Generative AI meets 3D: A Survey on Text-to-3D in AIGC Era

CHENGHAO LI, KAIST, South Korea
CHAONING ZHANG*, Kyung Hee University, South Korea
ATISH WAGHWASE, KAIST, South Korea
LIK-HANG LEE, Hong Kong Polytechnic University, Hong Kong (China)
FRANCOIS RAMEAU, State University of New York, Korea
YANG YANG, University of Electronic Science and technology, China
SUNG-HO BAE, Kyung Hee University, South Korea
CHOONG SEON HONG, Kyung Hee University, South Korea

Generative AI (AIGC, a.k.a. AI generated content) has made remarkable progress in the past few years, among which text-guided content generation is the most practical one since it enables the interaction between human instruction and AIGC. Due to the development in text-to-image as well 3D modeling technologies (like NeRF), text-to-3D has become a newly emerging yet highly active research field. Our work conducts the first yet comprehensive survey on text-to-3D to help readers interested in this direction quickly catch up with its fast development. First, we introduce 3D data representations, including both Euclidean data and non-Euclidean data. On top of that, we introduce various foundation technologies as well as summarize how recent works combine those foundation technologies to realize satisfactory text-to-3D. Moreover, we summarize how text-to-3D technology is used in various applications, including avatar generation, texture generation, shape transformation, and scene generation.

CCS Concepts: • **Computing methodologies → Reconstruction**; **Shape modeling**.

Additional Key Words and Phrases: text-to-3D, generative AI, AIGC, 3D generation, metaverse

ACM Reference Format:
Chenghao Li, Chaoning Zhang, Atish Waghwase, Lik-Hang Lee, Francois Rameau, Yang Yang, Sung-Ho Bae, and Choong Seon Hong. 2023. Generative AI meets 3D: A Survey on Text-to-3D in AIGC Era. 1, 1 (May 2023), 14 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding author: chaoningzhang1990@gmail.com

Authors' addresses: Chenghao Li, KAIST, South Korea, lch17692405449@gmail.com; Chaoning Zhang, Kyung Hee University, South Korea, chaoningzhang1990@gmail.com; Atish Waghwase, KAIST, South Korea, atishwaghwase@gmail.com; Lik-Hang Lee, Hong Kong Polytechnic University, Hong Kong (China), lik-hang.lee@polyu.edu.hk; Francois Rameau, State University of New York, Korea, lik-hang.lee@polyu.edu.hk; Yang Yang, University of Electronic Science and technology, China, dlyyang@gmail.com; Sung-Ho Bae, Kyung Hee University, South Korea, shbae@khu.ac.kr; Choong Seon Hong, Kyung Hee University, South Korea, cshong@khu.ac.kr.

XXXX-XXXX/2023/5-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

Contents

1 INTRODUCTION

Generative Artificial Intelligence as the main body generating high-quality and quantity content (also known as Artificial Intelligence Generated Content-AIGC) has aroused great attention in the past few years. The content generation paradigm guided and restrained by natural language, such as text-to-text (e.g. ChatGPT [Zhang et al. 2023c]) and text-to-image [Zhang et al. 2023d] (e.g. DALLE-2 [Ramesh et al. 2022]), is the most practical one, as it allows for a simple interaction between human guidance and generative AI [Zhang et al. 2023e]. The accomplishment of Generative AI in the field of text-to-image [Zhang et al. 2023d] is quite remarkable. As we are in a 3D world, it is necessary to extend AIGC to 3D domain. There is a great demand for 3D digital content in many application scenarios, including games, movies, virtual reality, architecture and robots, such as 3D character generation, 3D texture generation, 3D scene generation, etc. However, it requires a lot of artistic and aesthetic training, as well as professional knowledge in 3D modeling, to cultivate a professional 3D modeler. Given the current trend of 3D model development, it is essential to utilize generative AI to generate high-quality and large-scale 3D models. In addition, text-to-3D AI modeling can greatly assist both newbies and professionals to realize free creation of 3D contents.

Previous methods of text-to-3D shapes have attempted to learn a cross-modal mapping by directly learning from text-3D pairs [Achlioptas et al. 2019; Chen et al. 2019] and generate joint representations. Compared to text-to-image, the task of generating 3D shapes from text is more challenging. Firstly, unlike 2D images, 3D shapes are mostly unstructured and irregular non-Euclidean data, making it difficult to apply traditional 2D deep learning models to these data. Moreover, there are a large number of large-scale image-text pairs datasets available online to support text-to-image generation. However, to our knowledge, the largest text-to-3D dataset proposed in [Fu et al. 2022] contained only 369K text-3D pairs and is limited to a few object categories. This is significantly lower than the datasets which contain 5.85B text-image pairs [Schuhmann et al. 2022]. The lack of large-scale and high-quality training data makes the task of text-to-3D even more difficult.

Recently, the advent of some key technologies has enabled a new paradigm of text-to-3D tasks. Firstly, Neural Radiance Fields (NeRF) [Mildenhall et al. 2021] is an emergent 3D data representation approach. Initially, NeRF was found to perform well in the 3D reconstruction task, and recently NeRF and other neural 3D representations have been applied to new view synthesis tasks that can use real-world RGB photos. NeRF is trained to reconstruct images from multiple viewpoints. As the learned radiance fields are shared between viewpoints, NeRF can smoothly and consistently interpolate between viewpoints. Due to its neural representation, NeRF can sample with high spatial resolution, unlike voxel representations and point clouds, and is easier to optimize than meshes and other explicit geometric representations, since it is topology-free. The advent of NeRF breaks the stalemate of 3D data scarcity and serves as a soft bridge from 2D to 3D representation, elegantly solving the problem of 3D data scarcity. Secondly, with the remarkable progress of multimodal AI [Radford et al. 2021] and diffusion models [Ho et al. 2020], text-guided image content AI generation has made significant progress. The key driving factor is the large-scale datasets of billions of text-image pairs obtained from the Internet. Recent works have emerged that guide 3D modeling optimization by leveraging the prior of a pre-trained text-to-image generation model. In other words, they often text-guided 3D model generation with text-to-image priors.

Overall, this work conducts the first yet comprehensive survey on text-to-3D. The rest of this work is organized as follows. Sec 2 reviews the different representations of 3D. Sec 3 first introduces the technology behind text-to-3D, and then summarizes the recent papers. Sec 4 introduces the application of text-to-3D in various fields.

2 3D DATA REPRESENTATION

3D data can have different representations [Ahmed et al. 2018], divided into Euclidean and non-Euclidean. 3D Euclidean data has a potential grid structure, which allows global parameterization and a common coordinate system. These properties make extending existing 2D deep learning paradigms to 3D data a simple task, where convolution operations remain the same as 2D. On the other hand, 3D non-Euclidean data does not have a grid array structure and is not globally parameterized. Therefore, extending classical deep learning techniques to such representations is a challenging task. In real life, the research of deep learning techniques in non-Euclidean domains is of great significance. This is referred to as geometric deep learning [Cao et al. 2020].

2.1 Euclidean data

The Euclidean data preserves the attribute of the grid structure, with global parameterization and a common coordinate system. The major 3D data representations in this category include voxel grids and multi-view images.

Fig. 1. Voxel representation of Stanford bunny, the picture is obtained from [Shi et al. 2022].

2.1.1 *Voxel Grids.*

Voxels can be used to represent individual samples or data points on a regularly spaced three-dimensional grid, which is a Euclidean structured data structure similar to pixels [Blinn 2005] in 2D space. The data point can contain a single data, such as opacity, or multiple data, such as color and opacity. Voxels can also store high-dimensional feature vectors in data points, such as geometric occupancy [Mescheder et al. 2019], volumetric density [Minto et al. 2018], or signed distance values [Park et al. 2019]. A voxel only represents a point on this grid, not a volume; the space between voxels is not represented in the voxel-based dataset. Depending on the data type and the expected use of the dataset, this lost information can be reconstructed and/or approximated, for example, by interpolation. The representation of voxels is simple and the spatial structure is clear, which is highly extensible and can be easily applied to convolutional neural networks [Wang et al. 2019]. However, the efficiency is low, as it represents both the occupied parts of the scene and the unoccupied parts, which leads to a large amount of unnecessary storage requirements. This leads to voxels being unsuitable for representing high-resolution data. Voxel grids have many applications in rendering tasks[Hu et al. 2023; Rematas and Ferrari 2020]. Early methods store high-dimensional feature vectors in voxels to encode the geometry and appearance of the scene, usually referred to as a feature volume, which can be interpreted as a color image using projection and 2D cellular neural networks. This also includes volumetric imaging in medicine, as well as terrain representation in games and simulations.

