Text-to-3D using Gaussian Splatting

Zilong Chen, Feng Wang, Yikai Wang, Huaping Liu†
Beijing National Research Center for Information Science and Technology (BNRist),
Department of Computer Science and Technology, Tsinghua University
chenzl22@mails.tsinghua.edu.cn, hpliu@tsinghua.edu.cn
Project Page: gsgen3d.github.io

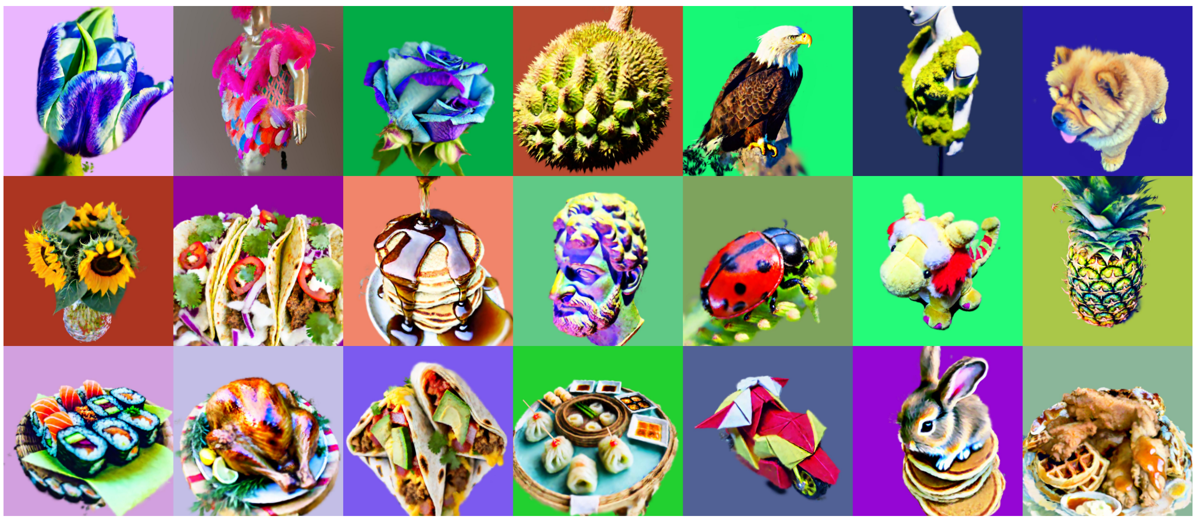

Figure 1. Delicate 3D assets generated by the proposed GSGEN. See the project page for videos.

Abstract

Automatic text-to-3D generation that combines Score Distillation Sampling (SDS) with the optimization of volume rendering has achieved remarkable progress in synthesizing realistic 3D objects. Yet most existing text-to-3D methods by SDS and volume rendering suffer from inaccurate geometry, e.g., the Janus issue, since it is hard to explicitly integrate 3D priors into implicit 3D representations. Besides, it is usually time-consuming for them to generate elaborate 3D models with rich colors. In response, this paper proposes GSGEN, *a novel method that adopts Gaussian Splatting, a recent state-of-the-art representation, to text-to-3D generation.* GSGEN *aims at generating high-quality 3D objects and addressing existing shortcomings by exploiting the explicit nature of Gaussian Splatting that enables the incorporation of 3D prior. Specifically, our method adopts a progressive optimization strategy, which includes a geometry optimization stage and an appearance refinement stage. In geometry optimization, a coarse representation is established under 3D point cloud diffusion prior along with the ordinary 2D SDS optimization, ensuring a sensible and 3D-consistent rough shape. Subsequently, the obtained Gaussians undergo an iterative appearance refinement to enrich texture details. In this stage, we increase the number of Gaussians by compactness-based densification to enhance continuity and improve fidelity. With these designs, our approach can generate 3D assets with delicate details and accurate geometry. Extensive evaluations demonstrate the effectiveness of our method, especially for capturing high-frequency components. Our code is available at* `https://github.com/gsgen3d/gsgen`.

1. Introduction

Diffusion model based text-to-image generation [1, 55, 57, 58] has achieved remarkable success in synthesizing photo-realistic images from textual prompts. Nevertheless, for high-quality text-to-3D content generation, the advancements lag behind that of image generation due to the inherent complex-

†Corresponding author

Stable DreamFusion [50, 69] threestudio ProlificDreamer [23, 77] GSGEN (Ours)

A DSLR photo of a panda

Fantasia3D [12] threestudio ProlificDreamer [23, 77] GSGEN (Ours)

A high quality photo of a furry corgi

threestudio Magic3D [23, 34] threestudio ProlificDreamer [23, 77] GSGEN (Ours)

