One-2-3-45++: Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion

Minghua Liu[1*] Ruoxi Shi[1*] Linghao Chen[1,2*†] Zhuoyang Zhang[3*†] Chao Xu[4*] Xinyue Wei[1]
Hansheng Chen[5] Chong Zeng[2†] Jiayuan Gu[1] Hao Su[1]
[1] UC San Diego [2] Zhejiang University [3] Tsinghua University [4] UCLA [5] Stanford University

Figure 1. One-2-3-45++ is capable of transforming a single RGB image of any object into a high-fidelity textured mesh **in under one minute**. The generated meshes closely mirror the input image. Input image (and text prompt), textured mesh, and normal map are shown.

Abstract

Recent advancements in open-world 3D object generation have been remarkable, with image-to-3D methods offering superior fine-grained control over their text-to-3D counterparts. However, most existing models fall short in simultaneously providing rapid generation speeds and high fidelity to input images - two features essential for practical applications. In this paper, we present One-2-3-45++, an innovative method that transforms a single image into a detailed 3D textured mesh in approximately one minute. Our approach aims to fully harness the extensive knowledge embedded in 2D diffusion models and priors from valuable yet limited 3D data. This is achieved by initially finetuning a 2D diffusion model for consistent multi-view image generation, followed by elevating these images to 3D with the aid of multi-view-conditioned 3D native diffusion models. Extensive experimental evaluations demonstrate that our method can produce high-quality, diverse 3D assets that closely mirror the original input image.

1. Introduction

Generating 3D shapes from a single image or text prompt is a long-standing problem in computer vision and is essential for numerous applications. While remarkable progress has been achieved in the field of 2D image generation due to advanced generative methods and large-scale image-text datasets, transferring this success to the 3D domain is hindered by the limited availability of 3D data. Although many works have introduced sophisticated 3D generative mod-

*Equal contribution.
†Work done during internship at UC San Diego.

els [8, 16, 37, 87], a majority rely solely on 3D shape datasets for training. Given the limited size of publicly available 3D datasets, these methods often struggle to generalize across unseen categories in open-world scenarios.

Another line of work, exemplified by DreamFusion [49], Magic3D [31], harnesses the expansive knowledge or robust generative potential of 2D prior models like CLIP [51] and Stable Diffusion [56]. They typically optimize a 3D representation (e.g., NeRF or mesh) from scratch for each input text or image. During the optimization process, the 3D representation is rendered into 2D images, and the 2D prior models are employed to calculate gradients for them. While these methods have yielded impressive outcomes, the per-shape optimization can be exceedingly time-intensive, requiring tens of minutes or even hours to generate a single 3D shape for each input. Moreover, they frequently encounter the "multi-face" or Janus problem, produce results with oversaturated colors and artifacts inherited from the NeRF or triplane representation, and face challenges in generating diverse results across different random seeds.

A recent work One-2-3-45 [33] presents an efficient feed-forward pipeline to leverage rich priors of 2D diffusion models for 3D generation. It initially predicts multi-view images via a view-conditioned 2D diffusion model, Zero123 [34]. These images are subsequently processed by a generalizable NeRF method [38] for 3D reconstruction. Although One-2-3-45 can produce 3D shapes in a single forward pass, its efficacy is often constrained by the inconsistent multi-view predictions of Zero123, leading to compromised 3D reconstruction results.

In this paper, we introduce One-2-3-45++, a novel method that effectively overcomes the shortcomings of One-2-3-45, delivering significantly enhanced robustness and quality. Taking a single image of any object as input, One-2-3-45++ also includes two primary stages: 2D multi-view generation and 3D reconstruction. During the initial phase, rather than employing Zero123 to predict each view independently, One-2-3-45++ predicts consistent multi-view images jointly. This is realized by tiling a concise set of six-view images into a single image and then finetuning a 2D diffusion model to generate this combined image conditioned on the input reference image. In this way, the 2D diffusion net is able to attend to each view during generation, ensuring more consistent results across views. In the second stage, One-2-3-45++ employs a multi-view conditioned 3D-diffusion-based module to predict the textured mesh in a coarse-to-fine fashion. The consistent multi-view conditional images act as a blueprint for 3D reconstruction, facilitating a zero-shot hallucination capability. Concurrently, the 3D diffusion network excels in lifting the multi-view images, thanks to its ability to harness a broad spectrum of priors extracted from the 3D dataset. Ultimately, One-2-3-45++ employs a lightweight optimization technique to enhance the texture quality efficiently, leveraging the consistent multi-view images for supervision.

As depicted in Fig. 1, One-2-3-45++ efficiently generates 3D meshes with realistic textures in under a minute, offering precise fine-grained control. Our comprehensive evaluations, including user studies and objective metrics across an extensive test set, highlight One-2-3-45++'s superiority in terms of robustness, visual quality, and, most importantly, fidelity to the input image.

2. Related Work

2.1. 3D Generation

3D generation has garnered significant attention in recent years. Before the advent of large-scale pre-trained 2D models, researchers often delved into 3D native generative models that learn directly from 3D synthetic data or real scans and generate various 3D representations such as point clouds [1, 15, 41, 47, 83], 3D voxels [9, 59, 74, 75], polygon meshes [16, 17, 26, 32, 37, 46, 68], parametric models [21], and implicit fields [8, 14, 19, 25, 30, 42, 48, 73, 78, 82, 84, 86, 87]. However, given the limited availability of 3D data, these models tended to focus on a select number of categories (e.g., chairs, cars, planes, humans, etc.), struggling to generalize to unseen categories in the open world.