2.1.2 *Multi-view Images.*

With the development of computer vision technology and the remarkable improvement in computing power,

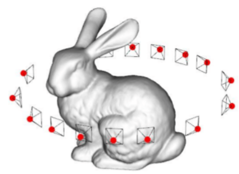

Fig. 2. Multi-view representation of Stanford bunny, the picture is obtained from [Park et al. 2016].

coupled with the latest developments in modern digital cameras, it is now possible to easily capture large amounts of high-resolution images. There is an urgent need to extract 3D structure from these images for many applications, such as 3D reconstruction [Jin et al. 2020]. A multi-view image dataset is an aggregation of multiple images, each representing an object or scene from different perspectives, such as the front, side and top, aggregated together to form a multi-view image dataset. Since collecting 3D data from the real world is time-consuming, and deep learning paradigms rely on large amounts of data for training, the availability of multi-view image datasets is its greatest advantage. Its drawback is that multi-view images cannot be strictly defined as 3D model data, but it provides a bridge between 2D and 3D visualizations. Recently, NeRF [Mildenhall et al. 2021] has emerged as a novel approach for 3D reconstruction, which is well-suited for the massive data requirement of the learning-based generalizable NeRF methods with large-scale multi-view datasets [Yu et al. 2023]. It can also be applicable to multi-view stereo [Furukawa et al. 2015] and view-consistent image understanding [Dong et al. 2022] tasks.

2.2 non-Euclidean data

The second type of 3D data representation is non-Euclidean data. This type of data does not have global parametrization or common coordinate systems, which makes it difficult to extend 2D deep learning paradigms. Much effort has been made in learning this data representation and applying DL techniques. Researching deep learning techniques in non-Euclidean domains is of great importance, this is referred to as Geometric Deep Learning [Cao et al. 2020]. The main types of non-Euclidean data are point clouds, 3D meshes and implicit.

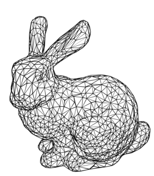

Fig. 3. Mesh representation of Stanford bunny, the picture is obtained from [Rossi et al. 2021].

2.2.1 Meshes. 3D meshes [Wang and Zhang 2022] are one of the most popular representations of 3D shapes. A 3D mesh structure is composed of a set of polygons, termed faces, which are described based on a set of vertices that describe the presence of coordinates in 3D space. These vertices are associated with a connectivity list that describes how these vertices are interconnected. As meshes only model the surface of the scene, they are more compact. The meshes provide connectivity of surface points for modeling point relationships. Due to these advantages, polygon meshes are widely used in traditional computer graphics [Zhou et al. 2021b] applications such as geometry processing, animation, and rendering. However, on a global level, meshes are non-Euclidean data, and the local geometry of the mesh can be represented as a subset of the Euclidean space where the known properties of the Euclidean space are not well-defined, such as shift-invariance, vector space operations, and global parameterization systems. Thus, deep learning of 3D meshes is a challenging task. However, with the development of graph neural networks [Wu et al. 2020], meshes can be seen as graphs. MeshCNN [Hanocka et al. 2019] specifically designs convolutional and pooling layers for mesh edges and extracts edge features for shape analysis. 3D meshes are important in many fields and industries, such as architecture and building, furniture and home living, gaming and entertainment, product design, medical and life sciences, etc. They can be used for designing, visualizing, analyzing architecture and products, creating character objects for games, movies, etc., designing new products, visualizing and analyzing anatomical structures, and helping to increase understanding of diseases and treatment methods.

Fig. 4. Point cloud representation of Stanford bunny model, the figure is obtained from [Agarwal and Prabhakaran 2009].

2.2.2 Point Clouds. With the trend of inexpensive and convenient point cloud acquisition equipment, point clouds have been widely used in modeling and rendering, augmented reality, autonomous vehicles, etc [Guo et al. 2020]. Point clouds are a disordered set of discrete samples of three-dimensional shapes on three-dimensional space. Traditionally, point clouds are non-Euclidean data because point cloud data is non-structured globally. However, point clouds can also be realized as a set of globally parametrized small Euclidean subsets. The definition of point cloud structure depends on whether to consider the global or local structure of the object. Since most applications strive to capture the global characteristics of the object to perform complex tasks, traditionally point clouds are non-Euclidean data. Point clouds are the direct output of depth sensors [Liu et al. 2019] and therefore are very popular in 3D scene understanding tasks. Despite being easily obtainable, the irregularity of point clouds makes them hard to process with traditional 2D neural networks. Numerous geometric deep learning [Cao et al. 2020] methods have been proposed to effectively analyze three-dimensional point clouds, such as PointNet [Qi et al. 2017], which is a deep learning network structure based on raw point cloud data and can directly use raw point cloud data as input and use a set of sparse keypoints to summarize the input point cloud, and can effectively process data and have robustness to small perturbations of the input and can achieve good performance in tasks such as shape classification, part segmentation, and scene segmentation. 3D point cloud technology can be applied to multiple fields such as architecture, engineering, civil building design, geological survey, machine vision, agriculture, space information, and automatic driving, and can provide more accurate modeling and analysis, as well as more accurate positioning and tracking.

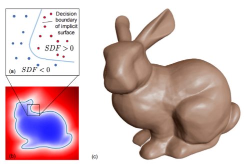

Fig. 5. Neural field representation of Stanford bunny, the picture is obtained from [Park et al. 2019].

2.2.3 Neural Fields. Neural fields are a domain that is either wholly or partially parameterized by neural networks and represented entirely or partially by neural networks for scenes or objects in 3D space.At each point in 3D space, a neural network can map its associated characteristics to attributes. Neural fields are capable of representing 3D scenes or objects in any resolution and unknown or complex topology due to their continuous representation. Additionally, compared to the above representations, only the parameters of the neural network need to be stored, resulting in lower memory consumption than other representations. The earliest work on neural fields was used for 3D shape representation [Peng and Shamsuddin 2004]. SDF [Park et al. 2019] is a classical approach to represent 3D shapes as neural fields. SDF is based on a continuous volumetric field, represented by the distance and sign at each point on surface. Several works [Gao et al. 2022b; Shen et al. 2021] use SDF to generate 3D shapes as representation. NeRF [Mildenhall et al. 2021], a recent emerging representation for 3D reconstruction in neural fields, has the advantages of high-quality and realistic 3D model generation, presenting realistic object surfaces and texture details at any angle and distance. Furthermore, it can generate 3D models from any number of input images without specific processing or labeling of the inputs. Another advantage of neural fields is that the neural network can be operated on low-power devices after it is trained. Polygon ray tracing renders high-resolution and realistic scenes at high frame rates, which requires expensive graphic cards, but high-quality neural fields can be rendered on mobile phones and even web browsers. However, there are also some drawbacks of neural field technology, such as the need for a large amount of computational resources and time for training, as well as difficulty in handling large-scale scenes and complex lighting conditions, and its inability to be structured data, which makes it difficult to be directly applied to 3D assets. Neural fields are a new emerging 3D representation technology with strong application prospects and can be used in 3D fields such as VR/AR and games.

3 TEXT-TO-3D TECHNOLOGIES

In the past few years, the success of deep generative models [Ho et al. 2020] in 2D images has been incredible. Training generative models in 2D space cannot meet the needs of some practical applications, as our physical world actually operates in 3D space. 3D data generation is of paramount importance. The success of Neural Radiance Fields [Mildenhall et al. 2021] has transformed the 3D reconstruction race, bringing 3D data to a whole new level. Combining prior knowledge from text-to-image models [Ramesh et al. 2021], many pioneers have achieved remarkable results in text-to-3D generation. In this section, we will first review the key techniques underlying text-to-3D generation. Secondly, we will survey recent text-to-3D models.

3.1 Foundation Technologies

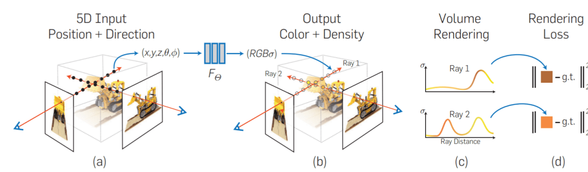

Fig. 6. An overview of our neural radiance field scene representation and differentiable rendering procedure, the picture is obtained from [Mildenhall et al. 2021].

3.1.1 NeRF. Neural Rediance Field (NeRF) [Gao et al. 2022a; Mildenhall et al. 2021] is a neural network-based implicit representation of 3D scenes, which can render projection images from a given viewpoint and a given position. Specifically, given a 3D point $\mathbf{x} \in \mathbb{R}^3$ and an observation direction unit vector $\mathbf{d} \in \mathbb{R}^2$, NeRF encodes the scene as a continuous volumetric radiance field f, yielding a differential density σ and an RGB color $\mathbf{c}$: $f(\mathbf{x}, \mathbf{d}) = (\sigma, \mathbf{c})$.

Rendering of images from desired perspectives can be achieved by integrating color along a suitable ray, $\mathbf{r}$, for each pixel in accordance with the volume rendering equation:

$$\hat{C}(r) = \int_{t_n}^{t_f} T(t)\sigma(t)\mathbf{c}(t)dt \tag{1}$$

$$T(t) = exp\left(-\int_{t_n}^{t} \sigma(s)ds\right) \tag{2}$$

The transmission coefficient $T(t)$ is defined as the probability that light is not absorbed from the near-field boundary t_n to t.

