an extensive real-world dataset and further fine-tuned on a synthetic dataset, to offer consistent guidance for geometry generation.

Another critical issue in text-to-3D generation is the inaccurate appearance modeling, where materials intermingle with the lighting effects like shadows and specular highlights, often resulting in imprecise relighting. In an attempt to address this, our method integrates an albedo diffusion model to regularize the albedo component of the materials, effectively separating the albedo from the influence of lighting artifacts.

3.1. Normal-Depth Diffusion Model

To endow the diffusion model with 3D geometric priors for 3D generation tasks, we introduce a novel Normal-Depth diffusion model. Different from existing methods that either learn a normal or a depth diffusion model [24, 69], our model captures the joint distribution of normal and depth, leveraging their intrinsic complementary nature-depths describe the macrostructure of the scene while normals provide local surface details.

Model Architecture We adapted the architecture of the publicly available text-to-image diffusion model, Stable Diffusion (SD) [61], with minor modifications. SD incorporates a variational auto-encoder (VAE) with KL-regularization [31], and a latent diffusion model (LDM). The VAE maps an image of size 512×512 to and from a latent space of size 64×64, and the LDM is a UNet denoiser that learns to denoise the latent feature guided by the text prompt.

For our purpose, we extended the input and output channel number of SD's VAE from three to four channels to encompass three for normals and one for depth, keeping other components unchanged.

Pre-training on Real-world Data The LAION-2B dataset, comprising billions of correlated image and text pairs, served as our foundational training resource [63]. We prepared our training set with text prompts paired with corresponding normal and depth maps, utilizing Normal-Bae [3] and Midas-3.1 [58], which are leading methods for monocular normal and depth estimation.

We first trained the Normal-Depth VAE to learn the joint distribution of normal and depth with the MSE reconstruction loss, adversarial loss, and the KL-regularization loss

[14, 61]. We then trained the LDM to enable text to Normal-Depth generation. Denoting x as the normal and depth data, $\mathcal{E}$ the encoder, z the latent feature, and ϵ_θ the UNet denoiser, the objective for training LDM can be written as:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z \sim \mathcal{E}(x), y, t, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon_\theta(z_t, y, t) - \epsilon \|_2^2 \right], \quad (1)$$

where z_t is the noised latent variable at a specific denoising timestep t, and y the text embedding obtained from the CLIP model [54]. Our results show that the pre-training on the real-world dataset is crucial to maintain the generalization ability on diverse prompts.

Fine-tuning on Synthetic 3D Data To enhance object-level 3D generation, we fine-tuned our Normal-Depth LDM on the Objaverse dataset [11], which features ground-truth 3D models. We render the ground-truth normal and depth maps with the provided object. In the fine-tuning stage, we employed a four-view diffusion technique proposed by MV-Dream [67]. The camera poses are mapped by a simple Multilayer Perceptron (MLP) to be the camera embeddings, which will be added to the time embedding to be accessed by the diffusion model. The training objective in the fine-tuning stage is:

$$\mathcal{L}'_{\text{LDM}} = \mathbb{E}_{z, y, c, t, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon_\theta(z_t, y, c, t) - \epsilon \|_2^2 \right], \quad (2)$$

where c is the camera condition.

Implementation Details For training on the LAION dataset, we follow the data filtering strategy used in the training of SD v2.1 to ensure high-quality data selection. The image is resized to 384×384 as input to Midas for estimating normal and depth. The computational expense for training the Normal-Depth VAE and LDM amounted to $1,344$ and $11,520$ GPU hours respectively on A100-80G GPUs. More details can be found in the supplementary materials

For fine-tuning on the Objaverse dataset, for each 3D object, we established camera positions within a radial distance of 1.4 to 2.0 units and an elevation angle spanning 5 to 30 degrees. We rendered 24 views per object, distributed uniformly across azimuth angles. To enhance the dataset quality, we discarded objects whose rendered images scored low in relevance to the object names as determined by CLIP scores, resulting in a pool of 270,000 training objects. The text prompts used in training were obtained by a hybrid approach: 30% stemmed from object tags and names, and the remaining 70% were from Cap3D [43]. The fine-tuning utilized a batch size of 512 with gradient accumulation performed every 8 batch. This was conducted on 8 GPUs for one week, reaching a total of $20,000$ iterations.