The advent of recent 2D generative models (e.g., DALL-E [53], Imagen [58], and Stable Diffusion [57]) and vision-language models (e.g., CLIP [51]) has equipped us with powerful priors about our 3D world, consequently fueling a surge of research in 3D generation. Notably, models like DreamFusion [49], Magic3D [31], and ProlificDreamer [71] have pioneered a line of approach for per-shape optimization [6, 7, 12, 23, 29, 40, 43–45, 50, 52, 60, 63, 64, 67, 76, 77, 81]. These models are designed to optimize a 3D representation for each unique input text or image, drawing on the 2D prior models for gradient guidance. While they have yielded impressive results, these methods tend to suffer from prolonged optimization times, the "multi-face problem," oversaturated colors, and a lack of diversity in results. Some works also concentrate on creating textures or materials for input meshes, utilizing the priors of 2D models [5, 55].

A new wave of studies, highlighted by works like Zero123 [34], has showcased the promise of using pre-trained 2D diffusion models for synthesizing novel views from singular images or texts, opening new doors for 3D generation. For instance, One-2-3-45 [33], using multi-view images predicted by Zero123, can produce a textured 3D mesh in a mere 45 seconds. Nevertheless, the multi-view images produced by Zero123 lack 3D consistency. Our research, along with several concurrent studies [36, 39, 61, 72, 80], is dedicated to enhancing the consistency of these multi-view images – a vital step for subsequent 3D reconstruction applications.

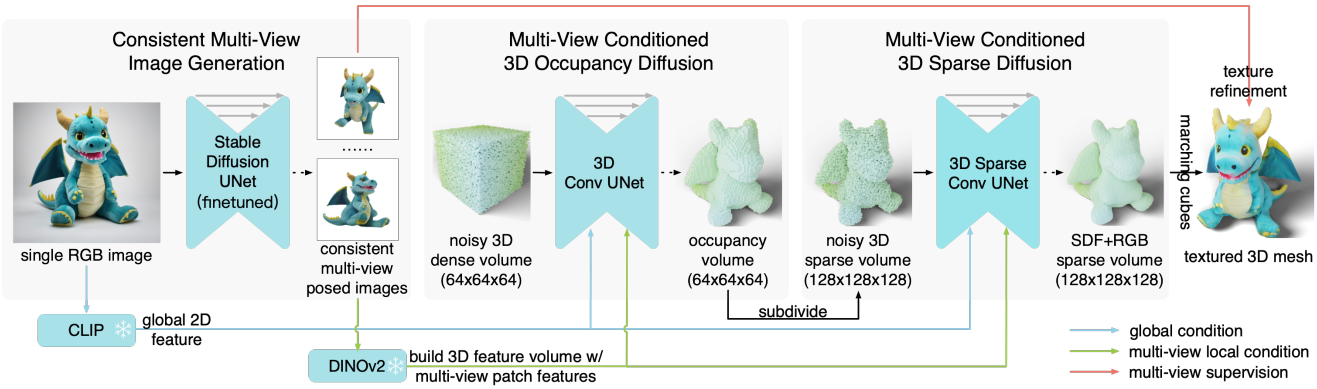

Figure 2. Starting with a single RGB image as input, we initially produce consistent multi-view images by fine-tuning a 2D diffusion model. These multi-view images are then elevated into 3D through a pair of 3D native diffusion networks. Throughout the 3D diffusion process, the generated multi-view images act as essential guiding conditions. After extracting the 3D mesh from the denoised volume, we further enhance the texture by employing a lightweight optimization with multi-view images as supervision. One-2-3-45++ is capable of producing an initial textured mesh **within 20 seconds** and delivering a refined one in **roughly one minute** using a single A100 GPU.

2.2. Sparse View Reconstruction

While traditional 3D reconstruction methods, such as multi-view stereo or NeRF-based techniques, often demand a dense collection of input images for accurate geometry inference, many of the latest generalizable NeRF solutions [3, 24, 28, 35, 38, 54, 66, 69, 70, 79] strive to learn priors across scenes. This enables them to infer NeRF from a sparse set of images and generalize to novel scenes. These methods typically ingest a few source views as input, leveraging 2D networks to extract 2D features. These pixel features are then unprojected and aggregated into 3D space, facilitating the inference of density (or SDF) and colors. However, these methods may either rely on consistent multi-view images with accurate correspondences or possess limited priors to generalize beyond training datasets.

Recently, some methods [2, 27, 65, 88] have employed diffusion models to aid sparse view reconstruction tasks. However, they generally frame the problem as novel view synthesis, necessitating additional processing, such as distillation using a 3D representation, to generate 3D content. Our work utilizes a multi-view conditioned 3D diffusion model for 3D generation. This model directly learns priors from 3D data and obviates the need for additional post-processing. Moreover, some concurrent works [36, 39, 61] employ NeRF-based per-scene optimization for reconstruction, leveraging specialized loss functions.

3. Method

In traditional game studios, the creation of 3D content encompasses a series of stages, including concept art, 3D modeling, and texturing, etc. Each stage demands distinct and complementary expertise. For instance, concept artists should possess creativity, a vivid imagination, and the skill to visualize 3D assets. In contrast, 3D modelers must be skilled in 3D modeling tools and capable of interpreting and translating multi-view concept drawings into life-like models, even when drawings contain inconsistencies or errors.

One-2-3-45++ aims to harness the rich 2D priors and the valuable yet limited 3D data following a similar philosophy. As shown in Fig. 2, with a single input image of an object, One-2-3-45++ starts by generating coherent multi-view images of the object. This is achieved by finetuning a pre-trained 2D diffusion model and acts akin to the role of a concept artist. These generated images are then input into a multi-view conditioned 3D diffusion model for 3D modeling. The 3D diffusion module, trained on extensive multi-view and 3D pairings, excels at converting multi-view images into 3D meshes. Finally, the produced meshes undergo a lightweight refinement module, guided by the multi-view images, to further enhance the texture quality.

