

Text-to-3D Shape Generation

H. Lee¹ M. Savva¹ A. X. Chang^{1,2}

¹Simon Fraser University ²Canada CIFAR AI Chair, Amii

3dlg-hcvc.github.io/tt3dstar

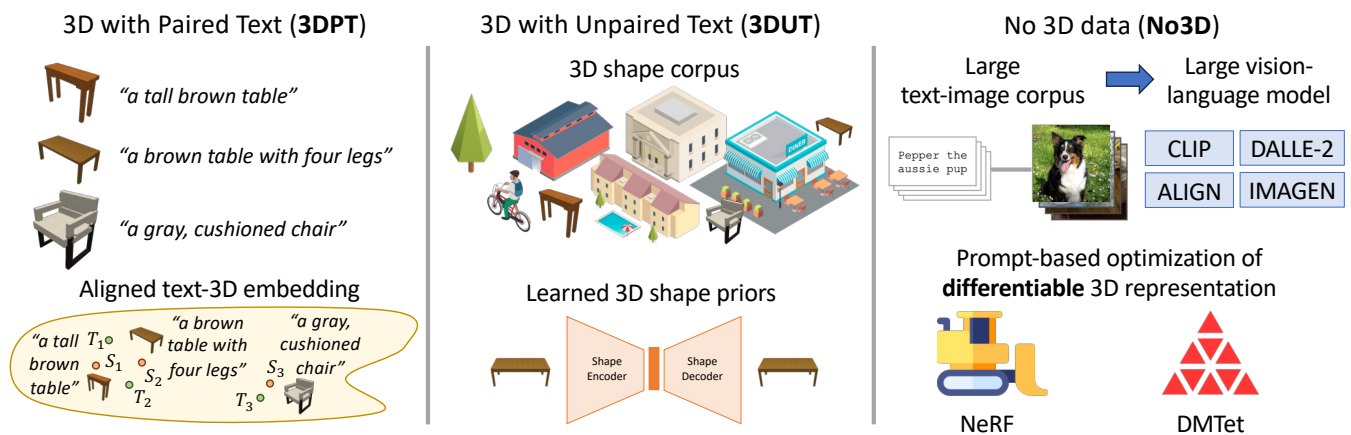


Figure 1: We survey and categorize methods for text-to-3D shape generation. Our categorization organizes methods into three families delineated by the use of 3D and text data. The first and second family rely on 3D data in combination with either paired text descriptions (3DPT) or with no text-3D pairing (3DUT). The third family does not rely on 3D data for training (No3D).

Abstract

Recent years have seen an explosion of work and interest in text-to-3D shape generation. Much of the progress is driven by advances in 3D representations, large-scale pretraining and representation learning for text and image data enabling generative AI models, and differentiable rendering. Computational systems that can perform text-to-3D shape generation have captivated the popular imagination as they enable non-expert users to easily create 3D content directly from text. However, there are still many limitations and challenges remaining in this problem space. In this state-of-the-art report, we provide a survey of the underlying technology and methods enabling text-to-3D shape generation to summarize the background literature. We then derive a systematic categorization of recent work on text-to-3D shape generation based on the type of supervision data required. Finally, we discuss limitations of the existing categories of methods, and delineate promising directions for future work.

1 Introduction

Text to 3D shape generation methods can revolutionize 3D content creation by allowing anyone to generate 3D content based on a simple text description. It is no wonder that there has been an explosion of interest in this research direction. Recent advances in generative models for text and images [RDN*22; SCS*22] enabled by large-scale language and vision-language models, as well as advances in learned 3D representations and 3D generative models have acted as catalysts for progress in text to 3D shape generation.

At the same time as this rapid progress, a number of open research challenges are coming into focus. There is currently a spar-

sity of available 3D data paired with natural language text descriptions, making it infeasible to rely purely on direct supervision from data pairs in both domains. Moreover, current text to 3D generation methods do not afford natural editability of the generated outputs in an intuitive way based on user inputs. Generation of larger-scale 3D outputs representing compositions of objects into natural scenes also remains challenging. Lastly, the complexity of underlying 3D generative models coupled with the complexity of the optimization problem when avoiding reliance on paired text and 3D data lead to a challenging learning problem with significant compute and training time requirements.

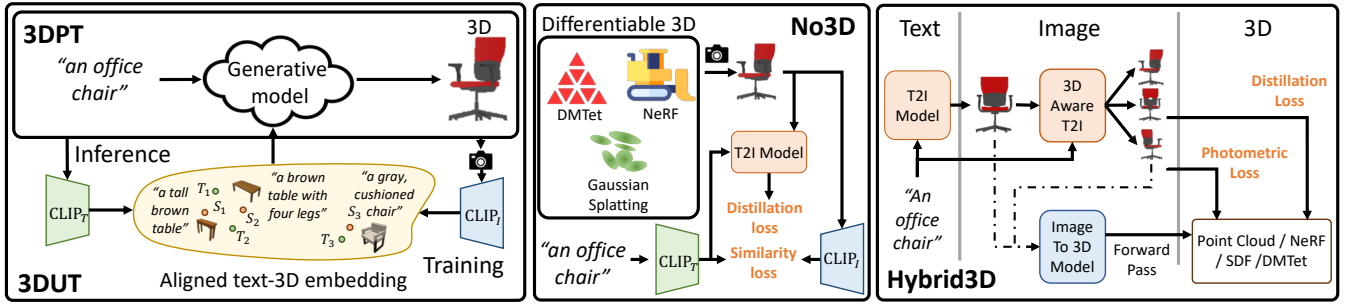


Figure 2: We illustrate the main components for the different families of methods in this survey. 3DPT methods involve training a generative model on paired text and 3D data. In 3DUT, data is limited to only 3D shapes, so methods in this group leverage the aligned text-3D embedding space of CLIP encoders allowing conditioning on rendered images during training and text prompts during inference. In No3D, no data is available so methods rely on pretrained guidance models such as CLIP and T2I diffusion models to optimize similarity and distillation respectively with differentiable 3D representations. Finally, Hybrid3D methods often combine techniques from No3D with priors learned from 3D assets to train 3D-aware T2I models or to enable text-to-image-to-3D using pretrained large text-to-image models.

There have been some recent related surveys addressing 3D generative models [CRW*20; SPX*22] and text-conditioned generation [CG23; LZW*23]. The former group of surveys on 3D generative models addresses what is typically one component of a complete text-to-3D shape system. Chao and Gingold [CG23] cover both text-to-image and text-to-3D generation and do not focus on a detailed categorization and discussion of 3D generation methods. Li et al. [LZW*23] do focus on text-to-3D generation specifically but address mainly application domains and do not derive a comprehensive categorization of the methods themselves, or pursue a systematic discussion of the method design decisions.

In this survey, we systematically summarize work on text-to-3D shape generation on the basis of four key components: training data, 3D representation type, generative model, and training setup. For the first, we further categorize by the amount and type of 3D data that is needed for supervision. Specifically, we categorize methods based on whether they: 1) use 3D data to train a generative model (Section 4); or 2) tackle text-to-3D without any 3D data (Section 5). Figure 1 and Table 1 provide a summary of this categorization. We note that overall techniques in 1) are similar to text-to-image methods, with the key differences being: a) the 3D generative model; and b) the link between text and 3D. For methods in 1), we further breakdown between methods that are supervised with paired text-3D data (Section 4.1) and methods that do not rely on paired data (Section 4.2). As other surveys have focused purely on 3D generative models, in this survey we focus particularly on methods for text-to-3D without any 3D data, and how the link between text and 3D can be established without paired text to 3D data (typically leveraging 2D image information). In addition, we discuss some of the key challenges of generating high quality 3D shapes without explicit 3D training data, and strategies that are employed to tackle these challenges. Recent work has started to address these challenges by learning shape priors from large 3D datasets, often coupled with techniques introduced for text-to-3D without 3D data. We describe these ‘hybrid’ approaches in Section 6.