3.2. Geometry Generation

Score Distillation Sampling (SDS) Existing 2D lifting approaches for text-to-3D typically employ either a NeRF

representation [46, 52] or the hybrid DMTet representation [35, 65] to represent the 3D content. Denoting ϕ as the parameters of a 3D representation and g as the differentiable rendering function, the rendered image can be expressed as $x = g(\phi)$. DreamFusion [52] introduces a Score Distillation Sampling (SDS) process that leverages gradient-based score functions to guide the optimization of parameters in 3D representation for object generation:

$$\nabla_\phi \mathcal{L}_{\text{SDS}}(\phi, x = g(\phi)) = \mathbb{E}_{t,\epsilon}\left[w(t)(\epsilon_\theta(z_t; y, t) - \epsilon)\frac{\partial x}{\partial \phi}\right],$$
(3)

where the expression $(\epsilon_\theta(z_t; y, t) - \epsilon)$ represents the difference between the actual noise ϵ and the noise estimated by the UNet ϵ_θ. The term $w(t)$ is a weighting term that depends on the timestep t, and y indicates the text embedding.

Fantasia3D [7] shows that the SDS loss derived from the image diffusion model (*e.g.*, Stable Diffusion [61]) can be applied to the rendered normal maps from a DMTet for geometry generation.

Normal-Depth Diffusion for 3D Generation Compared to the image diffusion model, our Normal-Depth diffusion model is specifically designed for modeling the joint distribution of normal and depth maps. It provides effective supervision for geometry optimization.

We utilize a DMTet representation and integrate our Normal-Depth diffusion model into the coarse-to-fine geometry generation pipeline of Fantasia3D [7]. The normal and depth of the object can be efficiently rendered using differentiable rasterization [32]. The geometry generation loss function is defined as:

$$\mathcal{L}_{\text{Geo}} = \lambda_{\text{SD}}\mathcal{L}_{\text{SDS-Normal}}^{\text{SD}} + \lambda_{\text{ND}}\mathcal{L}_{\text{SDS-ND}}^{\text{ND}},$$
(4)

where the first loss term is employed in Fantasia3D [7] to enforce SDS on the rendered normal maps using Stable Diffusion. The second loss term is enabled by our Normal-Depth diffusion model to impose SDS on the composite of the rendered normal and depth maps. By default, we initialize the DMTet as a Sphere.

Integration with NeRF Since normal and depth maps can be derived from the NeRF representation using volume rendering, our Normal-Depth diffusion model can also be utilized to optimize NeRF using the loss defined in Eq. (4). Given that the normal and depth maps derived from NeRF can be noisy at the start of the optimization, we impose SDS loss on the rendered RGB images with SD during the first 1,000 iterations to warm up the optimization.

Recognizing that the NeRF representation is more flexible in modeling complex structures during optimization, we also investigate the idea of converting the optimized NeRF to DMTet as the initialization for geometry refinement.

Optimization For geometry generation, we accelerate the SDF function in DMTet with an efficient hash encod-

ing [49]. The loss weights λ_{SD} and λ_{ND} are both set to 1. The optimization process takes approximately 1.5 hours on a single GPU with 30 GB of memory.

3.3. Appearance Modeling

Physically-based Rendering For DMTet representation, in line with prior studies [7, 50], we employ the Physically Based Rendering (PBR) Disney material model [4], which integrates a diffuse term with a specular GGX lobe [75]. The material property of a surface point is determined by the diffuse color $k_d \in \mathbb{R}^3$, roughness $k_r \in \mathbb{R}$, metallic term $k_m \in \mathbb{R}$, and normal variation in tangent space $k_n \in \mathbb{R}^3$. We parameterize the spatially-varying materials of the surface by a learnable MLP f_ψ with parameters ψ to predict material parameters for input 3D point p as:

$$(k_d, k_r, k_m, k_n) = f_\psi(p).$$
(5)

By specifying the environment lighting and the camera viewpoint, the image color can be computed using a differentiable renderer based on the surface geometry and materials [50].

