Advances in 3D Generation: A Short but Comprehensive Survey

Runpei Dong
University of Illinois Urbana-Champaign
runpeid2@illinois.edu

Jiahua Dong
University of Illinois Urbana-Champaign
jiahuad2@illinois.edu

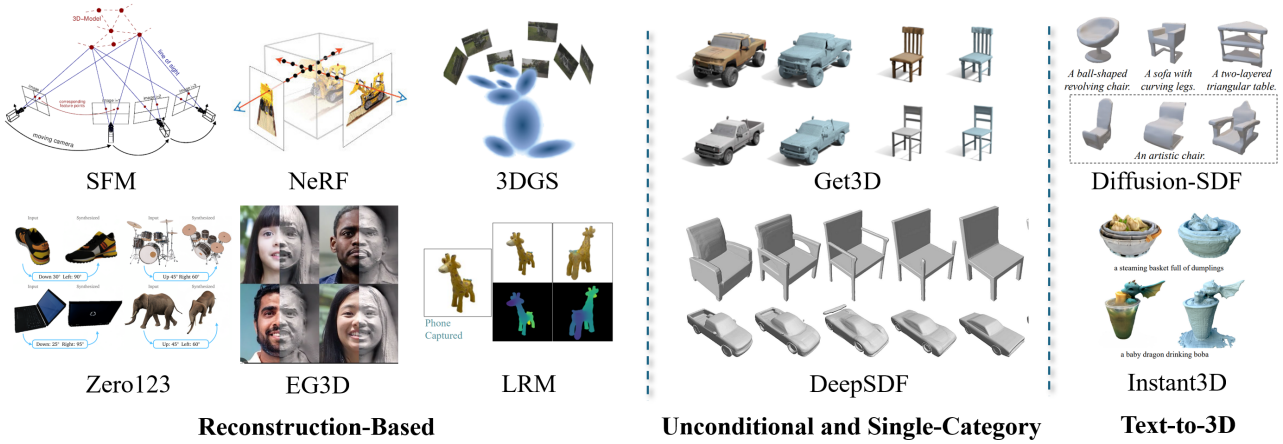

Figure 1. An overview of different 3D generation methods and results.

Abstract

*Generative modeling of 3D has extensive applications in 3D content creation and has become an active area of research. In this paper, we conduct a short but comprehensive survey of 3D generation methods, covering both classical and modern progress. To be specific, we present a novel taxonomy of method categorization and provide a thorough study of **100+ papers**, covering the **foundations**, **techniques**, and **evaluation systems** in 3D generation. In addition, we provide a discussion on promising future research and we hope this survey could spur more related research.*

1. Introduction

3D generation has long been a fundamental focus in computer vision and graphics. This task has gained significant attention due to its wide-ranging applications in video games, films, virtual characters, and immersive experiences, all of which demand a large supply of 3D assets. Moreover, text-to-3D AI tools can empower both beginners and experts, facilitating creative freedom in generating 3D content. Recently, major advancements in 3D content generation have been driven by the success of neural representations, particularly Neural Radiance Fields (NeRFs) [42, 56, 94], and generative models like diffusion models [28, 72].

In this paper, we conduct a short but comprehensive survey of 3D generation methods, showing the roadmap of how modern methods are developed and how results improve. To this end, we first introduce a **novel taxonomy** to categorize 3D generation into 3 groups: i) 3D generation by reconstruction from images, ii) unconditional and single-category generation, and iii) text-to-3D generation. Driven by this roadmap, we provide an in-depth survey of **100+ papers**, covering the major fundamentals, techniques, and evaluation systems in 3D generation. In addition, we highlight several promising future directions and discuss open challenges and we hope this survey could spur future research.

The paper is structured as follows: Sec. 2 covers the preliminaries, including 3D representations and generative models foundations. Sec. 3 present 3D generative methods, focusing on the three groups mentioned before. Sec. 4 shows the review of current and advanced evaluation systems for 3D generation. Sec. 5 concludes the survey and points out the outlooks for future 3D generation studies.