2 Scope of this Survey

2.1 Comparison with Related Surveys

Shi et al. [SPX*22] offers a comprehensive review of generative models for 3D data. They primarily concentrate on various 3D representations, including voxels, point clouds, neural fields, and meshes, and delve into different generative model types like GANs, VAEs, and energy-based models used by various works. While some papers that use text for conditional 3D generation are mentioned, it is not the main focus of their survey.

An earlier survey Chaudhuri et al. [CRW*20] in comparison emphasizes structure-aware generative models for 3D shapes and scenes. They focus on approaches that decompose 3D objects or scenes into smaller elements and introduces overarching structures using representations like hierarchies and programs. They showcase various generative models applied to these structures, ranging from classical probabilistic models to deep learning techniques.

Chao and Gingold [CG23] present a survey on text-conditioned editing, encompassing both 2D and 3D methodologies. Though the emphasis is on editing, the survey details the key text-to-3D generation techniques that underpin these editing methods. Notably, the study predominantly highlights works that leverage CLIP for shape generation and editing. However, recent advancements using diffusion distillation methods remain unaddressed.

A recent review by Li et al. [LZW*23] examines recent text-to-3D research, notably those incorporating CLIP and diffusion models. Although there is some overlap in the work discussed, our survey offers a more structured categorization for easier comparison and provides more in-depth examination of each work.

2.2 Organization of this Survey

As summarized in Table 1, our survey classifies recent text-to-3D research into three primary families: 1) Methods that utilize paired text with 3D data (3DPT), discussed in Section 4.1; 2) Methods reliant on 3D data but not requiring paired 3D and text data

3D Paired Text (3DPT): Requires paired text-3D data which is limited. Generation limited to observed data.

3D Unpaired Text (3DUT): Leverages 3D data to train 3D generative model. Bridges text and 3D using images. Can use vision-language models to generate captions for 3D data, reducing to “Paired” scenario.

No 3D Data (NO3D): No 3D data for training. Multi-view and structure consistency is an issue. Uses images as bridge, typically with differentiable rendering. Conceptually can generate arbitrary 3D content. Per-prompt optimization, slow.

Hybrid3D: Combine text-to-image and image-to-3D methods. Enforce 3D consistency using 3D-aware text-to-image models or multi-view images.

Table 1: Properties of the four families into which we categorize text-to-3D shape generation methods.

(3DUT), discussed in Section 4.2; 3) Methods that require no 3D data training data at all, and typically use CLIP or Diffusion models as guidance (NO3D), discussed in Section 5; and 4) Recent work that leverage a combination of text-to-image and image-to-3D, discussed in Section 6.

The main focus of this survey is primarily on the third family of works as they have not been addressed in detail by prior surveys. For these NO3D approaches we further divide into sections that discuss methods that: 1) leverage pre-trained text-image embeddings (Section 5.1) 2) formulate or improve upon ways to use diffusion models as a prior (Section 5.2); 3) use different 3D representations, rendering techniques, or improvements to the training setup to enhance the quality of the results. Figure 2 shows an overview of different methods covered in this survey.

We then discuss emerging work on generation of multi-object 3D scenes (Section 7), and on allowing editing of the output 3D shape in various ways (Section 8). The following section of the survey presents a brief overview of evaluation methods for text-to-3D shape generation (Section 9). We conclude the survey with a discussion of promising future directions (Section 10).

3 Preliminaries

In this section, we provide a brief background of fundamentals used in text-to-3D generation methods. In particular, we summarize choices for 3D representations (Section 3.1), the necessary background on deep generative models (Section 3.2), and guidance models for connecting 3D representations with text via images (Section 3.3).

3.1 3D Representations

There are a variety of representations that are possible when working with 3D data, each with specific properties and challenges. In their survey of 3D generative models, Shi et al. [SPX*22] include a good overview of the different choices. Here we provide a brief summary, with a focus on how appearance (color or texture) is encoded in 3D models as these attributes are commonly included in text descriptions. We also describe in detail neural radiance fields (NeRFs) and DMTet, as these are two popular representations used in text-to-3D generation.

Explicit Representations. In traditional computer graphics

pipelines, 3D shapes are predominantly represented as polygonal meshes (most commonly triangle meshes). Meshes encode both the geometry through the spatial coordinates of vertices on the shape surface, as well as the topology connecting those vertices. Material properties of the shape’s surface such as albedo color, specularity, reflectance etc. can be represented through values at the vertices or more densely through a texture map that is mapped onto the surface, typically through the use of barycentric coordinate-based interpolation on the triangles constituting the mesh. Given a 3D surface mesh representation, generation of a point cloud is possible through sampling of the mesh surface, where sampled point coordinates are in \mathbb{R}^3 , and additional information such as the surface normal vector, surface color value etc. can also be extracted. Through the process of voxelization, it is also possible to generate voxel grid representations of occupancy, color, or signed distance from the nearest surface. Point clouds and dense voxel grid representations are common choices for encoding 3D shape structure in neural architectures. In contrast, triangle meshes are the dominant representation for exchange and storage in 3D content databases, rendering, and other computer graphics pipeline operations.

NeRF [MST*20]. The Neural Radiance Field (NeRF) represents a 3D scene as a continuous function $\sigma, c = \text{NeRF}(xyz, \mathbf{d})$, where xyz is a query point within the 3D scene, \mathbf{d} consists of the camera viewing direction and σ, c are the density and color. To render a pixel from a ray r the following equation is used:

$$\hat{C}(r) = \sum_{i=1}^K Tr_i \alpha_i c_i, \text{ where } \alpha_i = 1 - \exp(-\sigma_i \delta_i), \text{ and } Tr_i = \prod_{j=1}^{i-1} (1 - \alpha_j) \quad (1)$$

Here Tr_i is the transmittance, α_i is the alpha compositing value, δ_i is the length between neighboring samples on the ray and $\hat{C}(r)$ is the rendered color for ray r . Here we mention some notable works that are used by methods in this survey. Voxel grid NeRF models [SSC22; CXG*22; FYT*22; KRWM22] store learnable explicit/implicit features in a voxel grid structure for fast learning due to the smaller or outright removal of the MLP predicting density and color. InstantNGP [MESK22] uses multi-resolution hash tables for storing learnable features. Mip-NeRF [BMT*21] uses conical frustums as opposed to rays for more accurate rendering. Mip-NeRF 360 [BMV*22] further add changes to the scene parameterization for better reconstruction of unbounded scenes. Xie et al. [XTS*22] offers a survey of neural fields and their applications.