A zoomed out DSLR photo of an amigurumi bulldozer

Figure 2. Compared to previous methods, GSGEN alleviates the Janus problem by representing the 3D scene using 3D Gaussian Splatting, which is capable of applying direct 3D geometry guidance and expressing content with delicate details. Note that the results of DreamFusion, Magic3D, and ProlificDreamer are obtained using Stable DreamFusion [69] and threestudio [23] since the official implementations have not been publicly available till the date of this work.

ity of real-world 3D scenes. Recently, DreamFusion [50] has made great progress in generating delicate assets by utilizing score distillation sampling with a pre-trained text-to-image diffusion prior. Its follow-up works further improve this paradigm in quality [12, 77], training speed [34, 41], and generating more reasonable geometry [2, 61, 85]. However, most existing text-to-3D methods still suffer greatly from collapsed geometry and limited fidelity, and are difficult to incorporate 3D priors due to the implicit nature of NeRF [43] and DMTET [62].

Recently, 3D Gaussian Splatting [31] has garnered significant attention in the field of 3D reconstruction, primarily due to its remarkable ability to represent intricate scenes and capability of real-time rendering. By modeling a scene using a set of 3D Gaussians, Kerbl et al. [31] adopt an explicit and object-centric approach that fundamentally diverges from implicit representations like NeRF and DMTET. This distinctive approach paves the way for the integration of explicit 3D priors into text-to-3D generation. Building upon this insight, instead of a straightforward replacement of NeRFs with Gaussians, we propose to guide the generation with an additional 3D point cloud diffusion prior to enhancing geometrical coherence. By adopting this strategy, we can better harness the inherent advantages of 3D Gaussians in the creation of complex and 3D-consistent assets.

Specifically, we propose to represent the generated 3D content with a set of Gaussians and optimize them progressively in two stages, namely geometry optimization and appearance refinement. In the geometry optimization stage, we optimize the Gaussians under the guidance of a 3D point cloud diffusion prior along with the ordinary 2D image prior. The incorporation of this extra 3D SDS loss ensures a 3D-consistent rough geometry. In the subsequent refinement stage, the Gaussians undergo an iterative enhancement to enrich delicate details. Due to the sub-optimal performance of the original adaptive control under SDS loss, we introduce an additional compactness-based densification technique to enhance appearance and fidelity. Besides, to prevent potential degeneration and break the symmetry in the early stage, the Gaussians are initialized with a coarse point cloud generated by a text-to-point-cloud diffusion model. As a result of these techniques, our approach can generate 3D assets with consistent geometry and exceptional fidelity. Fig. 2 illustrates a comparison between GSGEN and previous state-of-the-art methods on generating assets with asymmetric geometry. In summary, our contributions are:

- We propose GSGEN, a text-to-3D generation method using 3D Gaussians as representation. By incorporating direct geometric priors, we highlight the distinctive advantages of Gaussian Splatting in text-to-3D generation.
- We introduce a two-stage optimization strategy that first exploits joint guidance of 2D and 3D diffusion prior to shaping a coherent rough structure in geometry optimization; then enriches the details with compactness-based densification in appearance refinement.
- We validate GSGEN on various textual prompts. Experiments show that our method can generate 3D assets with accurate geometry and enhanced fidelity. Especially, GSGEN demonstrates superior performance in capturing *high-frequency components*, such as feathers, surfaces with intricate textures, animal fur, etc.

2. Related Work

2.1. 3D Scene Representations

Representing 3D scenes in a differentiable way has achieved remarkable success in recent years. NeRFs [43] demonstrates outstanding performance in novel view synthesis by representing 3D scenes with a coordinate-based neural network. After works have emerged to improve NeRF in reconstruction quality [7, 8, 76], handling large-scale [13, 40, 68, 81] and dynamic scenes [3, 47, 51, 60, 73], improving training [10, 44, 67, 79], rendering [25, 56, 80] speed and facilitating down-stream tasks [17, 46, 78, 83, 84]. Although great progress has been made, NeRF-based methods still suffer from low rendering speed and high training-time memory usage due to their implicit nature. To tackle these challenges, Kerbl et al. [31] propose to represent the 3D scene as a set of anisotropic Gaussians and render novel views using GPU-optimized tile-based rasterization. Gaussian Splatting could achieve better reconstruction results while being capable of real-time rendering. Our research highlights the distinctive advantages of Gaussian Splatting within text-to-3D by incorporating explicit 3D prior, generating 3D consistent and highly detailed assets.