2. Preliminaries

2.1. 3D Representations

Point Clouds Point clouds offer a straightforward approach to representing 3D structures, comprising discrete points in space. This method's popularity stems from its compatibility with data captured by depth sensors. Recent advancements have expanded point cloud capabilities, incorporating additional attributes like color and surface orientation [66]. Innovative rendering techniques, such as differentiable point cloud renderers [105], have emerged, facilitating their integration into machine learning pipelines for 3D content generation.

Meshes Mesh representations define 3D objects through interconnected vertices, edges, and faces. This approach is particularly effective in capturing intricate surface geometries and topological relationships. Due to their efficiency and seamless integration with content creation pipelines, meshes have become the dominant representation in computer graphics. Recent advances in research are pushing the boundaries of mesh generation via neural networks, alongside the development of sophisticated texture mapping techniques [24]. Furthermore, the introduction of differentiable mesh rendering algorithms has established meshes as a key component in AI-driven 3D generation [41].

Voxel Grids Voxel grids discretize 3D space into regular cubic elements, each storing properties such as density or color. This uniform structure integrates naturally with 3D convolutional neural networks, making voxel grids a preferred choice for deep learning applications in 3D vision tasks [53]. Recent advances have mitigated earlier challenges related to memory consumption and computational efficiency by incorporating adaptive resolution techniques and implicit encoding methods [54].

Neural Fields (NeRF) Neural Radiance Fields (NeRF) represent a transformative approach in 3D scene modeling by using neural networks to encode volumetric data [56]. This method maps 3D coordinates and viewing directions to color and density values, allowing for high-quality scene reconstruction and novel view synthesis. NeRF has significantly impacted not only computer graphics but also fields like computer vision and robotics, where accurate 3D scene understanding is crucial [90]. Current research efforts focus on improving NeRF's efficiency, generalization to new scenes, and adaptation to dynamic environments [5].

Gaussian Splatting Gaussian Splatting introduces an innovative method for 3D scene representation by modeling surfaces as collections of 3D Gaussian distributions [42]. This technique effectively bridges the gap between point-based and volumetric representations, offering a balance between rendering speed and detailed scene preservation. Its potential for real-time applications and integration with neural rendering pipelines has made it a focus of ongoing research in AI-generated 3D content [89].

Signed Distance Function (SDF) Signed Distance Functions (SDFs) implicitly define 3D geometries by representing surfaces as the zero-level set of a continuous scalar field. SDFs are particularly effective in capturing smooth surfaces and enabling various geometric operations, such as Boolean operations and collision detection. Recent research has integrated SDFs with neural networks, resulting in significant improvements in shape reconstruction and differentiable rendering [61]. These advancements have driven applications in 3D reconstruction from images, physics-based simulations, and robotics [4].

Hybrid Representation Hybrid representations combine the strengths of multiple 3D modeling approaches to address the limitations of individual methods. These techniques often integrate explicit and implicit representations, enabling more flexible and efficient 3D modeling [62]. Notable examples include multi-planar encodings, which project 3D data onto 2D planes, and tetrahedral decompositions that merge volumetric and surface-based representations [36]. Hybrid approaches hold significant potential for improving the efficiency and quality of 3D generation, particularly in applications that demand detailed geometry and fast rendering [88].

2.2. Generative Models

Generative Adversarial Networks (GANs) Generative Adversarial Networks (GANs) are a class of machine learning models designed for generating realistic data by training two competing neural networks: a generator and a discriminator. The generator produces synthetic data, attempting to mimic real data distributions, while the discriminator seeks to distinguish real data from generated data. The interaction between the two networks forms a min-max optimization problem. GANs have been widely used in applications such as image generation, style transfer, and data augmentation [23]. The loss function in GANs is expressed as:

$$\mathcal{L}_{\text{GAN}} = \min_G \max_D \mathbb{E}_{\mathbf{x}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}}[\log(1 - D(G(\mathbf{z})))] \tag{1}$$

where G is the generator, D is the discriminator, $\mathbf{x}$ is the real data sample, and $\mathbf{z}$ is a random noise vector sampled from a prior distribution.