3.1. Consistent Multi-View Generation

Recently, Zero123 has demonstrated the potential of fine-tuning a pretrained 2D diffusion network to incorporate camera view control, thereby synthesizing novel views of an object from a single reference image. While previous studies have employed Zero123 to generate multi-view images, they often suffer from inconsistencies across different views. This inconsistency arises because Zero123 models the conditional marginal distribution for each view in isolation, without considering inter-view communication during multi-view generation. In this work, we present an innovative method to produce consistent multi-view images, significantly benefiting downstream 3D reconstruction.

Multi-View Tiling To generate multiple views in a single diffusion process, we adopt a simple strategy by tiling a sparse set of 6 views into a single image with a 3×2 layout as shown in Fig. 3. Subsequently, we finetune a pre-trained 2D diffusion net to generate the composite image, conditioned on a single input image. This strategy enables multiple views to interact with each other during the diffusion.

It's nontrivial to define the camera poses of the multi-

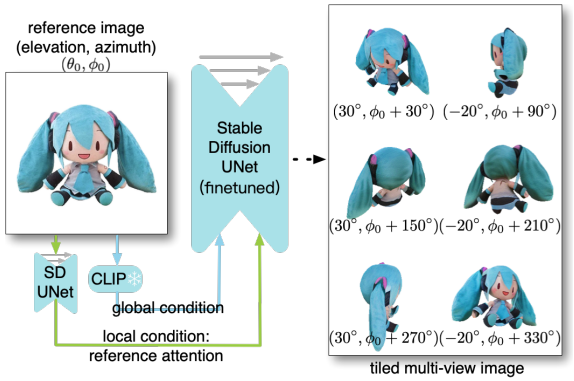

Figure 3. **Consistent multi-view generation:** We stitch multi-view images into a single frame and fine-tune the Stable Diffusion model to generate this composite image, using the input reference image as conditions. We utilize predetermined absolute elevation angles and relative azimuth angles. During 3D reconstruction, we do not need to infer the elevation angle of the input image.

view images. Given that the 3D shapes within the training dataset lack aligned canonical poses, employing absolute camera poses for the multi-view images could lead to ambiguities for the generative model. Alternatively, if we were to set the camera poses relative to the input view, as done in Zero123, downstream applications would then be required to infer the elevation angle of the input image to deduce the camera poses of the multi-view images. This additional step could introduce errors into the pipeline. To address these, we opt for fixed absolute elevation angles paired with relative azimuth angles to define the poses of multi-view images, effectively resolving the orientation ambiguity without necessitating further elevation estimation. To be more precise, the six poses are determined by alternating elevations of $30°$ and $-20°$, coupled with azimuths commencing at $30°$ and incrementing by $60°$ for each subsequent pose, as shown in Fig. 3.

Network and Training Details To fine-tune Stable Diffusion for adding image conditioning and generating coherent multi-view composite images, we employ three crucial network or training designs: (a) **Local Condition**: We adopt the reference attention technique [85] to incorporate the local condition of the image patch features. Specifically, we process the reference input image with the denoising UNet model and append the self-attention key and value matrices of the image tokens from the conditional reference image to the corresponding attention layers of the denoising multi-view image. (b) **Global Condition**: We leverage CLIP image embedding as a global condition, by replacing the text token features originally used in Stable Diffusion with the duplicated CLIP image features. These global image embeddings are multiplied by a set of learnable weights, providing the network with an overall semantic understanding of the object. (c) **Noise Schedule**: The original Stable Diffusion model was trained using a scaled-linear noise schedule. We found it necessary to switch to a linear noise scheme in our fine-tuning process.

We fine-tune the Stable Diffusion2 v-mode using 3D shapes from the Objaverse dataset [11]. For each shape, we generate three data points by randomly sampling the camera pose of the input image from a specified range, and selecting a random HDRI environment lighting from a curated set that offers uniform lighting. Initially, we fine-tuned only the self-attention layers along with the key and value matrices of the cross-attention layers using LoRA [22]. Subsequently, we fine-tuned the entire UNet using a conservative learning rate. The finetuning process was conducted using 16 A100 GPUs and took approximately 10 days.

3.2. 3D Diffusion with Multi-View Condition

While prior work utilizes generalizable NeRF methods for 3D reconstruction, it primarily depends on accurate local correspondence of multi-view images and possesses limited priors for 3D generation. This constrains their effectiveness in lifting intricate and inconsistent multi-view images generated by the 2D diffusion network. Instead, we propose an innovative way to lift the generated multi-view images to 3D by utilizing a multi-view conditioned 3D generative model. It seeks to learn a manifold of plausible 3D shapes conditioned on multi-view images by training expressive 3D native diffusion networks on extensive 3D data.

3D Volume Representations As shown in Fig. 2, we represent a textured 3D shape as two discrete 3D volumes, a signed distance function (SDF) volume, and a color volume. The SDF volume measures the signed distance from the center of each grid cell to the nearest shape surface, while the color volume captures the color of the closest surface points relative to the center of the grid cells. Additionally, we generate a discrete occupancy volume for the 3D shape, where each grid cell stores a binary occupancy based on whether the absolute value of its SDF is below a predefined threshold. The occupancy volume depicts the shell of the 3D shape.

Two-Stage Diffusion Capturing fine-grained details of 3D shapes necessitates the use of high-resolution 3D grids, which unfortunately entail substantial memory and computational costs. We thus follow LAS-Diffusion [87] to generate high-resolution volumes in a coarse-to-fine two-stage manner. Specifically, the initial stage generates a low-resolution (e.g., $n = 64$) full 3D occupancy volume $F \in \mathbb{R}^{n \times n \times n \times 1}$ to approximate the shell of the 3D shape. The second stage then focuses on the occupied shell region only and aims to generate a high-resolution (e.g., 128^3) four-channel sparse volume S, which predicts fine-grained SDF values and color for the sparsely occupied shell region.