2.2. Diffusion Models

Diffusion models have arisen as a promising paradigm for learning and sampling from a complex distribution. Inspired by the diffusion process in physics, these models involve a forward process to gradually add noise and an inverse process to denoise a noisy sample with a trained neural network. After DDPM [27, 66] highlights the effectiveness of diffusion models in capturing real-world image data, a plethora of research has emerged to improve the inherent challenges, including fast sampling [5, 39, 65] and architecture improvements [6, 20, 28, 36, 48, 49]. One of the most successful applications of diffusion models lies in text-to-image generation, where they have shown remarkable progress in generating realistic images from text prompts [1, 26, 55]. To produce high-resolution images, current methods utilize either a cascaded structure combining a low-resolution diffusion model with super-resolution models [1, 4, 58] or train a diffusion model in latent space using an auto-encoder [22, 57]. Our proposed GSGEN is built upon StableDiffusion [57], an open-source latent diffusion model that provides fine-grained guidance for delicate 3D content generation.

2.3. Text-to-3D Generation

Early efforts in text-to-3D generation, including CLIP-forge [59], Dream Fields [29], Text2Mesh [42], TANGO [14], CLIPNeRF [72], and CLIP-Mesh [32], harness CLIP [53] guidance to create 3D assets. To leverage the stronger diffusion prior, DreamFusion [50] introduces score distillation sampling that optimizes the 3D content by minimizing the difference between rendered images and the diffusion prior. This development sparked a surge of interest in text-to-3D generation through image diffusion prior [15, 38, 52, 54, 74, 85]. Magic3D [34] employs a coarse-to-fine strategy, optimizing a NeRF with a low-resolution diffusion prior and then enhancing texture under latent diffusion prior with a DMTET initialized with the coarse NeRF. Latent-NeRF [41] trains a NeRF within the latent space of StableDiffusion and introduces the Sketch-Shape method to guide the generation process. Fantasia3D [12] disentangles the learning of geometry and material, harnessing physics-based rendering techniques to achieve high-fidelity mesh generation. ProlificDreamer [77] introduces variational score distillation to improve SDS and facilitate the generation of high-quality and diverse 3D assets. Our concurrent work DreamGaussian [70] achieves fast image-to-3D by capitalizing on the rapid convergence of Gaussian Splatting, whose contribution is orthogonal to ours since we focus on incorporating 3D prior with more advanced representation. Another line of work lies in training or fine-tuning diffusion models directly on 3D datasets (e.g. ShapeNet [9] and Objaverse [18, 19]) to achieve more consistent results with advanced guidance [11, 16, 24, 30, 33, 35, 37, 63, 64, 75, 82]. Our approach builds upon Point-E [45], a text-to-point-cloud

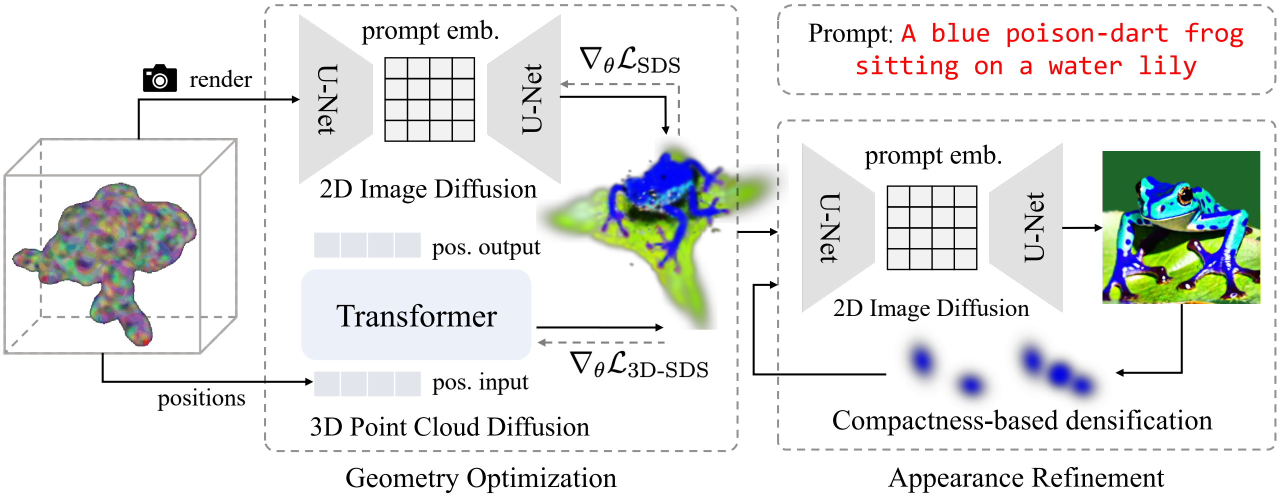

Figure 3. **Overview of the proposed GSGEN**. Our approach aims at generating 3D assets with accurate geometry and delicate appearance. GSGEN starts by utilizing Point-E to initialize the positions of the Gaussians (Sec 4.3). The optimization is grouped into geometry optimization (Sec 4.1) and appearance refinement (Sec 4.2) to meet a balance between coherent geometry structure and detailed texture.

diffusion model trained on millions of 3D models, which offers valuable 3D guidance and coarse initialization.