In order to train NeRF network and optimize the predicted color $\hat{C}$ to fit with the ray $\mathcal{R}$ corresponding to the pixel in the training images, gradient descent is used to optimize the network and match the target pixel color by loss:

$$\mathcal{L} = \sum_{\mathbf{r} \in \mathcal{R}} \|C(\mathbf{r}) - \hat{C}(\mathbf{r})\|_2^2 \tag{3}$$

3.1.2 CLIP. Recent advances in multimodal learning have enabled the development of cross-modal matching models such as CLIP [Radford et al. 2021](Contrastive Language-Image Pre-training) which learn shared representations from image-text pairs. These models are able to produce a scalar score that indicates whether an image and its associated caption match or not.

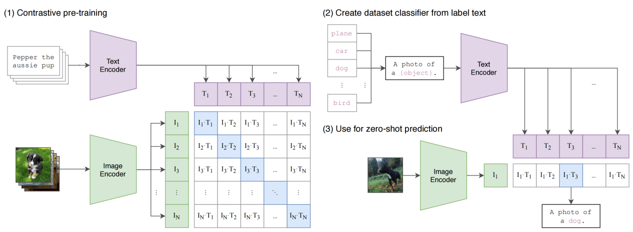

Fig. 7. CLIP structure, picture obtained from [Radford et al. 2021].

In training, the standard image model is used to jointly train the image feature extractor and linear classifier to predict some labels; CLIP jointly trains the image encoder and text encoder to predict the correct pairings of a batch of (image, text) training samples. The symmetric InfoNCE loss is used to train the image and text encoders, which can then be used for a number of downstream tasks. In testing, the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset categories. Building on this, a volume has been optimized to produce a high-scoring image rather than reranking. The CLIP structure is shown in Figure 7.

3.1.3 Diffusion model. In the past few years, the use of diffusion model [Ho et al. 2020] has seen a dramatic increase. Also known as denoising diffusion probabilistic models (DDPMs) or score-based generative models, these models generate new data that is similar to the data used to train them. Drawing inspiration from non-equilibrium thermodynamics, DDPMs are defined as a parameterized Markov chain of diffusion steps that adds random noise to the training data and learns to reverse the diffusion process to produce the desired data samples from the pure noise.

In the forward process, DDPM destroys the training data by gradually adding Gaussian noise. It starts from a data sample $\mathbf{x}_0$ and iteratively generates noisier samples $\mathbf{x}_T$ with $q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$, using a Gaussian diffusion kernel:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{4}$$

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{5}$$

where T and β_t are the steps and hyper-parameters, respectively. We can obtain noised image at arbitrary step t with Gaussian noise transition kernel as $\mathcal{N}$ in Eq. 5, by setting $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$:

$$q(x_t|x_0) := \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \tag{6}$$

The reverse denoising process of DDPM involves learning to undo the forward diffusion by performing iterative denoising, thereby generating data from random noise. This process is formally defined as a stochastic process, where the optimization objective is to generate $p_\theta(x_0)$ which follows the true data distribution $q(x_0)$ by starting from $p_\theta(T)$:

$$E_{t\sim\mathcal{U}(1,T),\mathbf{x}_0\sim q(\mathbf{x}_0),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\lambda(t)\,\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2 \tag{7}$$

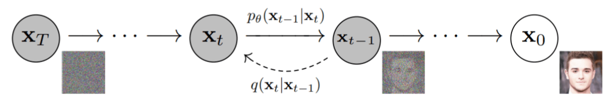

Fig. 8. Overview of DDPM, the picture is obtained from [Ho et al. 2020]

3.1.4 Pretrained text-guided image generation model. Recently, with the emergence of diffusion models, pre-trained models for text-to-image generation based on diffusion models have become good priors for text-to-3D tasks [Zhang et al. 2023d]. The pioneering works into image framework can be roughly categorized according to the diffusion priors in the pixel space or latent space. The first class of methods directly generates images from high-dimensional pixel-level, including GLIDE [Nichol et al. 2021] and Imagen [Saharia et al. 2022]. Another approach suggests compressing the image to a low-dimensional space first, and then training a diffusion model in this latent space. Representative methods of latent space include Stable Diffusion [Rombach et al. 2022] and DALLE-2 [Ramesh et al. 2022].

The training process of text-to-image generation can be roughly divided into three steps. Firstly, the CLIP model is employed to learn the correlations between text and visual semantics and map the text into the image representation space. Secondly, a prior model is used for inversion to map the image representation space back to the text space, and thus generate text-conditional images. Lastly, diffusion models are utilized to learn the mapping between text encoding and image encoding, providing prior models for text-conditional image generation. The structure of DALLE-2 is shown in Figure 9

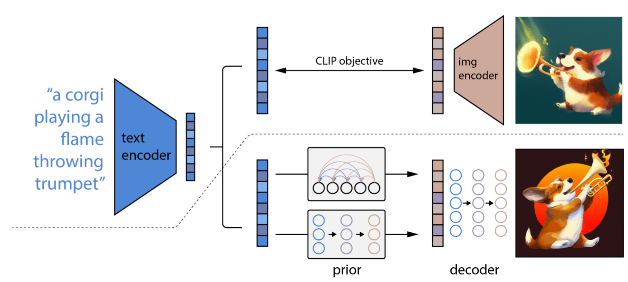

Fig. 9. Structure of DALLE-2, picture obtained from [Ramesh et al. 2022].

3.2 Successful attempts

Recent pioneering studies have demonstrated the utility of pre-trained text-to-image diffusion models to optimize neural radiance fields, achieving significant text-to-3D synthesis results. However, this paradigm also has some issues, and some followers aim to solve these problems. In this section, the first part presents the pioneers of this paradigm, and the second part presents some enhancement works of this paradigm.

Fig. 10. DreamFusion outcome. Picture obtained from [Poole et al. 2022].

3.2.1 *Poinners of CLIP-Based Text-Guided 3D Shape Generation.*

In recent years, with the success of text-to-image generation modelling [Zhang et al. 2023d], text-to-3D generation has also attracted the attention of the deep learning community [Han et al. 2023; Jain et al. 2022; Lin et al. 2022; Mohammad Khalid et al. 2022; Poole et al. 2022; Wang et al. 2022b; Xu et al. 2022b]. However, the scarcity of 3D data makes expansion with data challenging. DreamField [Jain et al. 2022] and CLIP-Mesh [Mohammad Khalid et al. 2022] rely on pre-trained image-text models [Radford et al. 2021] to optimize underlying 3D representations (RMS and meshes) in order to alleviate the training data problem, achieving high text-image alignment scores for all 2D renderings. Although these methods avoid the costly requirement of 3D training data and primarily rely on a large-scale pre-trained image-text model, they often yield less realistic 2D renderings. Recently, DreamFusion [Poole et al. 2022] and Magic3D [Lin et al. 2022] has demonstrated impressive capabilities in text-to-3D synthesis by leveraging a powerful pre-trained text-to-image diffusion model as a strong image prior.

CLIP-Forge [Sanghi et al. 2022] was the first to attempt to apply a pre-trained text-to-image model to 3D generation and was successful. CLIP-Forge is a novel approach to text-to-shape generation that involves no paired text-shape labels. It uses a pre-trained text encoder and an autoencoder to obtain a latent space for shapes. A normalizing flow network is then conditioned with text features to generate a shape embedding which is then converted into 3D shape. CLIP-Forge has an efficient generation process which requires no inference time optimization and can generate multiple shapes for a given text. See Figure 11 for details. This method also has the advantage of avoiding the expensive inference time optimizations as employed in existing text-to-shape generation models. Extensive evaluation of the method in various zero-shot generation settings is provided, both qualitatively and quantitatively.

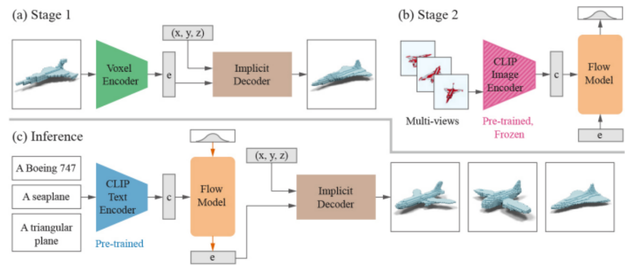

Fig. 11. Illustration of the main idea of CLIP-Forge, the picture obtained from [Sanghi et al. 2022].

Dream Fields [Jain et al. 2022] is a concurrent work to CLIP-Forge [Sanghi et al. 2022], which uses CLIP to synthesize and manipulate 3D object representations. However, the work differ from Dream Fields in terms of features, resolutions, and limitations. CLIP-Forge generalizes poorly outside of ShapeNet categories and requires ground-truth multi-view images and voxel data. Dream Fields proposes a method for synthesizing diverse three-dimensional objects based on natural language descriptions. This method combines neural rendering and multi-modal image and text representations, generating various categories of geometry and colors using an optimized multi-viewpoint neural radiation field without the need of three-dimensional supervision. To enhance the realism and visual quality, Dream Fields introduces simple geometric priors including sparsity-induced transmission regularization, scene boundaries and a new MLP architecture. See Figure 12 for details.Experimental results show that Dream Fields can generate realistic and multi-viewpoint consistent geometry and colors from various natural language descriptions.