We employ a separate diffusion network for each stage. For the first stage, normal 3D convolution is used within the UNet to produce the full 3D occupancy volume F, while for the second stage, we incorporate 3D sparse convolution [62]

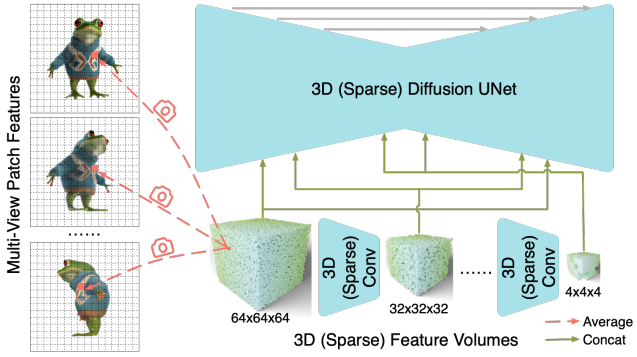

Figure 4. **Multi-view local condition:** We employ a pre-trained 2D backbone to extract 2D patch features for each view. These features are then aggregated using known projection matrices to construct a 3D feature volume. The volume is further processed by 3D convolutional neural networks, resulting in feature volumes of varying resolutions. Subsequently, these volumes are concatenated with the corresponding feature volumes within the diffusion U-Net to guide the 3D diffusion.

in the UNet to yield the 3D sparse volume S. Both diffusion networks are trained using the denoising loss [20]:

$$\mathcal{L}_{\boldsymbol{x}_0} = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,I),t\sim\mathcal{U}(0,1)} \left\| f\left(x_t, t, c\right) - x_0 \right\|_2^2$$

where ϵ and t are sampled noise and time step, x_0 is a data point (F or S) and x_t is its noised version, c is the multi-view condition, and f is the UNet. $\mathcal{N}$ and $\mathcal{U}$ denote Gaussian and uniform distribution, respectively.

Multi-View Condition Training a conventional 3D native diffusion network can be challenging to generalize due to the limited availability of 3D data. However, the use of generated multi-view images can provide a comprehensive guide, greatly simplifying the imagination difficulty of 3D generation. We integrate the multi-view images to guide the diffusion process by initially extracting local image features and subsequently constructing a conditional 3D feature volume, denoted as C. This strategy follows the rationale that local priors facilitate easier generalization [87].

As shown in Fig. 4, given m multi-view images, we first employ a pre-trained 2D backbone, DINOv2, to extract a set of local patch features for each image. We then build a 3D feature volume C by projecting each 3D voxel within the volume onto m multi-view images using the known camera poses. For each 3D voxel, we aggregate m associated 2D patch features through a shared-weight MLP, followed by max pooling. These aggregated features collectively form the feature volume C.

In the diffusion network, the UNet consists of several levels. For example, the occupancy UNet in the initial stage has five levels: 64^3, 32^3, 16^3, 8^3, and 4^3. Initially, we construct a conditional feature volume C that matches the starting resolution, as outlined earlier. A 3D convolution network is then applied to C, producing volumes for the subsequent resolutions. The resultant conditional volumes are then concatenated with the volumes inside the UNet to guide the diffusion process. For the second stage, we construct sparse conditional volumes and utilize 3D sparse convolution. To benefit the diffusion of color volume, we also concatenate 2D pixel-wise projected colors to the final layer of the diffusion UNet. Moreover, we integrate the CLIP feature of the input image as a global condition. For a detailed explanation, please refer to the supplementary materials.

Training and Inference Details We train the two diffusion networks using 3D shapes from the Objaverse dataset [11]. For each 3D shape, we first convert it to a watertight manifold before extracting its SDF volume. We unproject the multi-view renderings of the shape to get a 3D colored point cloud, which is used to build the color volume. During training, we utilize the ground truth renderings to serve as the multi-view conditions. Since two diffusion networks are trained separately, we introduced random perturbations to camera poses and infused random noises to the initial occupancy of the second stage to enhance robustness. We train the two diffusion nets using 8 A100 GPUs for about 10 days for each stage. Please refer to the supplementary materials for more details.

During inference, a 64^3 grid is first initialized with Gaussian noise and then denoised by the first diffusion net. Each predicted occupied voxel is further subdivided into 8 smaller voxels, used to construct a high-resolution sparse volume. The sparse volume is initialized with Gaussian noise and then denoised with the second diffusion net, resulting in predictions for the SDF and color of each voxel. The Marching Cubes algorithm is finally applied to extract a textured mesh.

3.3. Texture Refinement

Given that multi-view images possess higher resolution than the 3D color volume, we can refine the texture of the generated mesh through a lightweight optimization process. To achieve this, we fix the geometry of the generated mesh while optimizing a color field represented by a TensoRF [4]. In each iteration, the generated 3D mesh is rasterized, and the color network is queried to produce 2D renderings. We leverage the predicted consistent multi-view images to guide the texture optimization using a $l2$ loss. Lastly, we bake the optimized color field onto the mesh, with the surface normal serving as the viewing direction.