3. Preliminary

3.1. Score Distillation Sampling

Instead of directly generating 3D models, recent studies have achieved notable success by optimizing 3D representation with a 2D pre-trained image diffusion prior based on score distillation sampling, as proposed by Poole et al. [50]. In this paradigm, the scene is represented as a differentiable image parameterization (DIP) denoted as θ, where the image can be differentiably rendered based on the given camera parameters through a transformation function g. The DIP θ is iteratively refined to ensure that, for any given camera pose, the rendered image $\mathbf{x} = g(\theta)$ closely resembles a plausible sample derived from the guidance diffusion model. DreamFusion achieves this by leveraging Imagen [58] to provide a score estimation function denoted as $\epsilon_\phi(x_t; y, t)$, where x_t, y, and t represent the noisy image, text embedding, and timestep, respectively. This estimated score plays a pivotal role in guiding the gradient update, as expressed by the following equation:

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = \mathbb{E}_{\epsilon,t}\left[w(t)(\epsilon_\phi(x_t; y, t) - \epsilon)\frac{\partial \mathbf{x}}{\partial \theta}\right] \tag{1}$$

where ϵ is a Gaussian noise and $w(t)$ is a weighting function. Our approach combines score distillation sampling with 3D Gaussian Splatting at both 2D and 3D levels with different diffusion models to generate 3D assets with both detailed appearance and 3D-consistent geometry.

3.2. 3D Gaussian Splatting

Gaussian Splatting, as introduced in Kerbl et al. [31], presents a pioneering method for novel view synthesis and 3D reconstruction from multi-view images. Unlike NeRF, 3D Gaussian Splatting adopts a distinctive approach, where the underlying scene is represented through a set of anisotropic 3D Gaussians parameterized by their positions, covariances, colors, and opacities. When rendering, the 3D Gaussians are projected onto the camera's imaging plane [86]. Subsequently, the projected 2D Gaussians are assigned to individual tiles. The color of $\boldsymbol{p}$ on the image plane is rendered sequentially with point-based volume rendering technique [86]:

$$C(\boldsymbol{p}) = \sum_{i \in \mathcal{N}} c_i \alpha_i \prod_{j=1}^{i-1}(1 - \alpha_j) \tag{2}$$

where $\alpha_i = o_i e^{-\frac{1}{2}(\boldsymbol{p}-\mu_i)^T \Sigma_i^{-1}(\boldsymbol{p}-\mu_i)}$ refers to the opacity at point $\boldsymbol{p}$, c_i, o_i, μ_i, and Σ_i represent the color, opacity, position, and covariance of the i-th Gaussian respectively, $\mathcal{N}$ denotes the Gaussians in this tile. To maximize the utilization of shared memory, Gaussian Splatting further designs a GPU-friendly rasterization process where each thread block is assigned to render an image tile. These advancements enable Gaussian Splatting to achieve more detailed scene reconstruction, significantly faster rendering speed, and reduction of memory usage during training compared to NeRF-based methods. In this study, we expand the application of Gaussian Splatting into text-to-3D generation and introduce a novel approach that leverages the explicit nature of Gaussian Splatting by integrating direct 3D diffusion priors, highlighting the potential of 3D Gaussians as a fundamental representation for generative tasks.

4. Approach

Our goal is to generate 3D content with accurate geometry and delicate detail. To accomplish this, GSGEN exploits the 3D Gaussians as representation due to its flexibility to incorporate geometry priors and capability to represent high-frequency details. Based on the observation that a point cloud can be seen as a set of isotropic Gaussians, we propose to integrate a 3D SDS loss with a pre-trained point cloud diffusion model to shape a 3D-consistent geometry. With this additional geometry prior, our approach could mitigate the Janus problem and generate more sensible geometry. Subsequently, in appearance refinement, the Gaussians undergo an iterative optimization to gradually improve fine-grained details with a compactness-based densification strategy, while preserving the fundamental geometric information. The detailed GSGEN methodology is presented as follows.

4.1. Geometry Optimization

Many text-to-3D methods encounter the significant challenge of overfitting to several views, resulting in assets with multiple faces and collapsed geometry [12, 34, 50]. This issue, known as the Janus problem [2, 61], has posed a persistent hurdle in the development of such approaches. In our early experiments, we faced a similar challenge that relying solely on 2D guidance frequently led to flawed results. However, we noticed that the geometry of 3D Gaussians can be directly rectified with a point cloud prior, which is not feasible for previous text-to-3D methods using NeRFs as their geometries are represented in implicit density functions. Recognizing this distinctive advantage, we introduce a geometry optimization process to shape a reasonable structure. Concretely, in addition to the ordinary 2D image diffusion prior, we further optimize the positions of Gaussians using Point-E [45] guidance, a pre-trained text-to-point-cloud diffusion model. Instead of directly aligning the Gaussians with a Point-E generated point cloud, we apply a 3D SDS loss to lead the positions inspired by image diffusion SDS, which avoids challenges including registration, scaling, and potential degeneration. We summarize the loss in the geometry optimization stage as the following equation:

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\text{geometry}} &= \mathbb{E}_{\epsilon_I,t}\left[w_I(t)(\epsilon_\phi(x_t;y,t)-\epsilon_I)\frac{\partial \mathbf{x}}{\partial \theta}\right] \\ &+ \lambda_{\text{3D}}\cdot \mathbb{E}_{\epsilon_P,t}\left[w_P(t)(\epsilon_\psi(p_t;y,t)-\epsilon_P)\right],\end{aligned} \tag{3}$$

where p_t and x_t represent the noisy Gaussian positions and the rendered image, w_* and ϵ_* refer to the corresponding weighting function and Gaussian noise.

4.2. Appearance Refinement

While the introduction of 3D prior does help in learning a more reasonable geometry, we experimentally find it would

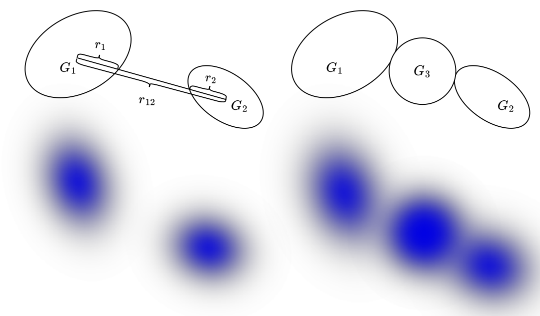

Figure 4. An illustration of the proposed compactness-based densification.

also disturb the learning of appearance, resulting in insufficiently detailed assets. Based on this observation, GSGEN employs another appearance refinement stage that iteratively refines and densifies the Gaussians utilizing only the 2D image prior.

To densify the Gaussians, Kerbl et al. [31] propose to split Gaussians with a large view-space spatial gradient. However, we encountered challenges in determining the appropriate threshold for this spatial gradient under score distillation sampling. Due to the stochastic nature of SDS loss, employing a small threshold is prone to be misled by some stochastic large gradient thus generating an excessive number of Gaussians, whereas a large threshold will lead to a blurry appearance, as illustrated in Fig. 7.

To tackle this, we propose compactness-based densification as a supplement to positional gradient-based split with a large threshold. Specifically, for each Gaussian, we first obtain its K nearest neighbors with a KD-Tree. Then, for each of the neighbors, if the distance between the Gaussian and its neighbor is smaller than the sum of their radius, a Gaussian will be added between them with a radius equal to the residual. As illustrated in Fig. 4, compactness-based densification could “fill the holes”, resulting in a more complete geometry structure. To prune unnecessary Gaussians, we add an extra loss to regularize opacity with a weight proportional to its distance to the center and remove Gaussians with opacity smaller than a threshold α_{min} periodically. Furthermore, we recognize the importance of ensuring the geometry consistency of the Gaussians throughout the refinement phase. With this concern, we penalize Gaussians which deviate significantly from their positions obtained during the preceding geometry optimization. The loss function in the appearance refinement stage is summarized as the following:

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\text{refine}} &= \lambda_{\text{SDS}}\mathbb{E}_{\epsilon_I,t}\left[w_I(t)(\epsilon_\phi(x_t;y,t)-\epsilon_I)\frac{\partial \mathbf{x}}{\partial \theta}\right] \\ &+ \lambda_{\text{mean}}\nabla_\theta \sum_i ||\mathbf{p}_i|| + \lambda_{\text{opacity}}\nabla_\theta \sum_i \mathtt{sg}(||\mathbf{p}_i||)\cdot o_i,\end{aligned} \tag{4}$$

where $\mathtt{sg}(\cdot)$ refers to the stop gradient operation, $\mathbf{p}_i$ and o_i represents the position and opacity of the i-th Gaussian

Figure 5. Qualitative comparison between the proposed GSGEN and state-of-the-art generation methods, including DreamFusion [50], Magic3D [34], Fantasia3D [12], and ProlificDreamer [77]. For more qualitative comparison results, please refer to the appendix. Videos of these images are provided in the project page.

respectively. λ_{SDS}, λ_{mean} and λ_{opacity} are loss weights.