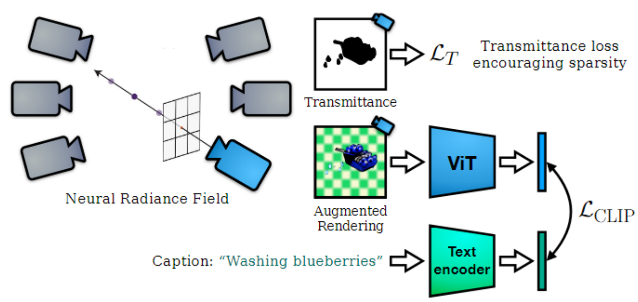

Fig. 12. Structure of Dream Fields, picture obtained from [Jain et al. 2022]

CLIP-NeRF [Wang et al. 2022a] is a novel disentangled conditional NeRF architecture that offers flexible control for editing NeRFs based on text prompts or reference images. It introduces a shape code and an appearance code to independently control the deformation of the volumetric field and the emitted colors. Two code mappers, fed by the pre-trained CLIP model, enable fast inference for editing different objects in the same category compared to the optimization-based editing method. In addition, an inversion method is proposed to infer the shape and appearance codes from a real image, enabling users to edit the shape and appearance of existing data. CLIP-NeRF is a contemporary work of Dream Fields [Jain et al. 2022], and unlike Dream Fields, the former offers greater freedom in shape manipulation and supports global deformation, introducing two intuitive NeRF editing methods: using short text prompts or sample images, both of which are more user-friendly to novice users. The structure of CLIP-NeRF is shown in Figure **??**

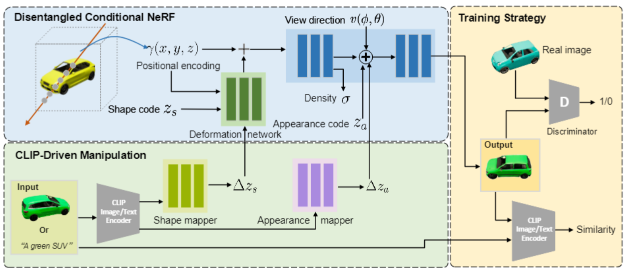

Fig. 13. Sturcute of CLIP-NeRF, picture obtained from [Wang et al. 2022a]

DreamFusion [Poole et al. 2022] employs a similar approach to Dream Fields [Jain et al. 2022] to train NeRF, using a frozen image-text joint embedding model from CLIP and an optimization-based approach, but replacing CLIP with a 2D Diffusion Model distilled loss. The DreamFusion architecture is shown in the Figure 14, where the scene is represented by a neural radiance field initialized and trained from scratch for each descriptive. With a pretrained text-to-image Diffusion Model, an image parameterized in the form of NeRF, and a loss function minimized towards good samples, DreamFusion has all the necessary components for text-to-3D synthesis without the use of 3D data. For each textual prompt, DreamFusion starts its training from scratch with a randomly initialized NeRF. Each iteration of DreamFusion's optimization performs the same step. For each optimization step, DreamFusion performs random sampling of camera and light sources, rendering of NeRF image from the camera with light occlusion, computation of SDS loss gradients relative to NeRF parameters, and updating of NeRF parameters with an optimizer. By combining SDS with a NeRF variant tailored to this 3D generation task, DreamFusion maximizes the fidelity and coherence of 3D generated shapes.

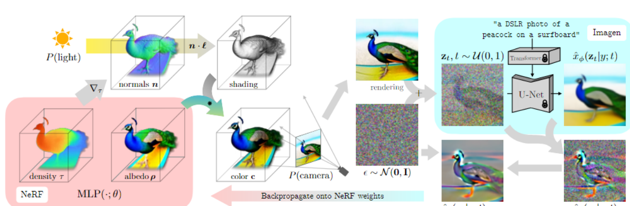

Fig. 14. Structure of DreamFusion, picture obtained from [Poole et al. 2022].

Magic3D [Lin et al. 2022] is a framework for high-quality 3D content composition with textual prompts, which optimizes the generation process by improving several major design choices from DreamFusion [Poole et al. 2022]. In DreamFusion, the monitoring signal runs on very low-resolution images of 64×64, and DreamFusion is unable to synthesize high-frequency 3D geometry and texture details. Because of its inefficient MLP architecture to represent NeRF, the actual high-resolution synthesis may even be impossible due to the rapid growth of memory usage and computation budget as the resolution increases. Magic3D proposes a two-stage optimization framework for optimizing from text to 3d synthesis results of NeRF. In the first stage, Magic3D optimizes a coarse neural field representation similar to DreamFusion, but with a hash grid-based memory and computation-efficient scene representation. In the second stage, Magic3D shifts to optimize the mesh representation, leveraging a diffusion prior at resolutions up to 512×512. Overview of Magic3D as shown in Figure 15. As 3D meshes fit well with fast graphics rendering solutions that can render high-resolution images in real-time, Magic3D also uses an efficient differentiable rasterizer to recover the high-frequency details in geometry and texture from camera close-ups. Magic3D synthesizes 3D contents with 8 times better resolution and 2 times faster speed than DreamFusion.

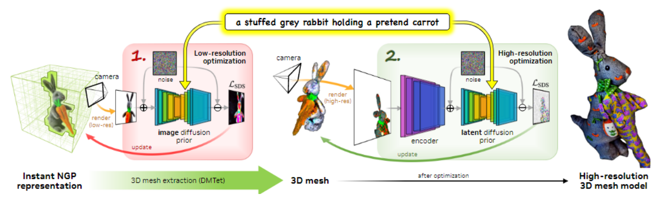

Fig. 15. Overview of Magic3D, picture obtained from [Lin et al. 2022].

Apart from the works showcased above, many other excellent contemporaneous works have served as pioneers in pioneering the text-to-3D generation. *CLIP-Mesh* [Mohammad Khalid et al. 2022] proposes a technique that can achieve zero-shot 3D model generation only using a target textual prompt. The motivation is that, without any 3D supervision, 3D resources corresponding to the input textual prompt can be obtained by adjusting the control shape of the restricted subdivision surface as well as its texture and normal maps, and can be easily deployed in games or modeling applications. *SJC* (Score Jacobian Chaining) [Wang et al. 2022b] proposes a method to promote 2D diffusion models to 3D by applying chain rules. The paper also pointed out the effectiveness of SJC in 3D text-driven generation tasks. *Dream3D* [Xu et al. 2022b] is the first attempt introducing explicit 3D shape priors into the CLIP-guided 3D optimization process. Specifically, Dream3D first generates high-quality 3D shapes as 3D shape priors in the text-to-shape phase, and then employs them as the initialization of neural radiance fields, and performs optimization with full hints. [Han et al. 2023] proposes a novel multi-class diffusion model for addressing the challenges in semantic-driven 3D shape generation. To solve the problems of single-class generation, low-frequency details, and the need of a large amount of paired data, the authors employ a pre-trained CLIP model to establish a bridge between text, 2D images, and 3D shapes, and apply a conditional flow model to generate shape vectors conditioned on the CLIP embeddings, as well as a latent diffusion model conditioned on multi-class shape vectors. Experimental results show that the proposed framework outperforms existing methods [Poole et al. 2022; Valsesia et al. 2019; Yang et al. 2019; Zhou et al. 2021a].

3.2.2 Followers of CLIP-Based Text-Guided 3D Shape Generation. Pioneers in the text-to-3D field have achieved new heights with emerging technologies, while also bringing new challenges, such as low textual 3D model matching, slow rendering speed, low resolution of generated 3D models, etc. Numerous research endeavors aim at addressing these issues, which are investigated in this section.

DITTO-NeRF [Seo et al. 2023b] is a novel pipeline that can generate high-quality 3D NeRF models from textual prompts or single images. It introduces a progressive 3D object reconstruction scheme, including scale, orientation and mask, which can propagate high-quality information from IB to OB. *Compared with the previous artworks from image/text to 3D such as DreamFusion [Poole et al. 2022] and NeuralLift-360 [Xu et al. 2022a], DITTO-NeRF achieves significant advantages in both quality and diversity, with shorter training time as well. 3D-CLFusion* [Li and Kitani 2023] presents a novel text-to-3D creation method which utilizes pre-trained latent variable-based NeRFs to rapidly complete 3D content creation within less than a minute. To tackle the challenges faced by NeRFs, 3D-CLFusion adopts a novel approach called view-invariant diffusion, which utilizes contrastive learning [He et al. 2020] to learn the latent variables and thus can generate high-resolution 3D content rapidly during the inference stage. *Experimental results demonstrate that 3D-CLFusion is up to 100 times faster than DreamFusion*, and can serve as a plug-and-play tool for text-to-3D with pre-trained NeRFs. *CLIP-Sculptor* [Sanghi et al. 2022] presents a 3D shape generation model under text conditions that improves shape diversity and fidelity solely with image-shape pairs as supervision, thus surpassing existing methods. The novelty of the CLIP-Sculptor lies in its multi-resolution, voxel-based conditioned generation scheme. Without text-shape pairs, CLIP-Sculptor learns to generate 3D shapes of common object categories from the joint embedding of CLIP's text-image. To achieve high-fidelity output, CLIP-Sculptor employs a multi-resolution approach. To generate diverse shapes, CLIP-Sculptor employs a discrete latent representation obtained with a vector quantization scheme. To further enhance shape fidelity and diversity, CLIP-Sculptor uses a mask transformer architecture.

