

Advances in 3D Generation: A Survey

Xiaoyu Li^{1*} Qi Zhang^{1*} Di Kang¹ Weihao Cheng² Yiming Gao²
 Jingbo Zhang³ Zhihao Liang⁴ Jing Liao³ Yan-Pei Cao^{1,2} Ying Shan^{1,2†}

¹Tencent AI Lab ²ARC Lab, Tencent PCG ³City University of Hong Kong ⁴South China University of Technology

*Equal contribution. †Corresponding author.

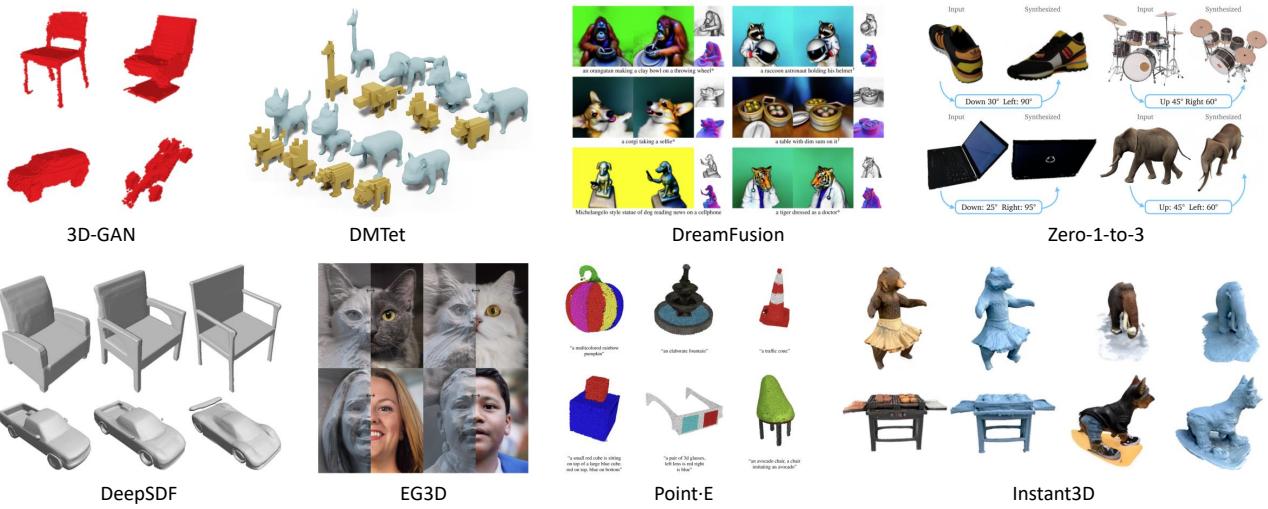


Figure 1: In this survey, we investigate a large variety of 3D generation methods. Over the past decade, 3D generation has achieved remarkable progress and has recently garnered considerable attention due to the success of generative AI in images and videos. 3D generation results from 3D-GAN [WZX*16], DeepSDF [PFS*19], DMTet [SGY*21], EG3D [CLC*22], DreamFusion [PJBM23], PointE [NJD*22], Zero-1-to-3 [LWH*23] and Instant3D [LTZ*23].

Abstract

Generating 3D models lies at the core of computer graphics and has been the focus of decades of research. With the emergence of advanced neural representations and generative models, the field of 3D content generation is developing rapidly, enabling the creation of increasingly high-quality and diverse 3D models. The rapid growth of this field makes it difficult to stay abreast of all recent developments. In this survey, we aim to introduce the fundamental methodologies of 3D generation methods and establish a structured roadmap, encompassing 3D representation, generation methods, datasets, and corresponding applications. Specifically, we introduce the 3D representations that serve as the backbone for 3D generation. Furthermore, we provide a comprehensive overview of the rapidly growing literature on generation methods, categorized by the type of algorithmic paradigms, including feedforward generation, optimization-based generation, procedural generation, and generative novel view synthesis. Lastly, we discuss available datasets, applications, and open challenges. We hope this survey will help readers explore this exciting topic and foster further advancements in the field of 3D content generation.

1. Introduction

Automatically generating 3D models using algorithms has long been a significant task in computer vision and graphics. This task has garnered considerable interest due to its broad applications in video games, movies, virtual characters, and immersive expe-

riences, which typically require a wealth of 3D assets. Recently, the success of neural representations, particularly Neural Radiance Fields (NeRFs) [MST*20, BMT*21, MESK22, KKLD23], and generative models such as diffusion models [HJA20, RBL*22a], has led to remarkable advancements in 3D content generation.

In the realm of 2D content generation, recent advancements in generative models have steadily enhanced the capacity for image generation and editing, leading to increasingly diverse and high-quality results. Pioneering research on generative adversarial networks (GANs) [GPAM^{*}14, AQW19], variational autoencoders (VAEs) [KPHL17, PGH^{*}16, KW13], and autoregressive models [RWC^{*}19, BMR^{*}20] has demonstrated impressive outcomes. Furthermore, the advent of generative artificial intelligence (AI) and diffusion models [HJA20, ND21, SCS^{*}22] signifies a paradigm shift in image manipulation techniques, such as Stable Diffusion [RBL^{*}22a], Imagen [SCS^{*}22], Midjourney [Mid], or DALL-E 3 [Ope]. These generative AI models enable the creation and editing of photorealistic or stylized images, or even videos [CZC^{*}24, HSG^{*}22, SPH^{*}23, GNL^{*}23], using minimal input like text prompts. As a result, they often generate imaginative content that transcends the boundaries of the real world, pushing the limits of creativity and artistic expression. Owing to their “emergent” capabilities, these models have redefined the limits of what is achievable in content generation, expanding the horizons of creativity and artistic expression.

The demand to extend 2D content generation into 3D space is becoming increasingly crucial for applications in generating 3D assets or creating immersive experiences, particularly with the rapid development of the metaverse. The transition from 2D to 3D content generation, however, is not merely a technological evolution. It is primarily a response to the demands of modern applications that necessitate a more intricate replication of the physical world, which 2D representations often fail to provide. This shift highlights the limitations of 2D content in applications that require a comprehensive understanding of spatial relationships and depth perception.

As the significance of 3D content becomes increasingly evident, there has been a surge in research efforts dedicated to this domain. However, the transition from 2D to 3D content generation is not a straightforward extension of existing 2D methodologies. Instead, it involves tackling unique challenges and re-evaluating data representation, formulation, and underlying generative models to effectively address the complexities of 3D space. For instance, it is not obvious how to integrate the 3D scene representations into 2D generative models to handle higher dimensions, as required for 3D generation. Unlike images or videos which can be easily collected from the web, 3D assets are relatively scarce. Furthermore, evaluating the quality of generated 3D models presents additional challenges, as it is necessary to develop better formulations for objective functions, particularly when considering multi-view consistency in 3D space. These complexities demand innovative approaches and novel solutions to bridge the gap between 2D and 3D content generation.

While not as prominently featured as its 2D counterpart, 3D content generation has been steadily progressing with a series of notable achievements. The representative examples shown in Fig. 1 demonstrate significant improvements in both quality and diversity, transitioning from early methods like 3D-GAN [WZX^{*}16] to recent approaches like Instant3D [LTZ^{*}23]. Therefore, This survey paper seeks to systematically explore the rapid advancements and multifaceted developments in 3D content generation. We present a structured overview and comprehensive roadmap of the many re-

cent works focusing on 3D representations, 3D generation methods, datasets, and applications of 3D content generation, and to outline open challenges.

Fig. 2 presents an overview of this survey. We first discuss the scope and related work of this survey in Sec. 2. In the following sections, we examine the core methodologies that form the foundation of 3D content generation. Sec. 3 introduces the primary scene representations and their corresponding rendering functions used in 3D content generation. Sec. 4 explores a wide variety of 3D generation methods, which can be divided into four categories based on their algorithmic methodologies: feedforward generation, optimization-based generation, procedural generation, and generative novel view synthesis. An evolutionary tree of these methodologies is also depicted to illustrate their primary branch. As data accumulation plays a vital role in ensuring the success of deep learning models, we present related datasets employed for training 3D generation methods. In the end, we include a brief discussion on related applications, such as 3D human and face generation, outline open challenges, and conclude this survey. We hope this survey offers a systematic summary of 3D generation that could inspire subsequent work for interested readers.

In this work, we present a comprehensive survey on 3D generation, with two main contributions:

- Given the recent surge in contributions based on generative models in the field of 3D vision, we provide a comprehensive and timely literature review of 3D content generation, aiming to offer readers a rapid understanding of the 3D generation framework and its underlying principles.
- We propose a multi-perspective categorization of 3D generation methods, aiming to assist researchers working on 3D content generation in specific domains to quickly identify relevant works and facilitate a better understanding of the related techniques.

2. Scope of This Survey

In this survey, we concentrate on the techniques for the generation of 3D models and their related datasets and applications. Specifically, we first give a short introduction to the scene representation. Our focus then shifts to the integration of these representations and the generative models. Then, we provide a comprehensive overview of the prominent methodologies of generation methods. We also explore the related datasets and cutting-edge applications such as 3D human generation, 3D face generation, and 3D editing, all of which are enhanced by these techniques.

This survey is dedicated to systematically summarizing and categorizing 3D generation methods, along with the related datasets and applications. The surveyed papers are mostly published in major computer vision and computer graphics conferences/journals as well as some preprints released on arXiv in 2023. While it's challenging to exhaust all methods related to 3D generation, we hope to include as many major branches of 3D generation as possible. We do not delve into detailed explanations for each branch, instead, we typically introduce some representative works within it to explain its paradigm. The details of each branch can be found in the related work section of these cited papers.

Related Survey. Neural reconstruction and rendering with scene

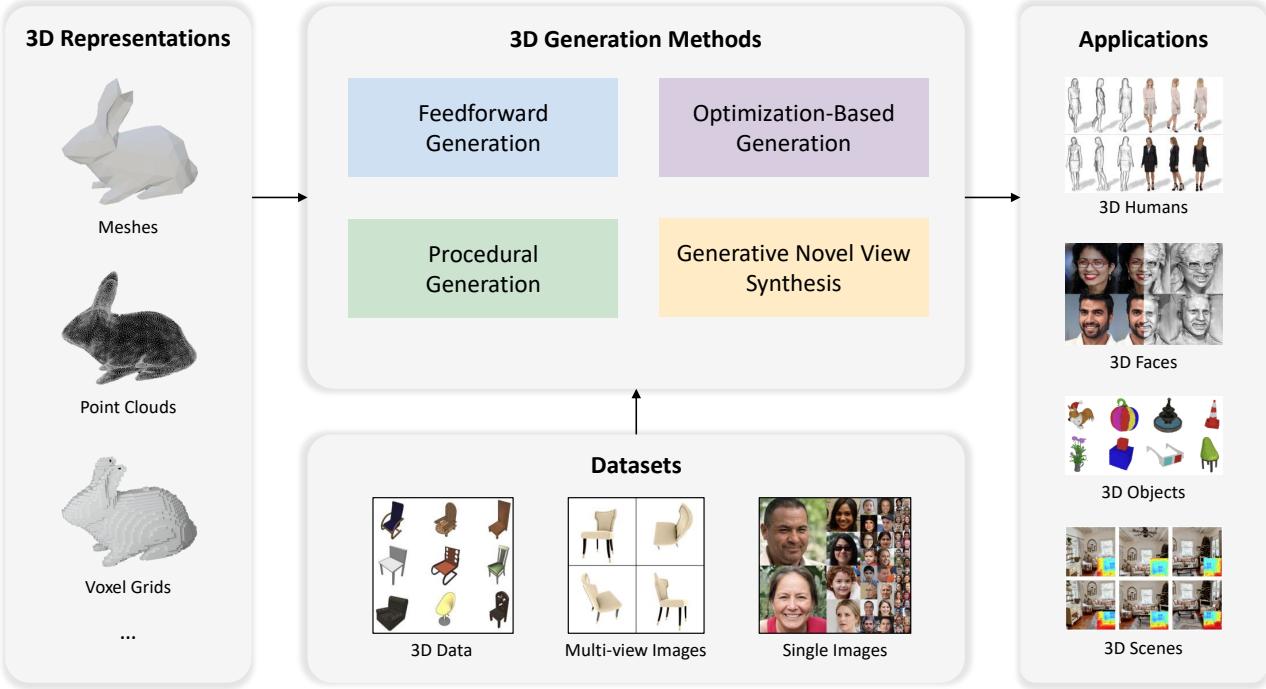


Figure 2: Overview of this survey, including 3D representations, 3D generation methods, datasets and applications. Specifically, we introduce the 3D representations that serve as the backbone for 3D generation. Furthermore, we provide a comprehensive overview of the rapidly growing literature on generation methods, categorized by the type of algorithmic paradigms, including feedforward generation, optimization-based generation, procedural generation, and generative novel view synthesis. Finally, we provide a brief discussion on popular datasets and available applications.

representations are closely related to 3D generation. However, we consider these topics to be outside the purview of this report. For a comprehensive discussion on neural rendering, we direct readers to [TFT^{*}20, TTM^{*}22], and for a broader examination of other neural representations, we recommend [KBM^{*}20, XTS^{*}22]. Our primary focus is on exploring techniques that generate 3D models. Therefore, this review does not encompass research on generation methods for 2D images within the realm of visual computing. For further information on a specific generation method, readers can refer to [Doe16] (VAEs), [GSW^{*}21] (GANs), [PYG^{*}23, CHIS23] (Diffusion) and [KNH^{*}22] (Transformers) for a more detailed understanding. There are also some surveys related to 3D generation that have their own focuses such as 3D-aware image synthesis [XX23], 3D generative models [SPX^{*}22], Text-to-3D [LZW^{*}23] and deep learning for 3D point clouds [GWH^{*}20]. In this survey, we give a comprehensive analysis of different 3D generation methods.

3. Neural Scene Representations

In the domain of 3D AI-generated content, adopting a suitable representation of 3D models is essential. The generation process typically involves a scene representation and a differentiable rendering algorithm for creating 3D models and rendering 2D images. Conversely, the created 3D models or 2D images could be supervised in the reconstruction domain or image domain, as illustrated in Fig. 3. Some methods directly supervise the 3D models of the scene rep-

resentation, while others render the scene representation into images and supervise the resulting renderings. In the following, we broadly classify the scene representations into three groups: explicit scene representations (Section 3.1), implicit representations (Section 3.2), and hybrid representations (Section 3.3). Note that, the rendering methods (*e.g.* ray casting, volume rendering, rasterization, *etc*), which should be differentiable to optimize the scene representations from various inputs, are also introduced.

3.1. Explicit Representations

Explicit scene representations serve as a fundamental module in computer graphics and vision, as they offer a comprehensive means of describing 3D scenes. By depicting scenes as an assembly of basic primitives, including point-like primitives, triangle-based meshes, and advanced parametric surfaces, these representations can create detailed and accurate visualizations of various environments and objects.

3.1.1. Point Clouds

A point cloud is a collection of elements in Euclidean space, representing discrete points with addition attributes (*e.g.* colors and normals) in three-dimensional space. In addition to simple points, which can be considered infinitesimally small surface patches, oriented point clouds with a radius (surfels) can also be used [PZVBG00]. Surfels are used in computer graphics for rendering

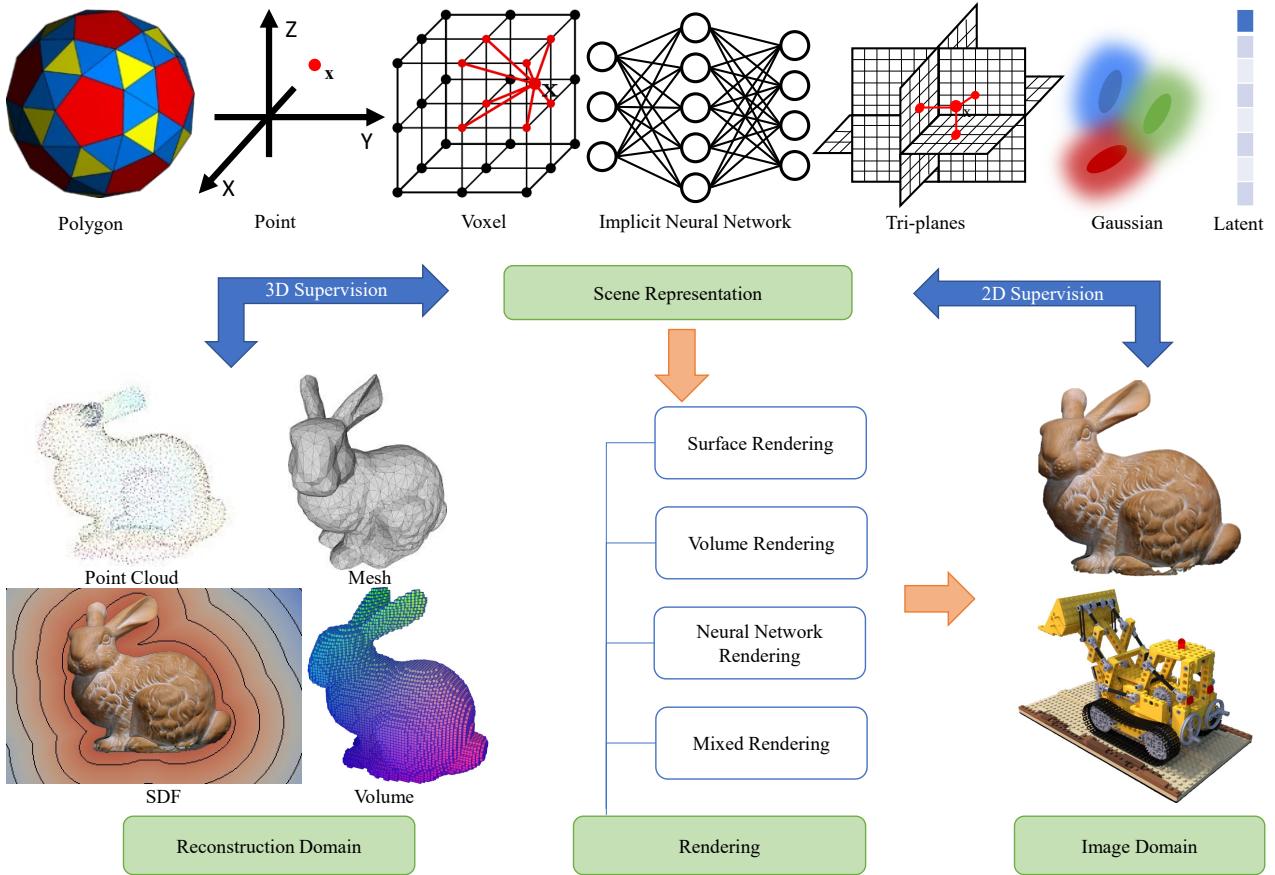


Figure 3: Neural scene representations used for 3D generation, including explicit, implicit, and hybrid representations. The 3D generation involves the use of scene representations and a differentiable rendering algorithm to create 3D models or render 2D images. On the flip side, these 3D models or 2D images can function as the reconstruction domain or image domain, overseeing the 3D generation of scene representations.

point clouds (called splitting), which are differentiable [YSW^{*}19, KKL^D23] and allow researchers to define differentiable rendering pipelines to adjust point cloud positions and features, such as radius or color. Techniques like Neural Point-based Rendering [ASK^{*}20, DZL^{*}20], SynSin [WGSJ20], Pulsar [LZ21, KPLD21] and ADOP [RFS22] leverage learnable features to store information about the surface appearance and shape, enabling more accurate and detailed rendering results. Several other methods, such as FVS [RK20], SVS [RK21], and FWD-Transformer [CRJ22], also employ learnable features to improve the rendering quality. These methods typically embed features into point clouds and warp them to target views to decode color values, allowing for more accurate and detailed reconstructions of the scene.

By incorporating point cloud-based differentiable renderers into the 3D generation process, researchers can leverage the benefits of point clouds while maintaining compatibility with gradient-based optimization techniques. This process can be generally categorized into two different ways: point splitting which blends the discrete samples with some local deterministic blurring kernels [ZPVBG02, LKL18, ID18, RROG18], and conventional point ren-

derer [ASK^{*}20, DZL^{*}20, KPLD21, RALB22]. These methods facilitate the generation and manipulation of 3D point cloud models while maintaining differentiability, which is essential for training and optimizing neural networks in 3D generation tasks.

3.1.2. Meshes

By connecting multiple vertices with edges, more complex geometric structures (*e.g.* wireframes and meshes) can be formed [BKP^{*}10]. These structures can then be further refined by using polygons, typically triangles or quadrilaterals, to create realistic representations of objects [SS87]. Meshes provide a versatile and efficient means of representing intricate shapes and structures, as they can be easily manipulated and rendered by computer algorithms. The majority of graphic editing toolchains utilize triangle meshes. This type of representation is indispensable for any digital content creation (DCC) pipeline, given its wide acceptance and compatibility. To align seamlessly with these pipelines, neural networks can be strategically trained to predict discrete vertex locations [BNT21, TZN19]. This ability allows for the direct importation of these locations into any DCC pipeline, facilitating a smooth

and efficient workflow. In contrast to predicting discrete textures, continuous texture methods optimized via neural networks are proposed, such as texture fields [OMN^{*}19] and NeRF-Tex [BGP^{*}22]. In this way, it could provide a more refined and detailed texture, enhancing the overall quality and realism of the generated 2D models.

Integrating mesh representation into 3D generation requires the use of mesh-based differentiable rendering methods, which enable meshes to be rasterized in a manner that is compatible with gradient-based optimization. Several such techniques have been proposed, including OpenDR [LB14], neural mesh renderer [KUH18], Paparazzi [LTJ18], and Soft Rasterizer [LLCL19]. Additionally, general-purpose physically based renderers like Mitsuba 2 [NDVZJ19] and Taichi [HLA^{*}19] support mesh-based differentiable rendering through automatic differentiation.

3.1.3. Multi-layer Representations

The use of multiple semi-transparent colored layers for representing scenes has been a popular and successful scheme in real-time novel view synthesis [ZTF^{*}18]. Using Layered Depth Image (LDI) representation [SGHS98] is a notable example, extending traditional depth maps by incorporating multiple layers of depth maps, each with associated color values. Several methods [PZ17, CGT^{*}19, SSKH20] have drawn inspiration from the LDI representation and employed deep learning advancements to create networks capable of predicting LDIs. In addition to LDIs, Stereomagnification [ZTF^{*}18] initially introduced the multiple image (MPI) representation. It describes scenes using multiple front-parallel semi-transparent layers, including colors and opacity, at fixed depth ranges through plane sweep volumes. With the help of volume rendering and homography projection, the novel view could be synthesized in real-time. Building on Stereomagnification [ZTF^{*}18], various methods [FBD^{*}19, MSOC^{*}19, STB^{*}19] have adopted the MPI representation to enhance rendering quality. The multi-layer representation has been further expanded to accommodate wider fields of view in [BFO^{*}20, ALG^{*}20, LXM^{*}20] by substituting planes with spheres. As research in this domain continues to evolve, we can expect further advancements in these methods, leading to more efficient and effective 3D generation techniques for real-time rendering.

3.2. Implicit Representations

Implicit representations have become the scene representation of choice for problems in view synthesis or shape reconstruction, as well as many other applications across computer graphics and vision. Unlike explicit scene representations that usually focus on object surfaces, implicit representations could define the entire volume of a 3D object, and use volume rendering for image synthesis. These representations utilize mathematical functions, such as radiance fields [MST^{*}20] or signed distance fields [PFS^{*}19, CZ19], to describe the properties of a 3D space.

3.2.1. Neural Radiance Fields

Neural Radiance Fields (NeRFs) [MST^{*}20] have gained prominence as a favored scene representation method for a wide range of applications. Fundamentally, NeRFs introduce a novel representation of 3D scenes or geometries. Rather than utilizing point clouds

and meshes, NeRFs depict the scene as a continuous volume. This approach involves obtaining volumetric parameters, such as view-dependent radiance and volume density, by querying an implicit neural network. This innovative representation offers a more fluid and adaptable way to capture the intricacies of 3D scenes, paving the way for enhanced rendering and modeling techniques.

Specifically, NeRF [MST^{*}20] represents the scene with a continuous volumetric radiance field, which utilizes MLPs to map the position \mathbf{x} and view direction \mathbf{r} to a density σ and color \mathbf{c} . To render a pixel's color, NeRF casts a single ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ and evaluates a series of points $\{t_i\}$ along the ray. The evaluated $\{(\sigma_i, \mathbf{c}_i)\}$ at the sampled points are accumulated into the color $C(\mathbf{r})$ of the pixel via volume rendering [Max95]:

$$C(\mathbf{r}) = \sum_i T_i \alpha_i \mathbf{c}_i, \quad \text{where } T_i = \exp \left(- \sum_{k=0}^{i-1} \sigma_k \delta_k \right), \quad (1)$$

and $\alpha_i = 1 - \exp(-\sigma_i \delta_i)$ indicates the opacity of the sampled point. Accumulated transmittance T_i quantifies the probability of the ray traveling from t_0 to t_i without encountering other particles, and $\delta_i = t_i - t_{i-1}$ denotes the distance between adjacent samples.