4. Experiments

4.1. Comparison on Image to 3D

Baselines: We evaluate One-2-3-45++ against both optimization-based and feed-forward methods. Within the optimization-based approaches, our baselines include DreamFusion [49] with Zero123 XL [34] as its backbone, as well as SyncDreamer [36], and DreamGaussian [63]. For feed-forward approaches, we compare with One-2-3-45 [33] and Shap-E [25]. We employ the ThreeStudio [18] implementation for Zero123 XL [18] and the original offi-

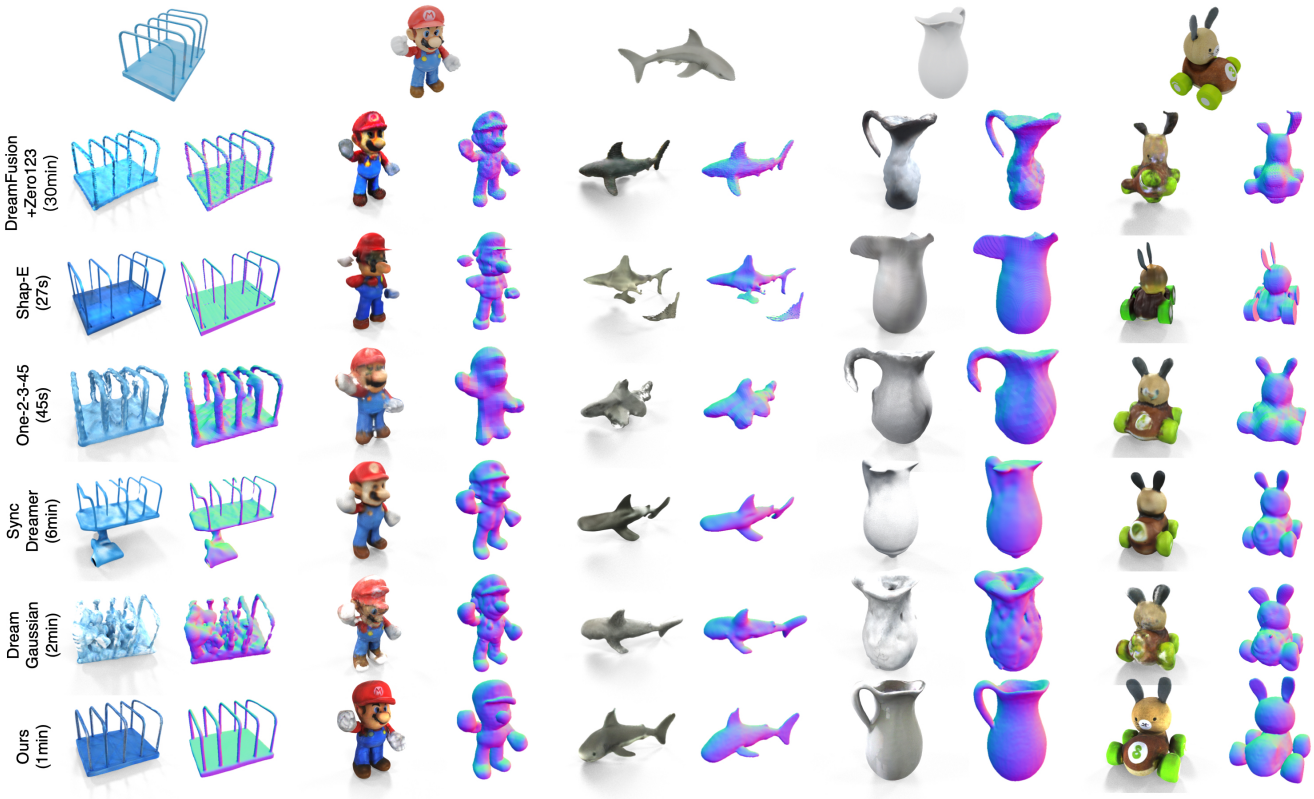

Figure 5. Qualitative results of various single image to 3D approaches. Input images, textured meshes, and normal maps are shown.

Table 1. Comparison on single image to 3D. Evaluated on the GSO [13] dataset, which contains 1,030 3D objects.

Method	F-Sco. (%)↑	CLIP-Sim↑	User-Pref. (%)↑	Time↓
Zero123 XL [10]	91.6	73.1	58.6	30min
One-2-3-45 [33]	90.4	70.8	52.7	45s
SyncDreamer [36]	84.8	68.9	28.4	6min
DreamGaussian [63]	81.0	68.4	31.5	2min
Shap-E [25]	91.8	73.1	40.8	27s
Ours	**93.6**	**81.0**	**87.6**	60s

cial implementations for the other methods.

Dataset and Metrics: We assess the performance of the methods using the entire set of 1,030 shapes from the GSO dataset [13], which were not exposed to any of the methods during training to the best of our knowledge. For each shape, we generate a frontal view image to serve as the input. In line with One-2-3-45 [33], we employ the F-Score and CLIP similarity as our evaluation metrics. The F-Score evaluates the geometric similarity between the predicted mesh and the ground truth mesh. For the CLIP similarity metric, we render 24 different views for each predicted and ground truth mesh, compute the CLIP similarity for each corresponding pair of images, and then average these values across all views. Prior to metric computation, we align the predicted mesh with the ground truth mesh using a combination of linear search and the ICP algorithm.

User Study: A user study was also carried out. For each participant, 45 shapes were randomly selected from the entire GSO dataset, and two methods were randomly sampled for each shape. Participants were asked to choose the result from each pair of comparative outcomes that exhibits superior quality and better aligns with the input image. The preference rate for all methods was then tallied based on these selections. In total, 2,385 evaluated pairs were collected from 53 participants.