3DFuse [Seo et al. 2023a] proposes a novel framework, incorporating 3D awareness into pre-trained 2D dispersion models to improve the robustness and 3D consistency of the 2D dispersion model-based approach. The view inconsistency problem in score-distilling text-to-3D generation, also known as *Janus problem* [Hong et al. 2023](as shown in Figure 16) is a big challenge that have to overcome. 3DFuse builds a rough 3D structure given the text prompt and then utilizes the projected specific depth map as the condition of the dispersion model. In addition, a training strategy is introduced to enable the 2D dispersion model to learn to handle the errors and sparsity in the rough 3D structure and a method to ensure semantic consistency among all viewpoints in the scene. Experimental results show that 3DFuse effectively solves the problem of 3D consistency and develops a new approach for 3D reconstruction with 2D dispersion models. In another work [Hong et al. 2023] proposes two de-biased methods to address the *Janus problem* in fractional distillation for 3D generation. These methods reduce artifacts and improve realism, while achieving a good balance between fidelity of 2D diffusion model and 3D consistency with low overhead.

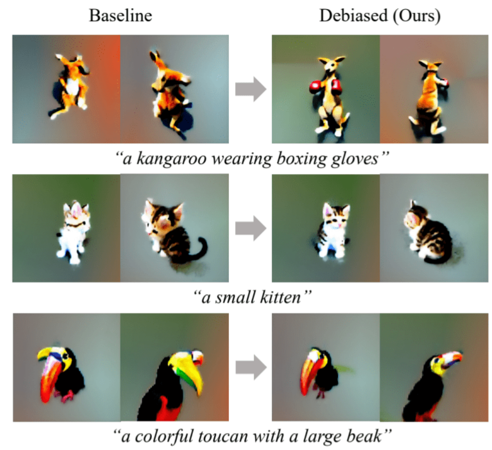

Fig. 16. Comparison of the model with *Janus problem* (left) [Wang et al. 2022b] and the upgraded model (right) [Hong et al. 2023], the picture is obtained from [Hong et al. 2023].

Although Text-to-3D can produce impressive results, it is essentially unconstrained and may lack the ability to guide or enforce 3D structure. *Latent-NeRF* [Metzer et al. 2022] incorporates both textual guidance and shape guidance for image generation and 3D model generation, as well as a latent-dispersal model for direct application of dispersed rendering on 3D meshes. *CompoNeRF* [Lin et al. 2023] proposes a novel framework that explicitly combines editable 3D scene layouts, providing effective guidance at both the local and global levels for NeRFs, to address the *Guidance collapse problem* faced in text-to-3D generation. CompoNeRF allows for flexible editing and recombination of trained local NeRFs into a new scene via manipulation of the 3D layout or textual hints, achieving faithful and editable text-to-3D results, as well as opening up potential directions for multi-object composition via editable 3D scene layouts guided by text.

DreamFusion generates volumetric representations instead of mesh representations, which makes it impractical in many downstream applications such as graphics, which require standard 3D representations such as meshes. *Text2Mesh* [Michel et al. 2022] presents a novel framework, which is capable of editing the style of 3D objects such as colors and local geometry details from textual descriptions. The framework uses a fixed mesh and a learned neural field to handle low-quality meshes without UV parameterization. Furthermore, Text2Mesh does not require any pre-trained generative models or specialized 3D mesh datasets. Thus, it can achieve style synthesis for various shapes of 3D meshes. *TextMesh* [Tsalicoglou et al. 2023] presents a new method for generating highly realistic 3D meshes from textual prompts, *which solves the problem of NeRF being infeasible for most practical applications.* To this end, the method extends NeRF to adopt a SDF framework, thus improving mesh extraction, and introduces a new method for fine-tuning mesh textures, eliminating over-saturation effects and enhancing the detail of the output 3D mesh. *Point·E* [Nichol et al. 2022] presents an alternative approach for fast 3D object generation which produces 3D models in only 1-2 minutes on a single GPU. The proposed method consists of two diffusion models, a text-to-image diffusion model and a point cloud diffusion model. Point·E's text-to-image model utilizes a large (text, image) corpus, allowing it to adhere to various complex cues. The image-to-3D model is trained on a smaller (image, 3D) pair dataset. To generate 3D objects from textual cues, Point·E first samples the images using its text-to-image model, followed by sampling 3D objects conditioned on the sampled images. Both steps can be completed within a few seconds and without the need for expensive optimization processes.

4 TEXT-TO-3D APPLICATIONS

With the emergence of text-to-3D models guided by text-to-image priors, more fine-grained application domains have been developed, including text-to-avatar, text-to-texture, text-to-scene, etc. This section surveys the text-guided 3D model generation models based on text-to-image priors.

4.1 Text Guided 3D Avatar Generation

In recent years, the creation of 3D graphical human models has drawn considerable attention due to its extensive applications in areas such as movie production, video gaming, AR/VR and human-computer interactions, and the creation of 3D avatars through natural language could save resources and holds great research prospects.

DreamAvatar [Cao et al. 2023] proposed a framework based on text and shape guidance for generating high-quality 3D human avatars with controllable poses. It utilizes a trainable NeRF to predict the density and color features of 3D points, as well as a pre-trained text-to-image diffusion model to provide 2D self-supervision. SMPL [Bogo et al. 2016] model is used to provide rough pose and shape guidance for generation, as well as a dual-space design, including a canonical space and an observation space, which are related by a learnable deformation field through NeRF, allowing optimized textures and geometries to be transferred from the canonical space to the target pose avatar with detailed geometry and textures. Experimental results demonstrate that DreamAvatar significantly outperforms the state of the art, setting a new technical level for 3D human generation based on text and shape guidance.

DreamFace [Zhang et al. 2023b] is a progressive scheme for personalized 3D facial generation guided by text. It enables ordinary users to naturally customize CG-pipe compatible 3D facial assets with desired shapes, textures and fine-grained animation capabilities. DreamFace introduces a coarse-to-fine scheme to generate a topologically unified neutral face geometry, utilizes Score Distillation Sampling (SDS) [Rombach et al. 2022] to optimize subtle translations and normals, adopts a dual-path mechanism to generate neutral appearance, and employs two-stage optimization to enhance compact priors for fine-grained synthesis, as well as to improve the animation capability for personalized deformation features. DreamFace can generate realistic 3D facial assets with physical rendering quality and rich animation capabilities from video materials, even for fashion icons, cartoons and fictional aliens in movies.

AvatarCraft [Jiang et al. 2023] utilizes a diffusion model to guide the learning of neural avatar geometry and texture based on a single text prompt, thereby addressing the challenge of creating 3D character avatars with specified identity and artistic style that can be easily animated. It also carefully designs an optimization framework of neural implicit fields, including coarse-to-fine multi-boundary box training strategy, shape regularization and diffusion-based constraints, to generate high-quality geometry and texture, and make the character avatars animatable, thus simplifying the animation and reshaping of the generated avatars. Experiments demonstrate the effectiveness and robustness of AvatarCraft in creating character avatars, rendering new views, poses, and shapes.

MotionCLIP [Tevet et al. 2022], a 3D human motion auto-encoder featuring a latent embedding that is disentangled, well behaved, and supports highly semantic textual descriptions. MotionCLIP is unique in that it aligns its latent space with that of the CLIP model, thus infusing the semantic knowledge of CLIP into the motion manifold. Furthermore, MotionCLIP leverages CLIP's visual understanding and self-supervised motion-to-frame alignment. The contributions of this paper are the text-to-motion capabilities it enables, out-of-domain actions, disentangled editing, and abstract language specification. In addition, MotionCLIP shows how the introduced latent space can be leveraged for motion interpolation, editing and recognition.

AvatarCLIP [Hong et al. 2022] introduces a text-driven framework for the production of 3D avatars and their motion generation. By utilizing the powerful vision-language model CLIP, AvatarCLIP enables non-expert users to craft customized 3D avatars with the shape and texture of their choice, and animate them with natural language instructions. Extensive experiments indicate that AvatarCLIP exhibits superior zero-shot performance in generating unseen avatars and novel animations.

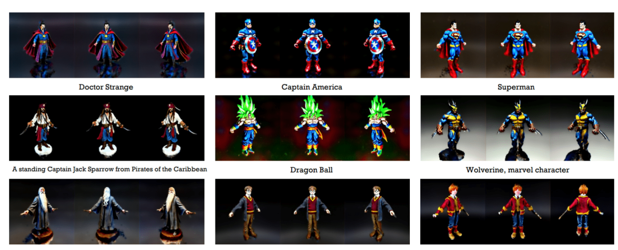

Fig. 17. 3D avatar created by text guided 3D generation model, picture obtained from [Cao et al. 2023]

4.2 Text Guided 3D Texture Generation

Recently, there have been a number of works on text-to-texture, inspired by text-to-3D. This summary lists these works.