NeRFs [MST^{*}20, NG21, BMT^{*}21, BMV^{*}22, VHM^{*}22, LWC^{*}23] have seen widespread success in problems such as edition [MBRS^{*}21, ZLLD21, CZL^{*}22, YSL^{*}22], joint optimization of cameras [LMTL21, WWX^{*}21, CCW^{*}23, TRMT23], inverse rendering [ZLW^{*}21, SDZ^{*}21, BBJ^{*}21, ZSD^{*}21, ZZW^{*}23, LZF^{*}23], generalization [YYTK21, WWG^{*}21, CXZ^{*}21, LFS^{*}21, JL22, HZF^{*}23b], acceleration [RPLG21, GKJ^{*}21, ZZZ^{*}23b], and free-viewpoint video [DZY^{*}21, LSZ^{*}22, PCPMMN21]. Apart from the above applications, NeRF-based representation can also be used for digit avatar generation, such as face and body reenactment [PDW^{*}21, GCL^{*}21, LHR^{*}21, WCS^{*}22, HPX^{*}22]. NeRFs have been extended to various fields such as robotics [KFH^{*}22, ZKW^{*}23, ACC^{*}22], tomography [RWL^{*}22, ZLZ^{*}22], image processing [HZF^{*}22, MLL^{*}22b, HZF^{*}23a], and astronomy [LSC^{*}22].

3.2.2. Neural Implicit Surfaces

Within the scope of shape reconstruction, a neural network processes a 3D coordinate as input and generates a scalar value, which usually signifies the signed distance to the surface. This method is particularly effective in filling in missing information and generating smooth, continuous surfaces. The implicit surface representation defines the scene's surface as a learnable function f that specifies the signed distance $f(\mathbf{x})$ from each point to the surface. The fundamental surface can then be extracted from the zero-level set, $S = \{\mathbf{x} \in \mathbb{R}^3 | f(\mathbf{x}) = 0\}$, providing a flexible and efficient way to reconstruct complex 3D shapes. Implicit surface representations offer numerous advantages, as they eliminate the need to define mesh templates. As a result, they can represent objects with unknown or changing topology in dynamic scenarios. Specifically, implicit surface representations recover signed distance fields for shape modeling using MLPs with coordinate inputs. These initial proposals sparked widespread enthusiasm and led to various improvements focusing on different aspects, such as enhancing training schemes [DZW^{*}20, YAK^{*}20, ZML^{*}22], leveraging global-local context [XWC^{*}19, EGO^{*}20, ZPL^{*}22], adopting specific parame-

terizations [GCV^{*}19, CTZ20, YRSh21, BSKG22], and employing spatial partitions [GCS^{*}20, TTG^{*}20, TLY^{*}21, WLG^{*}23].

NeuS [WLL^{*}21] and VolSDF [YGKL21] extend the basic NeRF formulation by integrating an SDF into volume rendering, which defines a function to map the signed distance to density σ . It attains a locally maximal value at surface intersection points. Specifically, accumulated transmittance $T(t)$ along the ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is formulated as a sigmoid function: $T(t) = \Phi(f(t)) = (1 + e^{sf(t)})^{-1}$, where s and $f(t)$ refer to a learnable parameter and the signed distance function of points at $\mathbf{r}(t)$, respectively. Discrete opacity values α_i can then be derived as:

$$\alpha_i = \max \left(\frac{\Phi_s(f(t_i)) - \Phi_s(f(t_{i+1}))}{\Phi_s(f(t_i))}, 0 \right). \quad (2)$$

NeuS employs volume rendering to recover the underlying SDF based on Eqs. (1) and (2). The SDF is optimized by minimizing the photometric loss between the rendering results and ground-truth images.

Building upon NeuS and VolSDF, NeuralWarp [DBD^{*}22], Geo-NeuS [FXOT22], MonoSDF [YPN^{*}22] leverage prior geometry information from MVS methods. IRON [ZLS22], MII [ZSH^{*}22], and WildLight [CLL23] apply high-fidelity shape reconstruction via SDF for inverse rendering. HF-NeuS [WSW22] and PET-Neus [WSW23] integrate additional displacement networks to fit the high-frequency details. LoD-NeuS [ZZF^{*}23] adaptively encodes Level of Detail (LoD) features for shape reconstruction.

3.3. Hybrid Representations

Implicit representations have indeed demonstrated impressive results in various applications as mentioned above. However, most of the current implicit methods rely on regression to NeRF or SDF values, which may limit their ability to benefit from explicit supervision on the target views or surfaces. Explicit representation could impose useful constraints during training and improve the user experience. To capitalize on the complementary benefits of both representations, researchers have begun exploring hybrid representations. These involve scene representations (either explicit or implicit) that embed features utilizing rendering algorithms for view synthesis.

3.3.1. Voxel Grids

Early work [WSK^{*}15, CXG^{*}16, MS15] depicted 3D shapes using voxels, which store coarse occupancy (inside/outside) values on a regular grid. This approach enabled powerful convolutional neural networks to operate natively and produce impressive results in 3D reconstruction and synthesis [DRB^{*}18, WZX^{*}16, BLW16]. These methods usually use explicit voxel grids as the 3D representation. Recently, to address the slow training and rendering speeds of implicit representations, the 3D voxel-based embedding methods [LGZL^{*}20, FKYT^{*}22, SSN^{*}22, SSC22] have been proposed. These methods encode the spatial information of the scene and decode the features more efficiently. Moreover, Instant-NGP [MESK22] introduces the multi-level voxel grids encoded implicitly via the hash function for each level. It facilitates rapid optimization and rendering while maintaining a compact model. These advancements

in 3D shape representation and processing techniques have significantly enhanced the efficiency and effectiveness of 3D generation applications.

3.3.2. Tri-plane

Tri-plane representation is an alternative approach to using voxel grids for embedding features in 3D shape representation and neural rendering. The main idea behind this method is to decompose a 3D volume into three orthogonal planes (e.g., XY, XZ, and YZ planes) and represent the features of the 3D shape on these planes. Specifically, TensoRF [CXG^{*}22] achieves similar model compression and acceleration by replacing each voxel grid with a tensor decomposition into planes and vectors. Tri-planes are efficient and capable of scaling with the surface area rather than volume and naturally integrate with expressive, fine-tuned 2D generative architectures. In the generative setting, EG3D [CLC^{*}22] proposes a spatial decomposition into three planes whose values are added together to represent a 3D volume. NFD [SCP^{*}23] introduces diffusion on 3D scenes, utilizing 2D diffusion model backbones and having built-in tri-plane representation.

3.3.3. Hybrid Surface Representation

DMTet, a recent development cited in [SGY^{*}21], is a hybrid three-dimensional surface representation that combines both explicit and implicit forms to create a versatile and efficient model. It segments the 3D space into dense tetrahedra, thereby forming an explicit partition. By integrating explicit and implicit representations, DMTet can be optimized more efficiently and transformed seamlessly into explicit structures like mesh representations. During the generation process, DMTet can be differentiably converted into a mesh, which enables swift high-resolution multi-view rendering. This innovative approach offers significant improvements in terms of efficiency and versatility in 3D modeling and rendering.

4. Generation Methods

In the past few years, the rapid development of generative models in 2D image synthesis, such as generative adversarial networks (GANs) [GPAM^{*}14, AQW19], variational autoencoders (VAEs) [KPHL17, PGH^{*}16, KW13], autoregressive models [RWC^{*}19, BMR^{*}20], diffusion models [HJA20, ND21, SCS^{*}22], etc., has led to their extension and combination with these scene representations for 3D generation. Tab. 1 shows well-known examples of 3D generation using generative models and scene representations. These methods may use different scene representations in the generation space, where the representation is generated by the generative models, and the reconstruction space, where the output is represented. For example, AutoSDF [MCST22a] uses a transformer-based autoregressive model to learn a feature voxel grid and decode this representation to SDF for reconstruction. EG3D [CLC^{*}22] employs GANs to generate samples in latent space and introduces a tri-plane representation for rendering the output. SSDNeRF [CGC^{*}23] uses the diffusion model to generate tri-plane features and decode them to NeRF for rendering. By leveraging the advantages of neural scene representations and generative models, these approaches have demonstrated remarkable potential in generating realistic and intricate 3D models while maintaining view consistency.

Table 1: Some examples of 3D generation methods. We first divide the methods according to the generative models and their corresponding representations in generation space. The representations in the reconstruction space determine how the 3D objects are formatted and rendered. We also list the main supervision and conditions of these methods. For the 2D supervision, a rendering technique is utilized to generate the images.

Method	Generative Model	Generation Space	Reconstruction Space	Rendering	Supervision	Condition
PointFlow [YHH*19a]	Normalizing Flow	Latent Code	Point Cloud	-	3D	Uncon
3dAAE [ZKZ*20]	VAE	Latent Code	Point Cloud	-	3D	Uncon
SDM-NET [GYW*19a]	VAE	Latent Code	Mesh	-	3D	Uncon
AutoSDF [MCST22a]	Autoregressive	Voxel	SDF	-	3D	Uncon.
PolyGen [NGB20a]	Autoregressive	Polygon	Mesh	-	3D	Uncon./Label/Image
PointGrow [SWL*20a]	Autoregressive	Point	Point Cloud	-	3D	Uncon./Label/Image
EG3D [CLC*22]	GAN	Latent Code	Tri-plane	Mixed Rendering	2D	Uncon.
GIRAFFE [NG21]	GAN	Latent Code	NeRF	Mixed Rendering	2D	Uncon.
BlockGAN [NPRM*20]	GAN	Latent Code	Voxel Grid	Network Rendering	2D	Uncon.
gDNA [CJS*22]	GAN	Latent Code	Occupancy Field	Surface Rendering	2D&3D	Uncon.
SurfGen [LLZL21]	GAN	Latent Code	SDF	-	3D	Uncon.
tree-GAN [SPK19]	GAN	Latent Code	Point Cloud	-	3D	Uncon.
HoloDiffusion [KVNMC23]	Diffusion	Voxel	NeRF	Volume Rendering	2D	Image
SSDNeRF [CGC*23]	Diffusion	Tri-plane	NeRF	Volume Rendering	2D	Uncon./Image
3DShape2VecSet [ZTNW23]	Diffusion	Latent Set	SDF	-	3D	Uncon./Text/Image
Point-E [NJD*22]	Diffusion	Point	Point Cloud	-	3D	Text
3DGen [GZN*23]	Diffusion	Tri-plane	Mesh	-	3D	Text/Image
DreamFusion [PJBM23]	Diffusion	-	NeRF	Volume Rendering	SDS	Text
Make-It-3D [TWZ*23]	Diffusion	-	Point Cloud	Network Rendering	SDS	Image
Zero-1-to-3 [LWVH*23]	Diffusion	Pixel	-	-	2D	Image
MVDream [SWY*23]	Diffusion	Pixel	-	-	2D	Image
DMV3D [XTL*23]	Diffusion	Pixel	Tri-plane	Volume Rendering	2D	Text/Image

In this section, we explore a large variety of 3D generation methods which are organized into four categories based on their algorithmic paradigms: Feedforward Generation (Sec. 4.1), generating results in a forward pass; Optimization-Based Generation (Sec. 4.2), necessitating a test-time optimization for each generation; Procedural Generation (Sec. 4.3), creating 3D models from sets of rules; and Generative Novel View Synthesis (Sec. 4.4), synthesizing multi-view images rather than an explicit 3D representation for 3D generation. An evolutionary tree of 3D generation methods is depicted in Fig. 4, which illustrates the primary branch of generation techniques, along with associated work and subsequent developments. A comprehensive analysis will be discussed in the subsequent subsection.

4.1. Feedforward Generation

A primary technical approach for generation methods is feedforward generation, which can directly produce 3D representations using generative models. In this section, we explore these methods based on their generative models as shown in Fig. 5, which include generative adversarial networks (GANs), diffusion Models, autoregressive models, variational autoencoders (VAEs) and normalizing flows.

4.1.1. Generative Adversarial Networks

Generative Adversarial Networks (GANs) [GPAM*14] have demonstrated remarkable outcomes in image synthesis tasks, con-

sisting of a generator $G(\cdot)$ and a discriminator $D(\cdot)$. The generator network G produces synthetic data by accepting latent code as input, while the discriminator network D differentiates between generated data from G and real data. Throughout the training optimization process, the generator G and discriminator D are jointly optimized, guiding the generator to create synthetic data as realistic as real data.

Building on the impressive results achieved by GANs in 2D image synthesis, researchers have begun to explore the application of these models to 3D generation tasks. The core idea is to marry GANs with various 3D representations, such as point clouds (l-GAN/r-GAN [ADMG18], tree-GAN [SPK19]), voxel grids (3D-GAN [WZX*16], Z-GAN [KKR18]), meshes (MeshGAN [CBZ*19]), or SDF (SurfGen [LLZL21], SDF-StyleGAN [ZLWT22]). In this context, the 3D generation process can be viewed as a series of adversarial steps, where the generator learns to create realistic 3D data from input latent codes, and the discriminator differentiates between generated data and real data. By iteratively optimizing the generator and discriminator networks, GANs learn to generate 3D data that closely resembles the realism of actual data.

For 3D object generation, prior GAN methodologies, such as l-GAN [ADMG18], 3D-GAN [WZX*16], and Multi-chart Generation [BHMK*18], directly utilize explicit 3D object representation of real data to instruct generator networks. Their discriminators employ 3D representation as supervision, directing the gener-

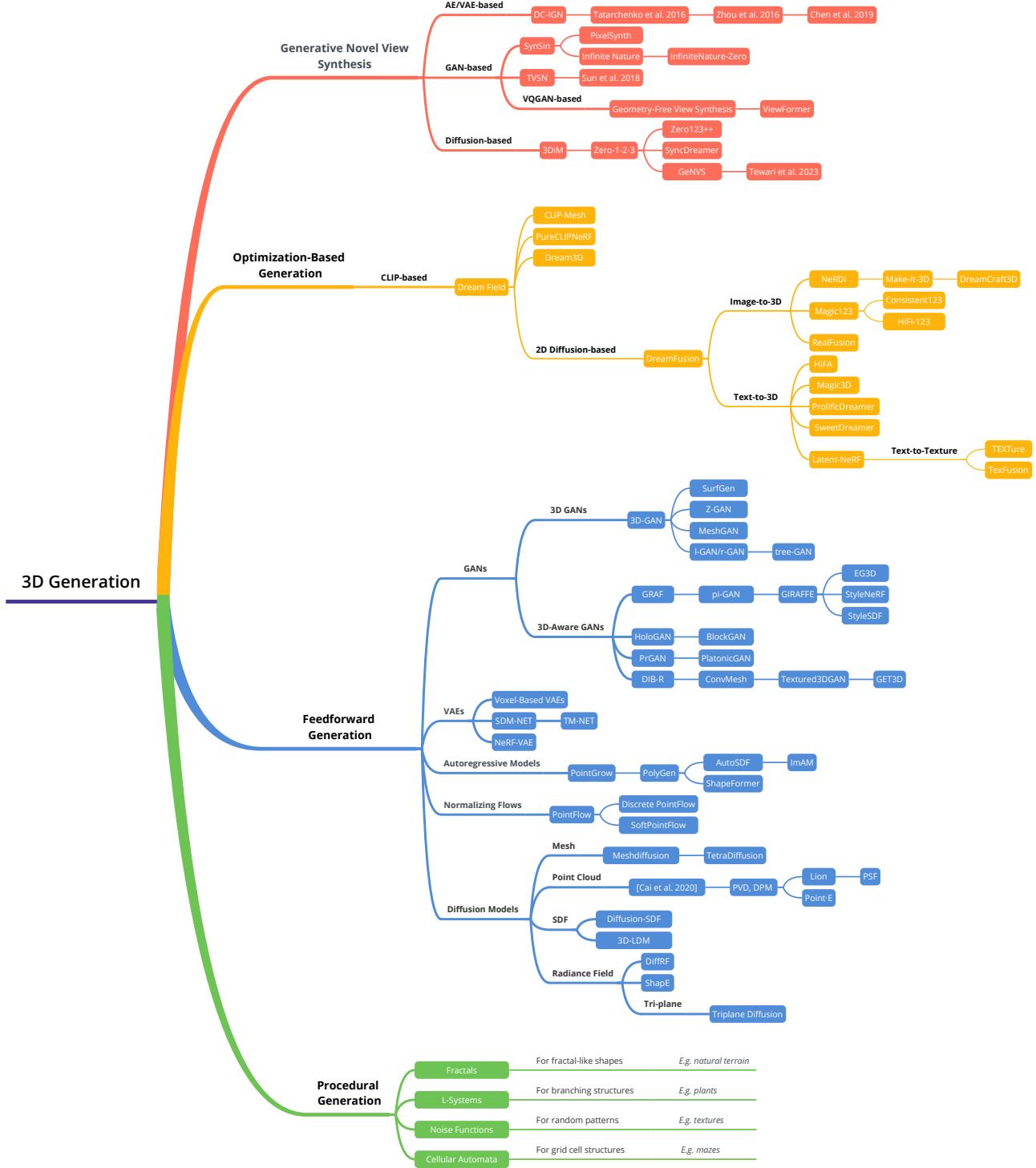


Figure 4: The evolutionary tree of 3D generation illustrates the primary branch of generation methods and their developments in recent years. Specifically, we provide a comprehensive overview of the rapidly growing literature on generation methods, categorized by the type of algorithmic paradigms, including feedforward generation, optimization-based generation, procedural generation, and generative novel view synthesis.

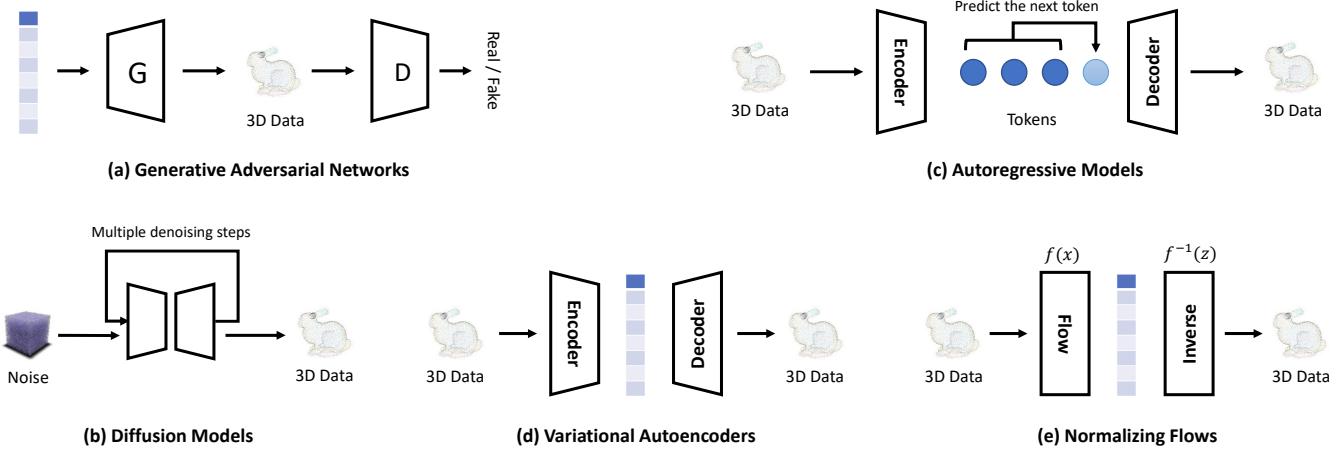


Figure 5: Exemplary feedforward 3D generation models. We showcase several representative pipelines of feedforward 3D generation models, including (a) generative adversarial networks, (b) diffusion models, (c) autoregressive models, (d) variational autoencoders and (e) normalizing flows.

ator to produce synthetic data that closely resembles the realism of actual data. During training, specialized generators generate corresponding supervisory 3D representations, such as point clouds, voxel grids, and meshes. Some studies, like SurfGen [LLZL21], have progressed further to generate intermediate implicit representations and then convert them to corresponding 3D representations instead of directly generating explicit ones, achieving superior performance. In particular, the generator of I-GAN [ADMG18], 3D-GAN [WZX*16], and Multi-chart Generation [BHMK*18] generate the position of point cloud, voxel grid, and mesh directly, respectively, taking latent code as input. SurfGen [LLZL21] generates implicit representation and then extracts explicit 3D representation.

In addition to GANs that directly generate various 3D representations, researchers have suggested incorporating 2D supervision through differentiable rendering to guide 3D generation, which is commonly referred to as 3D-Aware GAN. Given the abundance of 2D images, GANs can better understand the implicit relationship between 2D and 3D data than relying solely on 3D supervision. In this approach, the generator of GANs generates rendered 2D images from implicit or explicit 3D representation. Then the discriminators distinguish between rendered 2D images and real 2D images to guide the training of the generator.

Specifically, HoloGAN [NPLT*19] first learns a 3D representation of 3D features, which is then projected to 2D features by the camera pose. These 2D feature maps are then rendered to generate the final images. BlockGAN [NPRM*20] extends it to generate 3D features of both background and foreground objects and combine them into 3D features for the whole scene. In addition, PrGAN [GMW17] and PlatonicGAN [HMR19a] employ an explicit voxel grid structure to represent 3D shapes and use a render layer to create images. Other methods like DIB-R [CZ19], ConvMesh [PSH*20], Textured3DGAN [PKHL21] and GET3D [GSW*22] propose GAN frameworks for generating triangle meshes and textures using only 2D supervision.

Building upon representations such as NeRFs, GRAF [SLNG20] proposes generative radiance fields utilizing adversarial frameworks and achieves controllable image synthesis at high resolutions. pi-GAN [CMK*21a] introduces SIREN-based implicit GANs with FiLM conditioning to further improve image quality and view consistency. GIRAFFE [NG21] represents scenes as compositional generative neural feature fields to model multi-object scenes. Furthermore, EG3D [CLC*22] first proposes a hybrid explicit-implicit tri-plane representation that is both efficient and expressive and has been widely adopted in many following works.

4.1.2. Diffusion Models

Diffusion models [HJA20, RBL*22a] are a class of generative models that learn to generate data samples by simulating a diffusion process. The key idea behind diffusion models is to transform the original data distribution into a simpler distribution, such as Gaussian, through a series of noise-driven steps called the forward process. The model then learns to reverse this process, known as the backward process, to generate new samples that resemble the original data distribution. The forward process can be thought of as gradually adding noise to the original data until it reaches the target distribution. The backward process, on the other hand, involves iteratively denoising the samples from the distribution to generate the final output. By learning this denoising process, diffusion models can effectively capture the underlying structure and patterns of the data, allowing them to generate high-quality and diverse samples.

Building on the impressive results achieved by diffusion models in generating 2D images, researchers have begun to explore the applications of these models to 3D generation tasks. The core idea is to marry denoising diffusion models with various 3D representations. In this context, the 3D generation process can be viewed as a series of denoising steps, reversing the diffusion process from input 3D data to Gaussian noise. The diffusion models learn to generate 3D data from this noisy distribution through denoising.

Specifically, Cai et al. [CYAE^{*}20] build upon a denoising score-matching framework to learn distributions for point cloud generation. PVD [ZDW21] combines the benefits of both point-based and voxel-based representations for 3D generation. The model learns a diffusion process that transforms point clouds into voxel grids and vice versa, effectively capturing the underlying structure and patterns of the 3D data. Similarly, DPM [LH21] focuses on learning a denoising process for point cloud data by iterative denoising the noisy point cloud samples. Following the advancements made by PVD [ZDW21] and DPM [LH21], LION [ZVW^{*}22] builds upon the idea of denoising point clouds and introduces the concept of denoising in the latent space of point clouds, which is analogous to the shift in 2D image generation from denoising pixels to denoising latent space representations. To generate point clouds from text prompts, Point-E [NJD^{*}22] initially employs the GLIDE model [NDR^{*}21] to generate text-conditional synthetic views, followed by the production of a point cloud using a diffusion model conditioned on the generated image. By training the model on a large-scale 3D dataset, it achieves remarkable generalization capabilities.

In addition to point clouds, MeshDiffusion [LFB^{*}23], Tetrahedral Diffusion Models [KPWS22], and SLIDE [LWA^{*}23] explore the application of diffusion models to mesh generation. MeshDiffusion [LFB^{*}23] adopts the DMTet representation [SGY^{*}21] for meshes and optimizes the model by treating the optimization of signed distance functions as a denoising process. Tetrahedral Diffusion Models [KPWS22] extends diffusion models to tetrahedral meshes, learning displacement vectors and signed distance values on the tetrahedral grid through denoising. SLIDE [LWA^{*}23] explores diffusion models on sparse latent points for mesh generation.