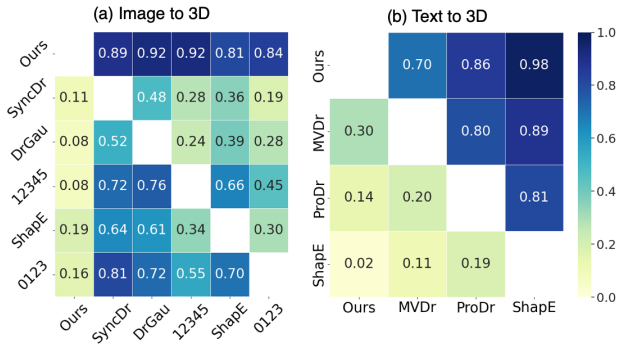

Figure 6. Results of a user study involving 53 participants. Each cell displays the probability or preference rate at which one method (row) outperforms another (column).

Results: As presented in Tab. 1, One-2-3-45++ surpasses all baseline methods regarding F-Score and CLIP similarity. The user preference scores further highlight a significant performance disparity, with our method outperform-

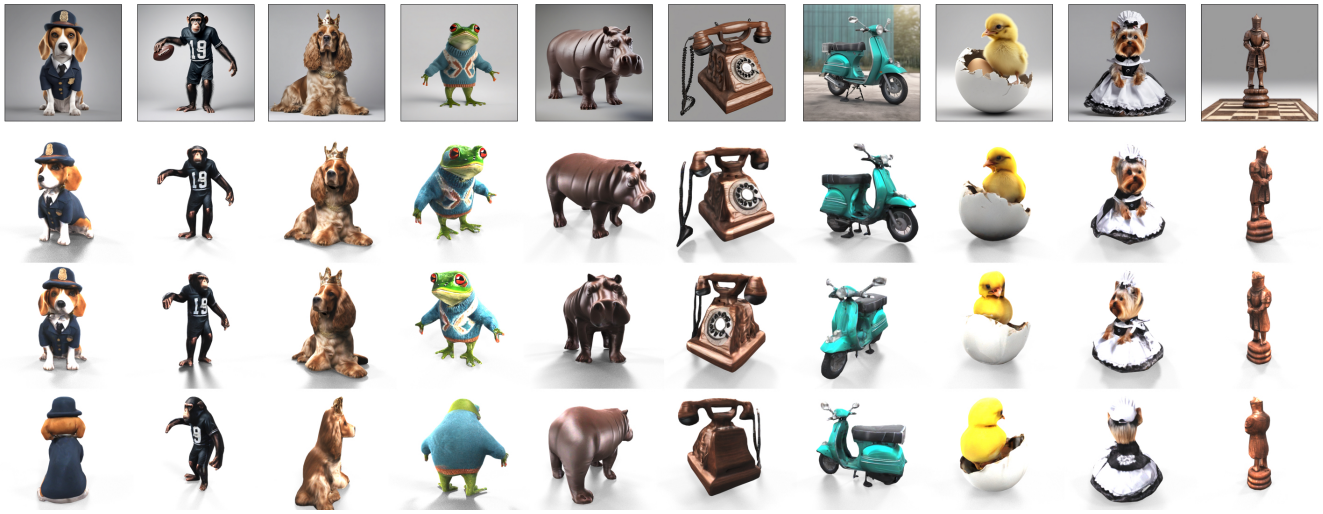

Figure 7. Our qualitative results: top row displays input images; subsequent rows show multi-view renderings of the generated meshes.

Table 2. Quantitative comparison with various text to 3D methods. Evaluated on 50 text prompts from DreamFusion [49].

Method	CLIP-Sim↑	User-Pref.↑	Runtime↓
ProlificDreamer [71]	25.7	39.5	10h+
MVDream [61]	24.8	66.2	2h
Shap-E [25]	22.3	11.1	27s
Ours	**26.8**	**84.1**	60s

ing competing approaches by a substantial margin. Refer to Fig. 6 for an in-depth confusion matrix, which illustrates that One-2-3-45++ outperforms One-2-3-45 92% of the time. Moreover, when compared to optimization-based methods, our approach demonstrates notable runtime advantages. Fig. 5 and 7 show qualitative results.

4.2. Comparison on Text to 3D

Baselines: We compared One-2-3-45++ with optimization-based methods, specifically ProlificDreamer [71] and MVDream [61], as well as a feed-forward approach, Shap-E [25]. For ProlificDreamer, we utilized the ThreeStudio implementation [18], while for the remaining methods, we employed their respective official implementations.

Dataset and Metrics: Given that many baseline approaches necessitate hours to produce a single 3D shape, our evaluation was conducted on 50 text prompts, sampled from DreamFusion [49]. We utilize CLIP similarity, calculated by comparing 24 rendered views of the predicted mesh against the input text prompt and then averaging the similarity scores across all views.

User Study: The user study, akin to the image-to-3D evaluation, involved 30 pairs of outcomes randomly selected for each participant. In total, 1,590 evaluation pairs were collected from 53 participants.

Results: As illustrated in Tab. 2, One-2-3-45++ outperforms all baseline methods in terms of CLIP similarity. This is further corroborated by user preference scores, with our method significantly outshining rival techniques. See Fig. 6 for an in-depth analysis. When directly comparing One-2-3-45++ with the second-best method, MVDream [61], our approach commands a 70% user preference rate. Moreover, while our method delivers prompt results, MVDream [61] requires about 2 hours to generate a single shape. Fig. 8 shows qualitative results.

Table 3. Ablation studies of different modules. Evaluated on the complete GSO [13] dataset. "MultiView", "Reconstruction", and "Texture" indicate multi-view generation, sparse view reconstruction, and texture refinement modules, respectively.