The existing method optimizes materials by imposing SDS loss on the final rendered RGB images [7]. However, this approach may lead to inaccuracies in material decomposition due to the inherent challenges in disentangling material components based solely on color.

To regularize the material generation, an ideal prior model should effectively regularize both the diffuse and specular components. However, due to the varied creation methods of existing 3D models and the lack of standardization [11], it is challenging to acquire a comprehensive dataset with consistent and accurate ground truth for the specular component. In light of this difficulty, we introduce an albedo diffusion model to decouple the albedo from complex lighting effects, serving as a preliminary approach to mitigate the challenge of mixed illumination.

Depth-Conditioned Albedo Generation A direct solution for the albedo diffusion model involves fine-tuning a text-to-image diffusion model using paired data of text prompts and albedo maps. While this method is effective for sampling, it falls short for 3D generation due to potential misalignments between the generated albedo and the geometry. To ensure the alignment of generated albedo maps with geometry, we employ the depth map from the corresponding viewpoint as a condition within the albedo Latent Diffusion Model (LDM). Specifically, we concatenate the depth map with the latent features to serve as input for the UNet denoiser [56]. We also employed the four-view diffusion strategy proposed by MV-Dream for the albedo diffusion model. We fine-tune the SD 2.1 on the Objaverse dataset [11] to capture the albedo distribution with the following training

objective:

$$\mathcal{L}_{\text{Albedo}} = \mathbb{E}_{z^a, y, c, t, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon_{\theta_a}(z_t^a, y, c, d, t) - \epsilon \|_2^2 \right], \tag{6}$$

where z^a represents the latent feature of the albedo map, and d is the depth condition.

Loss Function The loss function for appearance modeling is expressed as:

$$\mathcal{L}_{\text{App}} = \lambda_{\text{SD}} \mathcal{L}_{\text{SDS-RGB}}^{\text{SD}} + \lambda_{\text{Albedo}} \mathcal{L}_{\text{SDS-Albedo}}^{\text{Albedo}}, \tag{7}$$

where the first term reflects the SDS on the rendered RGB images using SD, and the latter term is the SDS imposed on the albedo component by our Albedo diffusion model.

Optimization For appearance modeling, the loss weights λ_{SD}' and λ_{ND}' are also set to 1. The optimization takes around 20 minutes on a GPU. After optimization, the material properties can be sampled at surface points and compiled into a 2D texture map [7, 50, 89], which can be directly imported into existing graphics engines for applications.

4. Experiments

In this section, we thoroughly evaluate the effectiveness of our proposed text-to-3D method by conducting a comprehensive comparison with state-of-the-art approaches.

Model Variants As discussed in Section 3.2, our Normal-Depth diffusion model can be applied to optimize DMTet and NeRF. To better verify the effectiveness of our Normal-Depth Diffusion model, we have designed two model variants for text-to-3D. The first model, denoted as *Ours (Sphere)*, initializes the DMTet with a Sphere for geometry generation. The second model, denoted as *Ours (NeRF)*, first optimizes a NeRF with our Normal-Depth diffusion model and then converts the NeRF to DMTet as an initialization for geometry generation.

Baselines We conducted extensive comparisons with a variety of baseline methods, including both DMTet-based and NeRF-based methods. For the DMTet-based methods, we compared our approach against the state-of-the-art Fantasia3D method [7], utilizing its publicly available official code with DMTet initialized as a Sphere. In the case of NeRF-based methods, we evaluated our approach against multiple competitors, including DreamFusion-IF [52], Magic3D-IF [35], TextMesh-IF [73], and Prolific-Dreamer [78]. As there is no publicly available code for these four methods, we used the implementation from three-studio [17], and IF indicates the DeepFloyd IF* diffusion model. We also compared our method with the state-of-the-art NeRF-based method, MVDream [67], using its publicly available official code.

---

* https://github.com/deep-floyd/IF

SweetDreamer [33] is a contemporaneous work that is compatible with DMTet and NeRF. Given the absence of a public implementation, we conducted a fair comparison by evaluating our results against those presented on its website using identical prompts in the supplementary materials.