Apart from applying diffusion operations on explicit 3D representations, some works focus on performing the diffusion process on implicit representations. SSDNeRF [CGC^{*}23], DiffRF [MSP^{*}23] and Shap-E [JN23] operate on 3D radiance fields, while SDF-Diffusion [SKJ23], LAS-Diffusion [ZPW^{*}23], Neural Wavelet-domain Diffusion [HLHF22], One-2-3-45++ [LXJ^{*}23], SDFusion [CLT^{*}23] and 3D-LDM [NKR^{*}22] focus on signed distance fields representations. Specifically, Diffusion-SDF [LDZL23] utilizes a voxel-shaped SDF representation to generate high-quality and continuous 3D shapes. 3D-LDM [NKR^{*}22] creates neural implicit representations of SDFs by initially using a diffusion model to generate the latent space of an auto-decoder. Subsequently, the latent space is decoded into SDFs to acquire 3D shapes. Moreover, Rodin [WZZ^{*}23] and Shue et al. [SCP^{*}23] adopt tri-plane as the representation and optimize the tri-plane features using diffusion methods. Shue et al. [SCP^{*}23] generates 3D shapes using occupancy networks, while Rodin [WZZ^{*}23] obtains 3D shapes through volumetric rendering.

These approaches showcase the versatility of diffusion models in managing various 3D representations, including both explicit and implicit forms. By tailoring the denoising process to different representation types, diffusion models can effectively capture the underlying structure and patterns of 3D data, leading to improved generation quality and diversity. As research in this area continues to advance, it is expected that diffusion models will play a crucial

role in pushing the boundaries of 3D shape generation across a wide range of applications.

4.1.3. Autoregressive Models

A 3D object can be represented as a joint probability of the occurrences of multiple 3D elements:

$$p(x_0, x_1, \dots, x_n), \quad (3)$$

where x_i is the i -th element which can be the coordinate of a point or a voxel. A joint probability with a large number of random variables is usually hard to learn and estimate. However, one can factorize it into a product of conditional probabilities:

$$p(x_0, x_1, \dots, x_n) = p(x_0) \prod_{i=1}^n p(x_i | x_{<i}), \quad (4)$$

which enables learning conditional probabilities and estimating the joint probability via sampling. Autoregressive models for data generation are a type of models that specify the current output depending on their previous outputs. Assuming that the elements x_0, x_1, \dots, x_n form an ordered sequence, a model can be trained by providing it with previous inputs x_0, \dots, x_{i-1} and supervising it to fit the probability of the outcome x_i :

$$p(x_i | x_{<i}) = f(x_0, \dots, x_{i-1}), \quad (5)$$

the conditional probabilities are learned by the model function f . This training process is often called teacher forcing. The model can be then used to autoregressively generate the elements step-by-step:

$$x_i = \text{argmax } p(x_i | x_{<i}). \quad (6)$$

State-of-the-art generative models such as GPTs [RWC^{*}19, BMR^{*}20] are autoregressive generators with Transformer networks as the model function. They achieve great success in generating natural languages and images. In 3D generation, several studies have been conducted based on autoregressive models. In this section, we discuss some notable examples of employing autoregressive models for 3D generation.

PointGrow [SWL^{*}20b] generates point clouds using an autoregressive network with self-attention context awareness operations in a point-by-point manner. Given its previously generated points, PointGrow reforms the points by axes and passes them into three branches. Each branch takes the inputs to predict a coordinate value of one axis. The model can also condition an embedding vector to generate point clouds, which can be a class category or an image. Inspired by the network from PointGrow, PolyGen [NGEB20b] generates 3D meshes with two transformer-based networks, one for vertices and one for faces. The vertex transformer autoregressively generates the next vertex coordinate based on previous vertices. The face transformer takes all the output vertices as context to generate faces. PolyGen can condition on a context of object classes or images, which are cross-attended by the transformer networks.

Recently, AutoSDF [MCST22b] generates 3D shapes represented by volumetric truncated-signed distance function (T-SDF). AutoSDF learns a quantized codebook regarding local regions of T-SDFs using VQ-VAE. The shapes are then presented by the codebook tokens and learned by a transformer-based network in a non-sequential autoregressive manner. In detail, given previous

tokens at arbitrary locations and a query location, the network predicts the token that is queried. AutoSDF is capable of completing shapes and generating shapes based on images or text. Concurrently with AutoSDF, ShapeFormer [YLM^{*}22] generates surfaces of 3D shapes based on incomplete and noisy point clouds. A compact 3D representation called vector quantized deep implicit function (VQDIF) is used to represent shapes using a feature sequence of discrete variables. ShapeFormer first encodes an input point cloud into a partial feature sequence. It then uses a transformer-based network to autoregressively sample out the complete sequence. Finally, it decodes the sequence to a deep implicit function from which the complete object surface can be extracted. Instead of learning in 3D volumetric space, Luo et al. proposes an improved auto-regressive model (ImAM) to learn discrete representation in a one-dimensional space to enhance the efficient learning of 3D shape generation. The method first encodes 3D shapes of volumetric grids into three axis-aligned planes. It uses a coupling network to further project the planes into a latent vector, where vector quantization is performed for discrete tokens. ImAM adopts a vanilla transformer to autoregressively learn the tokens with tractable orders. The generated tokens are decoded to occupancy values via a network by sampling spatial locations. ImAM can switch from unconditional generation to conditional generation by concatenating various conditions, such as point clouds, categories, and images.

4.1.4. Variational Autoencoders

Variational autoencoders (VAEs) [KW13] are probabilistic generative models that consist of two neural network components: the encoder and decoder. The encoder maps the input data point to a latent space that corresponds to the parameters of a variational distribution. In this way, the encoder can produce multiple different samples that all come from the same distribution. The decoder maps from the latent space to the input space, to produce or generate data points. Both networks are typically trained together with the usage of the reparameterization trick, although the variance of the noise model can be learned separately. VAEs have also been explored in 3D generation [KYLH21, GWY^{*}21, GYW^{*}19b, BLW16, KSZ^{*}21].

Brock et al. trains variational autoencoders directly for voxels using 3D ConvNet, while SDM-Net [GYW^{*}19b] focuses on the generation of structured meshes composed of deformable parts. The method uses one VAE network to model parts and another to model the whole object. The follow-up work TM-Net [GWY^{*}21] could generate texture maps of meshes in a part-aware manner. Other representations like point clouds [KYLH21] and NeRFs [KSZ^{*}21] are also explored in variational autoencoders. Owing to the reconstruction-focused objective of VAEs, their training is considerably more stable than that of GANs. However, VAEs tend to produce more blurred results compared to GANs.

4.1.5. Normalizing Flows

Normalizing flow models consist of a series of invertible transformations that map a simple distribution, such as Gaussian, to a target distribution, which represents the data to generation. These transformations are carefully designed to be differentiable and invertible, allowing one to compute the likelihood of the data under the



Figure 6: Results of text-guided 3D generation by DreamFusion [PJBM23] using SDS loss. * denotes a DSLR photo, † denotes a zoomed out DSLR photo.

model and optimize the model parameters using gradient-based optimization techniques.

In 3D generation, PointFlow [YHH^{*}19a] learns a distribution of shapes and a distribution of points using continuous normalizing flows. This approach allows for the sampling of shapes, followed by the sampling of an arbitrary number of points from a given shape. Discrete PointFlow (DPF) network [KBV20] improves PointFlow by replacing continuous normalizing flows with discrete normalizing flows, which reduces the training and sampling time. Soft-Flow [KLK^{*}20] is a framework for training normalizing flows on the manifold. It estimates a conditional distribution of the perturbed input data instead of learning the data distribution directly. Soft-Flow alleviates the difficulty of forming thin structures for flow-based models.

4.2. Optimization-based Generation

Optimization-based generation is employed to generate 3D models using runtime optimization. These methods usually leverage pre-trained multimodal networks to optimize 3D models based on user-specified prompts. The key lies in achieving alignment between the given prompts and the generated content while maintaining high fidelity and diversity. In this section, we primarily examine optimization-based generation methods that use texts and images, based on the types of prompts provided by users.

4.2.1. Text-to-3D

Language serves as the primary means of human communication and describing scenes, and researchers are dedicated to exploring the potential of text-based generation methods. These methods typically align the text with the images obtained through the differentiable rendering techniques, thereby guiding the generation of 3D content based on the text prompts. Given a fixed surface, TANGO [LZJ^{*}22] uses CLIP [RKH^{*}21a] to supervise differentiable physical-based rendering (PBR) images and obtain texture maps that align with the specified text prompt. Inspired by the success of NeRF [MST^{*}20] and diffusion models in modeling 3D static scenes and text-to-image tasks respectively, DreamFusion [PJBM23] (as shown in Fig. 6) combines the volumetric

Table 2: Quantitative comparison of image-to-3D methods on surface reconstruction. We summarize the Chamfer distance and volume IoU as the metrics to evaluate the quality of surface reconstruction.

Method	Chamfer Distance ↓	Volume IoU ↑
RealFusion [MKLRV23]	0.0819	0.2741
Magic123 [QMH*23]	0.0516	0.4528
Make-it-3D [TWZ*23]	0.0732	0.2937
One-2-3-45 [LXJ*23]	0.0629	0.4086
Point-E [NJD*22]	0.0426	0.2875
Shap-E [JN23]	0.0436	0.3584
Zero-1-to-3 [LWVH*23]	0.0339	0.5035
SyncDreamer [LLZ*23]	0.0261	0.5421

representation used in NeRF with the proposed Score Distillation Sampling (SDS) loss to achieve high-fidelity 3D content generation. SDS loss converts rendering error minimization into probability density distillation and enables 2D diffusion priors to optimize 3D representations (e.g., volumetric representation and triangle mesh) via image parameterization (e.g., differentiable rendering). As a concurrent work concurrent with SDS, Score Jacobian Chaining (SJC) [WDL*23] interprets predictions from pre-trained diffusion models as a score function of the data log-likelihood, similarly enabling 2D diffusion priors to optimize 3D representations via score matching. Based on DreamFusion, Magic3D [LGT*23] introduces a coarse-to-fine manner and extracts the underlying geometry of the volume as a mesh. It then combines differentiable neural rendering and SDS to refine the extracted mesh. Magic3D is capable of exporting high-quality textured meshes and seamlessly embedding them into the traditional graphics pipeline. Also as a two-stage method, Fantasia3D further combines DMSTet [SGY*21] and SDS in the first geometry modeling stage to explicitly optimize surface. In the second stage, it introduces the PBR material model and disentangle texture and environment illumination. ProlificDreamer [WLW*23] presents variational score distillation (VSD) to boost text-to-3D generation. VSD adopts particles to model the distribution of 3D scenes and derive a gradient-based optimization scheme from the Wasserstein gradient flow, narrowing the gap between the rendering results distribution of the modeling distribution and pre-trained diffusion distribution. Benefiting from the optimization of scene distribution rather than a single scene, VSD overcomes the over-saturated and over-smoothed results produced by SDS and improves diversities. MVDream [SWY*23] further fine-tunes a multi-view diffusion model and introduces multi-view consistent 3D priors, overcoming multi-face and content-drift problems. Text-to-3D has garnered significant attention recently, in addition to these, many other methods [ZZ23, LCCT23, MRP*23a] have been proposed in this field.

4.2.2. Image-to-3D

As the primary way to describe the visual effects of scenes, images can more intuitively describe the details and appearance of scenes at a finer-grained than language. Recent works thus are motivated to explore the image-to-3D techniques, which reconstruct remarkable and high-fidelity 3D models from specified im-

Table 3: Quantitative comparison of image-to-3D methods on novel view synthesis. We report the CLIP-Similarity, PSNR, and LPIPS as the metrics to evaluate the quality of view synthesis.

Method	CLIP-Similarity ↑	PSNR ↑	LPIPS ↓
RealFusion [MKLRV23]	0.735	20.216	0.197
Magic123 [QMH*23]	0.747	25.637	0.062
Make-it-3D [TWZ*23]	0.839	20.010	0.119
One-2-3-45 [LXJ*23]	0.788	23.159	0.096
Zero-1-to-3 [LWVH*23]	0.759	25.386	0.068
SyncDreamer [LLZ*23]	0.837	25.896	0.059

ages. These methods strive to maintain the appearance of the specified images and optimized 3D contents while introducing reasonable geometric priors. Similar to the text-to-3D methods, several image-to-3D methods leverage the volumetric representation used in NeRF to represent the target 3D scenes, which natively introduces multi-view consistency. NeuralLift-360 [XJW*23] uses estimated monocular depth and CLIP-guided diffusion prior to regularizing the geometry and appearance optimization respectively, achieving lift of a single image to a 3D scene represented by a NeRF. RealFusion [MKLRV23] and NeRDi [DJQ*23] leverage textual inversion [GAA*22] to extract text embeddings to condition a pre-trained image diffusion model [RBL*22b], and combine use the score distillation loss to optimize the volumetric representation. Based on Magic3D [LGT*23] that employs a coarse-to-fine framework as mentioned above, Magic123 [QMH*23] additionally introduces 3D priors from a pre-trained viewpoint-conditioned diffusion model Zero-1-to-3 [LWVH*23] in two optimization stage, yielding textured meshes that match the specified images. As another two-stage image-to-3D method, Make-it-3D [TWZ*23] enhances texture and geometric structure in the fine stage, producing high-quality textured point clouds as final results. Subsequent works [SZS*23, YYC*23] have been consistently proposed to enhance the previous results. Recently, 3D Gaussian Splatting (3DGS) [KKLD23] has emerged as a promising modeling as well as a real-time rendering technique. Based on 3DGS, DreamGaussian [TRZ*23] presents an efficient two-stage framework for both text-driven and image-driven 3D generation. In the first stage, DreamGaussian leverages SDS loss (i.e. 2D diffusion priors [LWVH*23] and CLIP-guided diffusion priors [PJBM23]) to generate target objects represented by 3D Gaussians. Then DreamGaussian extracts textured mesh from the optimized 3D Gaussians by querying the local density and refines textures in the UV space. For a better understanding of readers to various image-to-3D methods, we evaluate the performance of some open-source state-of-the-art methods. Tab. 2 shows the quantitative comparison of image-to-3D methods on surface reconstruction. We summarize the Chamfer distance and volume IoU as the metrics to evaluate the quality of surface reconstruction. Tab. 3 demonstrates the quantitative comparison of image-to-3D methods on novel view synthesis. We report the CLIP-Similarity, PSNR, and LPIPS as the metrics to evaluate the quality of view synthesis.

4.3. Procedural Generation

Procedural generation is a term for techniques that create 3D models and textures from sets of rules. These techniques often rely on predefined rules, parameters, and mathematical functions to generate diverse and complex content, such as textures, terrains, levels, characters, and objects. One of the key advantages of procedural generation is their ability to efficiently create various shapes from a relatively small set of rules. In this section, we mainly survey four most used techniques: fractal geometry, L-Systems, noise functions and cellular automata.

A fractal [Man67, MM82] is a geometric shape that exhibits detailed structure at arbitrarily small scales. A characteristic feature of many fractals is their similarity across different scales. This property of exhibiting recurring patterns at progressively smaller scales is referred to as self-similarity. A common application of fractal geometry is the creation of landscapes or surfaces. These are generated using a stochastic algorithm designed to produce fractal behavior that mimics the appearance of natural terrain. The resulting surface is not deterministic, but rather a random surface that exhibits fractal behavior.

An L-system [Lin68], or Lindenmayer system, is a type of formal grammar and parallel rewriting system. It comprises an alphabet of symbols that can be utilized to construct strings, a set of production rules that transform each symbol into a more complex string of symbols, a starting string for construction, and a mechanism for converting the produced strings into geometric structures. L-systems are used to create complex and realistic 3D models of natural objects like trees and plants. The string generated by the L-System can be interpreted as instructions for a “turtle” to move in 3D space. For example, certain characters might instruct the turtle to move forward, turn left or right, or push and pop positions and orientations onto a stack.

Noise functions, such as Perlin noise [Per85] and Simplex noise [Per02], are used to generate coherent random patterns that can be applied to create realistic textures and shapes in 3D objects. These functions can be combined and layered to create more complex patterns and are particularly useful in terrain generation, where they can be used to generate realistic landscapes with varying elevations, slopes, and features.

Cellular automata [VN*51, Neu66, Wol83] are a class of discrete computational models that consist of a grid of cells, each of which can be in one of a finite number of states. The state of each cell is determined by a set of rules based on the states of its neighboring cells. Cellular automata have been used in procedural generation to create various 3D objects and patterns, such as cave systems, mazes, and other structures with emergent properties.

4.4. Generative Novel View Synthesis

Recently, generative techniques have been utilized to tackle the challenge of novel view synthesis, particularly in predicting new views from a single input image. Compared to the conventional 3D generation methods, it does not explicitly utilize the 3D representation to enforce 3D consistency, instead, it usually employs a 3D-aware method by conditioning 3D information. In the field

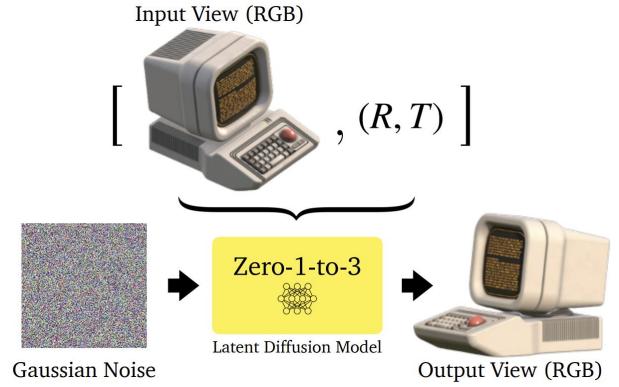


Figure 7: Zero-1-to-3 proposes a viewpoint-conditioned image diffusion model to generate the novel view of the input image. By training on a large-scale dataset, it achieves a strong generalization ability to in-the-wild images.

of novel view synthesis, a widely studied technical route will be regression-based methods [YYTK21, WWG*21, CXZ*21]. Different from them, generative novel view synthesis focuses more on generating new content rather than regressing the scenes from a few input images, which typically involves long-range view extrapolation.

With the development of image synthesis methods, significant progress has been made in generative novel view synthesis. Recently, 2D diffusion models have transformed image synthesis and therefore are also utilized in generative novel view synthesis [WCMB*22, TLK*23, LWVH*23, CNC*23, TYC*23, YGMG23]. Among these methods, 3DiM [WCMB*22] first introduces a geometry-free image-to-image diffusion model for novel view synthesis, taking the camera pose as the condition. Tseng et al. [TLK*23] designs epipolar attention layers to inject camera parameters into the pose-guided diffusion model for consistent view synthesis from a single input image. Zero-1-to-3 [LWVH*23] (as shown in Fig. 7) demonstrates the learning of the camera viewpoint in large-scale diffusion models for zero-shot novel view synthesis. [CNC*23, TYC*23, YGMG23] condition 2D diffusion models on pixel-aligned features extracted from input views to extend them to be 3D-aware. However, generating multiview-consistent images remains a challenging problem. To ensure consistent generation, [LHG*23, LLZ*23, SCZ*23, LGL*23] propose a multi-view diffusion model that could synthesize multi-view images simultaneously to consider the information between different views, which achieve more consistent results compared to the single view synthesis model like Zero-1-to-3 [LWVH*23].

Prior to that, the transformer which is a sequence-to-sequence model originally proposed in natural language processing, uses a multi-head attention mechanism to gather information from different positions and brings lots of attention in the vision community. Many tasks achieve state-of-the-art performance using the attention mechanism from the transformer including generative novel view synthesis [REO21, SMP*22, KDSB22]. Specifically, Geometry-free View Synthesis [REO21] learns the discrete repre-

Table 4: Selected datasets commonly used for 3D generation.

Dataset	Type	Year	Samples	Category
ShapeNet [CFG*15]	3D data	2015	51K	objects
Thingi10K [ZJ16]	3D data	2016	10K	objects
3D-Future [FJG*21]	3D data	2020	10K	furniture
GSO [DFK*22]	3D Data	2022	1K	household items
Objaverse [DSS*23]	3D data	2022	800K	
OmniObject3D [WZF*23]	3D data	2023	6K	objects
Objaverse-XL [DLW*23]	3D Data	2023	10.2M	objects
ScanNet [DCS*17]	multi-view images	2017	1.5K (2.5M images)	indoor scenes
CO3D [RSH*21]	multi-view images	2021	19K (1.5M images)	objects
MVImgNet [YXZ*23]	multi-view images	2023	219K (6.5M images)	objects
DeepFashion [LLQ*16]	single-view images	2016	800K	clothes
FFHQ [KLA19]	single-view images	2018	70K	human faces
AFHQ [CUYH20]	single-view images	2019	15K	animal faces
SHHQ [FLJ*22]	single-view images	2022	40K	human bodies

sentation via VQGAN to obtain an abstract latent space for training transformers. While ViewFormer [KDSB22] also uses a two-stage training consisting of a Vector Quantized Variational Autoencoder (VQ-VAE) codebook and a transformer model. And [SMP*22] employs an encoder-decoder model based on transformers to learn an implicit representation.

On the other hand, generative adversarial networks could produce high-quality results in image synthesis and consequently are applied to novel view synthesis [WGSJ20, KLY*21, RFJ21, LTJ*21, LWSK22]. Some methods [WGSJ20, KLY*21, RFJ21] maintain a 3D point cloud as the representation, which could be projected onto novel views followed by a GAN to hallucinate the missing regions and synthesize the output image. While [LTJ*21] and [LWSK22] focus on long-range view generation from a single view with adversarial training. At an earlier stage of deep learning methods when the auto-encoders and variational auto-encoders begin to be explored, it is also used to synthesize the novel views [WKWT15, ZTS*16, TDB16, CSH19].

In summary, generative novel view synthesis can be regarded as a subset of image synthesis techniques and continues to evolve alongside advancements in image synthesis methods. Besides the generative models typically included, determining how to integrate information from the input view as a condition for synthesizing the novel view is the primary issue these methods are concerned with.

5. Datasets for 3D Generation

With the rapid development of technology, the ways of data acquisition and storage become more feasible and affordable, resulting in an exponential increase in the amount of available data. As data accumulates, the paradigm for problem-solving gradually shifts from data-driven to model-driven approaches, which in turn contributes to the growth of "Big Data" and "AIGC". Nowadays, data plays a crucial role in ensuring the success of algorithms. A well-curated dataset can significantly enhance a model's robustness and performance. On the contrary, noisy and flawed data may cause model bias that requires considerable effort in algorithm design to rectify.

In this section, we will go over the common data used for 3D generation. Depending on the methods employed, it usually includes 3D data (Section 5.1), multi-view image data (Section 5.2), and single-view image data (Section 5.3), which are also summarized in Tab. 4.

5.1. Learning from 3D Data

3D data could be collected by RGB-D sensors and other technology for scanning and reconstruction. Apart from 3D generation, 3D data is also widely used for other tasks like helping improve classical 2D vision task performance by data synthesis, environment simulation for training embodied AI agents, 3D object understanding, etc. One popular and frequently used 3D model database in the early stage is The Princeton Shape Benchmark [SMKF04]. It contains about 1800 polygonal models collected from the World Wide Web. While [KXD12] constructs a special rig that contains a 3D digitizer, a turntable, and a pair of cameras mounted on a sled that can move along a bent rail to capture the kit object models database. To evaluate the algorithms to detect and estimate the objects in the image given 3D models, [LPT13] introduces a dataset of 3D IKEA models obtained from Google Warehouse. Some 3D model databases are presented for tasks like robotic manipulation [CWS*15, MCL20], 3D shape retrieval [LLL*14], 3D shape modeling from a single image [SWZ*18]. BigBIRD [SSN*14] presents a large-scale dataset of 3D object instances that also includes multi-view images and depths, camera pose information, and segmented objects for each image.

However, those datasets are very small and only contain hundreds or thousands of objects. Collecting, organizing, and labeling larger datasets in computer vision and graphics communities is needed for data-driven methods of 3D content. To address this, ShapeNet [CFG*15] is introduced to build a large-scale repository of 3D CAD models of objects. The core of ShapeNet covers 55 common object categories with about 51,300 models that are manually verified category and alignment annotations. Thingi10K [ZJ16] collects 10,000 3D printing models from an online repository Thingiverse. While PhotoShape [PRFS18] produces 11,000 photorealistic

tic, relightable 3D shapes based on online data. Other datasets such as 3D-Future [FJG^{*}21], ABO [CGD^{*}22], GSO [DFK^{*}22] and OmniObject3D [WZF^{*}23] try to improve the texture quality but only contain thousands of models. Recently, Objaverse [DSS^{*}23] presents a large-scale corpus of 3D objects that contains over 800K 3D assets for research in the field of AI and makes a step toward a large-scale 3D dataset. Objaverse-XL [DLW^{*}23] further extends Objaverse to a larger 3D dataset of 10.2M unique objects from a diverse set of sources. These large-scale 3D datasets have the potential to facilitate large-scale training and boost the performance of 3D generation.