MultiView	Reconstruction	Texture	F-Sc.↑	CLIP-Sim↑	Time↓
Zero123 XL [10]	Ours	w/o	92.9	71.9	14s
Ours	SparseNeuS [38]	w/o	81.2	67.2	15s
Ours	Ours	w/o	**93.6**	73.4	20s
Ours	Ours	w/	**93.6**	**81.0**	60s

4.3. Analyses

Ablation Studies of Overall Pipeline One-2-3-45++ is comprised of three key modules: consistent multi-view generation, multi-view conditioned 3D diffusion, and texture refinement. We conducted ablation studies on these modules using the complete GSO dataset [13], with results detailed in Tab. 3. Replacing our consistent multi-view generation module with Zero123XL [10] led to a noticeable performance decline. Furthermore, substituting our 3D diffusion module with the generalizable NeRF used in One-2-3-45 [33] resulted in an even more significant performance drop. However, the inclusion of our texture refinement module markedly improved texture quality, yielding higher CLIP similarity scores.

Ablation Studies of 3D Diffusion Tab. 4 presents the results of an ablation study of the 3D diffusion module. The study highlights the importance of multi-view images for the module's efficacy. When the module operates without

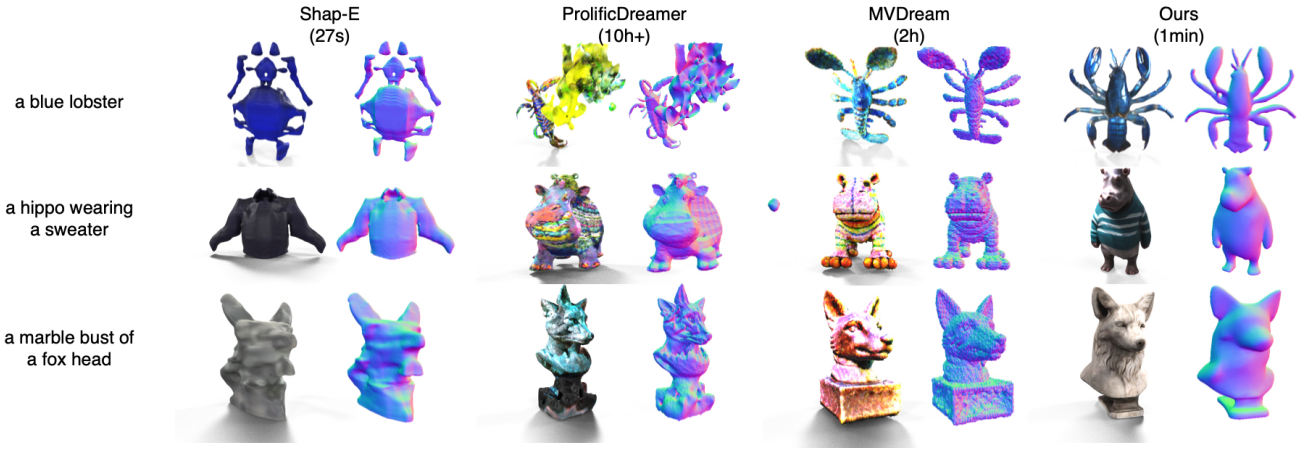

Figure 8. Qualitative results of various text to 3D approaches. Input images, textured meshes, and normal maps are shown.

Table 4. Ablation study of the 3D diffusion module. 3D IoU of the initial-stage occupancy prediction is reported. Note that the 3D IoU is computed for the 3D shell, excluding the solid interior.

id	multi-view cond.	global cond.	image source	proj. perturb.	3D IoU ↑
a	w/o	w/	rendering	N/A	18.3
b	global	w/	rendering	N/A	24.4
c	local	w/o	rendering	w/o	41.4
d	local	w/	prediction	w/o	41.9
e	local	w/	rendering	w/o	44.1
f	local	w/	rendering	w/	45.1

multi-view conditions, relying solely on the global CLIP feature from a single input view (rows a and f), there is a significant decline in performance. Conversely, the One-2-3-45++ approach leverages multi-view local features by constructing a 3D feature volume with known projection matrices. A mere concatenation of global CLIP features from multiple views also impairs performance (rows b and f), underlining the value of multi-view local conditions. Global CLIP features of the input view, however, provide global shape semantics; their removal results in decreased performance (rows c and e). Although One-2-3-45++ uses predicted multi-view images for 3D reconstruction, incorporating these predicted images during training of the 3D diffusion module can lead to a performance downturn (rows d and e) due to the potential mismatch between the predicted multi-view images and actual 3D ground truth meshes. To train the module effectively, we utilize ground truth renderings. Recognizing that predicted multi-view images may be flawed, we introduce random perturbations to projection matrices during training to enhance robustness when processing predicted multi-view images (rows e and f).

Comparison on Multi-View Generation We also evaluate our consistent multi-view generation module against existing approaches, namely Zero123 [34] and its scaled variant [10], alongside two concurrent works: SyncDreamer [36] and Wonder3D [39]. Our comparison utilizes the GSO [13] dataset, where for each object, we render a single input image and task the methods with producing multi-view images. For Zero123 and Zero123 XL, we utilize the same target poses as our approach. However, for Wonder3D and SyncDreamer, we employ the target poses preset by these methods, as they do not support altering camera positions during inference. As presented in Tab. 5, our approach surpasses current methodologies in PSNR, LPIPS, and foreground mask IoU. Notably, Wonder3D [39] employs orthographic projection in its training phase, which compromises its robustness when dealing with perspective images during inference. SyncDreamer [36] only generates views at an elevation of 30°, a simpler setting than ours. Moreover, since these metrics do not assess 3D consistency across views, please refer to supplementary for additional qualitative comparisons and discussions.

Table 5. Comparison of different multi-view generation methods. Evaluated on the complete GSO [13] dataset.