4.3. Initialization with Geometry Prior

Previous studies [12, 34, 41] have demonstrated the critical importance of starting with a reasonable geometry initialization. In our early experiments, we also found that initializing with a simple pattern could potentially lead to a degenerated 3D object. To overcome this, we opt for initializing the positions of the Gaussians either with a generated point cloud or with a 3D shape provided by the users (either a mesh or a point cloud). In the context of general text-to-3D generation, we employ a text-to-point-cloud diffusion model, *Point-E* [45], to generate a rough geometry according to the text prompt. While Point-E can produce colored point clouds, we opt for random color initialization based on empirical observations, as direct utilization of the generated colors has been found to have detrimental effects in early experiments (See the appendix for visualization). The scales and opaci-

Figure 6. Ablation study results on initialization and 3D prior.

ties of the Gaussians are assigned with fixed values, and the rotation matrix is set to the identity matrix. For user-guided generation, we convert the preferred shape to a point cloud. To avoid too many vertices in the provided shape, we use farthest point sampling [21] for point clouds and uniform surface sampling for meshes to extract a subset of the original shape instead of directly using all the vertices or points.

5. Experiments

In this section, we present our experiments on validating the effectiveness of the proposed approach. Specifically, we compare GSGEN with previous state-of-the-art methods in general text-to-3D generation. Additionally, we conduct several ablation studies to evaluate the importance of initialization, 3D guidance, and densification strategy. The detailed results are shown as follows.

5.1. Implementation Details

Guidance model setup. We implement the guidance model based on the publicly available diffusion model, StableDiffusion [57, 71]. For the guidance scale, we adopt 100 for *StableDiffusion* as suggested in DreamFusion and other works. We also exploit the view-dependent prompt technique proposed by DreamFusion. All the assets demonstrated in this section are obtained with StableDiffusion checkpoint *runwayml/stable-diffusion-v1-5*.

3D Gaussian Splatting setup. We implement the 3D Gaussian Splatting rendering pipeline with a PyTorch CUDA extension, and further add learnable background support to facilitate our application. For densification, we split the Gaussians by view-space position gradient every 500 iterations with a threshold $T_{pos} = 0.02$ and perform compactness-based densification every 1000 iterations which we empirically found effective for achieving a complete geometry. For pruning, we remove Gaussians with opacity lower than $\alpha_{min} = 0.05$, and excessively large world-space or view-space radius every 200 iterations.

Traning setup. We use the same focal length, elevation, and azimuth range as those of DreamFusion [50]. To sample more uniformly in the camera position, we employ a stratified sampling on azimuth. We choose the loss weight hyperparameters $\lambda_{\text{SDS}} = 0.1$ and $\lambda_{\text{3D}} = 0.01$ in geometry optimization stage, and $\lambda_{\text{SDS}} = 0.1$, $\lambda_{\text{mean}} = 1.0$ and $\lambda_{\text{opacity}} = 100.0$ in appearance refinement.

5.2. Text-to-3D Generation

We evaluate the performance of the proposed GSGEN in the context of general text-to-3D generation and present qualitative comparison against state-of-the-art methods. As illustrated in Fig. 2, our approach produces delicate 3D assets with more accurate geometry and intricate details. In contrast, previous state-of-the-art methods under SDS guidance [12, 23, 34, 50, 69] struggle in generating collapsed geometry under the same guidance and prompt, which underscores the effectiveness of our approach. While the VSD guidance proposed by ProlificDreamer [77] significantly improves the appearance of generated assets, it is still susceptible to the Janus problem, resulting in flawed geometry. We present more qualitative comparison results in Fig. 5, where our approach showcases notable enhancements in preserving high-frequency details such as the intricate patterns on sushi, the feathers of the peacock, and the thatched roof. In contrast, Magic3D and Fantasia3D yield over-smoothed geometry due to the limitation of mesh-based methods while ProlificDreamer is prone to the multi-face problem, making the generated assets less realistic. Furthermore, our GSGEN stands out for its efficiency, generating 3D assets in about 40 minutes, on par with Magic3D and Fantasia3D, but with improved fidelity and richer details. For more qualitative comparisons and the performance of GSGEN under more advanced guidance including MVDream [63] and DeepFloyd IF [1], please refer to the appendix.

5.3. Ablation Study

Initialization. To assess the impact of initialization, we introduce a variant that initiates the positions of the Gaussians with an origin-centered Gaussian distribution which emulates the initialization adopted in DreamFusion [50]. The qualitative comparisons are shown in Fig. 6a. It is evident that assets generated with DreamFusion-like initialization encounter severe degeneration issues, especially for prompts depicting asymmetric scenes, resulting in collapsed geometry. In contrast, Point-E initialization breaks the symmetry by providing an anisotropic geometry prior, leading to the creation of more 3D-consistent objects.

3D prior. We evaluate the necessity of incorporating 3D prior by generating assets without point cloud guidance during geometry optimization. The qualitative comparisons are visualized in Fig. 6b. Although achieved better geometry consistency compared to random initialization, relying solely on image diffusion prior still suffers from the Janus problem, which is particularly evident in cases with asymmetric geometries, such as the dog and the panda. In contrast, our approach effectively addresses this issue with the introduction of 3D prior, rectifying potentially collapsed structures in the geometry optimization stage and resulting in a 3D-consistent rough shape. Notably, we show in the appendix that GSGEN maintains great performance even when Point-E behaves sub-optimally. We attribute this to direct 3D prior provided by Point-E assisting in geometrical consistency by correcting major shape deviations in the early stage, without the need to guide fine-grained geometric details. For a comprehensive analysis, please refer to the appendix.