5.2. Learning from Multi-view Images

3D objects have been traditionally created through manual 3D modeling, object scanning, conversion of CAD models, or combinations of these techniques [DFK^{*}22]. These techniques may only produce synthetic data or real-world data of specific objects with limited reconstruction accuracy. Therefore, some datasets directly provide multi-view images in the wild which are also widely used in many 3D generation methods. ScanNet [DCS^{*}17] introduces an RGB-D video dataset containing 2.5M views in 1513 scenes and Objectron [AZA^{*}21] contains object-centric short videos and includes 4 million images in 14,819 annotated videos, of which only a limited number cover the full 360 degrees. CO3D [RSH^{*}21] extends the dataset from [HRL^{*}21] and increases the size to nearly 19,000 videos capturing objects from 50 MS-COCO categories, which has been widely used in the training and evaluations of novel view synthesis and 3D generation or reconstruction methods. Recently, MVImgNet [YXZ^{*}23] presents a large-scale dataset of multi-view images that collects 6.5 million frames from 219,188 videos by shooting videos of real-world objects in human daily life. Other lines of work provide the multi-view dataset in small-scale RGB-D videos [LBRF11, SHG^{*}22, CX^{*}23] compared with these works, large-scale synthetic videos [TME^{*}22], or egocentric videos [ZXA^{*}23]. A large-scale dataset is still a remarkable trend for deep learning methods, especially for generation tasks.

5.3. Learning from Single-view Images

3D generation methods usually rely on multi-view images or 3D ground truth to supervise the reconstruction and generation of 3D representation. Synthesizing high-quality multi-view images or 3D shapes using only collections of single-view images is a challenging problem. Benefiting from the unsupervised training of generative adversarial networks, 3D-aware GANs are introduced that could learn 3D representations in an unsupervised way from natural images. Therefore, several single-view image datasets are proposed and commonly used for these 3D generation methods. Although many large-scale image datasets have been presented for 2D generation, it is hard to directly use them for 3D generation due to the high uncertainty of this problem. Normally, these image datasets only contain a specific category or domain. FFHQ [KLA19], a real-world human face dataset consisting of 70,000 high-quality images at 1024^2 resolution, and AFHQ [CUYH20], an animal face dataset consisting of 15,000 high-quality images at 512^2 resolution, are introduced for 2D image synthesis and used a lot for 3D generation based on 3D-aware GANs. In the domain of the

Table 5: Recent 3D human generation techniques and their corresponding input-output formats.

Methods	Input Condition	Output Texture
ICON [XYTB22]	Single-Image	✗
ECON [XYC [*] 23]	Single-Image	✗
gDNA [CJS [*] 22]	Latent	✗
Chupa [KKL [*] 23]	Text/Latent	✗
ELICIT [HYL [*] 23]	Single-Image	✓
TeCH [HYX [*] 23]	Single-Image	✓
Get3DHuman [XKJ [*] 23]	Latent	✓
EVA3D [HCL [*] 22]	Latent	✓
AvatarCraft [JWZ [*] 23]	Text	✓
DreamHuman [KAZ [*] 23]	Text	✓
TADA [LYX [*] 24]	Text	✓

human body, SHHQ [FLJ^{*}22] and DeepFashion [LLQ^{*}16] have been adopted for 3D human generation. In terms of objects, many methods [LSMG20, GMW17, HMR19a, ZZZ^{*}18, WZX^{*}16] render synthetic single-view datasets using several major object categories of ShapeNet. While GRAF [SLNG20] renders 150k Chairs from Photoshapes [PRFS18]. Moreover, CelebA [LLWT15] and Cats [ZST08] datasets are also commonly used to train the models like HoloGAN [NPLT^{*}19] and pi-GAN [CMK^{*}21a]. Since the single-view images are easy to obtain, these methods could collect their own dataset for the tasks.

6. Applications

In this section, we introduce various 3D generation tasks (Sec. 6.1-6.3) and closely related 3D editing tasks (Sec. 6.4). The generation tasks are divided into three categories, including 3D human generation (Sec. 6.1), 3D face generation (Sec. 6.2), and generic object and scene generation (Sec. 6.3).

6.1. 3D Human Generation

With the emergence of the metaverse and the advancements in virtual 3D social interaction, the field of 3D human digitization and generation has gained significant attention in recent years. Different from general 3D generation methods that focus on category-free rigid objects with sample geometric structures [PJBM23, LXZ^{*}23], most 3D human generation methods aim to tackle the complexities of articulated pose changes and intricate geometric details of clothing. Tab. 5 presents a compilation of notable 3D human body generation methods in recent years, organized according to the input conditions and the output format of the generated 3D human bodies. Some results of these methods are shown in Fig. 8. Specifically, in terms of the input condition, current 3D human body generation methods can be categorized based on the driving factors including latent features randomly sampled from a pre-defined latent space [MYR^{*}20, CJS^{*}22, HCL^{*}22], a single reference image [APMTM19, CPA^{*}21, XYC^{*}23, HYX^{*}23, ZLZ^{*}23], or text prompts [KKL^{*}23, JWZ^{*}23, KAZ^{*}23, LYX^{*}24]. According to the form of the final output, these methods can be classified into two categories: textureless shape generation [APMTM19,

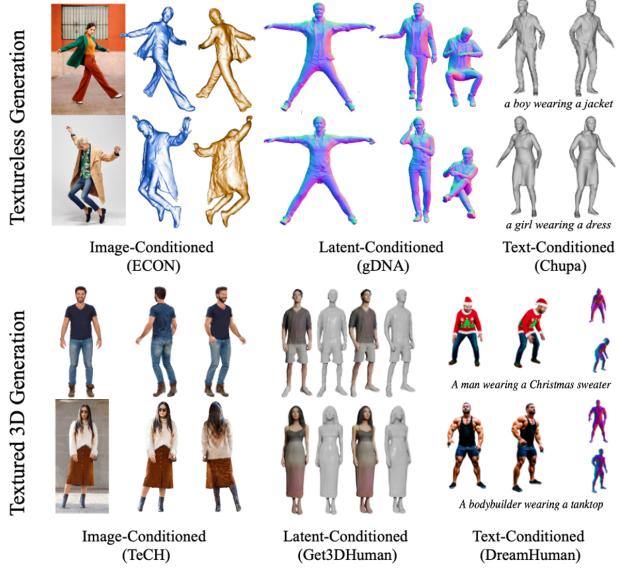


Figure 8: Examples of 3D human generation methods. 3D generation results source from ECON [XYC*23], gDNA [CJS*22], Chupa [KKL*23], TeCH [HYX*23], Get3DHuman [XKJ*23], and DreamHuman [KAZ*23].

XYTB22, XYC*23, CJS*22, MYR*20, CPA*21, KKL*23] and textured body generation [AZS22, LYX*24, HYL*23, KAZ*23, XKJ*23, HYX*23, ZLZ*23]. While the latter focuses on generating fully textured 3D clothed humans, the former aims to obtain textureless body geometry with realistic details.

In terms of textureless shape generation, early works [CPB*20, OBB20, LXC*21] attempt to predict SMPL parameters from the input image and infer a skinned SMPL mesh as the generated 3D representation of the target human. Nevertheless, such skinned body representation fails to represent the geometry of clothes. To overcome this issue, [APMTM19, XYTB22, XYC*23] leverage a pre-trained neural network to infer the normal information and combine the skinned SMPL mesh to deduce a clothed full-body geometry with details. In contrast to such methods, which require reference images as input, CAPE [MYR*20] proposes a generative 3D mesh model conditioned on latents of SMPL pose and clothing type to form the clothing deformation from the SMPL body. gDNA [CJS*22] introduces a generation framework conditioned on latent codes of shape and surface details to learn the underlying statistics of 3D clothing details from scanned human datasets via an adversarial loss. Different from the previous methods that generate an integrated 3D clothed human body geometry, SMPLicit [CPA*21] adopts an implicit model conditioned on shape and pose parameters to individually generate diverse 3D clothes. By combining the SMPL body and associated generated 3D clothes, SMPLicit enables to produce 3D clothed human shapes. To further improve the quality of the generated human shape, Chupa [KKL*23] introduces diffusion models to generate realistic human geometry and decompose the 3D generation task into 2D normal map generation and normal map-based 3D reconstruction.

Although these methods achieve the generation of detailed clothed human shapes, their application prospects are greatly restricted due to the lack of texture-generation capabilities. To generate textured clothed 3D human, lots of attempts have been made in previous work, including methods conditioned on latent codes [GII*21, BKY*22, ZJY*22, NSLH22, JJW*23, YLWD22, XKJ*23, CHB*23, HCL*22, XYB*23, AYS*23], single images [SHN*19, ZYLD21, AZS22, CMA*22, GLZ*23, HYL*23, YLX*23, HHP*23, AST*23, HYX*23, ZLZ*23], and text prompts [HZP*22, JWZ*23, CCH*23, HWZ*23, KAZ*23, ZCY*23, LYX*24, HSZ*23, ZZZ*23a, LZT*23]. Most latent-conditioned methods employ adversarial losses to restrict their latent space and generate 3D human bodies within the relevant domain of the training dataset. For example, StylePeople [GII*21] combines StyleGAN [KLA*20] and neural rendering to design a joint generation framework trained in an adversarial fashion on the full-body image datasets. Furthermore, GNARF [BKY*22] and AvatarGen [ZJY*22] employ tri-planes as the 3D representation and replace the neural rendering with volume rendering to enhance the view-consistency of rendered results. To improve editability, Get3DHuman [XKJ*23] divides the human body generation framework into shape and texture branches respectively conditioned on shape and texture latent codes, achieving re-texturing. EVA3D [HCL*22] divides the generated human body into local parts to achieve controllable human poses.

As text-to-image models [RKH*21b, RBL*22b, SCS*22] continue to advance rapidly, the field of text-to-3D has also reached its pinnacle of development. For the text-driven human generation, existing methods inject priors from pre-trained text-to-image models into the 3D human generation framework to achieve text-driven textured human generation, such as AvatarCLIP [HZP*22], AvatarCraft [JWZ*23], DreamHuman [KAZ*23], and TADA [LYX*24]. Indeed, text-driven human generation methods effectively address the challenge of limited 3D training data and significantly enhance the generation capabilities of 3D human assets. Nevertheless, in contrast to the generation of unseen 3D humans, it is also significant to generate a 3D human body from a specified single image in real-life applications. In terms of single-image-conditioned 3D human generation methods, producing generated results with textures and geometries aligned with the input reference image is widely studied. To this end, PIFu [SHN*19], PaMIR [ZYLD21], and PHORHUM [AZS22] propose learning-based 3D generators trained on scanned human datasets to infer human body geometry and texture from input images. However, their performance is constrained by the limitations of the training data. Consequently, they struggle to accurately infer detailed textures and fine geometry from in-the-wild input images, particularly in areas that are not directly visible in the input. To achieve data-free 3D human generation, ELICIT [HYL*23], Human-SGD [AST*23], TeCH [HYX*23], and HumanRef [ZLZ*23] leverage priors of pre-trained CLIP [RKH*21b] or image diffusion models [RBL*22b, SCS*22] to predict the geometry and texture based on the input reference image without the need for 3D datasets, and achieve impressive qualities in generated 3D clothed human.

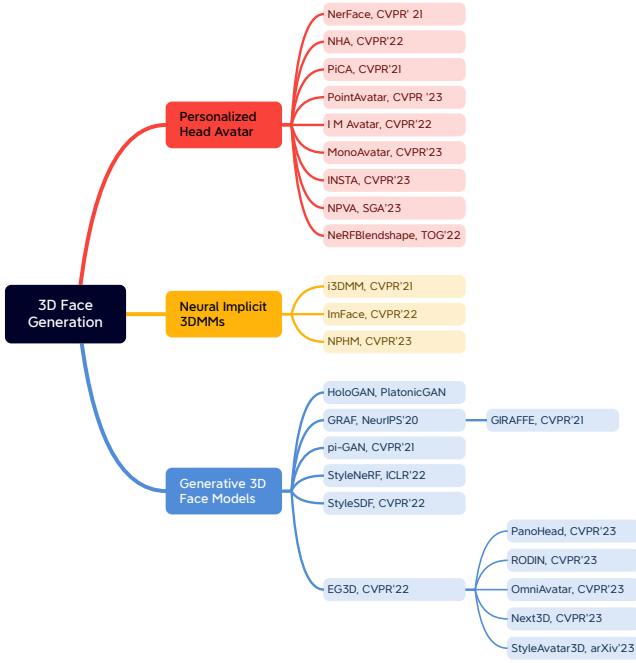


Figure 9: Representative applications and methods of 3D face generation.

6.2. 3D Face Generation

One essential characteristic of 3D face generation tasks is to generate high-quality human face images that can be viewed from different viewpoints. Popular tasks can be loosely classified into three major categories, including personalized head avatar creation (e.g. 3D talking head generation), neural implicit 3D morphable models (3DMMs), and generative 3D face models, which are shown in Fig. 9 and Fig. 10.

Personalized head avatar creation aims at creating an animatable avatar that can be viewed from different viewpoints of the target person, which has broad applications such as talking head generation. Most of the existing methods take as input a sequence of video frames (i.e. monocular video) [PSB*21, GTZN21, GPL*22, ZAB*22, ZBT23, ZYW*23, BTH*23, GZX*22]. Although convenient, the viewing angles of these avatars are limited in a relatively small range (i.e. near frontal) and their quality is not always satisfactory due to limited data. In contrast, another stream of works [LSSS18, MSS*21, LSS*21, WKC*23, KQG*23] aims at creating a very high-quality digital human that can be viewed from larger angles (e.g. side view). These methods usually require high-quality synchronized multi-view images under even illumination. However, both streams rely heavily on implicit or hybrid neural representations and neural rendering techniques. The quality and animation accuracy of the generated talking head video are usually measured with PSNR, SSIM, and LPIPS metrics.

Neural implicit 3DMMs. Traditional 3D morphable face models (3DMMs) assume a predefined template mesh (i.e. fixed topology) for the geometry and have explored various modeling methods in-

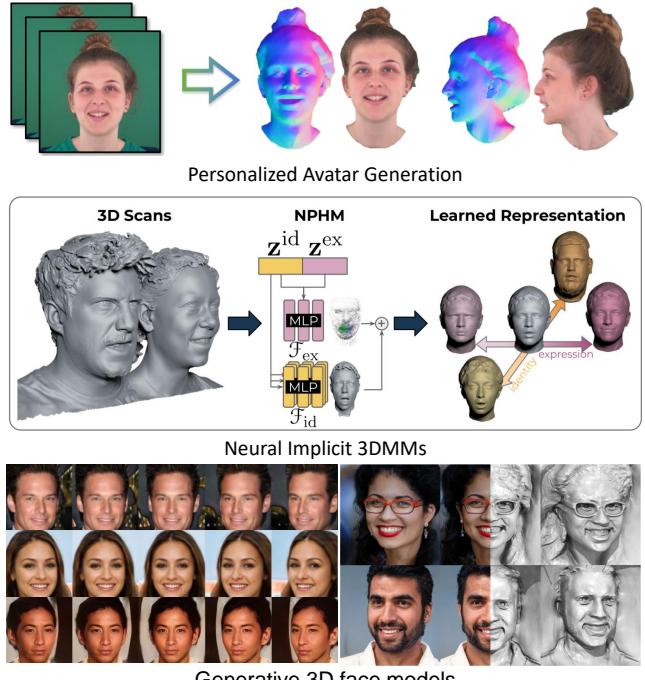


Figure 10: Representative 3D face generation tasks. Images adapted from NHA [GPL*22], NPHM [GKG*23], and EG3D [CLC*22].

cluding linear models (e.g. PCA-based 3DMMs) and non-linear models (e.g. network-based 3DMMs). A comprehensive survey of these methods has been discussed in [EST*20]. Recently, thanks to the rapid advances in implicit neural representations (INRs), several neural implicit 3DMMs utilizing INRs for face modeling emerges [YTB*21, ZYHC22, GKG*23] since continuous implicit neural representations do not face discretization error and can theoretically model infinite details. Indeed, NPHM [GKG*23] can generate more subtle expressions unseen in previous mesh-based 3DMMs. What's more, neural implicit 3DMMs can potentially model hair better since the complexity of different hairstyles varies drastically, which imposes a great challenge for fixed topology mesh-based traditional 3DMMs.

Generative 3D face models. One key difference from 2D generative face models (e.g. StyleGAN [KLA19, KLA*20]) is that 3D face models can synthesize multi-view consistent images (i.e. novel views) of the same target (identity and clothes). Early attempts towards this direction include HoloGAN [NPLT*19] and PlatonicGAN [HMR19b], which are both voxel-based methods and can only generate images in limited resolution. Quickly, methods [SLNG20, NG21, CMK*21b, OELS*22, GLWT22, CLC*22] utilizing neural radiance fields are proposed to increase the image resolution. For example, EG3D [CLC*22] proposes a hybrid tri-plane representation, which strikes a good trade-off to effectively address the memory and rendering inefficiency faced by previous generative 3D GANs and can produce high-quality images with good multi-view consistency.

Thanks to the success of various 3D GANs, many down-

Table 6: Applications of general scene generation methods.

Methods	Type	Condition	Texture Generation
PVD [ZDW21]	Object-Centered	Latent	✗
NFD [SCP*23]	Object-Centered	Latent	✗
Point-E [NJD*22]	Object-Centered	Text	✗
Diffusion-SDF [LDZL23]	Object-Centered	Text	✗
Deep3DSketch+ [CFZ*23]	Object-Centered	Sketch	✗
Zero-1-to-3 [LWVH*23]	Object-Centered	Single-Image	✓
Make-It-3D [TWZ*23]	Object-Centered	Single-Image	✓
GET3D [GSW*22]	Object-Centered	Latent	✓
EG3D [CLC*22]	Object-Centered	Latent	✓
CLIP-Mesh [MKXBP22]	Object-Centered	Text	✓
DreamFusion [PJBM23]	Object-Centered	Text	✓
ProlificDreamer [WLW*23]	Object-Centered	Text	✓
PixelSynth [RFJ21]	Outward-Facing	Single-Image	✓
DiffDreamer [CCP*23]	Outward-Facing	Single-Image	✓
Xiang <i>et al.</i> [XYHT23]	Outward-Facing	Latent	✓
CC3D [BPP*23]	Outward-Facing	Layout	✓
Text2Room [HCO*23]	Outward-Facing	Text	✓
Text2NeRF [ZLW*23]	Outward-Facing	Text	✓

stream applications (e.g. editing, talking head generation) are enabled or become less data-hungry, including 3D consistent editing [SWW*23, SWZ*22, SWS*22, LFLSY*23, JCL*22], 3D talking head generation [BFW*23, XSJ*23, WDY*22], etc.

6.3. General Scene Generation

Different from 3D human and face generation, which can use existing prior knowledge such as SMPL and 3DMM, general scene generation methods are more based on the similarity of semantics or categories to design a 3D model generation framework. Based on the differences in generation results, as shown in Fig. 11 and Tab. 6, we further subdivide general scene generation into object-centered asset generation and outward-facing scene generation.

6.3.1. Object-Centered Asset Generation

The field of object-centered asset generation has seen significant advancements in recent years, with a focus on both textureless shape generation and textured asset generation. For the textureless shape generation, early works use GAN-based networks to learn a mapping from latent space to 3D object space based on specific categories of 3D data, such as 3D-GAN [WZX*16], HoloGAN [NPLT*19], and PlatonicGAN [HMR19b]. However, limited by the generation capabilities of GANs, these methods can only generate rough 3D assets of specific categories. To improve the quality of generated results, SingleShapeGen [WZ22] leverages a pyramid of generators to generate 3D assets in a coarse to fine manner. Given the remarkable achievements of diffusion models in image generation, researchers are directing their attention towards the application of diffusion extensions in the realm of 3D generation. Thus, subsequent methods [LH21, ZDW21, HLHF22, SCP*23, EMS*23] explore the use of diffusion processes for 3D

shape generation from random noise. In addition to these latent-based methods, another important research direction is text-driven 3D asset generation [CCS*19, LWQF22]. For example, 3D-LDM [NKR*22], SDFusion [CLT*23], and Diffusion-SDF [LDZL23] achieve text-to-3D shape generation by designing the diffusion process in 3D feature space. Due to such methods requiring 3D datasets to train the diffusion-based 3D generators, they are limited to the training data in terms of the categories and diversity of generated results. By contrast, CLIP-Forge [SCL*22], CLIP-Sculptor [SFL*23], and Michelangelo [ZLC*23] directly employ the prior of the pre-trained CLIP model to constrain the 3D generation process, effectively improving the generalization of the method and the diversity of generation results. Unlike the above latent-conditioned or text-driven 3D generation methods, to generate 3D assets with expected shapes, there are some works [HMR19a, CFZ*23] that explore image or sketch-conditioned generation.

In comparison to textureless 3D shape generation, textured 3D asset generation not only produces realistic geometric structures but also captures intricate texture details. For example, HoloGAN [NPLT*19], GET3D [GSW*22], and EG3D [CLC*22] employ GAN-based 3D generators conditioned on latent vectors to produce category-specific textured 3D assets. By contrast, text-driven 3D generation methods rely on the prior knowledge of pre-trained large-scale text-image models to enable category-free 3D asset generation. For instance, CLIP-Mesh [MKXBP22], Dream Fields [JMB*22], and PureCLIPNeRF [LC22] employ the prior of CLIP model to constrain the optimization process and achieve text-driven 3D generation. Furthermore, DreamFusion [PJBM23] and SJC [WDL*23] propose a score distillation sampling (SDS) method to achieve 3D constraint which priors extracted from pre-trained 2D diffusion models. Then, some methods further improve the SDS-based 3D generation process in terms of generation quality, multi-face problem, and optimization efficiency, such as Magic3D [LGT*23], Latent-NeRF [MRP*23b], Fantasia3D [CCJJ23], DreamBooth3D [RKP*23], HiFA [ZZ23], ATT3D [LXZ*23], ProlificDreamer [WLW*23], IT3D [CZY*23], DreamGaussian [TRZ*23], and CAD [WPH*23]. On the other hand, distinct from text-driven 3D generation, single-image-conditioned 3D generation is also a significant research direction [LWVH*23, MKLRV23, CGC*23, WLY*23, KDJ*23].

6.3.2. Outward-Facing Scene Generation

Early scene generation methods often require specific scene data for training to obtain category-specific scene generators, such as GAUDI [BGA*22] and the work of Xiang *et al.* [XYHT23], or implement a single scene reconstruction based on the input image, such as PixelSynth [RFJ21] and Worldsheets [HRBP21]. However, these methods are either limited by the quality of the generation or by the extensibility of the scene. With the rise of diffusion models in image inpainting, various methods are beginning to use the scene completion capabilities of diffusion models to implement scene generation tasks [CCP*23, HCO*23, ZLW*23]. Recently, SceneScape [FAKD23], Text2Room [HCO*23], Text2NeRF [ZLW*23], and LucidDreamer [CLN*23] propose progressive inpainting and updating strategies for generating realistic 3D scenes using pre-trained diffusion models. SceneScape and Text2Room utilize explicit polygon meshes as their

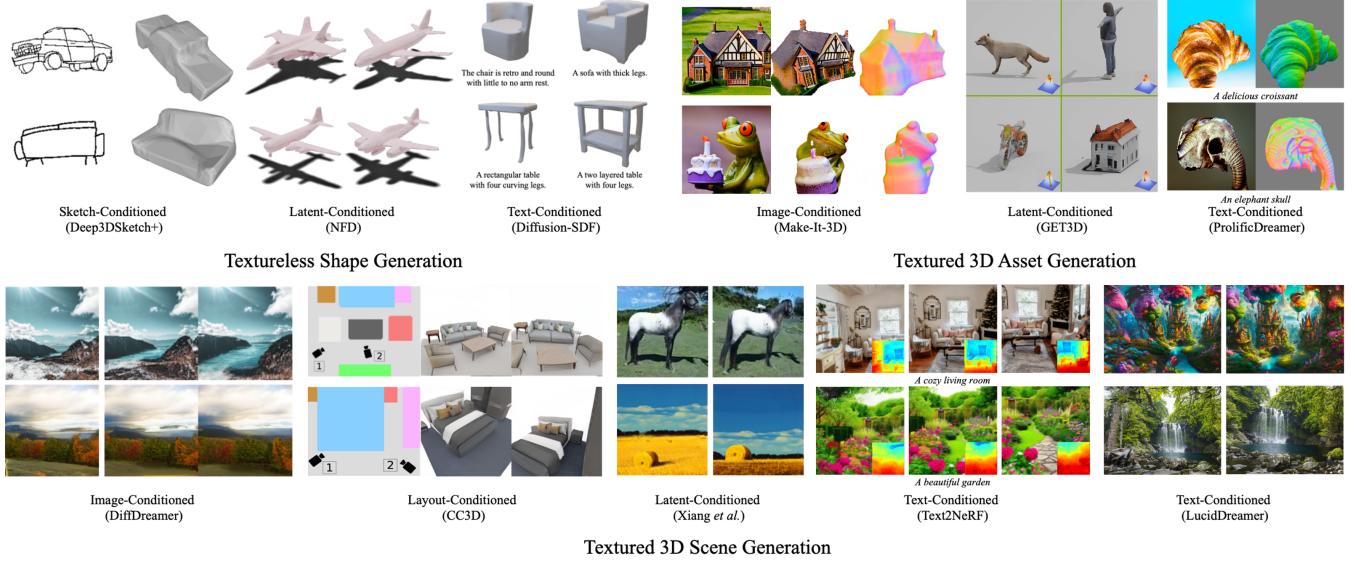


Figure 11: Some examples of general scene generation methods. 3D generation results source from Deep3DSketch+ [CFZ^{*}23], NFD [SCP^{*}23], Diffusion-SDF [LDZL23], Make-It-3D [TWZ^{*}23], GET3D [GSW^{*}22], ProlificDreamer [WLW^{*}23], Diff-Dreamer [CCP^{*}23], CC3D [BPP^{*}23], Xiang *et al.* [XYHT23], Text2NeRF [ZLW^{*}23], and LucidDreamer [CLN^{*}23].