4.1. Evaluation on Text-to-3D

We conducted evaluations in two key aspects: geometry generation and textured model generation.

Evaluation on Geometry Generation Evaluating the quality of generated geometry is a complex problem due to the lack of standard metrics. To objectively evaluate the geometry's quality, we employed a rendering-based approach. Specifically, to isolate the geometric attributes from the influence of texture, we rendered the generated geometry with a uniform albedo and then calculated the CLIP score [54] using the provided text prompts and CLIP model (vit-g-14). This process involved generating 16 different views for each object (a total of 113 objects) and computing the average score after removing the highest and lowest scores.

Table 1 (the first row) shows that two variants of our method achieve the top two values in the average CLIP scores in uniform rendering, demonstrating that our method outperforms existing methods in geometry generation. Visual results in Fig. 3 clearly show that our method can produce 3D content with exceptionally detailed geometry aligned with the text prompts, indicating the effectiveness of the proposed Normal-Depth diffusion model.

Evaluation on Textured Model Generation In parallel with the geometry evaluation, we assessed the quality of the generated textured models. To accomplish this, we computed CLIP scores for the rendered images of the textured models. As in the geometry evaluation, we rendered 16 distinct views and computed the average scores. Table 1 (the second row) shows that the two variants of our method achieve the second and third highest scores, outperforming most of the existing methods. Our result is slightly lower than that of the ProlificDreamer with a comparison of 31.7099 vs. 31.8022. The reason might be that the Prolific-Dreamer additionally fine-tunes the diffusion model with LoRA [23] during optimization, which might lead to a rendered image that better fits the text prompts. However, the visual comparison of textured model generation in Fig. 3 shows that our method generates much more accurate and detailed models. These results verify the design of our decoupled text-to-3D generation approach.

User Study To further assess the visual quality of the generated 3D models, we conducted a comprehensive user study. We separately compare the two variants of our method with existing methods. We collected a set of 87 prompts from DreamFusions, Sweetdreamer, and MV-Dream. 119 and 192 participants were involved for the comparison of "*Ours (NeRF)* vs. existing methods" and
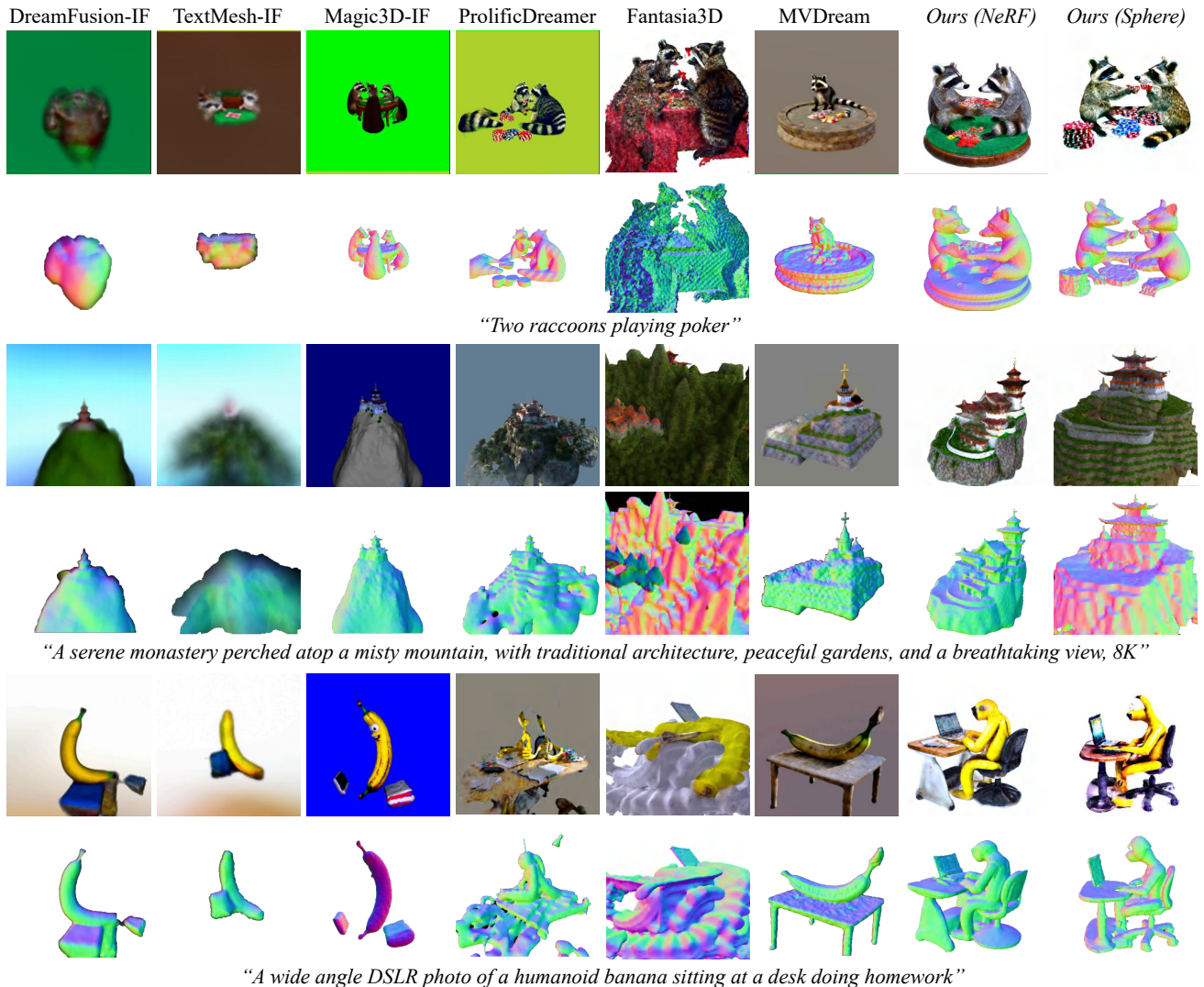
| DreamFusion-IF | TextMesh-IF | Magic3D-IF | ProlificDreamer | Fantasia3D | MVDream | *Ours (NeRF)* | *Ours (Sphere)* |