DMTet [SGY*21]. DMTet uses a hybrid representation scheme with an explicit tetrahedral grid [GCX*20] with signed distance field (SDF) values at each vertex to implicitly represent the underlying isosurface. Adopting their notation, the tetrahedral grid T containing the vertices V_T form (V_T, T) . The tetrahedrons in the grid $T_k \in T$ consists of four vertices $\{v_{a_k}, v_{b_k}, v_{c_k}, v_{d_k}\}$, where there are K tetrahedrons in total. Given the SDF values at each vertex $\{SDF(v_{a_k}), SDF(v_{b_k}), SDF(v_{c_k}), SDF(v_{d_k})\}$ we can calculate the zero crossing via interpolation on edges where sign changes occur. The surface within the tetrahedra can then be extracted, this process is referred to as Marching Tetrahedra. Deformation vectors are also encoded for each vertex, this allows for fitting 3D objects with even finer detail. As we will see for works using DMTet for text to 3D, the SDF and deformation values can either be optimized explicitly or predicted from a neural network. Along with color information

for each vertex the extracted mesh is often rendered with a differentiable rasterizer [LHK*20] for optimization.

3D Gaussians. Kerbl et al. [KKLD23] introduced blended 3D Gaussians for representing scenes. Each Gaussian is parameterized by a position, scaling vector and, rotation as a unit quaternion. The position corresponds to the mean of the Gaussian, and the covariance matrix is formed from the scale and rotation. Despite its simplicity, this representation has been shown to be effective at accurately and efficiently modeling complex scenes and has been adopted by more recent work [TRZ*23; YFW*23; CWL23].

Structured Representations. It is also possible to represent a 3D shape as a combination of primitive parts. The parts can be grouped in a hierarchical manner, or generated via a program. This constitutes a promising direction for future work on text-to-3D generation: leveraging structured representations as used by neurosymbolic methods (see Ritchie et al. [RGJ*23] for a recent survey).

3.2 Deep Generative Models

A goal of text-to-3D shape generation is to generate potentially a variety of different shapes given a single text query. Thus, a fundamental component in text-to-3D shape generation is the use of a generative model. A generative model models the joint probability distribution and can be sampled from to obtain diverse samples. In text-to-3D generation, the output space can be a 3D representation, a latent vector that is then transformed into a 3D representation via a shape decoder, or a 2D image that is used to generate a 3D shape. In addition to the above, it is also possible to use a text-to-image generative model to guide the optimization of a 3D representation. There are a number of popular families of generative models with different properties including auto-regressive models, GANs, VAEs, normalizing flow models, and diffusion models. For a more detailed discussion Shi et al. [SPX*22] provide an excellent overview of the types of generative models and how they can be used in generation of 3D output. Here, we elaborate on diffusion models since they are an important component of recent work on text-to-3D generation that does not require 3D data.

Diffusion Models. Sohl-Dickstein et al. [SWMG15] and Ho et al. [HJA20] are generative models that model the distribution of the training data $x_0 \sim q(x_0)$ by integrating latent variables $x_{1:T}$ by $p_\phi(x_0) = \int p_\phi(x_{0:T}) dx_{1:T}$. Here we follow the notation used by DDPM [HJA20]. The distribution of the latent variables is defined through a forward and reverse process using Markov Chains. The forward process involves corrupting a data sample x_0 at each time step according to a noise schedule until x_T is just Gaussian noise $x_T \sim \mathcal{N}(0, \mathbf{I})$. The formal definition is as follows:

$$q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1}) \quad (2)$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I}) \quad (3)$$

Here β_t is a predefined variance schedule. Note that we can obtain the latent variable x_t from x_0 by sampling $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0, (1 - \alpha_t)\mathbf{I})$, where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. The

reverse process can similarly be defined through Gaussian transitions:

$$p_\phi(x_{0:T}) = p(x_T) \prod_{t=1}^T p(x_{t-1}|x_t) \quad (4)$$

$$p_\phi(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\phi(x_t, t), \Sigma_\phi(x_t, t)) \quad (5)$$

In DDPM, by setting $\Sigma_\phi(x_t, t) = \sigma_t^2 \mathbf{I}$ and $\mu_\phi(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\phi(x_t, t))$, where ϵ_ϕ is a neural network used to predict the unit variance noise ϵ used to corrupt x_0 . We can supervise the training with a simplified noise residual objective:

$$\mathbb{E}_{t, x_0, \epsilon} [w_t \|\epsilon - \epsilon_\phi(\sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon, t)\|^2] \quad (6)$$

where w_t is a weighting term. The neural network ϵ_ϕ is usually implemented as a UNet [RFB15]. After training the network parameters ϕ with the above loss, we can generate samples by first sampling $x_T \sim \mathcal{N}(0, \mathbf{I})$ then using the reverse process to get x_0 . Note another parameterization of this loss in Karras et al. [KAAL22]:

$$\mathbb{E}_{\sigma_t, \mathbf{y}, \mathbf{n}} [\lambda(\sigma_t) \|D(\mathbf{y} + \mathbf{n}; \sigma_t) - \mathbf{y}\|_2^2] \quad (7)$$

where the denoiser is defined as $D(x; \sigma_t) = x - \sigma_t \epsilon_\phi(x, t)$, \mathbf{y} is a sampled point from the training data and $\mathbf{n} \sim \mathcal{N}(0, \sigma_t^2 \mathbf{I})$ is the noise. We encourage readers to read Karras et al. [KAAL22] for more details as it brings recent diffusion methods under the same framework and discuss different hyperparameters choices.

3.3 Guidance Models

CLIP [RKH*21]. The CLIP model consists of a text encoder $\text{Enc}_{\text{CLIP}}^T$ and image encoder $\text{Enc}_{\text{CLIP}}^I$ which project text and images onto an aligned latent space trained by minimizing the contrastive loss on 400 million image text pairs. The text and image embeddings are obtained through $e_{\text{CLIP}}^T = \text{Enc}_{\text{CLIP}}^T(x^T)$ and $e_{\text{CLIP}}^I = \text{Enc}_{\text{CLIP}}^I(x^I)$, where x^T and x^I are the input text and image respectively and $e_{\text{CLIP}} \in R^c$ is the embedding in the aligned latent space. Finally, we can calculate the similarity between a pair of image and text as $e_{\text{CLIP}}^T \cdot e_{\text{CLIP}}^I / \|e_{\text{CLIP}}^T\| \|e_{\text{CLIP}}^I\|$.

T2I Diffusion Models. Here we briefly discuss text to image (T2I) diffusion models used by the following works in this survey to provide a 2D prior for generating 3D objects. Earlier works [SWMG15; SSK*21; DN21] uses an additional classifier which is gradient conditioned on the noisy image and label to guide the sampling process. More recent T2I diffusion models use classifier-free guidance [HS22] which is formulated as:

$$\hat{\epsilon}_\phi(x_t, c) = (1 + \omega) \epsilon_\phi(x_t; t, c) - \omega \epsilon_\phi(x_t; t, \emptyset) \quad (8)$$

Here c is the conditioning variable, which is the input text in the case of T2I models. The predicted noise at each step involves the conditional and unconditional scores with weighting ω and \emptyset is a null prompt. The text conditioning can be incorporated through cross attention layers in the UNet network [RBL*22; SCS*22; BNH*22]. eDiff-I [BNH*22] and Imagen [SCS*22] follow a cascaded approach where the diffusion model is trained at a lower resolution, with cascading super resolution models to upsample the produced images. Stable Diffusion [RBL*22] uses a latent diffusion model (LDM) architecture that first trains to compress images

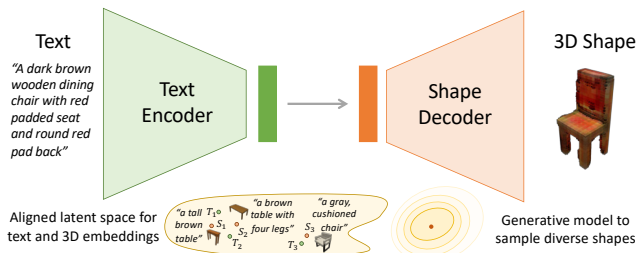


Figure 3: Illustration of key components in text to 3D generation models. A text encoder takes the text and produces an embedding in a latent space that is used to condition a generative model to sample a latent shape code. The latent shape code is then passed through a shape decoder to generate a 3D shape. Typically, the text and 3D latent spaces are aligned so the text embedding can be directly used as the shape embedding. The generative model is used to allow for sampling of diverse shapes. Text-to-3D models are characterized by the choice of text encoder, shape decoder, shape representation, text-3D alignment method, and how the generative model is integrated to allow sampling of diverse shapes.

to a lower resolution latent space with an autoencoder. Then, this approach fits a diffusion model in the latent space learned in the first stage. Due to its open source nature, Stable Diffusion is a popular choice used by the majority of works.