3D representation during the generation procedure. However, this choice of representation imposes limitations on the generation of outdoor scenes, resulting in stretched geometry and blurry artifacts in the fusion regions of mesh faces. In contrast, Text2NeRF and LucidDreamer adopt implicit representations, which offer the ability to model fine-grained geometry and textures without specific scene requirements. Consequently, Text2NeRF and LucidDreamer can generate both indoor and outdoor scenes with high fidelity.

6.4. 3D Editing

Based on the region where editing happens, we classify the existing works into global editing and local editing.

6.4.1. Global Editing

Global editing works aim at changing the appearance or geometry of the competing 3D scene globally. Different from local editing, they usually do not intentionally isolate a specific region from a complete and complicated scene or object. Most commonly, they only care if the resultant scene is in a desired new “style” and resembles (maintains some features of) the original scene. Most representative tasks falling into this category include stylization [HTS^{*}21, HHY^{*}22, FJW^{*}22, ZKB^{*}22, WJC^{*}23, HTE^{*}23], and single-object manipulation (e.g. re-texturing [MBOL^{*}22, LZJ^{*}22, MRP^{*}23b, CCJJ23]) as shown in Fig. 12.

Stylization. Early 3D scene stylization methods [HTS^{*}21, HHY^{*}22, FJW^{*}22, ZKB^{*}22] usually require style reference. The 3D scene is optimized either in the style feature space using a Gram matrix [GEB16] or nearest neighbor feature matching [ZKB^{*}22] loss or in the image space using the output color of a deep image style transfer network [HB17]. Later

methods [WJC^{*}23, HTE^{*}23] can support textual format style definition by utilizing the learned prior knowledge from large-scale language-vision models such as CLIP [RKH^{*}21a] and Stable Diffusion [RBL^{*}22a]. Other than commonly seen artistic style transfer, there also exist some special types of “style” manipulation tasks such as seasonal and illumination manipulations [LLF^{*}23, CZL^{*}22, HTE^{*}23, CYL^{*}22] and climate changes.

Single-Object Manipulation. There are many papers specifically aim at manipulating a single 3D object. For example, one representative task is texturing or painting a given 3D object (usually in mesh format) [MBOL^{*}22, LZJ^{*}22, MRP^{*}23b, CCJJ23, CSL^{*}23]. Except for diffuse albedo color and vertex displacement [MBOL^{*}22, MZS^{*}23, LYX^{*}24], other common property maps may be involved, including normal map [CCJJ23, LZJ^{*}22], roughness map [CCJJ23, LZJ^{*}22], specular map [LZJ^{*}22], and metallic map [CCJJ23], etc. A more general setting would be directly manipulating a NeRF-like object [WCH^{*}22, LZJ^{*}22, TLYCS22, YBZ^{*}22]. Notably, the human face/head is one special type of object that has drawn a lot of interest [ATDN23, ZQL^{*}23]. In the meanwhile, many works focus on fine-grained local face manipulation, including expression and appearance manipulation [SWZ^{*}22, SWS^{*}22, LFLSY^{*}23, JCL^{*}22, WDY^{*}22, XSJ^{*}23, MLL^{*}22a, ZLW^{*}22] and face swapping [LMY^{*}23] since human face related understanding tasks (e.g. recognition, parsing, attribute classification) have been extensively studied previously.

6.4.2. Local Editing

Local editing tasks intentionally modify only a specific region, either manually provided ([MPS^{*}23, LDS^{*}23, CYW^{*}23]) or automatically determined ([YZX^{*}21, WLC^{*}22, WWL^{*}23, KMS22, JKK^{*}23]), of the complete scene or object. Common

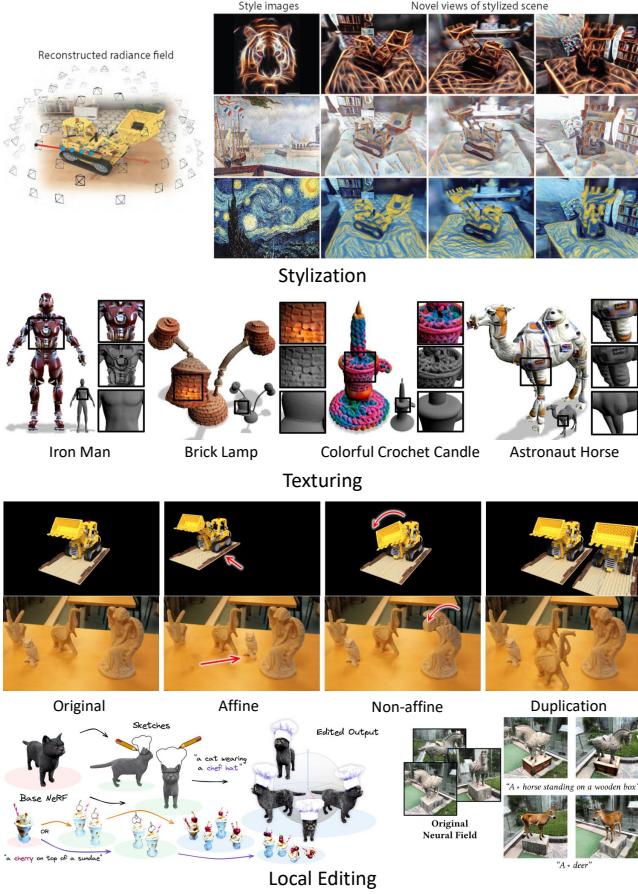


Figure 12: Representative 3D editing tasks. Images adapted from ARF [ZKB*22], Text2Mesh [MBOL*22], NeRFShop [JKK*23], SKED [MPS*23], and DreamEditor [ZWL*23].

local editing types include appearance manipulation [YBZ*22, ZWL*23], geometry deformation [JKK*23, PYL*22, YSL*22, TLYCS22], object-/semantic-level duplication/deletion and moving/removing [YZX*21, WLC*22, KMS22, WWL*23]. For example, NeuMesh [YBZ*22] supports several kinds of texture manipulation including swapping, filling, and painting since they distill a NeRF scene into a mesh-based neural representation. NeRF-Shop [JKK*23] and CageNeRF [PYL*22] transform/deform the volume bounded by a mesh cage, resulting in moved or deformed/articulated object. SINE [BZY*23] updates both the NeRF geometry and the appearance with geometry prior and semantic (image feature) texture prior as regularizations.

Another line of works (e.g. ObjectNeRF [YZX*21], ObjectSDF [WLC*22], DFF [KMS22]) focus on automatically decomposing the scene into individual objects or semantic parts during reconstruction, which is made possible by utilizing extra 2D image understanding networks (e.g. instance segmentation), and support subsequent object-level manipulations such as re-coloring, removal, displacement, duplication.

Recently, it is possible to create new textures and/or content

only according to text description in the existing 3D scenes due to the success of large-scale text-to-image models (e.g. Stable Diffusion [RBL*22a]). For example, instruct-NeRF2NeRF [HTE*23] iteratively updates the reference dataset images modified by a dedicated diffusion model [BHE23] and the NeRF model. DreamEditor [ZWL*23] performs local updates on the region located by text attention guided by score distillation sampling [PJBM23]. Focal-Dreamer [LDS*23] creates new geometries (objects) in the specified empty spaces according to the text input. SKED [MPS*23] supports both creating new objects and modifying the existing part located in the region specified by the provided multi-view sketches.

7. Open Challenges

The quality and diversity of 3D generation results have experienced significant progress due to advancements in generative models, 3D representations, and algorithmic paradigms. Considerable attention has been drawn to 3D generation recently as a result of the success achieved by large-scale models in natural language processing and image generation. However, numerous challenges remain before the generated 3D models can meet the high industrial standards required for video games, movies, or immersive digital content in VR/AR. In this section, we will explore some of the open challenges and potential future directions in this field.

Evaluation. Quantifying the quality of generated 3D models objectively is an important and not widely explored problem. Using metrics such as PSNR, SSIM, and F-Score to evaluate rendering and reconstruction results requires ground truth data on the one hand, but on the other hand, it can not comprehensively reflect the quality and diversity of the generated content. In addition, user studies are usually time-consuming, and the study results tend to be influenced by the bias and number of surveyed users. Metrics that capture both the quality and diversity of the results like FID can be applied to 3D data, but may not be always aligned with 3D domain and human preferences. Better metrics to judge the results objectively in terms of generation quality, diversity, and matching degree with the conditions still need further exploration.

Dataset. Unlike language or 2D image data which can be easily captured and collected, 3D assets often require 3D artists or designers to spend a significant amount of time using professional software to create. Moreover, due to the different usage scenarios and creators' personal styles, these 3D assets may differ greatly in scale, quality, and style, increasing the complexity of 3D data. Specific rules are needed to normalize this diverse 3D data, making it more suitable for generation methods. A large-scale, high-quality 3D dataset is still highly desirable in 3D generation. Meanwhile, exploring how to utilize extensive 2D data for 3D generation could also be a potential solution to address the scarcity of 3D data.

Representation. Representation is an essential part of the 3D generation, as we discuss various representations and the associated methods in Sec. 3. Implicit representation is able to model complex geometric topology efficiently but faces challenges with slow optimization; explicit representation facilitates rapid optimization convergence but struggles to encapsulate complex topology and demands substantial storage resources; Hybrid representation attempts to consider the trade-off between these two, but there are

still shortcomings. In general, we are motivated to develop a representation that balances optimization efficiency, geometric topology flexibility, and resource usage.

Controllability. The purpose of the 3D generation technique is to generate a large amount of user-friendly, high-quality, and diverse 3D content in a cheap and controllable way. However, embedding the generated 3D content into practical applications remains a challenge: most methods [PJBM23, CLC^{*}22, YHH^{*}19b] rely on volume rendering or neural rendering, and fail to generate content suitable for rasterization graphics pipeline. As for methods [CCJJ23, WLW^{*}23, TRZ^{*}23] that generate the content represented by polygons, they do not take into account layout (e.g. the rectangular plane of a table can be represented by two triangles) and high-quality UV unwrapping and the generated textures also face some issues such as baked shadows. These problems make the generated content unfavorable for artist-friendly interaction and editing. Furthermore, the style of generated content is still limited by training datasets. Furthermore, the establishment of comprehensive toolchains is a crucial aspect of the practical implementation of 3D generation. In modern workflows, artists use tools (e.g. LookDev) to harmonize 3D content by examining and contrasting the relighting results of their materials across various lighting conditions. Concurrently, modern Digital Content Creation (DCC) software offers extensive and fine-grained content editing capabilities. It is promising to unify 3D content produced through diverse methods and establish tool chains that encompass abundant editing capabilities.

Large-scale Model. Recently, the popularity of large-scale models has gradually affected the field of 3D generation. Researchers are no longer satisfied with using distillation scores that use large-scale image models as the priors to optimize 3D content, but directly train large-scale 3D models. MeshGPT [SAA^{*}23] follows large language models and adopts a sequence-based approach to autoregressively generate sequences of triangles in the generated mesh. MeshGPT takes into account layout information and generates compact and sharp meshes that match the style created by artists. Since MeshGPT is a decoder-only transformer, compared with the optimization-based generation, it gets rid of inefficient multi-step sequential optimization, achieving rapid generation. Despite this, MeshGPT's performance is still limited by training datasets and can only generate regular furniture objects. But there is no doubt that large-scale 3D generation models have great potential worth exploring.

8. Conclusion

In this work, we present a comprehensive survey on 3D generation, encompassing four main aspects: 3D representations, generation methods, datasets, and various applications. We begin by introducing the 3D representation, which serves as the backbone and determines the characteristics of the generated results. Next, we summarize and categorize a wide range of generation methods, creating an evolutionary tree to visualize their branches and developments. Finally, we provide an overview of related datasets, applications, and open challenges in this field. The realm of 3D generation is currently witnessing explosive growth and development, with new work emerging every week

or even daily. We hope this survey offers a systematic summary that could inspire subsequent work for interested readers.

References

- [ACC^{*}22] ADAMKIEWICZ M., CHEN T., CACCAVALE A., GARDNER R., CULBERTSON P., BOHG J., SCHWAGER M.: Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters* 7, 2 (2022), 4606–4613. 5
- [ADMG18] ACHLIOPAS P., DIAMANTI O., MITLIAGKAS I., GUIBAS L.: Learning representations and generative models for 3d point clouds. In *International conference on machine learning* (2018), PMLR, pp. 40–49. 7, 9
- [ALG^{*}20] ATTAL B., LING S., GOKASLAN A., RICHARDT C., TOMPKIN J.: Matryodshka: Real-time 6dof video view synthesis using multi-sphere images. In *European Conference on Computer Vision* (2020), Springer, pp. 441–459. 5
- [APMTM19] ALLDIECK T., PONS-MOLL G., THEOBALT C., MAGNOR M.: Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 2293–2303. 15, 16
- [AQW19] ABDAL R., QIN Y., WONKA P.: Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 4432–4441. 2, 6
- [ASK^{*}20] ALIEV K.-A., SEVASTOPOLSKY A., KOLOS M., ULYANOV D., LEMPITSKY V.: Neural point-based graphics. In *Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16* (2020), Springer, pp. 696–712. 4
- [AST^{*}23] ALBAHAR B., SAITO S., TSENG H.-Y., KIM C., KOPF J., HUANG J.-B.: Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers* (2023), pp. 1–11. 16
- [ATDN23] ANEJA S., THIES J., DAI A., NIESSNER M.: ClipFace: Text-guided editing of textured 3d morphable models. In *ACM SIGGRAPH 2023 Conference Proceedings* (2023). 19
- [AYS^{*}23] ABDAL R., YIFAN W., SHI Z., XU Y., PO R., KUANG Z., CHEN Q., YEUNG D.-Y., WETZSTEIN G.: Gaussian shell maps for efficient 3d human generation. *arXiv preprint arXiv:2311.17857* (2023). 16
- [AZA^{*}21] AHMADYAN A., ZHANG L., ABLAVATSKI A., WEI J., GRUNDMANN M.: Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 7822–7831. 15
- [AZS22] ALLDIECK T., ZANFIR M., SMINCHISESCU C.: Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 1506–1515. 16
- [BBJ^{*}21] BOSS M., BRAUN R., JAMPANI V., BARRON J. T., LIU C., LENSCHE H.: Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 12684–12694. 5
- [BFO^{*}20] BROXTON M., FLYNN J., OVERBECK R., ERICKSON D., HEDMAN P., DUVAL M., DOURGARIAN J., BUSCH J., WHALEN M., DEBEVEC P.: Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 86–1. 5
- [BFW^{*}23] BAI Y., FAN Y., WANG X., ZHANG Y., SUN J., YUAN C., SHAN Y.: High-fidelity facial avatar reconstruction from monocular video with generative priors. In *CVPR* (2023). 18

- [BGA*22] BAUTISTA M. A., GUO P., ABNAR S., TALBOTT W., TOSHEV A., CHEN Z., DINH L., ZHAI S., GOH H., ULRICHT D., ET AL.: Gaudi: A neural architect for immersive 3d scene generation. In *Advances in Neural Information Processing Systems 35* (2022), 25102–25116. [18](#)
- [BGP*22] BAATZ H., GRANSKOG J., PAPAS M., ROUSSELLE F., NOVÁK J.: Nerf-tex: Neural reflectance field textures. In *Computer Graphics Forum* (2022), vol. 41, Wiley Online Library, pp. 287–301. [5](#)
- [BHE23] BROOKS T., HOLYNSKI A., EFROS A. A.: InstructPix2Pix: Learning to follow image editing instructions. In *CVPR* (2023). [20](#)
- [BHMK*18] BEN-HAMU H., MARON H., KEZURER I., AVINERI G., LIPMAN Y.: Multi-chart generative surface modeling. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–15. [7, 9](#)
- [BKP*10] BOTSCHE M., KOBELT L., PAULY M., ALLIEZ P., LÉVY B.: *Polygon mesh processing*. CRC press, 2010. [4](#)
- [BKY*22] BERGMAN A., KELLNHOFER P., YIFAN W., CHAN E., LINDELL D., WETZSTEIN G.: Generative neural articulated radiance fields. In *Advances in Neural Information Processing Systems 35* (2022), 19900–19916. [16](#)
- [BLW16] BROCK A., LIM T., WESTON N.: Generative and discriminative voxel modeling with convolutional neural networks. *arXiv preprint arXiv:1608.04236* (2016). [6, 11](#)
- [BMR*20] BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SAstry G., ASKELL A., ET AL.: Language models are few-shot learners. *NeurIPS* (2020). [2, 6, 10](#)
- [BMT*21] BARRON J. T., MILDENHALL B., TANCIK M., HEDMAN P., MARTIN-BRUALLA R., SRINIVASAN P. P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 5855–5864. [1, 5](#)
- [BMV*22] BARRON J. T., MILDENHALL B., VERBIN D., SRINIVASAN P. P., HEDMAN P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 5470–5479. [5](#)
- [BNT21] BUROV A., NIESSNER M., THIES J.: Dynamic surface function networks for clothed human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 10754–10764. [4](#)
- [BPP*23] BAHMANI S., PARK J. J., PASCHALIDOU D., YAN X., WETZSTEIN G., GUIBAS L., TAGLIASACCHI A.: Cc3d: Layout-conditioned generation of compositional 3d scenes. *arXiv preprint arXiv:2303.12074* (2023). [18, 19](#)
- [BSKG22] BEN-SHABAT Y., KONEPUTUGODAGE C. H., GOULD S.: Digs: Divergence guided shape implicit neural representation for unoriented point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 19323–19332. [6](#)
- [BTH*23] BAI Z., TAN F., HUANG Z., SARKAR K., TANG D., QIU D., MEKA A., DU R., DOU M., ORTS-ESCOLANO S., ET AL.: Learning personalized high quality volumetric head avatars from monocular rgb videos. In *CVPR* (2023). [17](#)
- [BZY*23] BAO C., ZHANG Y., YANG B., FAN T., YANG Z., BAO H., ZHANG G., CUI Z.: SINE: Semantic-driven image-based nerf editing with prior-guided editing field. In *CVPR* (2023). [20](#)
- [CBZ*19] CHENG S., BRONSTEIN M., ZHOU Y., KOTSIA I., PANTIC M., ZAFEIRIOU S.: Meshgan: Non-linear 3d morphable models of faces. *arXiv preprint arXiv:1903.10384* (2019). [7](#)
- [CCH*23] CAO Y., CAO Y.-P., HAN K., SHAN Y., WONG K.-Y. K.: Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916* (2023). [16](#)
- [CCJJ23] CHEN R., CHEN Y., JIAO N., JIA K.: Fantasia3D: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV* (2023). [18, 19, 21](#)
- [CCP*23] CAI S., CHAN E. R., PENG S., SHAHBAZI M., OBUKHOV A., VAN GOOL L., WETZSTEIN G.: Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 2139–2150. [18, 19](#)
- [CCS*19] CHEN K., CHOY C. B., SAVVA M., CHANG A. X., FUNKHOUSER T., SAVARESE S.: Text2shape: Generating shapes from natural language by learning joint embeddings. In *Computer Vision-ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III* 14 (2019), Springer, pp. 100–116. [18](#)
- [CCW*23] CHEN Y., CHEN X., WANG X., ZHANG Q., GUO Y., SHAN Y., WANG F.: Local-to-global registration for bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 8264–8273. [5](#)
- [CFG*15] CHANG A. X., FUNKHOUSER T., GUIBAS L., HANRAHAN P., HUANG Q., LI Z., SAVARESE S., SAVVA M., SONG S., SU H., ET AL.: Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015). [14](#)
- [CFZ*23] CHEN T., FU C., ZANG Y., ZHU L., ZHANG J., MAO P., SUN L.: Deep3dsketch+: Rapid 3d modeling from single free-hand sketches. In *International Conference on Multimedia Modeling* (2023), Springer, pp. 16–28. [18, 19](#)
- [CGC*23] CHEN H., GU J., CHEN A., TIAN W., TU Z., LIU L., SU H.: Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. *arXiv preprint arXiv:2304.06714* (2023). [6, 7, 10, 18](#)
- [CGD*22] COLLINS J., GOEL S., DENG K., LUTHRA A., XU L., GUNDOGDU E., ZHANG X., VICENTE T. F. Y., DIDERIKSEN T., ARORA H., ET AL.: Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 21126–21136. [15](#)
- [CGT*19] CHOI I., GALLO O., TROCCHI A., KIM M. H., KAUTZ J.: Extreme view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 7781–7790. [5](#)
- [CHB*23] CHEN X., HUANG J., BIN Y., YU L., LIAO Y.: Veri3d: Generative vertex-based radiance fields for 3d controllable human image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 8986–8997. [16](#)
- [CHIS23] CROITORU F.-A., HONDRU V., IONESCU R. T., SHAH M.: Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023). [3](#)
- [CJS*22] CHEN X., JIANG T., SONG J., YANG J., BLACK M. J., GEIGER A., HILLIGES O.: gdna: Towards generative detailed neural avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 20427–20437. [7, 15, 16](#)
- [CLC*22] CHAN E. R., LIN C. Z., CHAN M. A., NAGANO K., PAN B., DE MELLO S., GALLO O., GUIBAS L. J., TREMBLAY J., KHAMIS S., ET AL.: Efficient geometry-aware 3d generative adversarial networks. In *CVPR* (2022). [1, 6, 7, 9, 17, 18, 21](#)
- [CLL23] CHENG Z., LI J., LI H.: Wildlight: In-the-wild inverse rendering with a flashlight. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 4305–4314. [6](#)
- [CLN*23] CHUNG J., LEE S., NAM H., LEE J., LEE K. M.: Lucid-dreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384* (2023). [18, 19](#)
- [CLT*23] CHENG Y.-C., LEE H.-Y., TULYAKOV S., SCHWING A. G., GUI L.-Y.: Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 4456–4465. [10, 18](#)
- [CMA*22] CHOI H., MOON G., ARMANDO M., LEROY V., LEE K. M., ROGEZ G.: Mononhr: Monocular neural human renderer. In *2022 International Conference on 3D Vision (3DV)* (2022), IEEE, pp. 242–251. [16](#)