"Two raccoons playing poker"

"A serene monastery perched atop a misty mountain, with traditional architecture, peaceful gardens, and a breathtaking view, 8K"

"A wide angle DSLR photo of a humanoid banana sitting at a desk doing homework"

Figure 3. Visual comparison between our method and existing methods.

Table 1. Quantitative comparison with existing methods. The geometry CLIP score is measured on the shading images rendered with uniform albedo, and the appearance CLIP score is measured on the images rendered with textured models (values the higher the better).

	DreamFusion-IF	Magic3D-IF	TextMesh-IF	ProlificDreamer	Fantasia3D (Sphere)	MVDream	*Ours (NeRF)*	*Ours (Shpere)*
Geometry CLIP Score	17.4548	20.1157	18.2222	23.3818	17.5398	24.8003	26.0570	25.8820
Appearance CLIP Score	24.1091	27.8231	25.1218	31.8022	26.4055	28.7331	31.3551	31.7099

"*Ours (Sphere)* vs. existing methods" respectively, with each participant undertaking 40 and 47 testing.

In each test case, participants simultaneously viewed the text prompt, textured models, and normal maps generated by various methods and were then asked to vote for the best textured model and geometry model. Figure 4 presents the results of our user study. Our method with NeRF representations received 75% and 70 of votes for "the best textured model" and "the best geometry". Our method with Sphere initialization received more than 59% and 58 of votes for the two comparisons. These results show that our method clearly outperforms existing methods in generating geome-

try and textured models.