4 Text-to-3D using 3D Data

At a high-level, text-to-3D methods require the following components (see Figure 3):

- *Text encoder.* Encodes text into an embedding space.
- *Shape decoder.* Generates a 3D shape from a latent vector. When paired text-3D data is available, this can also be trained to generate (or decode) from the encoded text embedding.
- *Joint text-shape embedding.* When paired text-3D data is available, this can be learned directly.
- *3D generative model.* This generative model can be used to generate a latent vector for the shape decoder, images from which shapes are then generated, or directly to generate diverse shapes.

Note that it is possible to train the text encoder and shape decoder separately on text only and 3D only data. The shape decoder is often trained as part of the shape autoencoder with just 3D data. When paired text-3D data is available, the joint text-shape embedding specifying the alignment can be trained directly. When there is no aligned text-3D data, image embeddings are typically used to bridge the two modalities. Specifically, pretrained vision-language models (where the text-image embeddings are already aligned) are used, and the shape encoder is trained to align shape embeddings into the same space, with the shape decoder then trained to decode from that space.

We first discuss paired text-3D data methods (3DPT) in Section 4.1, and then discuss unpaired text-3D data methods (3DUT) in Section 4.2.

Dataset	Contrastive	Shapes	Categories	Text	Text source
Text2Shape [CCS*19]	no	15K	2	75K	crowdsourced
PartIt [HLZH21]	yes	10K	4	10K	crowdsourced
Cap3D [LRLJ23]	no	550K-785K	many	—	generated
OpenShape [LSK*23]	no	876K	many	—	generated
Point-E [NJD*22]	no	>1M	many	—	—
ShapeGlot [AFH*19]	yes	5K	1	79K	crowdsourced
SNARE [TSB*22]	yes	8K	262	50K	crowdsourced
ShapeTalk [AHS*23]	yes	36K	30	536K	crowdsourced

Table 2: Datasets of paired text and shape. Gray text indicates dataset is not open-sourced. The ‘—’ symbol indicates the statistics are unavailable. Cap3D offers various 3D object-caption paired data versions of varying quality depending on filtering.

4.1 Paired Text to 3D (3DPT)

Traditional supervised approaches are based on the assumption that training data providing paired text and 3D samples is available. Recent work based on this assumption can produce models that are of very high quality. However, these approaches are typically unable to generate objects outside of the dataset and the quality of the 3D objects is highly dependent on the training dataset used. A summary of paired text-to-shape datasets can be found in Table 2.

Table 3 summarizes the methods in terms of text encoder, text-to-3D alignment approach, and generative model used. We categorize supervised methods into three primary classifications. The first encompasses conventional techniques [CCS*19; LWQF22] that learn an aligned space between text and 3D using modality-specific encoders. These latents are subsequently inputted into a 3D object decoder to produce the final shape. The subsequent two categories first employ an auto-encoder for 3D shapes. A separate network is then utilized to learn a prior based on the latent space derived from the initial stage. Text conditioning is integrated through various mechanisms within the prior network, producing latents conditioned on the input text. The key distinction between these latter two categories lies in their modeling approach: one leverages autoregressive models for prior learning [MCST22; FZC*22], while the other employs diffusion models [CLT*23; LDZL23; ZLC*23; LDC*23; JN23].

Much of the work in this area is focused on improving 3D generative modeling, with the text as an illustrative conditioning input along side other potential conditioning input.

4.1.1 Paired Text-to-shape Datasets

There are few datasets that provide both 3D shapes and natural language text descriptions. *Text2Shape* [CCS*19] provided the first paired text and 3D dataset based on the text and chair models from ShapeNet [CFG*15]. Descriptions provide information about both color and shape. *ShapeGlot* [AFH*19] and *ShapeTalk* [AHS*23] provided discriminative text that selected one object from multiple objects. However, as noted by Luo et al. [LLJ22], this style of text data omits important information that is shared between the three objects (e.g. the object category) and is not suitable for aligning text and 3D spaces for 3D shape generation. This data is however, appropriate for shape editing. Many of the shape datasets come from 3D model repositories that contain text information (such as the

name of the asset, product catalogue descriptions, etc). Such text information can be found in datasets such as ABO [CGD*22], Objaverse [DSS*23], and Objaverse-XL [DLW*23]. However, this text information is often noisy, uninformative, and for better or worse, in multiple languages. *OpenShape* [LSK*23] and Cap3D [LRLJ23] illustrated the use of large-language models (LLMs) for filtering and generating more informative text descriptions for shapes on the large scale Objaverse [DSS*23] dataset. Table 2 summarizes these paired datasets.

4.1.2 Notable Examples of 3DPT Methods

GAN-based. The pioneering work *Text2Shape* [CCS*19] was the first work to use deep generative models to generate shapes from text. They modeled the shapes as dense voxel grids representing occupancy and RGB color. Their text to 3D method involves two training stages. The first stage uses learning by association [HMC17] combined with metric learning [Soh16] to learn a shared representation between shapes and text descriptions. Then, a voxel-based conditional Wasserstein GAN [MO14; ACB17; GAA*17] model is used to learn to generate shapes from input text prompts. When training the GAN, the critic judges both whether the generated shape is ‘realistic’ and whether it matches the text.

Auto-encoder with IMLE. Liu et al. [LWQF22] proposed a three stage training process. The first stage learns an auto-encoder model to reconstruct the object occupancy and color grid, where the latent space is decomposed into shape and color features. In the second stage, a text encoder is learned to output features in the same decomposed latent space of the first stage from the paired text data. In addition, another decoder leveraging word and local point feature attention layers are learned for different local features based on point coordinates. Finally, to enable diverse generations from text prompts a generator outputs different shape and text features given input noise vectors and a source shape and text feature. IMLE [LM18] is used to guarantee that each ground truth shape has corresponding samples from the generator.

Autoregressive Prior. *AutoSDF* [MCST22] introduces the use of a Patch-wise encoding VQ-VAE [VV*17], termed P-VQ-VAE to encode 3D shapes characterized by truncated signed distance function (T-SDF) by locally encoding patches independently into discrete encodings, arranged in a grid structure, followed by a joint decoding process. They first train the P-VQ-VAE on 3D shapes from ShapeNet. Then a non-sequential autoregressive model is proposed to learn a prior over the discrete latent space learned from the first stage. For language conditioning, an auxiliary network is trained to predict latents based on a given text prompt, leveraging text-shape pairs from ShapeGlot [AFH*19] for training. Beyond this, their model is also capable of performing other tasks like shape completion and 3D reconstruction.