Variational Autoencoders (VAEs) Variational Autoencoders (VAEs) are generative models that assume the data is

generated by a latent variable model with continuous, multivariate distributions, typically Gaussian. VAEs consist of an encoder that transforms input data into a latent space distribution and a decoder that reconstructs the data from latent variables. The model is trained by maximizing a variational lower bound on the data likelihood, using a combination of reconstruction loss and regularization via the Kullback-Leibler (KL) divergence [43]. The loss function in VAEs is defined as:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \quad (2)$$

where KL represents the Kullback-Leibler divergence, $q_\phi(\mathbf{z}|\mathbf{x})$ is the approximate posterior distribution, and $p(\mathbf{z})$ is the prior distribution, typically assumed to be a standard Gaussian.

Diffusion Models Diffusion models are a class of generative models that learn to reverse a diffusion process, which progressively adds noise to the data. These models are capable of generating high-quality images by learning to denoise data and reverse the added perturbations. The process is typically modeled as a Gaussian process, and during training, the model minimizes the mean squared error between the added noise and the predicted noise [29, 86]. The loss function is given by:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{\mathbf{x},\mathbf{z},t}\left[\|\mathbf{z} - \epsilon_\theta(\mathbf{x}_t, t)\|^2\right] \quad (3)$$

where $\mathbf{x}_t$ is the noisy data at timestep t, $\mathbf{z}$ is the added noise, and ϵ_θ is the model's predicted noise for the reverse diffusion process.

3. 3D Generation Methods

3.1. 3D Generation by Reconstruction from Images

Compared to other 3D representations like SDF [14] and NeRF [56], 3D meshes remain the most utilized representation in modern 3D graphics tools (*e.g.*, Blender, Meshlab, and Maya). Reconstructing high-quality 3D content from single or multiple images is a challenging and fundamental task in graphics and 3D computer vision.

3.1.1 Multi-View Reconstruction

Driven by a complex photogrammetry and geometry objective, early approaches typically reconstruct 3D contents with a decoupled pipeline consisting of Structure-from-Motion (SfM) [2, 77, 85] and Multi-View Stereo (MVS) [22, 78] using depth maps [40] or 3D volumes [44]. A representative system that lies in this group is COLMAP [77, 78], which fuses 3D models from estimated multi-view depth maps. Since 2020, with the development of NeRF [56], great efforts have been devoted to 3D reconstruction by recovering

Figure 2. 3DGP [83] results.

the implicit scene representation using volume rendering. Notably, NeuS [94] proposes to improve NeRF by replacing 3D representation with SDF and achieves remarkable reconstruction quality. 3D Gaussian Splatting [42] sets another milestone that achieves higher reconstruction quality by an anisotropic 3D Gaussians neural field representation. Meanwhile, another line of work learns to reconstruct 3D from multi-view images captured in the temporal domain (*i.e.*, monocular videos), and impressive results have been obtained [25, 33].

3.1.2 Single-View Reconstruction

Different from multi-view reconstruction, a more challenging and ill-posed problem lies in 3D reconstruction/generation from a single-view observation. Notably, early approaches have been developed based on a shape or semantic prior [15, 106], driven by 3D primitives [59, 71]. Another line of works proposes to solve 3D volumetric reconstruction with space carving [44, 79, 84]. Further, Prisacariu et al. proposes to utilize mixed supervision of segmentation, pose, and geometry at the same time.

More recently, significant efforts have been devoted to learning single-view 3D reconstruction based on geometry-aware 3D deep learning. 3D-R2N2 [13] pioneers this direction by training image-conditional 3D generation on ShapeNet [8]. Fan et al. propose a novel Chamfer Distance that significantly improves 3D generation quality from images. Pixel2Mesh [93] proposes to use a multi-scale mesh generation method that achieves much better generation quality. AutoSDF [57] proposes to learn SDFs guided by a 3D shape prior. Alwala et al. propose a method that leverages pre-training and distillation from models trained on large image datasets. Wu et al. propose to learn unsupervised 3D reconstruction based on a symmetry prior. EG3D [7] proposes to learn a 3D GAN network driven by geometry guidance. EVA3D [30] proposes to learn a human 3D GAN network that targets the problem of 3D human generation. Skorokhodov et al. and Sargent et al. further improve the generalization of single-view reconstruction on the ImageNet dataset [18].