4.2. Method Analysis

Effects of the Normal-Depth Diffusion Model To explore the impact of the proposed Normal-Depth diffusion model in the context of text-to-3D, we show results of geometry generation without using the SDS loss from the SD model. Figure 5 and Table 2 show that only using our Normal-Depth model can robustly generate geometry with a coherent structure. When the SD model is incorporated alongside our Normal-Depth model, the resulting geometry exhibits finer details and an improved shape. These findings
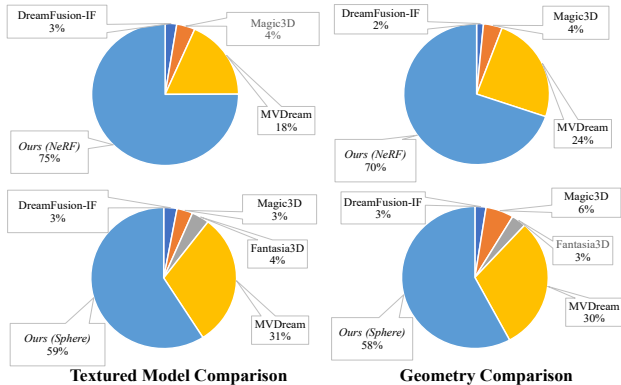
Figure 4. User study for text-to-3D.

Table 2. Ablation study for geometry generation.

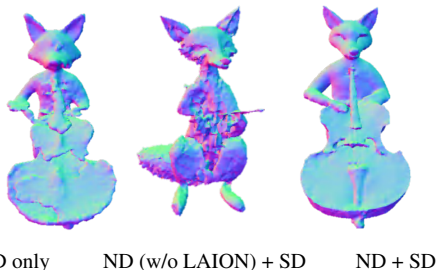|	ND Only	ND (w/o LAION) + SD	ND + SD
Geometry CLIP Score	24.1070	24.2601	25.8820
Appearance CLIP Score	29.5379	29.7522	31.7099

Figure 5. Ablation for the Normal-Depth (ND) diffusion model for geometry generation. Prompt: "*A fox plays a cello*".

suggest that our Normal-Depth model serves as a valuable 3D geometric prior for the overall structure, while the SD model excels in generating surface details.

Effects of Pre-training on LAION dataset To evaluate the impact of pre-training on the LAION-2B dataset using normal and depth generated by existing methods [3, 58], we conducted a comparison with a baseline model that directly fine-tunes on the synthetic Objaverse dataset [11]. The resulted baseline text-to-3D model is denoted as *ND (w/o LAION)+ SD*. Figure 5 illustrates that when the Normal-Depth model is fine-tuned solely on the synthetic dataset, its generalization ability significantly deteriorates. It struggles to generate content that aligns with the user prompts, and the quality of the generated geometry is notably inferior. In contrast, our method, which involves pre-training on the expansive LAION dataset, successfully preserves its generalization ability and produces superior results, which is also evidenced in Table 2.

Effects of Albedo Diffusion Model Figure 6 shows that the albedo diffusion model can effectively improve the generated texture and lead to a more accurate appearance. Without the depth condition, the generated texture fails to align with the underlying geometry, highlighting the importance of depth condition in albedo diffusion model. With
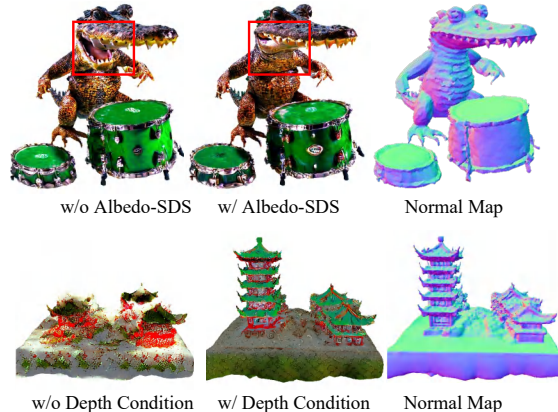

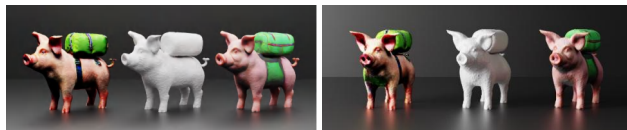Figure 6. Ablation results for the albedo diffusion model.

Figure 7. Relighting results. From left to right: results of model w/o albedo diffusion, shading, and model w/ albedo diffusion.

the inclusion of the albedo diffusion model, the generated albedo exhibits reduced shadows and specular highlights, leading to a more realistic relighting results (see Fig. 7).