ShapeCrafter [FZC*22] enhances the work of AutoSDF by introducing the Text2Shape++ dataset, an advancement of the original Text2Shape. This dataset breaks down text prompts into phrase sequences and calculates the similarity between these subsequences and shapes to establish a many-to-many shape correspondence. This approach supports recursive 3D shape modeling during generation. Although ShapeCrafter employs the same P-VQ-VAE from AutoSDF for latent shape representation learning, it

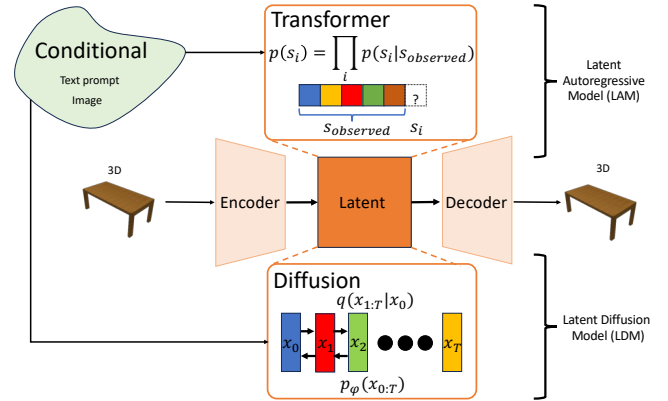


Figure 4: Illustration of the difference between latent autoregressive models (LAM) and latent diffusion models (LDM). LAM utilizes an autoregressive prior over a learned latent space whereas LDM uses a diffusion model. LDMs have demonstrated improved performance in text-to-3D shape generation.

uses a recursive shape generation training process. Here, the phrase sequence is recursively fed to the autoregressive model at each time step, leading to progressive shape generation.

Diffusion Prior. Recent advances in diffusion models have shifted attention from autoregressive models to diffusion-based techniques for modeling the latent space. This trend was notably driven by Latent Diffusion Models (LDMs) [RBL*22]. Li et al. [LDC*23] pinpointed several limitations of autoregressive models, including: 1) error accumulation during sampling; 2) suboptimal generation order due to the processing direction inherent in transformers; and 3) a tendency for transformers to mode-collapse without adequate noise injection, resulting in reduced diversity. Progress in text-to-image diffusion models has streamlined the integration of conditional text into the diffusion process. For instance, the use of domain-specific encoders and a cross-attention mechanism in Rombach et al. [RBL*22] can be adeptly employed for text-guided 3D generation. Similarly, techniques like Img2Img [MHS*21] have been used for text-driven editing and manipulation. Figure 4 illustrates the contrast between latent diffusion models (LDM) and latent autoregressive models (LAM).

SDFusion [CLT*23] similarly employs a two-stage approach for 3D shape generation. Initially, it trains a 3D VQ-VAE model on T-SDF representations using ShapeNet. Subsequently, a 3D UNet is utilized to learn the LDM for the latent space established in the first stage. Conditional generation with different modalities is incorporated following [RBL*22] with modality-specific encoders and a cross-attention mechanism within the 3D UNet. During inference, classifier-free guidance [HS22] is used for conditional shape sampling. Additionally, the model integrates Score Distillation Sampling (SDS) [PJBM23] to facilitate 3D shape texturing. Evaluation on the Text2Shape and ShapeGlot datasets reveals that SDFusion outperforms autoregressive methods such as AutoSDF.

Diffusion-SDF [LDZL23] employs a patch-based VAE [KW13] model to encode local patches derived from the SDF of shapes,

Method	Dataset	3D rep	Color	Images	Model				
					Text	Text-to-3D align	Shape	Gen model	Gen space
Text2Shape[CCS*19]	Text2Shape	voxels (32^3)	yes	no	CNN + GRU	metric + label by assoc.	—	WGAN	voxels
TITG3SG[LWQF22]	Text2Shape	implicit occ.	yes	no	BERT	cyclic loss	AE	IMLE	latent
AutoSDF[MCST22]	ShapeGlot	T-SDF (64^3)	no	no	BERT	—	VQ-VAE	Autoregressive	latent
SDFusion[CLT*23]	Text2Shape	T-SDF ($64^3/128^3$)	texture	no	BERT	—	VQ-VAE	Diffusion	latent
Diffusion-SDF[LDZL23]	Text2Shape	T-SDF (64^3)	no	no	CLIP	—	VAE	Diffusion	latent
3DQD[LDC*23]	ShapeGlot	T-SDF	no	no	CLIP	—	P-VQ-VAE	Diffusion	latent
Shap-E[JN23]	Internal Dataset	STF	yes	multi	CLIP	—	Transformer + NeRF	Diffusion	latent
Michelangelo[ZLC*23]	ShapeNet + templates	pt cloud / occ.	no	multi	CLIP	contrastive loss	SITA-VAE	Diffusion	latent
SALAD[KYNS23]	ShapeGlot	3D Gaussian + impl. occ	no	no	LSTM	—	AD	Casc. diffusion	exp. + latent
ShapeScaffolder[TYW23]	Text2Shape	structured impl. occ	yes	no	BERT + graph	MSE + part-node attn	AE	None - hierarchical decoding	

Table 3: Methods that use paired 3D data with text (3DPT). Early work [CCS*19] did not rely on pretrained models and trained everything from scratch. Later models initially leverage pretrained language models (e.g. BERT), and eventually pretrained vision-language models (e.g. CLIP/GLIDE) for encoding the text. We use ‘—’ to indicate that there was no extra text-to-3D alignment, and that the alignment is done as part of the training of conditional generation with text input. In the color column, ‘texture’ indicates the work proposes a way to apply texture to the generated 3D shape. The ‘Images’ column indicates whether the method uses single or multi-view images for training or aligning image to shape models. The ‘Shape’ column indicates how encoding-decoding into/from the 3D shape latent space is trained, typically with a 3D autoencoder (AE) or variational autoencoder (VAE), with a few works [KYNS23] using auto-decoding (AD) as introduced in DeepSDF [PFS*19] (where only the decoder is trained). Early work [CCS*19] did not train to encode into the shape latent space separately. In Shap-E [JN23], the latent code is used as parameters for a NeRF/STF MLP, and a transformer was trained to project point cloud and multi-view data into the latent space using rendering losses for NeRF and STF. We indicate the type of Generative model (‘Gen model’) used and whether the model is operating in the latent space or not (‘Gen space’). Note that WGAN and IMLE (used in [CCS*19; LWQF22]) require only a single forward pass for sampling and decoding the 3D shape during inference. Newer methods first require autoregressive or diffusion sampling in a latent space then decoding with a decoder model which usually means a longer inference time. We can succinctly refer to autoregressive models and diffusion models in latent space as LAM and LDM respectively (see Figure 4), but here we explicitly indicate the generative model and space for clarity.

utilizing a patch joint decoder for reconstruction. For its LDM, a 3D UNet-based architecture named UinU-Net is introduced. This architecture incorporates an inner network using $1 \times 1 \times 1$ convolutions to focus on individual patch information. The inner network is integrated with the primary UNet via skip connections. Text-conditioned generation parallels the approach of SDFusion. For text-driven shape completion, the model leverages inpainting mask-diffusion methods [LDR*22; RBL*22]. Furthermore, diffusion based image manipulation techniques [CG22; KKY22] facilitate text-guided shape completion.

3DQD [LDC*23] follows the use of P-VQ-VAE in AutoSDF to learn encodings for 3D shapes. Distinctly, 3DQD’s LDM model learns a discrete diffusion process directly on the one-hot encodings of the VQ-VAE’s codebook indexes with transition matrices during the diffusion process. Given the noise and categorical corruption introduced by this discrete diffusion, a Multi-frequency Fusion Module (MFM) is integrated at the end of the denoising transformer to mitigate high-frequency anomalies in token features.