- [CMK*21a] CHAN E. R., MONTEIRO M., KELLNHOFER P., WU J., WETZSTEIN G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 5799–5809. [9](#), [15](#)
- [CMK*21b] CHAN E. R., MONTEIRO M., KELLNHOFER P., WU J., WETZSTEIN G.: pi-GAN: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR* (2021). [17](#)
- [CNC*23] CHAN E. R., NAGANO K., CHAN M. A., BERGMAN A. W., PARK J. J., LEVY A., AIITALA M., DE MELLO S., KARRAS T., WETZSTEIN G.: Generative novel view synthesis with 3d-aware diffusion models. *arXiv preprint arXiv:2304.02602* (2023). [13](#)
- [CPA*21] CORONA E., PUMAROLA A., ALENYA G., PONS-MOLL G., MORENO-NOGUER F.: Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 11875–11885. [15](#), [16](#)
- [CPB*20] CHOUTAS V., PAVLAKOS G., BOLKART T., TZIONAS D., BLACK M. J.: Monocular expressive body regression through body-driven attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16* (2020), Springer, pp. 20–40. [16](#)
- [CRJ22] CAO A., ROCKWELL C., JOHNSON J.: Fwd: Real-time novel view synthesis with forward warping and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 15713–15724. [4](#)
- [CSH19] CHEN X., SONG J., HILLIGES O.: Monocular neural image based rendering with continuous view control. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 4090–4100. [14](#)
- [CSL*23] CHEN D. Z., SIDDIQUI Y., LEE H.-Y., TULYAKOV S., NIESSNER M.: Text2Tex: Text-driven texture synthesis via diffusion models. In *ICCV* (2023). [19](#)
- [CTZ20] CHEN Z., TAGLIASACCHI A., ZHANG H.: Bsp-net: Generating compact meshes via binary space partitioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 45–54. [6](#)
- [CUYH20] CHOI Y., UH Y., YOO J., HA J.-W.: Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 8188–8197. [14](#), [15](#)
- [CWS*15] CALLI B., WALSMAN A., SINGH A., SRINIVASA S., ABBEEL P., DOLLAR A. M.: Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols. *arXiv preprint arXiv:1502.03143* (2015). [14](#)
- [CX*23] CHAO Y.-W., XIANG Y., ET AL.: Fewsol: A dataset for few-shot object learning in robotic environments. In *2023 IEEE International Conference on Robotics and Automation (ICRA)* (2023), IEEE, pp. 9140–9146. [15](#)
- [CXG*16] CHOY C. B., XU D., GWAK J., CHEN K., SAVARESE S.: 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* (2016), Springer, pp. 628–644. [6](#)
- [CXG*22] CHEN A., XU Z., GEIGER A., YU J., SU H.: TensoRF: Tensorial radiance fields. In *European Conference on Computer Vision* (2022). [6](#)
- [CXZ*21] CHEN A., XU Z., ZHAO F., ZHANG X., XIANG F., YU J., SU H.: MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 14124–14133. [5](#), [13](#)
- [CYAE*20] CAI R., YANG G., AVERBUCH-ELOR H., HAO Z., BELONGIE S., SNAVELY N., HARIHARAN B.: Learning gradient fields for shape generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16* (2020), Springer, pp. 364–381. [10](#)
- [CYL*22] CHEN Y., YUAN Q., LI Z., LIU Y., WANG W., XIE C., WEN X., YU Q.: UPST-NeRF: Universal photorealistic style transfer of neural radiance fields for 3d scene. In *arXiv preprint arXiv:2208.07059* (2022). [19](#)
- [CYW*23] CHENG X., YANG T., WANG J., LI Y., ZHANG L., ZHANG J., YUAN L.: Progressive3D: Progressively local editing for text-to-3d content creation with complex semantic prompts. *arXiv preprint arXiv:2310.11784* (2023). [19](#)
- [CZ19] CHEN Z., ZHANG H.: Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 5939–5948. [5](#), [9](#)
- [CZC*24] CHEN H., ZHANG Y., CUN X., XIA M., WANG X., WENG C., SHAN Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. [arXiv:2401.09047](#). [2](#)
- [CZL*22] CHEN X., ZHANG Q., LI X., CHEN Y., FENG Y., WANG X., WANG J.: Hallucinated neural radiance fields in the wild. In *CVPR* (2022), pp. 12943–12952. [5](#), [19](#)
- [CZY*23] CHEN Y., ZHANG C., YANG X., CAI Z., YU G., YANG L., LIN G.: It3d: Improved text-to-3d generation with explicit view synthesis. *arXiv preprint arXiv:2308.11473* (2023). [18](#)
- [DBD*22] DARMON F., BASCLE B., DEVaux J.-C., MONASSE P., AUBRY M.: Improving neural implicit surfaces geometry with patch warping. In *CVPR* (2022). [6](#)
- [DCS*17] DAI A., CHANG A. X., SAVVA M., HALBER M., FUNKHOUSER T., NIESSNER M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 5828–5839. [14](#), [15](#)
- [DFK*22] DOWNS L., FRANCIS A., KOENIG N., KINMAN B., HICKMAN R., REYMANN K., MCHUGH T. B., VANHOUCKE V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)* (2022), IEEE, pp. 2553–2560. [14](#), [15](#)
- [DJQ*23] DENG C., JIANG C., QI C. R., YAN X., ZHOU Y., GUIBAS L., ANGUELOV D., 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* (2023), pp. 20637–20647. [12](#)
- [DLW*23] DEITKE M., LIU R., WALLINGFORD M., NGO H., MICHEL O., KUSUPATI A., FAN A., LAFORTE C., VOLETI V., GADRE S. Y., ET AL.: Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663* (2023). [14](#), [15](#)
- [Doe16] DOERSCH C.: Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908* (2016). [3](#)
- [DRB*18] DAI A., RITCHIE D., BOKELOH M., REED S., STURM J., NIESSNER M.: Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 4578–4587. [6](#)
- [DSS*23] DEITKE M., SCHWENK D., SALVADOR J., WEIHS L., MICHEL O., VANDERBILT E., SCHMIDT L., EHSANI K., KEMBAVI A., FARHADI A.: Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 13142–13153. [14](#), [15](#)
- [DZL*20] DAI P., ZHANG Y., LI Z., LIU S., ZENG B.: Neural point cloud rendering via multi-plane projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 7830–7839. [4](#)
- [DZW*20] DUAN Y., ZHU H., WANG H., YI L., NEVATIA R., GUIBAS L. J.: Curriculum deepsdf. In *European Conference on Computer Vision* (2020), Springer, pp. 51–67. [5](#)
- [DZY*21] DU Y., ZHANG Y., YU H.-X., TENENBAUM J. B., WU J.: Neural radiance flow for 4d view synthesis and video processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), IEEE Computer Society, pp. 14304–14314. [5](#)

- [EGO*20] ERLER P., GUERRERO P., OHRHALLINGER S., MITRA N. J., WIMMER M.: Points2surf learning implicit surfaces from point clouds. In *European Conference on Computer Vision* (2020), Springer, pp. 108–124. 5
- [EMS*23] ERKOÇ Z., MA F., SHAN Q., NIESSNER M., DAI A.: Hyper-diffusion: Generating implicit neural fields with weight-space diffusion. *arXiv preprint arXiv:2303.17015* (2023). 18
- [EST*20] EGGER B., SMITH W. A., TEWARI A., WUHRER S., ZOLLHOEFER M., BEELER T., BERNARD F., BOLKART T., KORTYLEWSKI A., ROMDHANI S., ET AL.: 3d morphable face models—past, present, and future. *ACM Trans. Graph.* (2020). 17
- [FAKD23] FRIDMAN R., ABECASIS A., KASTEN Y., DEKEL T.: Scenescape: Text-driven consistent scene generation. *arXiv preprint arXiv:2302.01133* (2023). 18
- [FBD*19] FLYNN J., BROXTON M., DEBEVEC P., DUVALL M., FYFFE G., OVERBECK R., SNAVELY N., TUCKER R.: Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 2367–2376. 5
- [FJG*21] FU H., JIA R., GAO L., GONG M., ZHAO B., MAYBANK S., TAO D.: 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision* 129 (2021), 3313–3337. 14, 15
- [FJW*22] FAN Z., JIANG Y., WANG P., GONG X., XU D., WANG Z.: Unified implicit neural stylization. In *ECCV* (2022), Springer. 19
- [FKYT*22] FRIDOVICH-KEIL S., YU A., TANCIK M., CHEN Q., RECHT B., KANAZAWA A.: Plenoxtels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 5501–5510. 6
- [FLJ*22] FU J., LI S., JIANG Y., LIN K.-Y., QIAN C., LOY C. C., WU W., LIU Z.: Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision* (2022), Springer, pp. 1–19. 14, 15
- [FXOT22] FU Q., XU Q., ONG Y. S., TAO W.: Geo-NeuS: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems* 35 (2022), 3403–3416. 6
- [GAA*22] GAL R., ALALUF Y., ATZMON Y., PATASHNIK O., BERMANO A. H., CHECHIK G., COHEN-OR D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022). 12
- [GCL*21] GUO Y., CHEN K., LIANG S., LIU Y.-J., BAO H., ZHANG J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 5784–5794. 5
- [GCS*20] GENOVA K., COLE F., SUD A., SARNA A., FUNKHOUSER T.: Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 4857–4866. 6
- [GCV*19] GENOVA K., COLE F., VLASIC D., SARNA A., FREEMAN W. T., FUNKHOUSER T.: Learning shape templates with structured implicit functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 7154–7164. 6
- [GEB16] GATYS L. A., ECKER A. S., BETHGE M.: Image style transfer using convolutional neural networks. In *CVPR* (2016). 19
- [GII*21] GRIGOREV A., ISKAKOV K., IANINA A., BASHIROV R., ZAKHARKIN I., VAKHITOV A., LEMPITSKY V.: Stylepeople: A generative model of fullbody human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 5151–5160. 16
- [GKG*23] GIEBENHAIN S., KIRSCHSTEIN T., GEORGOPoulos M., RÜNZ M., AGAPITO L., NIESSNER M.: Learning neural parametric head models. In *CVPR* (2023). 17
- [GKJ*21] GARBIN S. J., KOWALSKI M., JOHNSON M., SHOTTON J., VALENTIN J.: Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 14346–14355. 5
- [GLWT22] GU J., LIU L., WANG P., THEOBALT C.: StyleNeRF: A style-based 3d-aware generator for high-resolution image synthesis. In *Int. Conf. Learn. Represent.* (2022). 17
- [GLZ*23] GAO X., LI X., ZHANG C., ZHANG Q., CAO Y., SHAN Y., QUAN L.: Context-human: Free-view rendering of human from a single image with texture-consistent synthesis. *arXiv preprint arXiv:2311.17123* (2023). 16
- [GMW17] GADELHA M., MAJI S., WANG R.: 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)* (2017), IEEE, pp. 402–411. 9, 15
- [GNL*23] GE S., NAH S., LIU G., POON T., TAO A., CATANZARO B., JACOBS D., HUANG J.-B., LIU M.-Y., BALAJI Y.: Preserve your own correlation: A noise prior for video diffusion models. In *ICCV* (2023). 2
- [GPAM*14] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. *Advances in neural information processing systems* 27 (2014). 2, 6, 7
- [GPL*22] GRASSAL P.-W., PRINZLER M., LEISTNER T., ROTHER C., NIESSNER M., THIES J.: Neural head avatars from monocular RGB videos. In *CVPR* (2022). 17
- [GSW*21] GUI J., SUN Z., WEN Y., TAO D., YE J.: A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE transactions on knowledge and data engineering* 35, 4 (2021), 3313–3332. 3
- [GSW*22] GAO J., SHEN T., WANG Z., CHEN W., YIN K., LI D., LITANY O., GOJCIC Z., FIDLER S.: GET3D: A generative model of high quality 3d textured shapes learned from images. In *NeurIPS* (2022). 9, 18, 19
- [GTZN21] GAFNI G., THIES J., ZOLLHOFER M., NIESSNER M.: Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. In *CVPR* (2021). 17
- [GWH*20] GUO Y., WANG H., HU Q., LIU H., LIU L., BENNAMOUN M.: Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence* 43, 12 (2020), 4338–4364. 3
- [GWY*21] GAO L., WU T., YUAN Y.-J., LIN M.-X., LAI Y.-K., ZHANG H.: Tm-net: Deep generative networks for textured meshes. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–15. 11
- [GXN*23] GUPTA A., XIONG W., NIE Y., JONES I., OĞUZ B.: 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371* (2023). 7
- [GYW*19a] GAO L., YANG J., WU T., YUAN Y.-J., FU H., LAI Y.-K., ZHANG H.: Sdm-net: Deep generative network for structured deformable mesh. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–15. 7
- [GYW*19b] GAO L., YANG J., WU T., YUAN Y.-J., FU H., LAI Y.-K., ZHANG H.: Sdm-net: Deep generative network for structured deformable mesh. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–15. 11
- [GZX*22] GAO X., ZHONG C., XIANG J., HONG Y., GUO Y., ZHANG J.: Reconstructing personalized semantic facial nerf models from monocular video. *ACM Trans. Graph.* (2022). 17
- [HB17] HUANG X., BELONGIE S.: Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV* (2017). 19
- [HCL*22] HONG F., CHEN Z., LAN Y., PAN L., LIU Z.: Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888* (2022). 15, 16
- [HCO*23] HÖLLEIN L., CAO A., OWENS A., JOHNSON J., NIESSNER M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. *arXiv preprint arXiv:2303.11989* (2023). 18

- [HHP*23] HU S., HONG F., PAN L., MEI H., YANG L., LIU Z.: Sherf: Generalizable human nerf from a single image. *arXiv preprint arXiv:2303.12791* (2023). 16
- [HYH*22] HUANG Y.-H., HE Y., YUAN Y.-J., LAI Y.-K., GAO L.: Styliednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *CVPR* (2022). 19
- [HJA20] HO J., JAIN A., ABBEEL P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851. 1, 2, 6, 9
- [HLA*19] HU Y., LI T.-M., ANDERSON L., RAGAN-KELLEY J., DURAND F.: Taichi: a language for high-performance computation on spatially sparse data structures. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–16. 5
- [HLHF22] HUI K.-H., LI R., HU J., FU C.-W.: Neural wavelet-domain diffusion for 3d shape generation. In *SIGGRAPH Asia 2022 Conference Papers* (2022), pp. 1–9. 10, 18
- [HMR19a] HENZLER P., MITRA N. J., RITSCHEL T.: Escaping plato’s cave: 3D shape from adversarial rendering. In *ICCV* (2019). 9, 15, 18
- [HMR19b] HENZLER P., MITRA N. J., RITSCHEL T.: Escaping plato’s cave: 3d shape from adversarial rendering. In *ICCV* (2019). 17, 18
- [HPX*22] HONG Y., PENG B., XIAO H., LIU L., ZHANG J.: Head-NeRF: A real-time nerf-based parametric head model. In *CVPR* (2022). 5
- [HRBP21] HU R., RAVI N., BERG A. C., PATHAK D.: Worldsheets: Wrapping the world in a 3d sheet for view synthesis from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 12528–12537. 18
- [HRL*21] HENZLER P., REIZENSTEIN J., LABATUT P., SHAPOVALOV R., RITSCHEL T., VEDALDI A., NOVOTNY D.: Unsupervised learning of 3d object categories from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 4700–4709. 15
- [HSG*22] HO J., SALIMANS T., GRITSENKO A., CHAN W., NOROUZI M., FLEET D. J.: Video diffusion models. In *NeurIPS* (2022). 2
- [HSZ*23] HUANG X., SHAO R., ZHANG Q., ZHANG H., FENG Y., LIU Y., WANG Q.: Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. *arXiv preprint arXiv:2310.01406* (2023). 16
- [HTE*23] HAQUE A., TANCIK M., EFROS A. A., HOLYNSKI A., KANAZAWA A.: Instruct-NeRF2NeRF: Editing 3d scenes with instructions. In *ICCV* (2023). 19, 20
- [HTS*21] HUANG H.-P., TSENG H.-Y., SAINI S., SINGH M., YANG M.-H.: Learning to stylize novel views. In *ICCV* (2021). 19
- [HWZ*23] HUANG Y., WANG J., ZENG A., CAO H., QI X., SHI Y., ZHA Z.-J., ZHANG L.: Dreamwaltz: Make a scene with complex 3d animatable avatars. *arXiv preprint arXiv:2305.12529* (2023). 16
- [HYL*23] HUANG Y., YI H., LIU W., WANG H., WU B., WANG W., LIN B., ZHANG D., CAI D.: One-shot implicit animatable avatars with model-based priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 8974–8985. 15, 16
- [HYX*23] HUANG Y., YI H., XIU Y., LIAO T., TANG J., CAI D., THIES J.: Tech: Text-guided reconstruction of lifelike clothed humans. *arXiv preprint arXiv:2308.08545* (2023). 15, 16
- [HZF*22] HUANG X., ZHANG Q., FENG Y., LI H., WANG X., WANG Q.: Hdr-nerf: High dynamic range neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18398–18408. 5
- [HZF*23a] HUANG X., ZHANG Q., FENG Y., LI H., WANG Q.: Inverting the imaging process by learning an implicit camera model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 21456–21465. 5
- [HZF*23b] HUANG X., ZHANG Q., FENG Y., LI X., WANG X., WANG Q.: Local implicit ray function for generalizable radiance field representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 97–107. 5
- [HZP*22] HONG F., ZHANG M., PAN L., CAI Z., YANG L., LIU Z.: Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535* (2022). 16
- [ID18] INSAFUTDINOV E., DOSOVITSKIY A.: Unsupervised learning of shape and pose with differentiable point clouds. *Advances in neural information processing systems* 31 (2018). 4
- [JCL*22] JIANG K., CHEN S.-Y., LIU F.-L., FU H., GAO L.: NeRF-FaceEditing: Disentangled face editing in neural radiance fields. In *SIGGRAPH Asia Conference Papers* (2022). 18, 19
- [JJW*23] JIANG S., JIANG H., WANG Z., LUO H., CHEN W., XU L.: Humangen: Generating human radiance fields with explicit priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 12543–12554. 16
- [JKK*23] JAMBON C., KERBL B., KOPANAS G., DIOLATZIS S., DRETTAKIS G., LEIMKÜHLER T.: Nerfshop: Interactive editing of neural radiance fields. In *Proceedings of the ACM on Computer Graphics and Interactive Techniques* (2023). 19, 20
- [JLF22] JOHARI M. M., LEPOITTEVIN Y., FLEURET F.: GeoNeRF: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18365–18375. 5
- [JMB*22] JAIN A., MILDENHALL B., BARRON J. T., ABBEEL P., POOLE B.: Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 867–876. 18
- [JN23] JUN H., NICHOL A.: Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463* (2023). 10, 12
- [JWZ*23] JIANG R., WANG C., ZHANG J., CHAI M., HE M., CHEN D., LIAO J.: Avatacraft: Transforming text into neural human avatars with parameterized shape and pose control. *arXiv preprint arXiv:2303.17606* (2023). 15, 16
- [KAZ*23] KOLOTOUROS N., ALLDIECK T., ZANFIR A., BAZAVAN E. G., FIERARU M., SMINCHISESCU C.: Dreamhuman: Animatable 3d avatars from text. *arXiv preprint arXiv:2306.09329* (2023). 15, 16
- [KBM*20] KATO H., BEKER D., MORARIU M., ANDO T., MATSUOKA T., KEHL W., GAIDON A.: Differentiable rendering: A survey. *arXiv preprint arXiv:2006.12057* (2020). 3
- [KBV20] KLOKOV R., BOYER E., VERBEEK J.: Discrete point flow networks for efficient point cloud generation. In *European Conference on Computer Vision* (2020), Springer, pp. 694–710. 11
- [KDJ*23] KWAK J.-G., DONG E., JIN Y., KO H., MAHAJAN S., YI K. M.: Vivid-1-to-3: Novel view synthesis with video diffusion models. *arXiv preprint arXiv:2312.01305* (2023).
- [KDSB22] KULHÁNEK J., DERNER E., SATTLER T., BABUŠKA R.: Viewformer: Nerf-free neural rendering from few images using transformers. In *European Conference on Computer Vision* (2022), Springer, pp. 198–216. 13, 14
- [KFH*22] KERR J., FU L., HUANG H., AVIGAL Y., TANCIK M., ICHNOWSKI J., KANAZAWA A., GOLDBERG K.: Evo-nerf: Evolving nerf for sequential robot grasping of transparent objects. In *6th Annual Conference on Robot Learning* (2022). 5
- [KKL*23] KIM B., KWON P., LEE K., LEE M., HAN S., KIM D., JOO H.: Chupa: Carving 3d clothed humans from skinned shape priors using 2d diffusion probabilistic models. *arXiv preprint arXiv:2305.11870* (2023). 15, 16
- [KKLD23] KERBL B., KOPANAS G., LEIMKÜHLER T., DRETTAKIS G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)* 42, 4 (2023), 1–14. 1, 4, 12

- [KKR18] KNYAZ V. A., KNAZ V. V., REMONDINO F.: Image-to-voxel model translation with conditional adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (2018), pp. 0–0. 7
- [KLA19] KARRAS T., LAINE S., AILA T.: A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 4401–4410. 14, 15, 17
- [KLA*20] KARRAS T., LAINE S., AITTALA M., HELLSTEN J., LEHTINEN J., AILA T.: Analyzing and improving the image quality of StyleGAN. In *CVPR* (2020). 16, 17
- [KLK*20] KIM H., LEE H., KANG W. H., LEE J. Y., KIM N. S.: Soft-flow: Probabilistic framework for normalizing flow on manifolds. *Advances in Neural Information Processing Systems* 33 (2020), 16388–16397. 11
- [KLY*21] KOH J. Y., LEE H., YANG Y., BALDRIDGE J., ANDERSON P.: Pathdreamer: A world model for indoor navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 14738–14748. 14
- [KMS22] KOBAYASHI S., MATSUMOTO E., SITZMANN V.: Decomposing nerf for editing via feature field distillation. In *NeurIPS* (2022). 19, 20
- [KNH*22] KHAN S., NASEER M., HAYAT M., ZAMIR S. W., KHAN F. S., SHAH M.: Transformers in vision: A survey. *ACM computing surveys (CSUR)* 54, 10s (2022), 1–41. 3
- [KPHL17] KUSNER M. J., PAIGE B., HERNÁNDEZ-LOBATO J. M.: Grammar variational autoencoder. In *International conference on machine learning* (2017), PMLR, pp. 1945–1954. 2, 6
- [KPLD21] KOPANAS G., PHILIP J., LEIMKÜHLER T., DRETTAKIS G.: Point-based neural rendering with per-view optimization. In *Computer Graphics Forum* (2021), vol. 40, Wiley Online Library, pp. 29–43. 4
- [KPWS22] KALISCHEK N., PETERS T., WEGNER J. D., SCHINDLER K.: Tetrahedral diffusion models for 3d shape generation. *arXiv preprint arXiv:2211.13220* (2022). 10
- [KQG*23] KIRSCHSTEIN T., QIAN S., GIEBENHAIN S., WALTER T., NEISSNER M.: NeRSemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.* (2023). 17
- [KSZ*21] KOSIOREK A. R., STRATHMANN H., ZORAN D., MORENO P., SCHNEIDER R., MOKRÁ S., REZENDE D. J.: Nerf-vae: A geometry aware 3d scene generative model. In *ICML* (2021). 11
- [KUH18] KATO H., USHIKU Y., HARADA T.: Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 3907–3916. 5
- [KVNM23] KARNEWAR A., VEDALDI A., NOVOTNY D., MITRA N. J.: Holodiffusion: Training a 3d diffusion model using 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 18423–18433. 7
- [KW13] KINGMA D. P., WELLING M.: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013). 2, 6, 11
- [KWKT15] KULKARNI T. D., WHITNEY W. F., KOHLI P., TENENBAUM J.: Deep convolutional inverse graphics network. *Advances in neural information processing systems* 28 (2015). 14
- [KXD12] KASPER A., XUE Z., DILLMANN R.: The kit object models database: An object model database for object recognition, localization and manipulation in service robotics. *The International Journal of Robotics Research* 31, 8 (2012), 927–934. 14
- [KYLH21] KIM J., YOO J., LEE J., HONG S.: Setvae: Learning hierarchical composition for generative modeling of set-structured data. In *CVPR* (2021). 11
- [LB14] LOPER M. M., BLACK M. J.: OpenDr: An approximate differentiable renderer. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VII 13* (2014), Springer, pp. 154–169. 5
- [LBRF11] LAI K., BO L., REN X., FOX D.: A large-scale hierarchical multi-view rgb-d object dataset. In *2011 IEEE international conference on robotics and automation* (2011), IEEE, pp. 1817–1824. 15
- [LC22] LEE H.-H., CHANG A. X.: Understanding pure clip guidance for voxel grid nerf models. *arXiv preprint arXiv:2209.15172* (2022). 18
- [LCCT23] LI W., CHEN R., CHEN X., TAN P.: Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arXiv preprint arXiv:2310.02596* (2023). 12
- [LDS*23] LI Y., DOU Y., SHI Y., LEI Y., CHEN X., ZHANG Y., ZHOU P., NI B.: FocalDreamer: Text-driven 3d editing via focal-fusion assembly. *arXiv preprint arXiv:2308.10608* (2023). 19, 20
- [LDZL23] LI M., DUAN Y., ZHOU J., LU J.: Diffusion-sdf: Text-to-shape via voxelized diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 12642–12651. 10, 18, 19
- [LFB*23] LIU Z., FENG Y., BLACK M. J., NOWROUZEZHARAI D., PAULL L., LIU W.: Meshdiffusion: Score-based generative 3d mesh modeling. *arXiv preprint arXiv:2303.08133* (2023). 10
- [LFLSY*23] LIN G., FENG-LIN L., SHU-YU C., KAIWEN J., CHUN-PENG L., LAI Y., HONGBO F.: SketchFaceNeRF: Sketch-based facial generation and editing in neural radiance fields. *ACM Trans. Graph.* (2023). 18, 19
- [LFS*21] LI J., FENG Z., SHE Q., DING H., WANG C., LEE G. H.: Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 12578–12588. 5
- [LGL*23] LONG X., GUO Y.-C., LIN C., LIU Y., DOU Z., LIU L., MA Y., ZHANG S.-H., HABERMANN M., THEOBALT C., ET AL.: Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008* (2023). 13
- [LGT*23] LIN C.-H., GAO J., TANG L., TAKIKAWA T., ZENG X., HUANG X., KREIS K., FIDLER S., LIU M.-Y., LIN T.-Y.: Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 300–309. 12, 18
- [LGZL*20] LIU L., GU J., ZAW LIN K., CHUA T.-S., THEOBALT C.: Neural sparse voxel fields. *Advances in Neural Information Processing Systems* 33 (2020), 15651–15663. 6
- [LH21] LUO S., HU W.: Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 2837–2845. 10, 18
- [LHG*23] LIN Y., HAN H., GONG C., XU Z., ZHANG Y., LI X.: Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. *arXiv preprint arXiv:2309.17261* (2023). 13
- [LHR*21] LIU L., HABERMANN M., RUDNEV V., SARKAR K., GU J., THEOBALT C.: Neural actor: Neural free-view synthesis of human actors with pose control. *ACM transactions on graphics (TOG)* 40, 6 (2021), 1–16. 5
- [Lin68] LINDENMAYER A.: Mathematical models for cellular interactions in development i. filaments with one-sided inputs. *Journal of theoretical biology* 18, 3 (1968), 280–299. 13
- [LKL18] LIN C.-H., KONG C., LUCEY S.: Learning efficient point cloud generation for dense 3d object reconstruction. In *proceedings of the AAAI Conference on Artificial Intelligence* (2018), vol. 32. 4
- [LLC19] LIU S., LI T., CHEN W., LI H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 7708–7717. 5
- [LLF*23] LI Y., LIN Z.-H., FORSYTH D., HUANG J.-B., WANG S.: ClimateNeRF: Physically-based neural rendering for extreme climate synthesis. In *ICCV* (2023). 19
- [LLL*14] LI B., LU Y., LI C., GODIL A., SCHRECK T., AONO M., CHEN Q., CHOWDHURY N. K., FANG B., FURUYA T., ET AL.:

- Shrec'14 track: Large scale comprehensive 3d shape retrieval. In *Eurographics Workshop on 3D Object Retrieval* (2014), vol. 2, . 14
- [LLQ*16] LIU Z., LUO P., QIU S., WANG X., TANG X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 1096–1104. 14, 15
- [LLWT15] LIU Z., LUO P., WANG X., TANG X.: Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 3730–3738. 15
- [LLZ*23] LIU Y., LIN C., ZENG Z., LONG X., LIU L., KOMURA T., WANG W.: Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453* (2023). 12, 13
- [LLZL21] LUO A., LI T., ZHANG W.-H., LEE T. S.: Surfgen: Adversarial 3d shape synthesis with explicit surface discriminators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 16238–16248. 7, 9
- [LMTL21] LIN C.-H., MA W.-C., TORRALBA A., LUCEY S.: Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 5741–5751. 5
- [LMY*23] LI Y., MA C., YAN Y., ZHU W., YANG X.: 3d-aware face swapping. In *CVPR* (2023). 19
- [LPT13] LIM J. J., PIRSIAVASH H., TORRALBA A.: Parsing ikea objects: Fine pose estimation. In *Proceedings of the IEEE international conference on computer vision* (2013), pp. 2992–2999. 14
- [LSC*22] LEVIS A., SRINIVASAN P. P., CHAEL A. A., NG R., BOUMAN K. L.: Gravitationally lensed black hole emission tomography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 19841–19850. 5
- [LSMG20] LIAO Y., SCHWARZ K., MESCHDER L., GEIGER A.: Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 5871–5880. 15
- [LSS*21] LOMBARDI S., SIMON T., SCHWARTZ G., ZOLLHOEFER M., SHEIKH Y., SARAGIH J.: Mixture of volumetric primitives for efficient neural rendering. *ACM Trans. Graph.* (2021). 17
- [LSSS18] LOMBARDI S., SARAGIH J., SIMON T., SHEIKH Y.: Deep appearance models for face rendering. *ACM Trans. Graph.* (2018). 17
- [LSZ*22] LI T., SLAVCHEVA M., ZOLLHOEFER M., GREEN S., LASSNER C., KIM C., SCHMIDT T., LOVEGROVE S., GOESELE M., NEWCOMBE R., ET AL.: Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 5521–5531. 5
- [LTJ18] LIU H.-T. D., TAO M., JACOBSON A.: Paparazzi: surface editing by way of multi-view image processing. *ACM Trans. Graph.* 37, 6 (2018), 221–1. 5
- [LTJ*21] LIU A., TUCKER R., JAMPANI V., MAKADIA A., SNAVELY N., KANAZAWA A.: Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 14458–14467. 14
- [LTZ*23] LI J., TAN H., ZHANG K., XU Z., LUAN F., XU Y., HONG Y., SUNKAVALLI K., SHAKHNAROVICH G., BI S.: Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214* (2023). 1, 2
- [LWA*23] LYU Z., WANG J., AN Y., ZHANG Y., LIN D., DAI B.: Controllable mesh generation through sparse latent point diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 271–280. 10
- [LWC*23] LI Z., WANG Q., COLE F., TUCKER R., SNAVELY N.: Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 4273–4284. 5
- [LWQF22] LIU Z., WANG Y., QI X., FU C.-W.: Towards implicit text-guided 3d shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 17896–17906. 18
- [LWSK22] LI Z., WANG Q., SNAVELY N., KANAZAWA A.: Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *European Conference on Computer Vision* (2022), Springer, pp. 515–534. 14
- [LWVH*23] LIU R., WU R., VAN HOORICK B., TOKMAKOV P., ZAKHAROV S., VONDRIK C.: Zero-1-to-3: Zero-shot one image to 3d object. *arXiv preprint arXiv:2303.11328* (2023). 1, 7, 12, 13, 18
- [LXC*21] LI J., XU C., CHEN Z., BIAN S., YANG L., LU C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 3383–3393. 16
- [LXJ*23] LIU M., XU C., JIN H., CHEN L., XU Z., SU H., 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). 10, 12
- [LXM*20] LIN K.-E., XU Z., MILDENHALL B., SRINIVASAN P. P., HOLD-GEOFFROY Y., DiVERDI S., SUN Q., SUNKAVALLI K., RAMAMOORTHI R.: Deep multi depth panoramas for view synthesis. In *European Conference on Computer Vision* (2020), Springer, pp. 328–344. 5
- [LXZ*23] LORRAINE J., XIE K., ZENG X., LIN C.-H., TAKIKAWA T., SHARP N., LIN T.-Y., LIU M.-Y., FIDLER S., LUCAS J.: Att3d: Amortized text-to-3d object synthesis. *arXiv preprint arXiv:2306.07349* (2023). 15, 18
- [LYX*24] LIAO T., YI H., XIU Y., TANG J., HUANG Y., THIES J., BLACK M. J.: TADA! text to animatable digital avatars. In *3DV* (2024). 15, 16, 19
- [LZ21] LASSNER C., ZOLLHOFER M.: Pulsar: Efficient sphere-based neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1440–1449. 4
- [LZF*23] LIANG Z., ZHANG Q., FENG Y., SHAN Y., JIA K.: Gsir: 3d gaussian splatting for inverse rendering. *arXiv preprint arXiv:2311.16473* (2023). 5
- [LZJ*22] LEI J., ZHANG Y., JIA K., ET AL.: TANGO: Text-driven photorealistic and robust 3d stylization via lighting decomposition. In *NeurIPS* (2022). 11, 19
- [LZT*23] LIU X., ZHAN X., TANG J., SHAN Y., ZENG G., LIN D., LIU X., LIU Z.: Humangaussian: Text-driven 3d human generation with gaussian splatting. *arXiv preprint arXiv:2311.17061* (2023). 16
- [LZW*23] LI C., ZHANG C., WAGHWASE A., LEE L.-H., RAMEAU F., YANG Y., BAE S.-H., HONG C. S.: Generative ai meets 3d: A survey on text-to-3d in aigc era. *arXiv preprint arXiv:2305.06131* (2023). 3
- [Man67] MANDELBROT B.: How long is the coast of britain? statistical self-similarity and fractional dimension. *science* 156, 3775 (1967), 636–638. 13
- [Max95] MAX N.: Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics* 1, 2 (1995), 99–108. 5
- [MBOL*22] MICHEL O., BAR-ON R., LIU R., BENAÏM S., HANOCKA R.: Text2mesh: Text-driven neural stylization for meshes. In *CVPR* (2022). 19, 20
- [MBRS*21] MARTIN-BRUALLA R., RADWAN N., SAJJADI M. S., BARRON J. T., DOSOVITSKIY A., DUCKWORTH D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR* (2021), pp. 7210–7219. 5
- [MCL20] MORRISON D., CORKE P., LEITNER J.: Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation. *IEEE Robotics and Automation Letters* 5, 3 (2020), 4368–4375. 14

- [MCST22a] MITTAL P., CHENG Y.-C., SINGH M., TULSIANI S.: Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 306–315. 6, 7
- [MCST22b] MITTAL P., CHENG Y.-C., SINGH M., TULSIANI S.: Autosdf: Shape priors for 3d completion, reconstruction and generation. In *CVPR* (2022). 10
- [MESK22] MÜLLER T., EVANS A., SCHIED C., KELLER A.: Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)* 41, 4 (2022), 1–15. 1, 6
- [Mid] MIDJOURNEY: Midjourney. <https://www.midjourney.com/>. 2
- [MKLVRV23] MELAS-KYRIAZI L., LAINA I., RUPPRECHT C., VEDALDI A.: Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 8446–8455. 12, 18
- [MKXBP22] MOHAMMAD KHALID N., XIE T., BELIOVSKY E., POPA T.: Clip-mesh: Generating textured meshes from text using pre-trained image-text models. In *SIGGRAPH Asia 2022 conference papers* (2022), pp. 1–8. 18
- [MLL*22a] MA L., LI X., LIAO J., WANG X., ZHANG Q., WANG J., SANDER P. V.: Neural parameterization for dynamic human head editing. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–15. 19
- [MLL*22b] MA L., LI X., LIAO J., ZHANG Q., WANG X., WANG J., SANDER P. V.: Deblur-nerf: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 12861–12870. 5
- [MM82] MANDELBROT B. B., MANDELBROT B. B.: *The fractal geometry of nature*, vol. 1. WH freeman New York, 1982. 13
- [MPS*23] MIKAEILI A., PEREL O., SAFAEE M., COHEN-OR D., MAHDAMI-AMIRI A.: SKED: Sketch-guided text-based 3d editing. In *ICCV* (2023). 19, 20
- [MRP*23a] METZER G., RICHARDSON E., PATASHNIK O., Giryes R., COHEN-OR D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 12663–12673. 12
- [MRP*23b] METZER G., RICHARDSON E., PATASHNIK O., Giryes R., COHEN-OR D.: Latent-NeRF for shape-guided generation of 3d shapes and textures. In *CVPR* (2023). 18, 19
- [MS15] MATURANA D., SCHERER S.: Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (2015), IEEE, pp. 922–928. 6
- [MSOC*19] MILDENHALL B., SRINIVASAN P. P., ORTIZ-CAYON R., KALANTARI N. K., RAMAMOORTHI R., NG R., KAR A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–14. 5
- [MSP*23] MÜLLER N., SIDDIQUI Y., PORZI L., BULO S. R., KONTSCHIEDER P., NIESSNER M.: Diffrr: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 4328–4338. 10
- [MSS*21] MA S., SIMON T., SARAGIH J., WANG D., LI Y., DE LA TORRE F., SHEIKH Y.: Pixel codec avatars. In *CVPR* (2021). 17
- [MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision* (2020), Springer, pp. 405–421. 1, 5, 11
- [MYR*20] MA Q., YANG J., RANJAN A., PUJADES S., PONS-MOLL G., TANG S., BLACK M. J.: Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 6469–6478. 15, 16
- [MZS*23] MA Y., ZHANG X., SUN X., JI J., WANG H., JIANG G., ZHUANG W., JI R.: X-Mesh: Towards fast and accurate text-driven 3d stylization via dynamic textual guidance. In *ICCV* (2023). 19
- [ND21] NICHOL A. Q., DHARIWAL P.: Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning* (2021), PMLR, pp. 8162–8171. 2, 6
- [NDR*21] NICHOL A., DHARIWAL P., RAMESH A., SHYAM P., MISHKIN P., MCGREW B., SUTSKEVER I., CHEN M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021). 10
- [NDVZJ19] NIMIER-DAVID M., VICINI D., ZELTNER T., JAKOB W.: Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–17. 5
- [Neu66] NEUMANN J. v.: Theory of self-reproducing automata. *Edited by Arthur W. Burks* (1966). 13
- [NG21] NIEMEYER M., GEIGER A.: GIRAFFE: Representing scenes as compositional generative neural feature fields. In *CVPR* (2021). 5, 7, 9, 17
- [NGEB20a] NASH C., GANIN Y., ESLAMI S. A., BATTAGLIA P.: Polypgen: An autoregressive generative model of 3d meshes. In *International conference on machine learning* (2020), PMLR, pp. 7220–7229. 7
- [NGEB20b] NASH C., GANIN Y., ESLAMI S. A., BATTAGLIA P.: Polypgen: An autoregressive generative model of 3d meshes. In *ICML* (2020). 10
- [NJD*22] NICHOL A., JUN H., DHARIWAL P., MISHKIN P., CHEN M.: Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751* (2022). 1, 7, 10, 12, 18
- [NKR*22] NAM G., KHLIFI M., RODRIGUEZ A., TONO A., ZHOU L., GUERRERO P.: 3d-ldm: Neural implicit 3d shape generation with latent diffusion models. *arXiv preprint arXiv:2212.00842* (2022). 10, 18
- [NPPLT*19] NGUYEN-PHUOC T., LI C., THEIS L., RICHARDT C., YANG Y.-L.: HoloGAN: Unsupervised learning of 3d representations from natural images. In *ICCV Workshop* (2019). 9, 15, 17, 18
- [NPRM*20] NGUYEN-PHUOC T. H., RICHARDT C., MAI L., YANG Y., MITRA N.: Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in neural information processing systems* 33 (2020), 6767–6778. 7, 9
- [NSLH22] NOGUCHI A., SUN X., LIN S., HARADA T.: Unsupervised learning of efficient geometry-aware neural articulated representations. In *European Conference on Computer Vision* (2022), Springer, pp. 597–614. 16
- [OBB20] OSMAN A. A., BOLKART T., BLACK M. J.: Star: Sparse trained articulated human body regressor. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16* (2020), Springer, pp. 598–613. 16
- [OELS*22] OR-EL R., LUO X., SHAN M., SHECHTMAN E., PARK J. J., KEMELMACHER-SHLIZERMAN I.: StyleSDF: High-resolution 3d-consistent image and geometry generation. In *CVPR* (2022). 17
- [OMN*19] OECHSLE M., MESCHEDER L., NIEMEYER M., STRAUSS T., GEIGER A.: Texture fields: Learning texture representations in function space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 4531–4540. 5
- [Ope] OPENAI: Dall-e 3. <https://openai.com/dall-e-3>
- [PCPMN21] PUMAROLA A., CORONA E., PONS-MOLL G., MORENO-NOGUER F.: D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 10318–10327. 5
- [PDW*21] PENG S., DONG J., WANG Q., ZHANG S., SHUAI Q., ZHOU X., BAO H.: Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 14314–14323. 5
- [Per85] PERLIN K.: An image synthesizer. *ACM Siggraph Computer Graphics* 19, 3 (1985), 287–296. 13
- [Per02] PERLIN K.: Improving noise. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques* (2002), pp. 681–682. 13

- [PFS*19] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 165–174. 1, 5
- [PGH*16] PU Y., GAN Z., HENAO R., YUAN X., LI C., STEVENS A., CARIN L.: Variational autoencoder for deep learning of images, labels and captions. *Advances in neural information processing systems* 29 (2016). 2, 6
- [PJBM23] POOLE B., JAIN A., BARRON J. T., MILDENHALL B.: DreamFusion: Text-to-3d using 2d diffusion. In *Int. Conf. Learn. Represent.* (2023). 1, 7, 11, 12, 15, 18, 20, 21
- [PKHL21] PAVLLO D., KOHLER J., HOFMANN T., LUCCHI A.: Learning generative models of textured 3d meshes from real-world images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 13879–13889. 9
- [PRFS18] PARK K., REMATAS K., FARHADI A., SEITZ S. M.: Photoshape: Photorealistic materials for large-scale shape collections. *arXiv preprint arXiv:1809.09761* (2018). 14, 15
- [PSB*21] PARK K., SINHA U., BARRON J. T., BOUAZIZ S., GOLDMAN D. B., SEITZ S. M., MARTIN-BRUALLA R.: Nerfies: Deformable neural radiance fields. In *ICCV* (2021). 17
- [PSH*20] PAVLLO D., SPINKS G., HOFMANN T., MOENS M.-F., LUCCHI A.: Convolutional generation of textured 3d meshes. *Advances in Neural Information Processing Systems* 33 (2020), 870–882. 9
- [PYG*23] PO R., YIFAN W., GOLYANIK V., ABERMAN K., BARRON J. T., BERMANO A. H., CHAN E. R., DEKEL T., HOLYNSKI A., KANAZAWA A., ET AL.: State of the art on diffusion models for visual computing. *arXiv preprint arXiv:2310.07204* (2023). 3
- [PYL*22] PENG Y., YAN Y., LIU S., CHENG Y., GUAN S., PAN B., ZHAI G., YANG X.: CageNeRF: Cage-based neural radiance field for generalized 3d deformation and animation. In *NeurIPS* (2022). 20
- [PZ17] PENNER E., ZHANG L.: Soft 3d reconstruction for view synthesis. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 1–11. 5
- [PZVBG00] PFISTER H., ZWICKER M., VAN BAAR J., GROSS M.: Surfels: Surface elements as rendering primitives. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques* (2000), pp. 335–342. 3
- [QMH*23] QIAN G., MAI J., HAMDI A., REN J., SIAROHIN A., LI B., LEE H.-Y., SKOROKHODOV I., WONKA P., TULYAKOV S., ET AL.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843* (2023). 12
- [RALB22] RAKHIMOV R., ARDELEAN A.-T., LEMPITSKY V., BURNAEV E.: Npbg++: Accelerating neural point-based graphics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 15969–15979. 4
- [RBL*22a] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *CVPR* (2022). 1, 2, 9, 19, 20
- [RBL*22b] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 10684–10695. 12, 16
- [REO21] ROMBACH R., ESSER P., OMMER B.: Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 14356–14366. 13
- [RFJ21] ROCKWELL C., FOHEY D. F., JOHNSON J.: Pixelsynth: Generating a 3d-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 14104–14113. 14, 18
- [RFS22] RÜCKERT D., FRANKE L., STAMMINGER M.: Adop: Approximate differentiable one-pixel point rendering. *ACM Transactions on Graphics (ToG)* 41, 4 (2022), 1–14. 4
- [RK20] RIEGLER G., KOLTUN V.: Free view synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX* 16 (2020), Springer, pp. 623–640. 4
- [RK21] RIEGLER G., KOLTUN V.: Stable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 12216–12225. 4
- [RKH*21a] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SAstry G., ASKELL A., MISHKIN P., CLARK J., ET AL.: Learning transferable visual models from natural language supervision. In *ICML* (2021). 11, 19
- [RKH*21b] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SAstry G., ASKELL A., MISHKIN P., CLARK J., ET AL.: Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (2021), PMLR, pp. 8748–8763. 16
- [RKP*23] RAJ A., KAZA S., POOLE B., NIEMEYER M., RUIZ N., MILDENHALL B., ZADA S., ABERMAN K., RUBINSTEIN M., BARRON J., ET AL.: Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508* (2023). 18
- [RPLG21] REISER C., PENG S., LIAO Y., GEIGER A.: Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 14335–14345. 5
- [RROG18] ROVERI R., RAHMANN L., OZTIRELI C., GROSS M.: A network architecture for point cloud classification via automatic depth images generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 4176–4184. 4
- [RSH*21] REIZENSTEIN J., SHAPOVALOV R., HENZLER P., SBORDONE L., LABATUT P., NOVOTNY D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 10901–10911. 14, 15
- [RWC*19] RADFORD A., WU J., CHILD R., LUAN D., AMODEI D., SUTSKEVER I., ET AL.: Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9. 2, 6, 10
- [RWL*22] RÜCKERT D., WANG Y., LI R., IDOUGHI R., HEIDRICH W.: Neat: Neural adaptive tomography. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13. 5
- [SAA*23] SIDDIQUI Y., ALLIEGRO A., ARTEMOV A., TOMMASI T., SIRIGATTI D., ROSOV V., DAI A., NIESSNER M.: Meshgpt: Generating triangle meshes with decoder-only transformers. *arXiv preprint arXiv:2311.15475* (2023). 21
- [SCL*22] SANGHI A., CHU H., LAMBOURNE J. G., WANG Y., CHENG C.-Y., FUMERO M., MALEKSHAN K. R.: Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18603–18613. 18
- [SCP*23] SHUE J. R., CHAN E. R., PO R., ANKNER Z., WU J., WETZSTEIN G.: 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 20875–20886. 6, 10, 18, 19
- [SCS*22] SAHARIA C., CHAN W., SAXENA S., LI L., WHANG J., DENTON E. L., GHASEMIPOUR K., GONTIJO LOPES R., KARAGOL AYAN B., SALIMANS T., ET AL.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494. 2, 6, 16
- [SCZ*23] SHI R., CHEN H., ZHANG Z., LIU M., XU C., WEI X., CHEN L., ZENG C., SU H.: Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110* (2023). 13
- [SDZ*21] SRINIVASAN P. P., DENG B., ZHANG X., TANCIK M., MILDENHALL B., BARRON J. T.: Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the*