However, a critical issue of lacking data for a 3D generation still remains. To address this, Objaverse [17], ObjaverseXL [16], and MVImgnet [107] have been proposed.

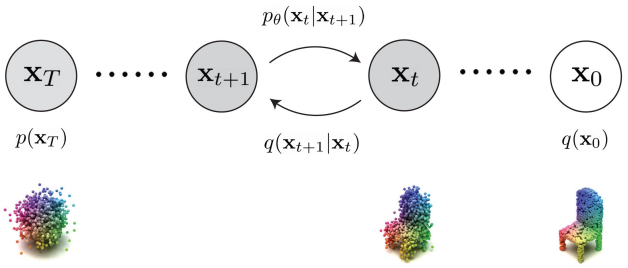

Figure 3. Diffusion process on 3D point clouds [113].

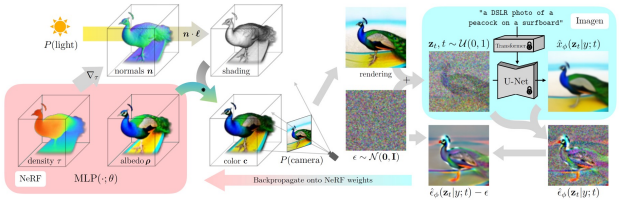

Figure 4. Score distillation process of DreamFusion [67].

Afterward, a lot of works like LRM [31], Unique3D [99], and CRM [96] have been developed based on scaling up the training data to Objaverse-level datasets.

3.2. Unconditional and Single-Category Generation

Before the booming era of conditional 3D generation that typically requires scaled-up training on broad datasets or the help of 2D or language foundation models, there are a lot of research working on unconditional and single-category 3D generation. These works typically utilize ShapeNet [8] and ModelNet [102] as the training corpora and a generative model is trained for each category. The mainstream methods lie in this direction can be divided into three groups: **i) GAN-based methods** that train a 3D GAN network that is guided by 3D geometry prior or distance learning [1, 11, 34, 98]; **ii) Diffusion-based methods** that learn a 3D diffusion network on representations such as SDF [46, 112], point clouds [32, 108], or hybrid representations [113]. **iii) Implicit representation-based methods** that train to learn a latent implicit neural representation like occupancy fields [55, 63], latent grids [37, 109], and deep implicit functions [12, 52, 64].

3.3. Text-to-3D Generation

Conditional 3D synthesis is now becoming one of the most commonly used techniques for creative content generation that follows human's fabulous imaginations through free-form languages. [9] introduced a text-shape paired dataset based on ShapeNet [8] focusing on chair and table categories, paving the way for text-conditioned shape generation for specific object types. Building on this human-curated 3D data, ShapeCrafter [21] expanded the approach by enabling recursive text-guided shape editing. However, methods like ShapeCrafter are not generalizable to arbitrary language inputs.

3.3.1 CLIP-based Methods

CLIP To address the limitation of non-generalization of single-category 3D generation, one line of research has proposed to utilize 2D foundation models like CLIP [70] for open-world 3D generation [19, 110]. This line of methods leverages the pretrained property of CLIP, *i.e.*, the semantic alignment of 2D images and languages in CLIP during the contrastive pretraining [6, 70].

CLIP-Mesh and CLIP-Forge [74] are pioneering works that first leverage the easily rendered 2D images and train a 3D generative model conditioned the CLIP image features, and then during inference, the CLIP text features replace the image features and serve as the text condition. Following these two methods, CLIP-Sculptor [75] and ISS [51] are proposed by improving the text-image alignment for generation conditioning. Besides rendered images, Point-E [60], Shape-E [39], and VPP [69] propose to use real-world text-image-shape triplets from a large-scale in-house dataset, where images are used as the representation bridge.

3.3.2 Learning from 2D Prior with Score Distillation

However, collecting 3D data is both scarce and costly [19]. DreamFields [35] pioneered text-only training using NeRF [56], where 2D rendered images are weakly aligned with text inputs via CLIP.