Shap-E’s 3D encoder [JN23] accepts both point cloud data and multi-view renderings of a 3D object. The input data is integrated through cross-attention and subsequently processed by a transformer backbone. This backbone produces a latent representation which acts as the weight matrices for an MLP. Notably, this transformer functions as a hypernetwork [HDL17]. The MLP is designed to predict STF outputs. These outputs include three distinct branches: an SDF branch, a color branch, and a density branch. The latter two are used for the NeRF representation. The SDF values facilitate the generation of a mesh using the marching cubes algorithm, which can then be rendered with the color data. As a result, Shap-e’s decoder can produce both NeRF and mesh outputs.

For training purposes, photometric losses are used with the images rendered from both NeRF and mesh representations. The training dataset comes from Nichol et al. [NJD*22], boasting several million 3D assets paired with textual descriptions. The authors expand upon this dataset with roughly 1 million more 3D assets and an additional 120K captions. Furthermore, their LDM model is trained to predict the MLP weight matrices, which are based on the latents obtained from the 3D encoders. To incorporate text conditioning, the text embedding is prepended to the input of the transformer diffusion model. Thanks to the large scale of the training dataset, Shap-E can generate a rich array of diverse 3D objects. However, regrettably, this training dataset is not publicly available to the research community.

Michelangelo [ZLC*23] introduces a methodology to learn an alignment between 3D shapes, text, and images prior to generation. Initially, a transformer is used to encode the 3D point cloud, resulting in shape tokens. Using frozen CLIP encoders [RKH*21], similar encoding processes are applied to both text and images, producing text and image tokens, respectively. To ensure alignment across these three modalities, modality-specific projectors are used to process these tokens into distinct image, text, and shape embeddings. The alignment of these embeddings is then done through a contrastive loss between the (shape, image) pairs and the (shape, text) pairs. A reconstruction loss, based on the occupancy grid generated by the 3D decoder, is also employed. The overarching network architecture is termed the Shape-Image-Text Aligned Variational Auto-Encoder (SITA-VAE). After alignment, an Aligned Shape LDM (ASLDM) is trained to model the shape tokens. The process of conditioning on text and image tokens follows the methodology presented by Rombach et al. [RBL*22].

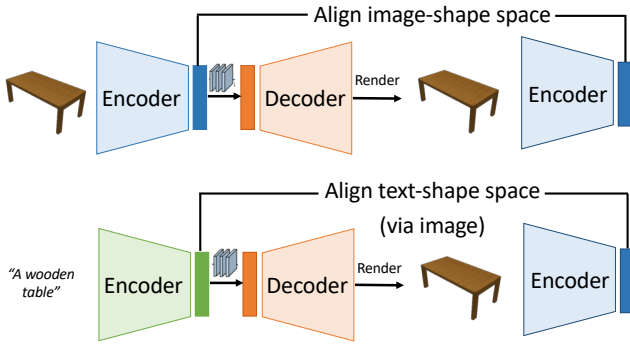


Figure 5: In text-to-shape generation with unpaired data (3DUT), an image-to-3D generative model is combined with pretrained text-image embeddings. Typically, 3D assets are first used to train a 3D autoencoder that takes a shape embedding to an 3D shape. Large pretrained text-image embeddings are then used to align the text to 3D space, with a generative model that can sample from the 3D latent space conditioned on an input embedding. The shape space is aligned to the pretrained text-image space by: 1) aligning the shape decoder to render from image embeddings e.g., as in CLIP-Forge and CLIP-Sculptor (top); or 2) ensuring 3D shapes generated from a text embedding have rendered embeddings that match the text embedding, e.g., as in TAP3D (bottom).

4.1.3 Structure aware text to shape generation

Most methods we described so far treat the output shape as an unstructured 3D representation. An emerging line of work attempts to produce structured 3D shapes based on parts and part connectivity [LLJ22; TYW23]. ShapeScaffolder [TYW23] first uses shape data to pretrain a structured shape decoder that can take a global latent vector encoding (consisting of both shape and color latents) and hierarchically decode it into parts. For text conditioning paired text-shape data is used to align a global text embedding to the shape and color latent spaces via MSE loss. The text is also converted to a graph representing parts, their attributes, and relations. Features for the parts and relations are extracted and attention between the text-based graph features and part features guides the hierarchical decoding process. Another recent work, SALAD [KYNS23] proposed a two-phase diffusion approach. In the first phase, the part structure is generated as a mixture of 3D Gaussians, and in the second phase a latent vector representing the shape details for each part is generated. To condition the model on text, they concatenate the language feature and train the model using paired text-shape data from ShapeGlot [AFH*19]. There is also growing interest in generating shapes based on programs [JGMR23; YXP*23]. A recent work that connects text to programs is Neural Shape Compiler [LLJ22]. Translating text to programs that can generate 3D shapes is an interesting direction for future investigation.

4.2 Unpaired 3D Data (3DUT)

In this section, we discuss methods where a database of 3D shapes is available for training a 3D generative model, but there is no paired text-to-3D data. These methods are summarized in Table 4. In this case, pretrained text-to-image models are used as a bridge

to align the text and 3D embeddings. Without paired 3D data, but access to a 3D dataset, it is possible to render 2D images from the 3D data and use the generated images as training data to train a generative model that goes from 2D image to 3D shapes. Work in this space can either train their own image-to-3D model or leverage existing pretrained models. The generative model will typically make use of latent image embeddings that comes from a pretrained vision language model (e.g. CLIP). Since CLIP has aligned text-and-image embeddings, this family of work assume using CLIP text-embedding directly will be sufficient. However, recent work has identified that the pretrained embeddings have shortcomings and proposed ways to compensate. Figure 5 provides an illustrative summary of this family of approaches.

4.2.1 Notable Examples of 3DUT Methods

CLIP-Forge [SCL*22]. Initially, a shape autoencoder is trained on the ShapeNet dataset to reconstruct 3D voxels. This involves the encoder transforming voxels into embeddings, represented as $e^V = \text{Enc}^V(x^V)$. Subsequently, both the embedding and voxel positions are inputted into a decoder occupancy network [MON*19], denoted as Dec^V , to recreate the original shape.

The following stage uses a flow model [DSB16]. This model is trained to process the embeddings, e^V , in conjunction with CLIP image embeddings derived from multi-view renderings of the 3D objects. These CLIP embeddings are represented as $e_{\text{CLIP}}^I = \text{Enc}_{\text{CLIP}}^I(x^I)$, where x^I are the rendered images. The flow model maps inputs onto a Gaussian distribution.

During inference, multi-view images are substituted with text prompts. These prompts are then encoded using CLIP to obtain the associated embeddings, represented as $e_{\text{CLIP}}^T = \text{Enc}_{\text{CLIP}}^T(x^T)$. Given that the CLIP encoders are trained to map to a unified latent space, both the text embedding and a sample from the Gaussian distribution are inputted into the flow model in reverse to obtain the shape embedding, e^V . Finally, the occupancy network decodes the final shape.