TEXTure [Richardson et al. 2023] presents a novel text guided 3D shape texture generation, editing and transmission method. It utilizes a pre-trained deep-to-image topology model and iterative schemes to colorize the 3D models from different viewpoints, and proposes a novel detailed topology sampling procedure to generate seamless textures from different viewpoints using the three-step segmentation map. Additionally, it presents a method for transferring the generated texture maps to new 3D geometry without explicit surface-to-surface mapping and a method for extracting semantic textures from a set of images without any explicit reconstruction, and provides a way to edit and refine existing textures with text hints or user-provided doodle.

TANGO [Chen et al. 2022] proposes a novel method for programmatic rendering of realistic appearance effects on arbitrary topology given surface meshes. Based on CLIP model, the model is used to decompose the appearance style into spatially varying bidirectional reflectance distribution functions, local geometric variations, and illumination conditions. This enables realistic 3D style transfer through the automatic prediction of reflectance effects, even for bare, low-quality meshes, without the need for training on particular task-specific datasets. Numerous experiments demonstrate that TANGO outperforms the existing text-driven 3D style transfer methods in terms of realism, 3D geometry consistency, and robustness for stylizing low-quality meshes.

Fantasia3D [Chen et al. 2023a] presents a novel approach for text-to-high-quality 3D content creation. The method decouples geometry and appearance modeling and learning, and uses a hybrid scene representation and Spatially-varying Bidirectional Reflectance Distribution Function (BRDF) learning for surface material to achieve photorealistic rendering of the generated surface. Experimental results show that the method outperforms existing approaches [Lin et al. 2022; Poole et al. 2022] and supports physically plausible simulations of relit, edited, and generated 3D assets. *X-Mesh* [Ma et al. 2023] presents a novel text-driven 3D stylization framework, containing a novel text-guided dynamic attention module (TDAM) for more accurate attribute prediction and faster convergence speed. Additionally, a new standard text-mesh benchmark, MIT-30, and two automatic metrics standards are introduced for future research to achieve fair and objective comparison.

Text2Tex [Chen et al. 2023b] proposes a novel approach for generating high-quality textures for 3D meshes given text prompts. The goal of this method is to address the accumulation inconsistency and stretching artifacts in text-driven texture generation. The method integrates repair and merging into a pre-trained deep-perceptual image diffusion model to synthesize high-resolution local textures progressively from multiple perspectives. Experiments show that Text2Tex significantly outperforms existing text-driven and GAN-based methods.

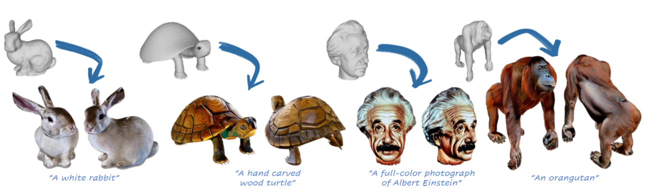

Fig. 18. Texturing results generated by text guided 3D texture model, picture obtained by [Richardson et al. 2023]

4.3 Text Guided 3D Scene Generation

3D scene modeling is a time-consuming task, usually requiring professional 3D designers to complete. To make 3D scene modeling easier, 3D generation should be simple and intuitive to operate while retaining enough controllability to meet users' precise requirements. Recent works [Cohen-Bar et al. 2023; Fridman et al. 2023; Höllein et al. 2023; Po and Wetzstein 2023] in text-to-3D generation has made 3D scene modeling easier.

Set-the-Scene [Cohen-Bar et al. 2023] proposes an agent-based global-local training framework for synthesizing 3D scenes, thus filling an important gap from controllable text to 3D synthesis. It can learn a complete representation of each object while also creating harmonious scenes with style and lighting matching. The framework allows various editing options such as adjusting the placement of each individual object, deleting objects from scenes, or refining objects.

[Po and Wetzstein 2023] proposes a local condition diffusion-based text-to-3D scene synthesis approach, which aims to make the generation of complex 3D scenes more intuitive and controllable. By providing control on the semantic parts via text hints and bounding boxes, the method ensures seamless transitions between these parts. Experiments show that the proposed method achieves higher fidelity in the composition of 3D scenes than the related baselines [Liu et al. 2022; Wang et al. 2022b].

SceneScape [Fridman et al. 2023] proposes a novel text-driven approach to generate permanent views, which is capable of synthesizing long videos of arbitrary scenes solely based on input texts describing the scene and camera positions. The framework combines the generative capacity of a pre-trained text-to-image model [Rombach et al. 2022] with the geometry priors learned from a pre-trained monocular depth prediction model [Ranftl et al. 2021, 2020], to generate videos in an online fashion, and achieves 3D consistency through online testing time training to generate videos with geometry-consistent scenes. Compared with the previous works that are limited to a restricted domain, this framework is able to generate various scenes including walking through a spaceship, a cave, or an ice city.

Text2Room [Höllein et al. 2023] proposes a method of generating room-scale 3D meshes with textures from given text prompts. It is the first method to generate attention-grabbing textured room-scale 3D geometries solely from text inputs, which is different from existing methods that focus on generating single objects [Lin et al. 2022; Poole et al. 2022] or scaling trajectories(SceneScape) [Fridman et al. 2023] from text.

Text2NeRF [Zhang et al. 2023a] proposes a text-driven realistic 3D scene generation framework combining diffusion model with NeRF representations to support zero-shot generation of various indoor/outdoor scenes from a variety of natural language prompts. Additionally, a progressive inpainting and updating (PIU) strategy is introduced to generate view-consistent novel contents for 3D scenes, and a support set is built to provide multi-view constraints for the NeRF model during view-by-view updating. Moreover, a depth loss is employed to achieve depth-aware NeRF optimization, and a two-stage depth alignment strategy is introduced to eliminate estimated depth misalignment in different views. Experimental results demonstrate that the proposed Text2NeRF outperforms existing methods [Höllein et al. 2023; Mohammad Khalid et al. 2022; Poole et al. 2022; Wang et al. 2022b] in producing photo-realistic, multiview consistent, and diverse 3D scenes from a variety of natural language prompts.

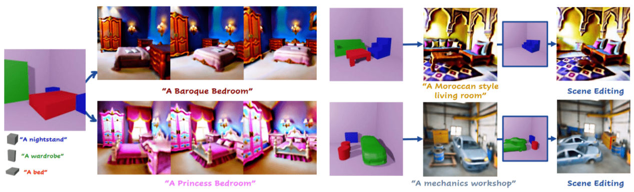

Fig. 19. Controllable scenes generation from text prompts, the picture is obtained by [Cohen-Bar et al. 2023]

MAV3D (Make-A-Video3D) [Singer et al. 2023] is a method for generating 3D dynamic scenes from text descriptions. The motivation of MAV3D is to provide a method for generating dynamic videos without 3D or 4D data, thus saving a lot of time and money. This method adopts 4D dynamic NeRF and optimizes the scene appearance, density, and motion consistency by querying the Text-to-Video (T2V) [Singer et al. 2022] model based on diffusion. Quantitative and qualitative experiments show that MAV3D improves over the internal baselines established previously. MAV3D is the first method to generate 3D dynamic scenes given text descriptions.

4.4 Text Guided 3D Shape Transformation

The traditional process of editing 3D models involves dedicated tools and years of training in order to manually carve, extrude, and re-texture given objects. This process is laborious and costly in terms of resources, and recently some research works have attempted to use text-guided 3D model editing, which is demonstrated in this section.

Instruct-NeRF2NeRF [Haque et al. 2023] presents a novel text-instructable editing of NeRF scenes, which uses an iterative image-based diffusion model (InstructPix2Pix) [Brooks et al. 2022] to edit the input image while optimizing the underlying scene, thereby generating an optimized 3D scene that follows the edits instructions. Experimental results show that the method is capable of editing large-scale real-world scenes, achieving more realistic and targeted edits than previous works.

Instruct 3D-to-3D [Kamata et al. 2023] presents a novel 3D-to-3D transformation method, which uses a pre-trained image-to-image diffusion model to achieve 3D-to-3D transformation. Furthermore, the method also proposes dynamic scaling, as well as explicitly conditioning on the input source 3D scene, to enhance 3D consistency and controllability. Quantitative and qualitative evaluations demonstrate that Instruct 3D-to-3D achieves higher-quality 3D-to-3D transformation than the baseline methods [Poole et al. 2022; Wang et al. 2022a].

SKED [Mikaeili et al. 2023] presents a sketch-based technique for editing 3D shapes represented by NeRF. The motivation is to introduce interactive editing into the text-to-3D pipeline, enabling users to edit from a more intuitive control perspective. Results show that SKED can effectively modify the existing neural fields and generate outputs that satisfy user sketches.