Score Distillation Following DreamFields, DreamFusion [67] incorporates distillation loss from a pretrained diffusion model [72], which is called *score distillation*. The overall process is shown in Fig. 4. To be specific, DreamFusion learns a parameterized NeRF by randomly sampling rendered images and trains it by using them as the input of the score distillation loss. Score distillation learns by distilling from 2D diffusion models, which can be viewed as pretrained score functions [28]. Formally, let θ be the NeRF parameters and g be the differentiable rendering transformation [58], the gradient of the score distillation loss $\nabla_\theta \mathcal{L}_{\text{Diff}}$ can be given as,

$$\nabla_\theta \mathcal{L}_{\text{Diff}}(\phi, \mathbf{x} = g(\theta)) = \mathbb{E}_{t,\epsilon}\left[w(t) \underbrace{\left(\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) - \epsilon\right)}_{\text{Noise Residual}} \underbrace{\frac{\partial \hat{\epsilon}_\phi(\mathbf{z}_t; y, t)}{\mathbf{z}_t}}_{\text{U-Net Jacobian}} \underbrace{\frac{\partial \mathbf{x}}{\partial \theta}}_{\text{Generator Jacobian}} \right], \tag{4}$$

where $\mathbf{z}_t$ denotes the noise latents, t is the timestamp, ϵ is the noise, $\mathbf{x}$ is the data point, and $\hat{\epsilon}_\phi$ is the U-Net parameterized with ϕ. It can be proven that optimizing Eq. (4) is equivalent

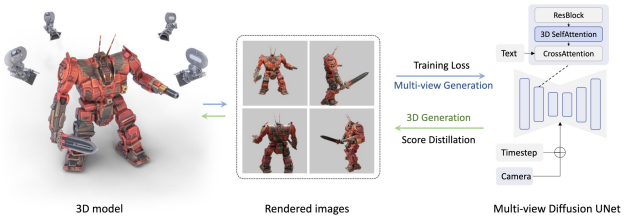

Figure 5. Multi-view diffusion score distillation [81].

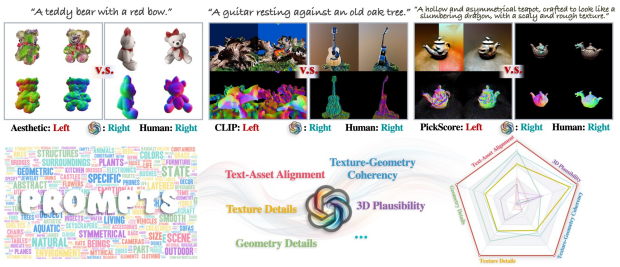

Figure 6. GPT-based human-aligned 3D benchmark [101].

to optimizing a denoising score matching process with the pretrained score function (*i.e.*, pretrained diffusion model).

Following DreamFusion, Dream3D [103] improves shape initialization using prompt-based guidance. Despite their impressive results, these methods incur high computational costs due to case-specific NeRF training at inference time. Recently, TAPS3D utilizes DMTet [80] to align rendered images with text prompts for conditional generation, though its generalizability is restricted to only four categories. 3DGen [26] achieves high-quality results by employing Triplane-based diffusion models [7, 82] and the Objaverse dataset [17]. Zero-1-to-3 [48] introduces relative camera viewpoints as conditions for diffusion, leading to One-2-3-45 [47], which achieved rapid open-vocabulary mesh generation through multi-view synthesis. Fantasia3D [10] uses DMTet for differentiable SDF rendering, offering a more direct alternative to NeRF-based mesh generation. SJC [92] and ProlificDreamer [95] achieve remarkable results by introducing variational diffusion distillation.

3.3.3 Hierarchical Generation

Besides the previously mentioned family of methods, a lot of approaches have emerged by using a hierarchical generation method that combines methods like multi-view image generation and multi-view reconstruction. Notable efforts are as follows, SyncDreamer [50], MVDream [81], DreamComposer [104], and HarmonyView [97] propose to first generate multi-view images using a multi-view diffusion model and then reconstruct 3D by multi-view fusion; SV3D [91] and Animate3D [38] propose to first generate a video and then fuse video to a 3D object.