- [IEEECVF Conference on Computer Vision and Pattern Recognition (2021), pp. 7495–7504. 5
- [SFL*23] SANGHI A., FU R., LIU V., WILLIS K. D., SHAYANI H., KHASAHMADI A. H., SRIDHAR S., RITCHIE D.: Clip-sculptor: Zero-shot generation of high-fidelity and diverse shapes from natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 18339–18348. 18
- [SGHS98] SHADE J., GORTLER S., HE L.-W., SZELISKI R.: Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques* (1998), pp. 231–242. 5
- [SGY*21] SHEN T., GAO J., YIN K., LIU M.-Y., FIDLER S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 6087–6101. 1, 6, 10, 12
- [SHG*22] SHRESTHA R., HU S., GOU M., LIU Z., TAN P.: A real world dataset for multi-view 3d reconstruction. In *European Conference on Computer Vision* (2022), Springer, pp. 56–73. 15
- [SHN*19] SAITO S., HUANG Z., NATSUME R., MORISHIMA S., KANAZAWA A., LI H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 2304–2314. 16
- [SKJ23] SHIM J., KANG C., JOO K.: Diffusion-based signed distance fields for 3d shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 20887–20897. 10
- [SLNG20] SCHWARZ K., LIAO Y., NIEMEYER M., GEIGER A.: GRAF: Generative radiance fields for 3d-aware image synthesis. In *NeurIPS* (2020). 9, 15, 17
- [SMKF04] SHILANE P., MIN P., KAZHDAN M., FUNKHOUSER T.: The princeton shape benchmark. In *Proceedings Shape Modeling Applications, 2004.* (2004), IEEE, pp. 167–178. 14
- [SMP*22] SAJJADI M. S., MEYER H., POT E., BERGMANN U., GREFF K., RADWAN N., VORA S., LUČIĆ M., DUCKWORTH D., DOSOVITSKIY A., ET AL.: Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 6229–6238. 13, 14
- [SPH*23] SINGER U., POLYAK A., HAYES T., YIN X., AN J., ZHANG S., HU Q., YANG H., ASHUAL O., GAFNI O., ET AL.: Make-a-video: Text-to-video generation without text-video data. In *Int. Conf. Learn. Represent.* (2023). 2
- [SPK19] SHU D. W., PARK S. W., KWON J.: 3d point cloud generative adversarial network based on tree structured graph convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 3859–3868. 7
- [SPX*22] SHI Z., PENG S., XU Y., GEIGER A., LIAO Y., SHEN Y.: Deep generative models on 3d representations: A survey. *arXiv preprint arXiv:2210.15663* (2022). 3
- [SS87] SHIRMAN L. A., SEQUIN C. H.: Local surface interpolation with bézier patches. *Computer Aided Geometric Design* 4, 4 (1987), 279–295. 4
- [SSC22] SUN C., SUN M., CHEN H.-T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 5459–5469. 6
- [SSKH20] SHIH M.-L., SU S.-Y., KOPF J., HUANG J.-B.: 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 8028–8038. 5
- [SSN*14] SINGH A., SHA J., NARAYAN K. S., ACHIM T., ABBEEL P.: Bigbird: A large-scale 3d database of object instances. In *2014 IEEE international conference on robotics and automation (ICRA)* (2014), IEEE, pp. 509–516. 14
- [SSN*22] SCHWARZ K., SAUER A., NIEMEYER M., LIAO Y., GEIGER A.: Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *Advances in Neural Information Processing Systems* 35 (2022), 33999–34011. 6
- [STB*19] SRINIVASAN P. P., TUCKER R., BARRON J. T., RAMAMOORTHI R., NG R., SNAVELY N.: Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 175–184. 5
- [SWL*20a] SUN Y., WANG Y., LIU Z., SIEGEL J., SARMA S.: Pointgrow: Autoregressively learned point cloud generation with self-attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2020), pp. 61–70. 7
- [SWL*20b] SUN Y., WANG Y., LIU Z., SIEGEL J., SARMA S.: Pointgrow: Autoregressively learned point cloud generation with self-attention. In *WACV* (2020). 10
- [SWS*22] SUN J., WANG X., SHI Y., WANG L., WANG J., LIU Y.: IDE-3D: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Trans. Graph.* (2022). 18, 19
- [SWW*23] SUN J., WANG X., WANG L., LI X., ZHANG Y., ZHANG H., LIU Y.: Next3D: Generative neural texture rasterization for 3d-aware head avatars. In *CVPR* (2023). 18
- [SWY*23] SHI Y., WANG P., YE J., LONG M., LI K., YANG X.: Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512* (2023). 7, 12
- [SWZ*18] SUN X., WU J., ZHANG X., ZHANG Z., ZHANG C., XUE T., TENENBAUM J. B., FREEMAN W. T.: Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 2974–2983. 14
- [SWZ*22] SUN J., WANG X., ZHANG Y., LI X., ZHANG Q., LIU Y., WANG J.: Fenerf: Face editing in neural radiance fields. In *CVPR* (2022). 18, 19
- [SZS*23] SUN J., ZHANG B., SHAO R., WANG L., LIU W., XIE Z., LIU Y.: Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818* (2023). 12
- [TDB16] TATARCHENKOV M., DOSOVITSKIY A., BROX T.: Multi-view 3d models from single images with a convolutional network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14* (2016), Springer, pp. 322–337. 14
- [TFT*20] TEWARI A., FRIED O., THIES J., SITZMANN V., LOMBARDI S., SUNKAVALLI K., MARTIN-BRULLA R., SIMON T., SARAGIH J., NIESSNER M., ET AL.: State of the art on neural rendering. In *Computer Graphics Forum* (2020), vol. 39, Wiley Online Library, pp. 701–727. 3
- [TLK*23] TSENG H.-Y., LI Q., KIM C., ALSISAN S., HUANG J.-B., KOPF J.: Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 16773–16783. 13
- [TLY*21] TAKIKAWA T., LITALIEN J., YIN K., KREIS K., LOOP C., NOWROUZEZHRAI D., JACOBSON A., MCGUIRE M., FIDLER S.: Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 11358–11367. 6
- [TLYCS22] TSENG W.-C., LIAO H.-J., YEN-CHEN L., SUN M.: CLA-NeRF: Category-level articulated neural radiance field. In *International Conference on Robotics and Automation (ICRA)* (2022). 19, 20
- [TME*22] TREMBLAY J., MESHRY M., EVANS A., KAUTZ J., KELLER A., KHAMIS S., MÜLLER T., LOOP C., MORRICAL N., NAGANO K., ET AL.: Rtmv: A ray-traced multi-view synthetic dataset for novel view synthesis. *arXiv preprint arXiv:2205.07058* (2022). 15
- [TRMT23] TRUONG P., RAKOTOSONA M.-J., MANHARDT F., TOMBARI F.: Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 4190–4200. 5

- [TRZ*23] TANG J., REN J., ZHOU H., LIU Z., ZENG G.: Dream-gaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653* (2023). 12, 18, 21
- [TTG*20] TRETSCHK E., TEWARI A., GOLYANIK V., ZOLLMÖFER M., STOLL C., THEOBALT C.: Patchnets: Patch-based generalizable deep implicit 3d shape representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16* (2020), Springer, pp. 293–309. 6
- [TTM*22] TEWARI A., THIES J., MILDENHALL B., SRINIVASAN P., TRETSCHK E., YIFAN W., LASSNER C., SITZMANN V., MARTIN-BRUALLA R., LOMBARDI S., ET AL.: Advances in neural rendering. In *Computer Graphics Forum* (2022), vol. 41, Wiley Online Library, pp. 703–735. 3
- [TWZ*23] TANG J., WANG T., ZHANG B., ZHANG T., YI R., MA L., CHEN D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184* (2023). 7, 12, 18, 19
- [TYC*23] TEWARI A., YIN T., CAZENAVETTE G., REZCHIKOV S., TENENBAUM J. B., DURAND F., FREEMAN W. T., SITZMANN V.: Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *arXiv preprint arXiv:2306.11719* (2023). 13
- [TZN19] THIES J., ZOLLMÖFER M., NIESSNER M.: Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)* 38, 4 (2019), 1–12. 4
- [VHM*22] VERBIN D., HEDMAN P., MILDENHALL B., ZICKLER T., BARRON J. T., SRINIVASAN P. P.: Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), IEEE, pp. 5481–5490. 5
- [VN*51] VON NEUMANN J., ET AL.: The general and logical theory of automata. 1951 (1951), 1–41. 13
- [WCH*22] WANG C., CHAI M., HE M., CHEN D., LIAO J.: CLIP-NeRF: Text-and-image driven manipulation of neural radiance fields. In *CVPR* (2022). 19
- [WCMB*22] WATSON D., CHAN W., MARTIN-BRUALLA R., HO J., TAGLIASACCHI A., NOROUZI M.: Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628* (2022). 13
- [WCS*22] WENG C.-Y., CURLESS B., SRINIVASAN P. P., BARRON J. T., KEMELMACHER-SHLIZERMAN I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition* (2022), pp. 16210–16220. 5
- [WDL*23] WANG H., DU X., LI J., YEH R. A., SHAKHNAROVICH G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 12619–12629. 12, 18
- [WDY*22] WU Y., DENG Y., YANG J., WEI F., CHEN Q., TONG X.: AniFaceGAN: Animatable 3d-aware face image generation for video avatars. In *NeurIPS* (2022). 18, 19
- [WGSJ20] WILES O., GKIOXARI G., SZELISKI R., JOHNSON J.: Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 7467–7477. 4, 14
- [WJC*23] WANG C., JIANG R., CHAI M., HE M., CHEN D., LIAO J.: NeRF-Art: Text-driven neural radiance fields stylization. *IEEE Trans. Vis. Comput. Graph.* (2023). 19
- [WKC*23] WANG C., KANG D., CAO Y., BAO L., SHAN Y., ZHANG S.-H.: Neural point-based volumetric avatar: Surface-guided neural points for efficient and photorealistic volumetric head avatar. In *ACM SIGGRAPH Asia 2023 Conference Proceedings* (2023). 17
- [WLC*22] WU Q., LIU X., CHEN Y., LI K., ZHENG C., CAI J., ZHENG J.: Object-compositional neural implicit surfaces. In *ECCV* (2022). 19, 20
- [WLG*23] WANG M., LIU Y.-S., GAO Y., SHI K., FANG Y., HAN Z.: Lp-dif: Learning local pattern-specific deep implicit function for 3d objects and scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 21856–21865. 6
- [WLL*21] WANG P., LIU L., LIU Y., THEOBALT C., KOMURA T., WANG W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems* 34 (2021), 27171–27183. 6
- [WLW*23] WANG Z., LU C., WANG Y., BAO F., LI C., SU H., ZHU J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213* (2023). 12, 18, 19, 21
- [WLY*23] WU T., LI Z., YANG S., ZHANG P., PAN X., WANG J., LIN D., LIU Z.: Hyperdreamer: Hyper-realistic 3d content generation and editing from a single image. In *SIGGRAPH Asia 2023 Conference Papers* (2023), pp. 1–10. 18
- [Wol83] WOLFRAM S.: Statistical mechanics of cellular automata. *Reviews of modern physics* 55, 3 (1983), 601. 13
- [WPH*23] WAN Z., PASCHALIDOU D., HUANG I., LIU H., SHEN B., XIANG X., LIAO J., GUIBAS L.: Cad: Photorealistic 3d generation via adversarial distillation. *arXiv preprint arXiv:2312.06663* (2023). 18
- [WSK*15] WU Z., SONG S., KHOSLA A., YU F., ZHANG L., TANG X., XIAO J.: 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1912–1920. 6
- [WSW22] WANG Y., SKOROKHODOV I., WONKA P.: HF-NeuS: Improved surface reconstruction using high-frequency details. *Advances in Neural Information Processing Systems* (2022). 6
- [WSW23] WANG Y., SKOROKHODOV I., WONKA P.: PET-NeuS: Positional encoding triplanes for neural surfaces. In *CVPR* (2023). 6
- [WWG*21] WANG Q., WANG Z., GENOVA K., SRINIVASAN P. P., ZHOU H., BARRON J. T., MARTIN-BRUALLA R., SNAVELY N., FUNKHOUSER T.: IBRNet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 4690–4699. 5, 13
- [WWL*23] WU Q., WANG K., LI K., ZHENG J., CAI J.: ObjectSDF++: Improved object-compositional neural implicit surfaces. In *ICCV* (2023). 19, 20
- [WWX*21] WANG Z., WU S., XIE W., CHEN M., PRISACARIU V. A.: Nerf-: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064* (2021). 5
- [WZ22] WU R., ZHENG C.: Learning to generate 3d shapes from a single example. *arXiv preprint arXiv:2208.02946* (2022). 18
- [WZF*23] WU T., ZHANG J., FU X., WANG Y., REN J., PAN L., WU W., YANG L., WANG J., QIAN C., ET AL.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 803–814. 14, 15
- [WZX*16] WU J., ZHANG C., XUE T., FREEMAN B., TENENBAUM J.: Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems* (2016). 1, 2, 6, 7, 9, 15, 18
- [WZZ*23] WANG T., ZHANG B., ZHANG T., GU S., BAO J., BALTRUSAITIS T., SHEN J., CHEN D., WEN F., CHEN Q., ET AL.: RODIN: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR* (2023). 10
- [XJW*23] XU D., JIANG Y., WANG P., FAN Z., WANG Y., WANG Z.: 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* (2023), pp. 4479–4489. 12
- [XKJ*23] XIONG Z., KANG D., JIN D., CHEN W., BAO L., CUI S., HAN X.: Get3dhuman: Lifting stylegan-human into a 3d generative model using pixel-aligned reconstruction priors. In *Proceedings of*

- the IEEE/CVF International Conference on Computer Vision (2023), pp. 9287–9297. [15](#), [16](#)
- [XSJ*23] XU H., SONG G., JIANG Z., ZHANG J., SHI Y., LIU J., MA W., FENG J., LUO L.: OmniAvatar: Geometry-guided controllable 3d head synthesis. In *CVPR* (2023). [18](#), [19](#)
- [XTL*23] XU Y., TAN H., LUAN F., BI S., WANG P., LI J., SHI Z., SUNKAVALLI K., WETZSTEIN G., XU Z., ET AL.: Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217* (2023). [7](#)
- [XTS*22] XIE Y., TAKIKAWA T., SAITO S., LITANY O., YAN S., KHAN N., TOMBARI F., TOMPKIN J., SITZMANN V., SRIDHAR S.: Neural fields in visual computing and beyond. In *Computer Graphics Forum* (2022), vol. 41, Wiley Online Library, pp. 641–676. [3](#)
- [XWC*19] XU Q., WANG W., CEYLAN D., MECH R., NEUMANN U.: Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in neural information processing systems* 32 (2019). [5](#)
- [XX23] XIA W., XUE J.-H.: A survey on deep generative 3d-aware image synthesis. *ACM Computing Surveys* 56, 4 (2023), 1–34. [3](#)
- [XYB*23] XU Y., YIFAN W., BERGMAN A. W., CHAI M., ZHOU B., WETZSTEIN G.: Efficient 3d articulated human generation with layered surface volumes. *arXiv preprint arXiv:2307.05462* (2023). [16](#)
- [XYC*23] XIU Y., YANG J., CAO X., TZIONAS D., BLACK M. J.: Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 512–523. [15](#), [16](#)
- [XYHT23] XIANG J., YANG J., HUANG B., TONG X.: 3d-aware image generation using 2d diffusion models. *arXiv preprint arXiv:2303.17905* (2023). [18](#), [19](#)
- [XYTB22] XIU Y., YANG J., TZIONAS D., BLACK M. J.: Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), IEEE, pp. 13286–13296. [15](#), [16](#)
- [YAK*20] YIFAN W., AIGERMAN N., KIM V. G., CHAUDHURI S., SORKINE-HORNUNG O.: Neural cages for detail-preserving 3d deformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 75–83. [5](#)
- [YBZ*22] YANG B., BAO C., ZENG J., BAO H., ZHANG Y., CUI Z., ZHANG G.: NeuMesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *ECCV* (2022), Springer. [19](#), [20](#)
- [YGKL21] YARIV L., GU J., KASTEN Y., LIPMAN Y.: Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems* 34 (2021), 4805–4815. [6](#)
- [YGMG23] YOO P., GUO J., MATSUO Y., GU S. S.: Dreampsparse: Escaping from plato’s cave with 2d frozen diffusion model given sparse views. *CoRR* (2023). [13](#)
- [YHH*19a] YANG G., HUANG X., HAO Z., LIU M.-Y., BELONGIE S., HARIHARAN B.: Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 4541–4550. [7](#), [11](#)
- [YHH*19b] YANG G., HUANG X., HAO Z., LIU M.-Y., BELONGIE S., HARIHARAN B.: Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 4541–4550. [21](#)
- [YLM*22] YAN X., LIN L., MITRA N. J., LISCHINSKI D., COHEN-OR D., HUANG H.: Shapeformer: Transformer-based shape completion via sparse representation. In *CVPR* (2022). [11](#)
- [YLWD22] YANG Z., LI S., WU W., DAI B.: 3dhumangan: Towards photo-realistic 3d-aware human image generation. *arXiv preprint arXiv:2212.07378* (2022). [16](#)
- [YXL*23] YANG X., LUO Y., XIU Y., WANG W., XU H., FAN Z.: Dif: Uncertainty-aware human digitization via implicit distribution field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 9122–9132. [16](#)
- [YPN*22] YU Z., PENG S., NIEMEYER M., SATTLER T., GEIGER A.: Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems* 35 (2022), 25018–25032. [6](#)
- [YRSh21] YIFAN W., RAHMANN L., SORKINE-HORNUNG O.: Geometry-consistent neural shape representation with implicit displacement fields. In *International Conference on Learning Representations* (2021). [6](#)
- [YSL*22] YUAN Y.-J., SUN Y.-T., LAI Y.-K., MA Y., JIA R., GAO L.: NeRF-Editing: geometry editing of neural radiance fields. In *CVPR* (2022). [5](#), [20](#)
- [YSW*19] YIFAN W., SERENA F., WU S., ÖZTIRELI C., SORKINE-HORNUNG O.: Differentiable surface splatting for point-based geometry processing. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–14. [4](#)
- [YTB*21] YENAMANDRA T., TEWARI A., BERNARD F., SEIDEL H.-P., ELGHARIB M., CREMERS D., THEOBALT C.: i3DMM: Deep implicit 3d morphable model of human heads. In *CVPR* (2021). [17](#)
- [YXZ*23] YU X., XU M., ZHANG Y., LIU H., YE C., WU Y., YAN Z., ZHU C., XIONG Z., LIANG T., ET AL.: Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 9150–9161. [14](#), [15](#)
- [YYC*23] YU W., YUAN L., CAO Y.-P., GAO X., LI X., QUAN L., SHAN Y., TIAN Y.: Hifi-123: Towards high-fidelity one image to 3d content generation. *arXiv preprint arXiv:2310.06744* (2023). [12](#)
- [YYTK21] YU A., YE V., TANCIK M., KANAZAWA A.: pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 4578–4587. [5](#), [13](#)
- [YZX*21] YANG B., ZHANG Y., XU Y., LI Y., ZHOU H., BAO H., ZHANG G., CUI Z.: Learning object-compositional neural radiance field for editable scene rendering. In *ICCV* (2021). [19](#), [20](#)
- [ZAB*22] ZHENG Y., ABREVAYA V. F., BÜHLER M. C., CHEN X., BLACK M. J., HILLIGES O.: IMAvatar: Implicit morphable head avatars from videos. In *CVPR* (2022). [17](#)
- [ZBT23] ZIELONKA W., BOLKART T., THIES J.: Instant volumetric head avatars. In *CVPR* (2023). [17](#)
- [ZCY*23] ZHANG H., CHEN B., YANG H., QU L., WANG X., CHEN L., LONG C., ZHU F., DU K., ZHENG M.: Avatarverse: High-quality & stable 3d avatar creation from text and pose. *arXiv preprint arXiv:2308.03610* (2023). [16](#)
- [ZDW21] ZHOU L., DU Y., WU J.: 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 5826–5835. [10](#), [18](#)
- [ZJ16] ZHOU Q., JACOBSON A.: Thingi10k: A dataset of 10,000 3d-printing models. *arXiv preprint arXiv:1605.04797* (2016). [14](#)
- [ZJY*22] ZHANG J., JIANG Z., YANG D., XU H., SHI Y., SONG G., XU Z., WANG X., FENG J.: Avatargen: a 3d generative model for animatable human avatars. In *European Conference on Computer Vision* (2022), Springer, pp. 668–685. [16](#)
- [ZKB*22] ZHANG K., KOLKIN N., BI S., LUAN F., XU Z., SHECHTMAN E., SNAVELY N.: ARF: Artistic radiance fields. In *ECCV* (2022), Springer. [19](#), [20](#)
- [ZKW*23] ZHOU A., KIM M. J., WANG L., FLORENCE P., FINN C.: Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 17907–17917. [5](#)
- [ZLC*23] ZHAO Z., LIU W., CHEN X., ZENG X., WANG R., CHENG P., FU B., CHEN T., YU G., GAO S.: Michelangelo: Conditional 3d

- shape generation based on shape-image-text aligned latent representation. *arXiv preprint arXiv:2306.17115* (2023). 18
- [ZLLD21] ZHI S., LAIDLLOW T., LEUTENEGGER S., DAVISON A. J.: In-place scene labelling and understanding with implicit scene representation. In *ICCV* (2021), pp. 15838–15847. 5
- [ZLLS22] ZHANG K., LUAN F., LI Z., SNAVELY N.: Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 5565–5574. 6
- [ZLW*21] ZHANG K., LUAN F., WANG Q., BALA K., SNAVELY N.: Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 5453–5462. 5
- [ZLW*22] ZHANG J., LI X., WAN Z., WANG C., LIAO J.: Fdnerf: Few-shot dynamic neural radiance fields for face reconstruction and expression editing. In *SIGGRAPH Asia 2022 Conference Papers* (2022), pp. 1–9. 19
- [ZLW*23] ZHANG J., LI X., WAN Z., WANG C., LIAO J.: Text2nerf: Text-driven 3d scene generation with neural radiance fields. *arXiv preprint arXiv:2305.11588* (2023). 18, 19
- [ZLWT22] ZHENG X., LIU Y., WANG P., TONG X.: Sdf-stylegan: Implicit sdf-based stylegan for 3d shape generation. In *Computer Graphics Forum* (2022), vol. 41, Wiley Online Library, pp. 52–63. 7
- [ZLZ*22] ZHU H., LIU Z., ZHOU Y., MA Z., CAO X.: DNF: Diffractive neural field for lensless microscopic imaging. *Optics Express* 30, 11 (2022), 18168–18178. 5
- [ZLZ*23] ZHANG J., LI X., ZHANG Q., CAO Y., SHAN Y., LIAO J.: Humanref: Single image to 3d human generation via reference-guided diffusion. *arXiv preprint arXiv:2311.16961* (2023). 15, 16
- [ZML*22] ZHOU J., MA B., LIU Y.-S., FANG Y., HAN Z.: Learning consistency-aware unsigned distance functions progressively from raw point clouds. *Advances in Neural Information Processing Systems* 35 (2022), 16481–16494. 5
- [ZPL*22] ZHU Z., PENG S., LARSSON V., XU W., BAO H., CUI Z., OSWALD M. R., POLLEFEYS M.: Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 12786–12796. 5
- [ZPVBG02] ZWICKER M., PFISTER H., VAN BAAR J., GROSS M.: Ewa splatting. *IEEE Transactions on Visualization and Computer Graphics* 8, 3 (2002), 223–238. 4
- [ZPW*23] ZHENG X.-Y., PAN H., WANG P.-S., TONG X., LIU Y., SHUM H.-Y.: Locally attentional sdf diffusion for controllable 3d shape generation. *arXiv preprint arXiv:2305.04461* (2023). 10
- [ZQL*23] ZHANG L., QIU Q., LIN H., ZHANG Q., SHI C., YANG W., SHI Y., YANG S., XU L., YU J.: DreamFace: Progressive generation of animatable 3d faces under text guidance. *ACM Trans. Graph.* (2023). 19
- [ZSD*21] ZHANG X., SRINIVASAN P. P., DENG B., DEBEVEC P., FREEMAN W. T., BARRON J. T.: Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)* 40, 6 (2021), 1–18. 5
- [ZSH*22] ZHANG Y., SUN J., HE X., FU H., JIA R., ZHOU X.: Modeling indirect illumination for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18643–18652. 6
- [ZST08] ZHANG W., SUN J., TANG X.: Cat head detection-how to effectively exploit shape and texture features. In *Computer Vision-ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV* 10 (2008), Springer, pp. 802–816. 15
- [ZTF*18] ZHOU T., TUCKER R., FLYNN J., FYFFE G., SNAVELY N.: Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817* (2018). 5
- [ZTNW23] ZHANG B., TANG J., NIESSNER M., WONKA P.: 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *arXiv preprint arXiv:2301.11445* (2023). 7
- [ZTS*16] ZHOU T., TULSIANI S., SUN W., MALIK J., EFROS A. A.: View synthesis by appearance flow. In *Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV* 14 (2016), Springer, pp. 286–301. 14
- [ZVW*22] ZENG X., VAHDAT A., WILLIAMS F., GOJCIC Z., LITANY O., FIDLER S., KREIS K.: Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978* (2022). 10
- [ZWL*23] ZHUANG J., WANG C., LIU L., LIN L., LI G.: Dreameditor: Text-driven 3d scene editing with neural fields. In *SIGGRAPH Asia Conference Papers* (2023). 20
- [ZXZ*23] ZHU C., XIAO F., ALVARADO A., BABAEI Y., HU J., ELMOHRI H., CULATANA S., SUMBALY R., YAN Z.: Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 20110–20120. 15
- [ZYHC22] ZHENG M., YANG H., HUANG D., CHEN L.: ImFace: A nonlinear 3d morphable face model with implicit neural representations. In *CVPR* (2022). 17
- [ZYLD21] ZHENG Z., YU T., LIU Y., DAI Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence* 44, 6 (2021), 3170–3184. 16
- [ZYW*23] ZHENG Y., YIFAN W., WETZSTEIN G., BLACK M. J., HILLIGES O.: PointAvatar: Deformable point-based head avatars from videos. In *CVPR* (2023). 17
- [ZZ23] ZHU J., ZHUANG P.: Hifa: High-fidelity text-to-3d with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766* (2023). 12, 18
- [ZZF*23] ZHUANG Y., ZHANG Q., FENG Y., ZHU H., YAO Y., LI X., CAO Y.-P., SHAN Y., CAO X.: Anti-aliased neural implicit surfaces with encoding level of detail. *arXiv preprint arXiv:2309.10336* (2023). 6
- [ZZK*20] ZAMORSKI M., ZIEBA M., KLUKOWSKI P., NOWAK R., KURACH K., STOKOWIEC W., TRZCIŃSKI T.: Adversarial autoencoders for compact representations of 3d point clouds. *Computer Vision and Image Understanding* 193 (2020), 102921. 7
- [ZZW*23] ZHUANG Y., ZHANG Q., WANG X., ZHU H., FENG Y., LI X., SHAN Y., CAO X.: Neai: A pre-convoluted representation for plug-and-play neural ambient illumination. *arXiv preprint arXiv:2304.08757* (2023). 5
- [ZZZ*18] ZHU J.-Y., ZHANG Z., ZHANG C., WU J., TORRALBA A., TENENBAUM J., FREEMAN B.: Visual object networks: Image generation with disentangled 3d representations. *Advances in neural information processing systems* 31 (2018). 15
- [ZZZ*23a] ZHANG J., ZHANG X., ZHANG H., LIEW J. H., ZHANG C., YANG Y., FENG J.: Avatarstudio: High-fidelity and animatable 3d avatar creation from text. *arXiv preprint arXiv:2311.17917* (2023). 16
- [ZZZ*23b] ZHU J., ZHU H., ZHANG Q., ZHU F., MA Z., CAO X.: Pyramid nerf: Frequency guided fast radiance field optimization. *International Journal of Computer Vision* (2023), 1–16. 5