5. Conclusion

In this work, we presented a generalizable approach to 3D generation through a Normal-Depth diffusion model, trained extensively on real-world data before undergoing fine-tuning with synthetic datasets. We also introduced a depth-conditioned albedo diffusion model that facilitates the separation of material attributes and lighting effects. Our models seamlessly integrate into current text-to-3D pipelines and demonstrate compatibility with the NeRF and DMTet representations. Extensive experiments show that our method achieves state-of-the-art text-to-3D results in both geometry and appearance modeling.

References

[1] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *CVPR*, 2023. 2, 3

[2] Mohammadreza Armandpour, Huangjie Zheng, Ali Sadeghian, Amir Sadeghian, and Mingyuan Zhou. Reimagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023. 3

[3] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *ICCV*, 2021. 2, 3, 4, 8

[4] Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. In *SIGGRAPH*. vol. 2012, 2012. 5

[5] 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. 3

[6] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *ICCV*, 2023. 2

[7] 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. 2, 3, 4, 5, 6

[8] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin. It3d: Improved text-to-3d generation with explicit view synthesis. *arXiv preprint arXiv:2308.11473*, 2023. 3

[9] Xinhua Cheng, Tianyu Yang, Jianan Wang, Yu Li, Lei Zhang, Jian Zhang, and Li Yuan. Progressive3d: Progressively local editing for text-to-3d content creation with complex semantic prompts. *arXiv preprint arXiv:2310.11784*, 2023. 3

[10] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *CVPR*, 2023. 2

[11] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023. 2, 3, 4, 5, 8

[12] 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. 3

[13] Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. *arXiv preprint arXiv:2303.17015*, 2023. 2

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 4

[15] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *NeurIPS*, 2022. 2

[16] 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. 3

[17] 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. 3, 6

[18] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023. 2

[19] Xiao Han, Yukang Cao, Kai Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, Tao Xiang, and Kwan-Yee K Wong. Headsculpt: Crafting 3d head avatars with text. *arXiv preprint arXiv:2306.03038*, 2023. 3

[20] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *ICCV*, 2019. 2

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 1

[22] Susung Hong, Donghoon Ahn, and Seungryong Kim. Debiasing scores and prompts of 2d diffusion for robust text-to-3d generation. *arXiv preprint arXiv:2303.15413*, 2023. 3

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

[24] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. *arXiv preprint arXiv:2310.01406*, 2023. 3, 4

[25] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 3

[26] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *arXiv preprint arXiv:2305.12529*, 2023. 3

[27] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, 2022. 2

[28] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 2

[29] Animesh Karnewar, Niloy J Mitra, Andrea Vedaldi, and David Novotny. Holofusion: Towards photo-realistic 3d generative modeling. In *ICCV*, 2023. 2

[30] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *TOG*, 2023. 3

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[32] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *TOG*, 2020. 5

[33] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweet-dreamer: Aligning geometric priors in 2d diffusion for con-sistent text-to-3d. *arXiv preprint arXiv:2310.02596*, 2023. 3, 6

[34] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxaing Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. *arXiv preprint arXiv:2308.10899*, 2023. 3

[35] 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. 2, 3, 5, 6

[36] 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. *arXiv preprint arXiv:2306.16928*, 2023. 3

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

[38] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Gen-erating multiview-consistent images from a single-view im-age. *arXiv preprint arXiv:2309.03453*, 2023. 3

[39] Zhen Liu, Yao Feng, Michael J Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffu-sion: Score-based generative 3d mesh modeling. In *ICLR*, 2023. 2

[40] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 3

[41] Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, and James Lucas. Att3d: Amortized text-to-3d object synthesis. *arXiv preprint arXiv:2306.07349*, 2023. 3

[42] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, 2021. 2

[43] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin John-son. Scalable 3d captioning with pretrained models. *arXiv preprint arXiv:2306.07279*, 2023. 4

[44] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *CVPR*, 2023. 3

[45] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *CVPR*, 2023. 2, 3

[46] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view syn-thesis. In *ECCV*, 2020. 2, 5