	Target Elevations	PSNR ↑	LPIPS ↓	Mask IoU ↑
Zero123 [34]	30° and −20°	20.32	**0.110**	0.856
Zero123 XL [10]		20.11	0.113	0.869
Ours		**22.12**	**0.110**	**0.878**
SyncDreamer [36]	30°	21.67	0.095	0.894
Wonder3D [39]	0°	18.67	0.130	0.635

5. Conclusion

In this paper, we introduced One-2-3-45++, an innovative approach for transforming a single image of any object into a 3D textured mesh. This method stands out by offering more precise control compared to existing text-to-3D models, and it is capable of delivering high-quality meshes swiftly—typically in under 60 seconds. Additionally, the generated meshes exhibit a high fidelity to the original input image. Looking ahead, there is potential to enhance the robustness and detail of the geometry by incorporating additional guiding conditions from 2D diffusion models, alongside RGB images.

Acknowledgments

We would like to thank Google Cloud and Lambda Labs for their invaluable support in providing computing resources.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 2

[2] 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. Genvs: Generative novel view synthesis with 3d-aware diffusion models, 2023. 3

[3] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 3

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022. 5

[5] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023. 2

[6] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 2

[7] Zilong Chen, Feng Wang, and Huaping Liu. Text-to-3d using gaussian splatting. *arXiv preprint arXiv:2309.16585*, 2023. 2

[8] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023. 2

[9] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer, 2016. 2

[10] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 6, 7, 8

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 4, 5

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20637–20647, 2023. 2

[13] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 6, 7, 8

[14] 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

[15] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 2

[16] 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. *Advances In Neural Information Processing Systems*, 35:31841–31854, 2022. 2

[17] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 2

[18] 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. 5, 7

[19] 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

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[21] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 2

[22] 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. 4

[23] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 2

[24] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022. 3

[25] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 2, 5, 6, 7

[26] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 2

[27] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18423–18433, 2023. 3

[28] Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. Viewformer: Nerf-free neural rendering from few images using transformers. In *European Conference on Computer Vision*, pages 198–216. Springer, 2022. 3

[29] Han-Hung Lee and Angel X Chang. Understanding pure clip guidance for voxel grid nerf models. *arXiv preprint arXiv:2209.15172*, 2022. 2

[30] Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Diffusion-sdf: Text-to-shape via voxelized diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12642–12651, 2023. 2

[31] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 2

[32] Minghua Liu, Minhyuk Sung, Radomir Mech, and Hao Su. Deepmetahandles: Learning deformation meta-handles of 3d meshes with biharmonic coordinates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12–21, 2021. 2

[33] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 2, 5, 6, 7

[34] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 2, 5, 8

[35] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7824–7833, 2022. 3

[36] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 3, 5, 6, 8

[37] Zhen Liu, Yao Feng, Michael J Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling. *arXiv preprint arXiv:2303.08133*, 2023. 2

[38] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *European Conference on Computer Vision*, pages 210–227. Springer, 2022. 2, 3, 7

[39] 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, 2023. 2, 3, 8

[40] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2023. 2

[41] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12923–12932, 2023. 2

[42] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2

[43] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023. 2

[44] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022.

[45] 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 2022 conference papers*, pages 1–8, 2022. 2

[46] Charlie Nash, Yaroslav Ganin, SM Ali Eslami, and Peter Battaglia. Polygen: An autoregressive generative model of 3d meshes. In *International conference on machine learning*, pages 7220–7229. PMLR, 2020. 2

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

[48] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[49] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 5, 7

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

2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 2

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[52] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023. 2

[53] 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 Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[54] Yufan Ren, Tong Zhang, Marc Pollefeys, Sabine Süsstrunk, and Fangjinhua Wang. Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16685–16695, 2023. 3

[55] 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

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[58] 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. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[59] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022. 2

[60] 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

[61] 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, 7

[62] Haotian Tang, Shang Yang, Zhijian Liu, Ke Hong, Zhongming Yu, Xiuyu Li, Guohao Dai, Yu Wang, and Song Han. Torchsparse++: Efficient training and inference framework for sparse convolution on gpus. In *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2023. 4

[63] 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. 2, 5, 6

[64] 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. *arXiv preprint arXiv:2303.14184*, 2023. 2

[65] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Joshua B Tenenbaum, Frédo Durand, William T Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *arXiv preprint arXiv:2306.11719*, 2023. 3

[66] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 3

[67] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 2

[68] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 2

[69] Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, Zhangyang Wang, et al. Is attention all nerf needs? *arXiv preprint arXiv:2207.13298*, 2022. 3

[70] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 3

[71] 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, 7

[72] 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. 2

[73] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9065–9075, 2023. 2

[74] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. *Advances in neural information processing systems*, 30, 2017. 2

[75] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d

reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2690–2698, 2019. 2

[76] 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 360deg views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4479–4489, 2023. 2

[77] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20918, 2023. 2

[78] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in neural information processing systems*, 32, 2019. 2

[79] Hao Yang, Lanqing Hong, Aoxue Li, Tianyang Hu, Zhenguo Li, Gim Hee Lee, and Liwei Wang. Contranerf: Generalizable neural radiance fields for synthetic-to-real novel view synthesis via contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16508–16517, 2023. 3

[80] 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. 2

[81] Chaohui Yu, Qiang Zhou, Jingliang Li, Zhe Zhang, Zhibin Wang, and Fan Wang. Points-to-3d: Bridging the gap between sparse points and shape-controllable text-to-3d generation. *arXiv preprint arXiv:2307.13908*, 2023. 2

[82] 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

[83] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022. 2

[84] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *arXiv preprint arXiv:2301.11445*, 2023. 2

[85] Lyumin Zhang. Reference-only control. https://github.com/Mikubill/sd-webui-controlnet/discussions/1236, 2023. 4

[86] 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. 2

[87] 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. *arXiv preprint arXiv:2305.04461*, 2023. 2, 4, 5

[88] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12588–12597, 2023. 3