TextDeformer [Gao et al. 2023] proposes an automatic technique for generating input triangle mesh deformations guided entirely by text prompts. The framework is capable of generating large, low-frequency shape changes as well as small, high-frequency details, relying on differentiable rendering to connect geometry to powerful pre-trained image encoders such as CLIP[Radford et al. 2021] and DINO [Caron et al. 2021]. In order to overcome the problems of artifacts, TextDeformer proposes to use the Jacobian matrix to represent the mesh deformation and encourages deep features to be computed on 2D encoded rendering to ensure shape coherence from all 3D viewpoints. Experimental results show that the method can smoothly deform various source meshes and target text prompts to achieve large modifications and add details.

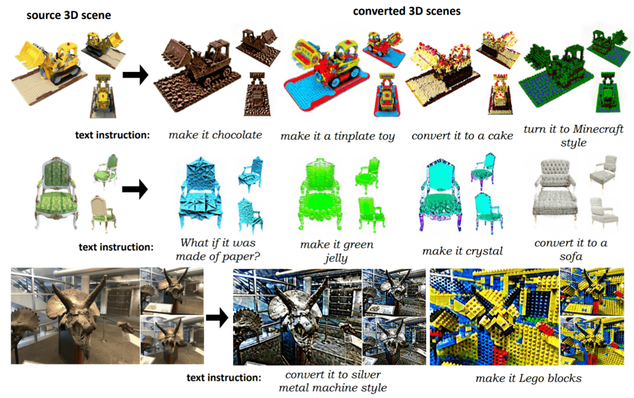

Fig. 20. Converting input 3D scenes according to the text instructions, the picture is obtained from [Kamata et al. 2023]

5 DISCUSSION

Combining NeRF and textual-to-image-prior textual-to-3D generation paradigm is an emerging research direction with the strong advantage of generating diverse content. However, there are also a lot of issues with it such as long inference time, 3D consistency issues, poor controllability and the generated content that cannot be applied well to industrial needs.

5.1 Fidelity

For 3D asset generation, constrained by the weakly supervised and low-resolution of CLIP, the upscaled results are not perfect. In addition, it is difficult to generate a more varied portrait under the same prompt. For the same prompt, CLIP text features are always the same. Fidelity and speed are two indicators often requiring trade-offs which often require the increase in inference speed to improve fidelity. At the same time, downstream application requirements should also be taken into account. Films require high-precision models, while games often require more quantity than film-level precision.

5.2 Inference velocity

A fatal issue of generating 3D content by leveraging pre-training models based on diffusion models as a powerful prior and learning objective is that the inference process is too slow. Even at a resolution of just 64×64, DreamFusion [Poole et al. 2022] would take 1.5 hours to infer for each prompt using a TPUv4. The inference time is further rising quickly along with the increase in resolution. This is mainly due to the fact that the inference process for generating 3D content is actually starting from scratch to train a Neural Radiance Field [Mildenhall et al. 2021]. Notably, NeRF models are renowned for their slow training and inference speeds, and training a deep network takes a lot of time. Magic3D [Lin et al. 2022] has addressed the time issue by using a two-phase optimization framework. Firstly, a coarse model is obtained by leveraging a low-resolution diffusion prior, and secondly, acceleration is performed by using a sparse 3D hash grid structure. 3D-CLFusion[Li and Kitani 2023] utilizes a pre-trained latent NeRf and performs a fast 3D content creation in less than a minute.

5.3 Consistency

Distortion and ghost images are often encountered in the 3D scenes generated by DreamFusion [Poole et al. 2022], and unstable 3D scenes are observed when text prompts and random seeds are changed. This issue is mainly caused by the lack of perception of 3D information from 2D prior diffusion models, as the transmission model has no knowledge of which direction the object is observed from, leading to the serious distortion of 3D scenes by generating the front-view geometry features from all viewpoints, including sides and the backs, which is usually referred to as the Janus problem. [Hong et al. 2023] proposed two debiasing methods to address such issues: score debiasing, which involves gradually increasing the truncation value of the 2D diffusion model's estimation throughout the entire optimization process; and prompt debiasing, which employs a language model to recognize the conflicts between the user prompts and the view prompts, and adjust the difference between the view prompts and the spatial camera pose of objects. 3DFuse [Seo et al. 2023a] optimized the training process to make the 2D diffusion model learn to process the wrong and sparse 3D structures for robust generation, as well as a way to ensure that the semantic consistency of all viewpoints in the scene is ensured.

5.4 Controllability

Although Text-to-3D can generate impressive results, as the text-to-image diffusion models are essentially unconstrained, they generally tend to suffer from guiding collapse. This makes them less capable of accurately associating object semantics with specific 3D structures. The issue of poor controllability has long been mainstream in the text-to-image generation task, with ControlNet [Zhang and Agrawala 2023] addressing it by adding extra input conditions to make the generation process even more controllable for large-scale text-to-image models, such as with the addition of canny edge, hough lines and depth maps. This unique combination of text and shape guidance allows for increased control over the generation process. LatentNeRF [Metzer et al. 2022] allows for increased control over the 3D generation process through its unique combination of text and shape guidance. CompoNeRF [Lin et al. 2023] is capable of precisely associating guidance with particular structures via its integration of editable 3D layouts and multiple local NeRFs, addressing the guidance failure issue when generating multiple object 3D scenes.

5.5 Applicability

Although NeRF, a novel 3D representation, cannot be directly applied to traditional 3D application scenarios, its powerful representation capabilities enable it to possess unlimited application prospects. The greatest advantage of NeRF is that it can be trained with 2D images. Google has already begun to use NeRFs to transform street map images into immersive views on Google Maps. In the future, NeRFs can supplement other technologies to more efficiently, accurately, and realistically represent 3D objects in the metaverse, augmented reality, and digital twins. To further improve these applications, future research may focus on extracting 3D meshes, point-clouds, or SDFs from the density MLP, and integrating faster NeRF models. It remains to be seen whether the Paradigm of Pure Text Guiding Shape Generation can cope with all scenarios. Perhaps incorporating a more intuitive guidance mechanism, such as sketch guidance or picture guidance, might be a more reasonable choice.

6 CONCLUSION

This work conducts the first yet comprehensive survey on text-to-3D. Specifically, we summarize text-to-3D from three aspects: data representations, technologies and applications. We hope this survey can help readers quickly understand the field of text-to-3D and inspire more future works to explore text-to-3D.

REFERENCES

Panos Achlioptas, Judy Fan, Robert Hawkins, Noah Goodman, and Leonidas J Guibas. 2019. ShapeGlot: Learning language for shape differentiation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8938–8947.

Parag Agarwal and Balakrishnan Prabhakaran. 2009. Robust blind watermarking of point-sampled geometry. *IEEE Transactions on Information Forensics and Security* 4, 1 (2009), 36–48.

Eman Ahmed, Alexandre Saint, Abd El Rahman Shabayek, Kseniya Cherenkova, Rig Das, Gleb Gusev, Djamila Aouada, and Bjorn Ottersten. 2018. A survey on deep learning advances on different 3D data representations. *arXiv preprint arXiv:1808.01462* (2018).

James F Blinn. 2005. What is a pixel? *IEEE computer graphics and applications* 25, 5 (2005), 82–87.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 561–578.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2022. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800* (2022).

Wenming Cao, Zhiyue Yan, Zhiquan He, and Zhihai He. 2020. A comprehensive survey on geometric deep learning. *IEEE Access* 8 (2020), 35929–35949.

Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. 2023. DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models. *arXiv preprint arXiv:2304.00916* (2023).

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.

Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. 2023b. Text2Tex: Text-driven Texture Synthesis via Diffusion Models. *arXiv preprint arXiv:2303.11396* (2023).

Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. 2019. 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*. Springer, 100–116.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023a. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. *arXiv preprint arXiv:2303.13873* (2023).

Yongwei Chen, Rui Chen, Jiabao Lei, Yabin Zhang, and Kui Jia. 2022. Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. *arXiv preprint arXiv:2210.11277* (2022).

Dana Cohen-Bar, Elad Richardson, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. 2023. Set-the-Scene: Global-Local Training for Generating Controllable NeRF Scenes. *arXiv preprint arXiv:2303.13450* (2023).

Yinpeng Dong, Shouwei Ruan, Hang Su, Caixin Kang, Xingxing Wei, and Jun Zhu. 2022. ViewFool: Evaluating the Robustness of Visual Recognition to Adversarial Viewpoints. *arXiv preprint arXiv:2210.03895* (2022).

Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. 2023. Scenescape: Text-driven consistent scene generation. *arXiv preprint arXiv:2302.01133* (2023).

Rao Fu, Xiao Zhan, Yiwen Chen, Daniel Ritchie, and Srinath Sridhar. 2022. Shapecrafter: A recursive text-conditioned 3d shape generation model. *arXiv preprint arXiv:2207.09446* (2022).

Yasutaka Furukawa, Carlos Hernández, et al. 2015. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision* 9, 1-2 (2015), 1–148.

Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. 2022b. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems* 35 (2022), 31841–31854.

Kyle Gao, Yina Gao, Hongjie He, Denning Lu, Linlin Xu, and Jonathan Li. 2022a. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379* (2022).