CLIP-Sculptor [SFL*23]. The concept of leveraging CLIP encoders that map to a shared latent space is also seen in CLIP-Sculptor. However, CLIP-Sculptor employs a more powerful autoregressive model in lieu of a flow model. During the initial training phase, two VQ-VAEs are trained at varying resolutions for reconstruction. For the 3D shape, the voxel encoder transforms it into $e_r^V = \text{VQ}(\text{Enc}_r^V(x_r^V))$, where e_r^V is the vector-quantized grid encoding for the input voxel and r indicates the voxel resolution. The lower resolution voxel grid is configured at 32^3 , while the higher resolution operates at 64^3 . The shape is then reconstructed with the decoder $\hat{x}_r^V = \text{Dec}_r^V(e_r^V)$.

Similarly to CLIP-Forge, multi-view renderings of the 3D objects are encoded using CLIP, expressed as $e_{\text{CLIP}}^I = \text{Enc}_{\text{CLIP}}^I(x^I)$. Noise is introduced to these embeddings, resulting in the conditional vector c for variation. This vector is then processed by an MLP, which subsequently predicts the affine parameters of the layer normalization layers present in a coarse transformer with dropout. This transformer is designed to autoregressively decode masked encodings under the constraint of conditionals, as denoted

Method	Dataset	3D representation	Color	Images	Model		
					Text	Text-to-3D alignment	Generative model
CLIP-Forge[SCL*22]	ShapeNet	voxels	no	multi	CLIP	CLIP + image / 3D align	Flow (latent)
CLIP-Sculptor[SFL*23]	ShapeNet	voxels (32 ³ /64 ³)	no	multi	CLIP	CLIP	LAM
ISS[LDL*23]	ShapeNet, CO3D	DVR	yes	single	CLIP	CLIP + mappers + test time consistency loss	DVR
TAP3D[WWF*23]	ShapeNet	DMTet	yes	multi	CLIP	CLIP + mappers + align loss	GAN

Table 4: Methods that use unpaired 3D data (3DUT). The ‘Images’ column indicates whether the method uses single or multi-view images for training or aligning image to shape models. For ‘Text-to-3D alignment’, all models use CLIP to align 3D objects during training with text prompts during inference. Additional components may be further leveraged to better align image, 3D or text to bridge domain gaps in the CLIP latent space. Note that TAP3D[WWF*23] (in light gray) trains an aligned text-shape space using generated templated text.

by $P(e_{32}^V | \text{Mask}(e_{32}^V), c)$. In the subsequent training stage, a super-resolution (fine) transformer is trained to unmask high resolution encodings $\text{Mask}(e_{64}^V)$ conditioned on the predicted lower resolution encodings e_{32}^V from the coarse transformer through cross attention.

During the inference phase, the image embeddings are supplanted by text embeddings derived from user prompts. Following this, the masked encoding undergoes an iterative decoding process, first by the coarse transformer and subsequently by the fine transformer, to produce the final encoding representing the desired shape. Drawing inspiration from classifier-free guidance prevalent in diffusion techniques, CLIP-Sculptor frames the sampling with the equation: $\hat{P}_t(c) = P_t(0) + a(t)(P_t(c) - P_t(0))$. Here, $P_t(c)$ is represented by $P(e_{32,t}^V | \text{Mask}(e_{32,t-1}^V), c)$ and $P_t(0)$ is given by $P(e_{32,t}^V | \text{Mask}(e_{32,t-1}^V), 0)$. The function $a(t)$ starts with a high guidance scale, ensuring the shape adheres to the text-based constraints. However, as the process advances, its value decreases to introduce greater diversity in the output.

ISS [LDL*23]. The two aforementioned papers operate with the assumption that the CLIP text and image spaces align seamlessly. However, the ISS authors find otherwise. Specifically, they discovered significant distances between the embeddings of paired text and images. When directly swapping the CLIP image encoder for the text encoder during inference, this discrepancy can lead to inconsistencies between the generated shape and its corresponding text. Additionally, learning to generate shapes straight from the CLIP feature space causes a loss of details in the final 3D shape. To counteract these issues, ISS introduces a two-phase method for aligning the feature spaces, enabling 3D shape generation guided by text without assuming perfectly matched text-shape pairs.

The initial alignment phase uses a pretrained single-view reconstruction model, specifically DVR [NMOG20]. This comprises an encoder converting images to latent embeddings, and a decoder responsible for shape reconstruction. They align the CLIP image features with the DVR network’s latent features using a mapping network M . This is achieved by minimizing the function $L_2(M(\text{Enc}_{\text{CLIP}}^I(x^I)), \text{Enc}_{\text{DVR}}^T(x^I))$. Here, $\text{Enc}_{\text{DVR}}^T$ represents the DVR decoder, x^I are images rendered from the 3D object, and the CLIP features are normalized. The decoder loss and other regularizing losses are optimized to refine the decoder network.

At the inference stage, to narrow the disparity between CLIP’s image and text encoders, the mapping network M undergoes additional fine-tuning. This is to minimize the CLIP consistency loss

represented by $\langle \text{Enc}_{\text{CLIP}}^I(x^I) \cdot \text{Enc}_{\text{CLIP}}^T(x^T) \rangle$. Here, x^I denotes images rendered from the 3D shape using the DVR decoder, while x^T is the provided text prompt. All CLIP features are also normalized. Acknowledging the constraints of the DVR model, they allow for further fine-tuning of the DVR decoder with the CLIP consistency loss for text-driven stylization.

TAP3D [WWF*23] proposes a simple approach of adding text controls to an existing unconditional 3D generation model. They utilize CLIP similarity as a loss to fine-tune parts of the original network to accept CLIP text embeddings as input. Specifically, they use a GET3D [GSW*22] model pretrained on ShapeNet, a DMTet-based GAN. The GET3D model takes as input two noise vectors and transforms them into intermediate latent codes using mapping networks for modeling the geometry and texture. The mapping networks are the only parts of the model being fine-tuned. First, to obtain text prompts for 3D shapes, they generate pseudo-captions by calculating nouns and adjectives that have high similarity with rendered images of the 3D objects. They then construct the captions with template sentences. The mappers are trained by maximizing the CLIP similarity between the pseudo-captions and rendered images of the generated object from GET3D. An additional regularization loss aims to maximize the similarity between rendered images of generated objects and ground-truth object renderings.

5 Text-to-3D without 3D data (NO3D)

Here we discuss the family of approaches designed for scenarios when either no 3D data is available, or reliance on such 3D data is avoided. In this scenario, the typical strategy uses per-prompt optimization together with a differentiable renderer to optimize an underlying 3D representation such that 2D images that are compatible with the text prompt can be generated. In this unsupervised setting, there are currently two popular sets of approaches: 1) maximizing the similarity of the prompt and rendered images using a pretrained vision-language joint embedding (Section 5.1); or 2) using a pretrained text-to-image diffusion model to guide updates to the parameters of the 3D representation (Section 5.2). Figure 6 illustrates these two sets of approaches.

A key design decision applying to both sets of approaches is the choice of 3D representation. There are several popular choices that permit differentiable rendering: 1) NeRF-based model; 2) deforming triangular meshes; and 3) deep marching tetrahedra (DMTet).