Densification strategy. To validate the effectiveness of the proposed densification strategy, we propose two variants for comparison: (1) The original densification strategy that split Gaussians with an average view-space gradient larger than $T_{pos} = 0.0002$. (2) With larger $T_{pos} = 0.02$ that avoids too many new Gaussians. While effective in 3D reconstruction, the original densification strategy that relies only on view-space gradient encounters a dilemma in the context of score distillation sampling: within limited times of densification, a large threshold tends to generate an over-smoothed appearance while a small threshold is easily affected by unstable gradients. As shown in Fig. 7, the proposed compactness-based densification is an effective supplement to the original densification strategy under SDS guidance.

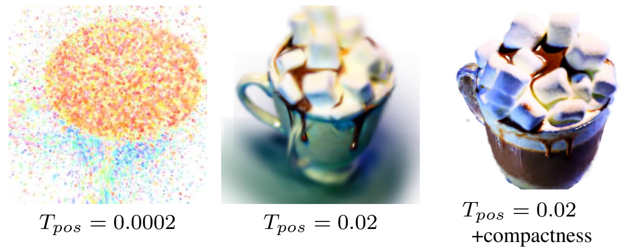

Figure 7. Ablation study on densification strategy.

6. Limitations and Conclusion

Limitations. GSGEN tends to generate unsatisfying results when the provided text prompt contains a complex description of the scene or with complicated logic due to the limited language understanding ability of Point-E and the CLIP text encoder used in *StableDiffusion*. Moreover, although incorporating 3D prior mitigates the Janus problem, it is far from eliminating the potential degenerations, especially when the textual prompt is extremely biased in the guidance diffusion models. Concrete failure cases and corresponding analyses are illustrated in the appendix.

Conclusion. In this paper, we propose GSGEN, a novel method for generating highly detailed and 3D consistent assets using Gaussian Splatting. In particular, we adopt a two-stage optimization strategy including geometry optimization and appearance refinement. In the geometry optimization stage, a rough shape is established under the joint guidance of a point cloud diffusion prior along with the common image SDS loss. In appearance refinement, the Gaussians are further optimized to enrich details and densified to achieve better continuity and fidelity with compactness-based densification. We conduct comprehensive experiments to validate the effectiveness of the proposed method, demonstrating its ability to generate 3D consistent assets and superior performance in capturing high-frequency components. We hope our method can serve as an efficient and powerful approach for high-quality text-to-3D generation and could pave the way for more extensive applications of Gaussians Splatting and direct incorporation of 3D prior.

7. Acknowledgement

This work was supported in part by the National Natural Science Fund for Distinguished Young Scholars under Grant 62025304.

References

[1] Alex, Misha Konstantinov, apolinário, Daria Bakshandaeva, Ksenia Ivanova, Sayak Paul, Will Berman, and Emad. deepfloyd/if, 2023. 1, 3, 8

[2] Mohammadreza Armandpour, Huangjie Zheng, Ali Sadeghian, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023. 2, 5

[3] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim. Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. *arXiv preprint arXiv:2301.02238*, 2023. 3

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *CoRR*, abs/2211.01324, 2022. 3

[5] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3

[6] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023. 3

[7] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 3

[8] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *ICCV*, 2023. 3

[9] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015. 3

[10] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *arXiv preprint arXiv:2203.09517*, 2022. 3

[11] 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. 3

[12] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 5, 6, 7

[13] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12952, 2022. 3

[14] Yongwei Chen, Rui Chen, Jiabao Lei, Yabin Zhang, and Kui Jia. TANGO: text-driven photorealistic and robust 3d stylization via lighting decomposition. In *NeurIPS*, 2022. 3

[15] 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, 2023. 3

[16] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander Schwing, and Liangyan Gui. SDFusion: Multimodal 3d shape completion, reconstruction, and generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[17] Qiyu Dai, Yan Zhu, Yiran Geng, Ciyu Ruan, Jiazhao Zhang, and He Wang. Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 1757–1763. IEEE, 2023. 3

[18] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *CoRR*, abs/2307.05663, 2023. 3

[19] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 13142–13153. IEEE, 2023. 3

[20] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021. 3

[21] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y. Zeevi. The farthest point strategy for progressive image sampling. *IEEE Trans. Image Process.*, 6(9):1305–1315, 1997. 7

[22] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10686–10696. IEEE, 2022. 3

[23] 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. 2, 7

[24] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oguz. 3dgen: Triplane latent diffusion for textured mesh generation. *CoRR*, abs/2303.05371, 2023. 3