4. Benchmarking 3D Generation

4.1. Traditional Metrics

CLIP Score [70] is used to measure the text-image consistency, which is by directly computing the feature distance.

Inception Score (IS) [73] is used to measure the image quality based on a statistic of an ImageNet-pre-trained Inception V3 network [87] outputs when applied to generated images.

Fréchet Inception Distance (FID) [27] is another metric that measures the image quality by computing the feature distribution statistics between the generated images and the ground truth images.

Human Evaluation is to conduct benchmarks and involve human evaluators to judge the comprehensive performance of the generated results [45].

4.2. GPT-based Human-aligned Evaluation

Though effective as the traditional metrics to some extent, similar to the evaluation of image generation, benchmarking 3D generation is a long-standing challenge. The key problem lies in the probable misalignment with human evaluations. However, utilizing human evaluations is both time-consuming and expensive. To address this issue, recent efforts have been devoted to design human-aligned 3D benchmarks using GPT models [65, 101, 111].

5. Conclusions and Outlooks

This paper presents a comprehensive review of the advances in 3D generation. We first introduce the background of 3D representations and generative algorithms and then show the developments of both classical and modern 3D generation methods. In addition, we point out how the methods are progressively developed, and we hope this survey could spur more research in the community.

Outlooks In the future, there are some possible directions and trends that should be followed, **i) 3D generation with video generation foundation models.** There are an increasing number of works in physical-driven video generation [49], which could be very helpful for 3D geometry consistency and generation. **ii) 4D generation and its applications.** Another trending research direction is to generate 3D content with dynamics, which is very helpful for applications such as movie production and robotics.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas J. Guibas. Learning representations and generative models for 3d point clouds. In *Int. Conf. Mach. Learn. (ICML)*, 2018. 4

[2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M. Seitz, and Richard Szeliski. Building rome in a day. *Commun. ACM*, 54(10):105–112, 2011. 3

[3] Kalyan Vasudev Alwala, Abhinav Gupta, and Shubham Tulsiani. Pretrain, self-train, distill: A simple recipe for supersizing 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3763–3772. IEEE, 2022. 3

[4] Matan Atzmon and Yaron Lipman. Sald: Sign agnostic learning with derivatives. *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020. 2

[5] 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. 2

[6] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. 4

[7] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 3, 5

[8] 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, 4

[9] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian Conf. Comput. Vis. (ACCV)*, 2019. 4

[10] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Int. Conf. Comput. Vis. (ICCV)*, 2023. 5

[11] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019. 4

[12] Zhang Chen, Yinda Zhang, Kyle Genova, Sean Ryan Fanello, Sofien Bouaziz, Christian Häne, Ruofei Du, Cem Keskin, Thomas A. Funkhouser, and Danhang Tang. Multiresolution deep implicit functions for 3d shape representation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 13067–13076. IEEE, 2021. 4

[13] 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 *Eur. Conf. Comput. Vis. (ECCV)*, 2016. 3

[14] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996*, pages 303–312. ACM, 1996. 3

[15] Amaury Dame, Victor A. Prisacariu, Carl Y. Ren, and Ian Reid. Dense reconstruction using 3d object shape priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 3

[16] 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

[17] 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 Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 3, 5

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2009. 3

[19] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 4

[20] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017. 3

[21] Rao Fu, Xiao Zhan, Yiwen Chen, Daniel Ritchie, and Srinath Sridhar. Shapecrafter: A recursive text-conditioned 3d shape generation model. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 4

[22] Yasutaka Furukawa and Carlos Hernández. Multi-view

stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 3

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 27, 2014. 2

[24] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 216–224, 2018. 2

[25] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023. 3

[26] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oguz. 3dgen: Triplane latent diffusion for textured mesh generation. *CoRR*, abs/2303.05371, 2023. 5

[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Adv. Neural Inform. Process. Syst. (NIPS)*, 2017. 5

[28] 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. 1, 4

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 33:6840–6851, 2020. 3

[30] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. EVA3D: compositional 3d human generation from 2d image collections. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3