William Gao, Noam Aigerman, Thibault Groueix, Vladimir G Kim, and Rana Hanocka. 2023. TextDeformer: Geometry Manipulation using Text Guidance. *arXiv preprint arXiv:2304.13348* (2023).

Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. 2020. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence* 43, 12 (2020), 4338–4364.

Bo Han, Yitong Liu, and Yixuan Shen. 2023. Zero3D: Semantic-Driven Multi-Category 3D Shape Generation. *arXiv preprint arXiv:2301.13591* (2023).

Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. 2019. Meshcnn: a network with an edge. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–12.

Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. 2023. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. *arXiv preprint arXiv:2303.12789* (2023).

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. 2023. Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models. *arXiv preprint arXiv:2303.11989* (2023).

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

Susung Hong, Donghoon Ahn, and Seungryong Kim. 2023. Debiasing Scores and Prompts of 2D Diffusion for Robust Text-to-3D Generation. *arXiv preprint arXiv:2303.15413* (2023).

Dongting Hu, Zhenkai Zhang, Tingbo Hou, Tongliang Liu, Huan Fu, and Mingming Gong. 2023. Multiscale Representation for Real-Time Anti-Aliasing Neural Rendering. *arXiv preprint arXiv:2304.10075* (2023).

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

Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2023. AvatarCraft: Transforming Text into Neural Human Avatars with Parameterized Shape and Pose Control. *arXiv preprint arXiv:2303.17606* (2023).

Yiwei Jin, Diqiong Jiang, and Ming Cai. 2020. 3d reconstruction using deep learning: a survey. *Communications in Information and Systems* 20, 4 (2020), 389–413.

Hiromichi Kamata, Yuiko Sakuma, Akio Hayakawa, Masato Ishii, and Takuya Narihira. 2023. Instruct 3D-to-3D: Text Instruction Guided 3D-to-3D conversion. *arXiv preprint arXiv:2303.15780* (2023).

Yu-Jhe Li and Kris Kitani. 2023. 3D-CLFusion: Fast Text-to-3D Rendering with Contrastive Latent Diffusion. *arXiv preprint arXiv:2303.11938* (2023).

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2022. Magic3D: High-Resolution Text-to-3D Content Creation. *arXiv preprint arXiv:2211.10440* (2022).

Yiqi Lin, Haotian Bai, Sijia Li, Haonan Lu, Xiaodong Lin, Hui Xiong, and Lin Wang. 2023. CompoNeRF: Text-guided Multi-object Compositional NeRF with Editable 3D Scene Layout. *arXiv preprint arXiv:2303.13843* (2023).

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*. Springer, 423–439.

Weiping Liu, Jia Sun, Wanyi Li, Ting Hu, and Peng Wang. 2019. Deep learning on point clouds and its application: A survey. *Sensors* 19, 19 (2019), 4188.

Yiwei Ma, Xiaioqing Zhang, Xiaoshuai Sun, Jiayi Ji, Haowei Wang, Guannan Jiang, Weilin Zhuang, and Rongrong Ji. 2023. X-Mesh: Towards Fast and Accurate Text-driven 3D Stylization via Dynamic Textual Guidance. *arXiv preprint arXiv:2303.15764* (2023).

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

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. 2022. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. *arXiv preprint arXiv:2211.07600* (2022).

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

Aryan Mikaeili, Or Perel, Daniel Cohen-Or, and Ali Mahdavi-Amiri. 2023. SKED: Sketch-guided Text-based 3D Editing. *arXiv preprint arXiv:2303.10735* (2023).

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Ludovico Minto, Pietro Zanuttigh, and Giampaolo Pagnutti. 2018. Deep Learning for 3D Shape Classification based on Volumetric Density and Surface Approximation Clues.. In *VISIGRAPP (5: VISAPP)*. 317–324.

Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. 2022. CLIP-Mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 Conference Papers*. 1–8.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. *arXiv preprint arXiv:2212.08751* (2022).

Jaesik Park, Sudipta N Sinha, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. 2016. Robust multiview photometric stereo using planar mesh parameterization. *IEEE transactions on pattern analysis and machine intelligence* 39, 8 (2016), 1591–1604.

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

Lim Wen Peng and Siti Mariyam Shamsuddin. 2004. 3D object reconstruction and representation using neural networks. In *Proceedings of the 2nd international conference on Computer graphics and interactive techniques in Australasia and South East Asia*. 139–147.

Ryan Po and Gordon Wetzstein. 2023. Compositional 3D Scene Generation using Locally Conditioned Diffusion. *arXiv preprint arXiv:2303.12218* (2023).

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 652–660.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12179–12188.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence* 44, 3 (2020), 1623–1637.

Konstantinos Rematas and Vittorio Ferrari. 2020. Neural voxel renderer: Learning an accurate and controllable rendering tool. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5417–5427.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

Alessandro Rossi, Marco Barbiero, Paolo Scremin, and Ruggero Carli. 2021. Robust Visibility Surface Determination in Object Space via Plücker Coordinates. *Journal of Imaging* 7, 6 (2021), 96.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

Aditya Sanghi, Rao Fu, Vivian Liu, Karl Willis, Hooman Shayani, AmirHosein Khasahmadi, Srinath Sridhar, and Daniel Ritchie. 2022. CLIP-Sculptor: Zero-Shot Generation of High-Fidelity and Diverse Shapes from Natural Language. *arXiv preprint arXiv:2211.01427* (Nov 2022).

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402* (2022).

Hoigi Seo, Hayeon Kim, Gwanghyun Kim, and Se Young Chun. 2023b. DITTO-NeRF: Diffusion-based Iterative Text To Omni-directional 3D Model. *arXiv preprint arXiv:2304.02827* (2023).

Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. 2023a. Let 2D Diffusion Model Know 3D-Consistency for Robust Text-to-3D Generation. *arXiv preprint arXiv:2303.07937* (2023).

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 6087–6101.

Zifan Shi, Sida Peng, Yinghao Xu, Yiyi Liao, and Yujun Shen. 2022. Deep generative models on 3d representations: A survey. *arXiv preprint arXiv:2210.15663* (2022).

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).

Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. 2023. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280* (2023).

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*. Springer, 358–374.

Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. 2023. TextMesh: Generation of Realistic 3D Meshes From Text Prompts. *arXiv preprint arXiv:2304.12439* (2023).

Diego Valsesia, Giulia Fracastoro, and Enrico Magli. 2019. Learning localized generative models for 3d point clouds via graph convolution. In *International conference on learning representations*.

Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2022a. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3835–3844.

Cheng Wang, Ming Cheng, Ferdous Sohel, Mohammed Bennamoun, and Jonathan Li. 2019. NormalNet: A voxel-based CNN for 3D object classification and retrieval. *Neurocomputing* 323 (2019), 139–147.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. 2022b. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. *arXiv preprint arXiv:2212.00774* (2022).

He Wang and Juyong Zhang. 2022. A Survey of Deep Learning-Based Mesh Processing. *Communications in Mathematics and Statistics* 10, 1 (2022), 163–194.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* 32, 1 (2020), 4–24.

Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. 2022a. NeuralLift-360: Lifting An In-the-wild 2D Photo to A 3D Object with 360° Views. *arXiv e-prints* (2022), arXiv–2211.

Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. 2022b. Dream3D: Zero-Shot Text-to-3D Synthesis Using 3D Shape Prior and Text-to-Image Diffusion Models. *arXiv preprint arXiv:2212.14704* (2022).

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. 2019. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4541–4550.

Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. 2023. MVImgNet: A Large-scale Dataset of Multi-view Images. *arXiv preprint arXiv:2303.06042* (2023).

Chaoning Zhang, Chenshuang Zhang, Chenghao Li, Yu Qiao, Sheng Zheng, Sumit Kumar Dam, Mengchun Zhang, Jung Uk Kim, Seong Tae Kim, Jinwoo Choi, et al. 2023c. One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era. *arXiv preprint arXiv:2304.06488* (2023).

Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. 2023d. Text-to-image Diffusion Model in Generative AI: A Survey. *arXiv preprint arXiv:2303.07909* (2023).

Chaoning Zhang, Chenshuang Zhang, Sheng Zheng, Yu Qiao, Chenghao Li, Mengchun Zhang, Sumit Kumar Dam, Chu Myaet Thwal, Ye Lin Tun, Le Luang Huy, et al. 2023e. A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need? *arXiv preprint arXiv:2303.11717* (2023).

Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. 2023a. Text2NeRF: Text-Driven 3D Scene Generation with Neural Radiance Fields. *arXiv preprint arXiv:2305.11588* (2023).

Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023).

Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu. 2023b. DreamFace: Progressive Generation of Animatable 3D Faces under Text Guidance. *arXiv preprint arXiv:2304.03117* (2023).

Hang Zhou, Weiming Zhang, Kejiang Chen, Weixiang Li, and Nenghai Yu. 2021b. Three-dimensional mesh steganography and steganalysis: a review. *IEEE Transactions on Visualization and Computer Graphics* (2021).

Linqi Zhou, Yilun Du, and Jiajun Wu. 2021a. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5826–5835.