[25] Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. *ICCV*, 2021. 3

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. 3

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3

[28] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 13213–13232. PMLR, 2023. 3

[29] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 857–866. IEEE, 2022. 3

[30] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *CoRR*, abs/2305.02463, 2023. 3

[31] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 3, 4, 5

[32] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Popa Tiberiu. Clip-mesh: Generating textured meshes from text using pretrained image-text models. *SIGGRAPH Asia 2022 Conference Papers*, 2022. 3

[33] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arxiv:2310.02596*, 2023. 3

[34] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 5, 6, 7

[35] 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. *https://arxiv.org/abs/2303.11328*, 2023. 3

[36] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. *arXiv preprint arXiv:2309.06380*, 2023. 3

[37] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Learning to generate multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3

[38] 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. In *International Conference on Computer Vision ICCV*, 2023. 3

[39] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 3

[40] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 3

[41] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 2, 3, 6

[42] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13482–13492. IEEE, 2022. 3

[43] 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. 2, 3

[44] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 3

[45] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *CoRR*, abs/2212.08751, 2022. 3, 5, 6

[46] Tolga Özer and Ömer Türkmen. Low-cost ai-based solar panel detection drone design and implementation for solar power systems. *Robotic Intelligence and Automation*, 43(6): 605–624, 2023. 3

[47] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 3

[48] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 3

[49] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023. 3

[50] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2, 3, 4, 5, 6, 7, 8

[51] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 3

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

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen

Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 3

[54] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, Yuanzhen Li, and Varun Jampan. DreamBooth3D: Subject-driven text-to-3d generation. In *International Conference on Computer Vision ICCV*, 2023. 3

[55] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. 1, 3

[56] Christian Reiser, Richard Szeliski, Dor Verbin, Pratul P. Srinivasan, Ben Mildenhall, Andreas Geiger, Jonathan T. Barron, and Peter Hedman. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *SIGGRAPH*, 2023. 3

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 1, 3, 7

[58] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1, 3, 4

[59] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, and Marco Fumero. Clip-forge: Towards zero-shot text-to-shape generation. *arXiv preprint arXiv:2110.02624*, 2021. 3

[60] Sara Fridovich-Keil and Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023. 3

[61] 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. 2, 5

[62] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[63] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 3, 8

[64] Jaehyeok Shim, Changwoo Kang, and Kyungdon Joo. Diffusion-based signed distance fields for 3d shape generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 20887–20897. IEEE, 2023. 3

[65] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[66] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[67] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 3

[68] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 3

[69] Jiaxiang Tang. Stable-dreamfusion: Text-to-3d with stable-diffusion, 2022. https://github.com/ashawkey/stable-dreamfusion. 2, 7

[70] 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

[71] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 7

[72] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3825–3834. IEEE, 2022. 3

[73] Feng Wang, Sinan Tan, Xinghang Li, Zeyue Tian, and Huaping Liu. Mixed neural voxels for fast multi-view video synthesis. *arXiv preprint arXiv:2212.00190*, 2022. 3

[74] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 12619–12629. IEEE, 2023. 3

[75] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. RODIN: A generative model for sculpting 3d digital avatars using diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 4563–4573. IEEE, 2023. 3

[76] Zhongshu Wang, Lingzhi Li, Zhen Shen, Li Shen, and Liefeng Bo. 4k-nerf: High fidelity neural radiance fields at ultra high resolutions. *arXiv preprint arXiv:2212.04701*, 2022. 3

[77] 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, 7

[78] Jianglong Ye, Jiashun Wang, Binghao Huang, Yuzhe Qin, and Xiaolong Wang. Learning continuous grasping function with a dexterous hand from human demonstrations. *IEEE Robotics and Automation Letters*, 8(5):2882–2889, 2023. 3

[79] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021. 3

[80] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 3

[81] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 3

[82] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. *ACM Transactions on Graphics (SIGGRAPH)*, 42(4), 2023. 3

[83] Shaohong Zhong, Alessandro Albini, Oiwi Parker Jones, Perla Maiolino, and Ingmar Posner. Touching a nerf: Leveraging neural radiance fields for tactile sensory data generation. In *Conference on Robot Learning*, pages 1618–1628. PMLR, 2023. 3

[84] Yang Zhou, Long Wang, Yongbin Lai, and Xiaolong Wang. The general method of the tanker car mouth pose measurement. *Robotic Intelligence and Automation*, 43(6):625–636, 2023. 3

[85] Junzhe Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance. *CoRR*, abs/2305.18766, 2023. 2, 3

[86] Matthias Zwicker, Hanspeter Pfister, Jeroen van Baar, and Markus H. Gross. EWA volume splatting. In *12th IEEE Visualization Conference, IEEE Vis 2001, San Diego, CA, USA, October 24-26, 2001, Proceedings*, pages 29–36. IEEE Computer Society, 2001. 4