[31] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: large reconstruction model for single image to 3d. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 4

[32] Jingyu Hu, Ka-Hei Hui, Zhengzhe Liu, Ruihui Li, and Chi-Wing Fu. Neural wavelet-domain diffusion for 3d shape generation, inversion, and manipulation. *ACM Trans. Graph.*, 43(2):16:1–16:18, 2024. 4

[33] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, New York, NY, USA, 2022. Association for Computing Machinery. 3

[34] Moritz Ibing, Isaak Lim, and Leif Kobbelt. 3d shape generation with grid-based implicit functions. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021. 4

[35] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 4

[36] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 6001–6010, 2020. 2

[37] Chiyu Max Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas A. Funkhouser. Local implicit grid representations for 3d scenes. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6000–6009. Computer Vision Foundation / IEEE, 2020. 4

[38] Yanqin Jiang, Chaohui Yu, Chenjie Cao, Fan Wang, Weiming Hu, and Jin Gao. Animate3d: Animating any 3d model with multi-view video diffusion. *arXiv preprint arXiv:2407.11398*, 2024. 5

[39] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *CoRR*, abs/2305.02463, 2023. 4

[40] Sing Bing Kang, Richard Szeliski, and Jinxiang Chai. Handling occlusions in dense multi-view stereo. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pages I–I. IEEE, 2001. 3

[41] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3907–3916, 2018. 2

[42] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. 1, 2, 3

[43] Diederik P Kingma. Auto-encoding variational bayes. *Int. Conf. Learn. Represent. (ICLR)*, 2011. 3

[44] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal of computer vision*, 38:199–218, 2000. 3

[45] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Fei-Fei Li, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic evaluation of text-to-image models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 5

[46] Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Diffusion-sdf: Text-to-shape via voxelized diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 12642–12651. IEEE, 2023. 4

[47] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *CoRR*, abs/2306.16928, 2023. 5

[48] 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 *Int. Conf. Comput. Vis. (ICCV)*, 2023. 5

[49] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *Eur. Conf. Comput. Vis. (ECCV)*, 2024. 5

[50] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 5

[51] Zhengzhe Liu, Peng Dai, Ruihui Li, Xiaojuan Qi, and Chi-Wing Fu. Iss: Image as stetting stone for text-guided 3d shape generation. In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 4

[52] Luca De Luigi, Adriano Cardace, Riccardo Spezialetti, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Deep learning on implicit neural representations of shapes. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 4

[53] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IEEE/RSJ Int. Conf. Intell. Robot. and Syst. (IROS)*, pages 922–928, 2015. 2

[54] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4460–4470, 2019. 2

[55] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019. 4

[56] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2, 3, 4

[57] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 3

[58] Alexander Mordvintsev, Nicola Pezzotti, Ludwig Schubert, and Chris Olah. Differentiable image parameterizations. *Distill*, 3(7):e12, 2018. 4

[59] Ramakant Nevatia and Thomas O Binford. Description and recognition of curved objects. *Artificial intelligence*, 8(1): 77–98, 1977. 3

[60] 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. 4

[61] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 165–174, 2019. 2

[62] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 523–540. Springer, 2020. 2

[63] Songyou Peng, Michael Niemeyer, Lars M. Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, pages 523–540. Springer, 2020. 4

[64] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9054–9063. Computer Vision Foundation / IEEE, 2021. 4

[65] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024. 5

[66] Hanspeter Pfister, Matthias Zwicker, Jeroen van Baar, and Markus H. Gross. Surfels: surface elements as rendering primitives. *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000. 2

[67] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 4

[68] Victor Adrian Prisacariu, Aleksandr V Segal, and Ian Reid. Simultaneous monocular 2d segmentation, 3d pose recovery and 3d reconstruction. In *Asian conference on computer vision*, pages 593–606. Springer, 2012. 3

[69] Zekun Qi, Muzhou Yu, Runpei Dong, and Kaisheng Ma. VPP: efficient conditional 3d generation via voxel-point progressive representation. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 4

[70] 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 *Int. Conf. Mach. Learn. (ICML)*, 2021. 4, 5