[47] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia Conference Papers*, 2022. 2

[48] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *CVPR*, 2023. 2

[49] Thomas Müller, Alex Evans, Christoph Schied, and Alexan-der Keller. Instant neural graphics primitives with a multires-olution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 3, 5

[50] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fi-dler. Extracting Triangular 3D Models, Materials, and Light-ing From Images. In *CVPR*, 2022. 5, 6

[51] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generat-ing 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 2

[52] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Milden-hall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 2, 5, 6

[53] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Sko-rokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 3

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learn-ing transferable visual models from natural language super-vision. In *ICML*, 2021. 2, 4, 6

[55] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aber-man, Michael Rubinstein, Jonathan Barron, et al. Dream-booth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023. 3

[56] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Confer-ence on Machine Learning*. PMLR, 2021. 1, 5

[57] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vi-sion transformers for dense prediction. In *CVPR*, 2021. 2

[58] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 2, 3, 4, 8

[59] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Com-mon objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *CVPR*, 2021. 2

[60] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023. 2

[61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image syn-thesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 4, 5

[62] 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. 2

[63] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 1, 2, 4

[64] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023. 3

[65] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *NeurIPS*, 2021. 5

[66] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[67] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2, 3, 4, 6

[68] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[69] Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, et al. Ldm3d: Latent diffusion model for 3d. *arXiv preprint arXiv:2305.10853*, 2023. 3, 4

[70] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. *arXiv preprint arXiv:2306.07881*, 2023. 3

[71] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 3

[72] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *ICCV*, 2023. 3

[73] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. *arXiv preprint arXiv:2304.12439*, 2023. 3, 6

[74] 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. 3

[75] Bruce Walter, Stephen R Marschner, Hongsong Li, and Kenneth E Torrance. Microfacet models for refraction through rough surfaces. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, 2007. 5

[76] 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. 2, 3

[77] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR*, 2023. 2

[78] 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. *arXiv preprint arXiv:2305.16213*, 2023. 2, 3, 6

[79] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3

[80] 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. 3

[81] Jinbo Wu, Xiaobo Gao, Xing Liu, Zhengyang Shen, Chen Zhao, Haocheng Feng, Jingtuo Liu, and Errui Ding. Hd-fusion: Detailed text-to-3d generation leveraging multiple noise estimation. *arXiv preprint arXiv:2307.16183*, 2023. 3

[82] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *NeurIPS*, 2016. 2

[83] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *CVPR*, 2023. 2

[84] 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. *arXiv e-prints*, 2022. 3

[85] 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. 2

[86] Xudong Xu, Zhaoyang Lyu, Xingang Pan, and Bo Dai. Matlaber: Material-aware text-to-3d via latent brdf auto-encoder. *arXiv preprint arXiv:2308.09278*, 2023. 3

[87] Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. *arXiv preprint arXiv:2310.03020*, 2023. 3

[88] Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023. 3

[89] Jonathan Young. xatlas. https://github.com/jpcy/xatlas, 2021. 6

[90] Wang Yu, Xuelin Qian, Jingyang Huo, Tiejun Huang, Bo Zhao, and Yanwei Fu. Pushing the limits of 3d shape generation at scale. *arXiv preprint arXiv:2306.11510*, 2023. 2

[91] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, Guanying Chen,

Shuguang Cui, and Xiaoguang Han. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, 2023. 2

[92] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. In *NeurIPS*, 2022. 2

[93] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. In *SIGGRAPH*, 2023. 2

[94] Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu. Dreamface: Progressive generation of animatable 3d faces under text guidance. *arXiv preprint arXiv:2304.03117*, 2023. 3

[95] Minda Zhao, Chaoyi Zhao, Xinyue Liang, Lincheng Li, Zeng Zhao, Zhipeng Hu, Changjie Fan, and Xin Yu. Efficientdreamer: High-fidelity and robust 3d creation via orthogonal-view diffusion prior. *arXiv preprint arXiv:2308.13223*, 2023. 2, 3

[96] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *arXiv preprint arXiv:2306.17115*, 2023. 3

[97] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 3