[71] Lawrence G Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963. 3

[72] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 1, 4

[73] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Adv. Neural Inform. Process. Syst. (NIPS)*, 2016. 5

[74] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malek-

shan. Clip-forge: Towards zero-shot text-to-shape generation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 4

[75] Aditya Sanghi, Rao Fu, Vivian Liu, Karl Willis, Hooman Shayani, Amir Hosein Khasahmadi, Srinath Sridhar, and Daniel Ritchie. Clip-sculptor: Zero-shot generation of high-fidelity and diverse shapes from natural language. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 4

[76] Kyle Sargent, Jing Yu Koh, Han Zhang, Huiwen Chang, Charles Herrmann, Pratul Srinivasan, Jiajun Wu, and Deqing Sun. VQ3D: learning a 3d-aware generative model on imagenet. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 4217–4227. IEEE, 2023. 3

[77] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4104–4113, 2016. 3

[78] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision – ECCV 2016*, pages 501–518, Cham, 2016. Springer International Publishing. 3

[79] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International journal of computer vision*, 35:151–173, 1999. 3

[80] 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 *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2021. 5

[81] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 5

[82] J. Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 5

[83] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3d generation on imagenet. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3

[84] Gregory G Slabaugh, W Bruce Culbertson, Thomas Malzbender, Mark R Stevens, and Ronald W Schafer. Methods for volumetric reconstruction of visual scenes. *International Journal of Computer Vision*, 57:179–199, 2004. 3

[85] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM SIGGRAPH 2006 Papers*, page 835–846, New York, NY, USA, 2006. Association for Computing Machinery. 3

[86] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Int. Conf. Mach. Learn. (ICML)*, pages 2256–2265. PMLR, 2015. 3

[87] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016. 5

[88] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 11358–11367, 2021. 2

[89] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *Int. Conf. Learn. Represent. (ICLR)*, 2023. 2

[90] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, pages 703–735. Wiley Online Library, 2022. 2

[91] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024. 5

[92] 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 Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 5

[93] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single RGB images. In *Eur. Conf. Comput. Vis. (ECCV)*, 2018. 3

[94] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27171–27183, 2021. 1, 3

[95] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 5

[96] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *Eur. Conf. Comput. Vis. (ECCV)*, 2024. 4

[97] Sangmin Woo, Byeongjun Park, Hyojun Go, Jin-Young Kim, and Changick Kim. Harmonyview: Harmonizing consistency and diversity in one-image-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10574–10584, 2024. 5

[98] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Adv. Neural Inform. Process. Syst. (NIPS)*, 2016. 4

[99] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *CoRR*, abs/2405.20343, 2024. 4

[100] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild (invited paper). *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):5268–5281, 2023. 3

[101] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas J. Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22227–22238. IEEE, 2024. 5

[102] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2015. 4

[103] 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 *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 5

[104] Yunhan Yang, Yukun Huang, Xiaoyang Wu, Yuan-Chen Guo, Song-Hai Zhang, Hengshuang Zhao, Tong He, and Xihui Liu. Dreamcomposer: Controllable 3d object generation via multi-view conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8111–8120, 2024. 5

[105] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Trans. Graph.*, 38 (6):1–14, 2019. 2

[106] Sid Yingze Bao, Manmohan Chandraker, Yuanqing Lin, and Silvio Savarese. Dense object reconstruction with semantic priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 3

[107] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Mvimgnet: A large-scale dataset of multi-view images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 9150–9161. IEEE, 2023. 3

[108] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. LION: latent point diffusion models for 3d shape generation. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 4

[109] Biao Zhang, Matthias Nießner, and Peter Wonka. 3dilg: Irregular latent grids for 3d generative modeling. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 4

[110] Junbo Zhang, Runpei Dong, and Kaisheng Ma. Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2048–2059, 2023. 4

[111] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks. *CoRR*, abs/2311.01361, 2023. 5

[112] 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 Trans. Graph.*, 42(4):91:1–91:13, 2023. 4

[113] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Int. Conf. Comput. Vis. (ICCV)*, 2021. 4