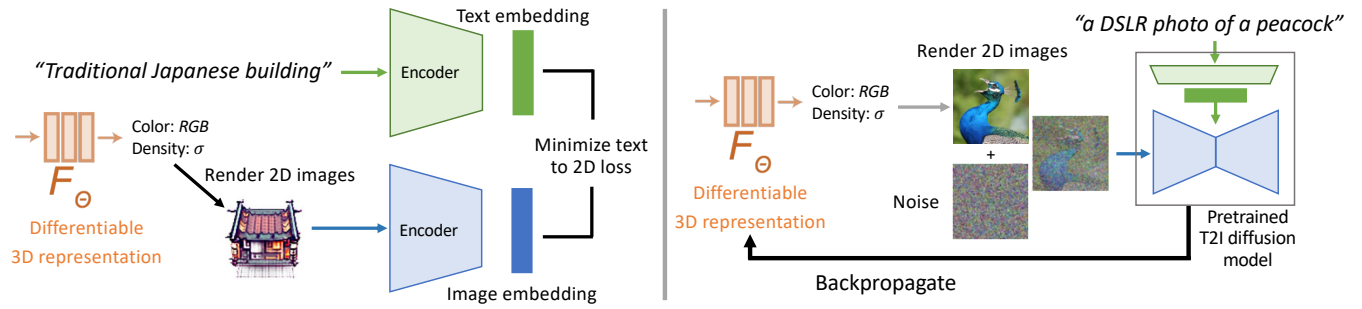


Figure 6: Illustration of the text-to-3D shape generation strategies that do not rely on 3D data. Left: methods minimizing text to 2D loss using a pretrained vision-text embedding space such as CLIP. Right: methods using pretrained image-to-text diffusion-based models to guide the training of a 3D representation. These are per-prompt optimization strategies, leading to significant compute and time required per shape.

Method	3D Rep	Augmentation	Additional loss
DreamFields [JMB*22]	MipNeRF [BMT*21]	BG (GN, CH, FT)	Transmittance
CLIP-Mesh [MXBP22]	mesh + normal + texture	BG	Diff. prior + Lap reg
PureCLIPNeRF [LC22]	explicit/implicit NeRF	BG + diff + persp	Tr + TV + KL + Bg

Table 5: Summary of unsupervised CLIP guidance methods. These methods optimize the 3D representation so that the CLIP similarity between rendered images and text is maximized. They also make use of additional augmentation and loss terms to help enforce the generation of more geometrically plausible objects. For instance, DreamFields [JMB*22] includes various background (BG) augmentations such as Gaussian noise (GN), checkerboard pattern (CH), and random Fourier textures (FT).

5.1 Unsupervised CLIP Guidance

In this class of methods, the parameters of a differentiable 3D representation are updated so that rendered images have high similarity to text prompt embeddings, as evaluated by a pretrained vision-language joint embedding (CLIP). This approach is challenging as training by just optimizing the CLIP similarity can be tricky and does not provide any signal for geometric consistency. To counter this, work in this area proposes a variety of regularization and augmentation techniques. Table 5 provides a summary of methods and categorization along the axes of 3D representation, augmentation strategies, and additional regularization.

5.1.1 Notable Examples

Dream Fields [JMB*22]. Dream Fields is a pioneering work in this direction, first demonstrating the possibility of creating 3D shapes without relying on any 3D data. While previous research utilized CLIP to bridge the language-shape gap in scenarios with available 3D data but lacking pairing with text, Dream Fields showcased the capability to guide the training of a 3D representation directly using pre-trained text-image CLIP embeddings.

To achieve this Dream Fields uses a NeRF as the backbone 3D representation. Given a camera pose described by azimuth and elevation (θ, ϕ) , an image $x^I = \text{NeRF}(\theta, \phi)$ is generated using volumetric rendering. The specific NeRF used is MipNeRF [BMT*21]. Then an input text prompt x^T guides the model training using the CLIP similarity loss, expressed as $\mathcal{L}_{\text{CLIP}} =$

$-\text{Enc}_{\text{CLIP}}^I(x^I)^T \text{Enc}_{\text{CLIP}}^T(x^T)$. A key challenge when using CLIP in this fashion is that the embeddings minimize a contrastive loss between images and text, and thus may prioritize texture learning over structural or spatial information. Dream Fields shows that solely relying on this loss can lead to the generation of implausible 3D objects. To counteract this, they propose multiple regularization techniques. Initially, they incorporate a transmittance loss to eliminate floating density anomalies. They also experiment with various background augmentations (e.g., Gaussian noise, checkerboard patterns, random Fourier textures). These are alpha-composited onto the NeRF-rendered images, enhancing the coherence of the resulting 3D models.

CLIP-Mesh [MXBP22]. CLIP-Mesh similarly employs CLIP guidance. However, they use a mesh representation. Notably, their training begins with a primitive geometric shape, such as a sphere. This differentiates the approach from Text2Mesh [MBL*22], which requires an initial object aligned with the semantic class of the input text prompt.

The primitive geometry’s vertices dictate a subdivision surface, with both the normal map and texture maps initialized randomly. The object is rendered from various viewpoints using a differentiable renderer, and these images are then processed through the CLIP image encoder. The similarity loss with CLIP is subsequently calculated against the text embeddings. Additionally, the CLIP-Mesh approach integrates a diffusion prior akin to DALL-E 2 [RDN*22]. Here, the text embeddings are mapped to image embeddings. This is also leveraged to calculate similarity between the rendered images. Their findings indicate that jointly employing these losses enhances generation quality. Training is regularized using a Laplacian regularizer and random background augmentations (mirroring the Dream Fields approach).

PureCLIPNeRF [LC22]. In the works discussed thus far, optimizing images under CLIP guidance often leads to incoherent results. In PureCLIPNeRF [LC22], various elements are examined to mitigate the adversarial generations caused by CLIP guidance. These elements include different image augmentations, pretrained CLIP backbones, and both explicit and implicit NeRF architectures. This work finds that the combination and selection of these factors influences the quality of the final output 3D shape results significantly. By integrating augmentation strategies, CLIP backbones, and 3D



Figure 7: Qualitative results from DreamFusion [PJBM23] comparing against methods utilizing CLIP guidance (Dream Fields, CLIP-Mesh). DreamFusion produces higher quality 3D shapes. Visuals reproduced from Poole et al. [PJBM23].

representation design choices effectively, one can produce more consistent 3D objects.

Dream3D [XWC*23]. While the aforementioned regularizations aim to counteract adversarial generations, CLIP guidance remains susceptible to emphasizing textures or patterns from the input text prompt. This often leads to inconsistencies in the geometry of the generated 3D object. To address this, Dream3D [XWC*23] introduces a 3D prior for initializing the CLIP guidance optimization. This is achieved by fine-tuning a Stable Diffusion [RBL*22] model using text and renderings from ShapeNet. Another diffusion model is trained to translate the produced images into a 3D shape latent code. This latent input is processed by an SDF-StyleGAN [ZLWT22] network to generate the 3D shape. Ultimately, CLIP guidance refines this 3D model, ensuring it aligns more closely with the provided text.

5.1.2 Discussion

While methods in this section are not bounded to generating objects within the limits of a dataset like prior sections, the geometric quality of the output objects is still lacking in many cases. This limitation is primarily because CLIP is trained with a contrastive objective, which can cause the image encoder to prioritize texture pattern extraction. As a result, using CLIP for guidance may not effectively generate fine geometric detail on objects and can yield overly simplified geometry. To address this issue, prior work has explored various augmentations that help regularize the training process, different CLIP architectures, as well as different 3D shape priors. While these improvements can improve generated outcomes, the subse-

quent section will demonstrate that the alternative strategy using text-to-image diffusion models as guidance helps to increase the quality of the output 3D shape geometry as can be seen in Figure 7. This strategy leads to objects that not only align more closely with more complex text prompts but also showcase detailed geometric and texture details.

5.2 Unsupervised Diffusion Guidance

Much like the approach with CLIP guidance, we can bridge the gap between 3D shape being generated and the 2D guidance model by rendering images of the 3D shape from various camera perspectives. Instead of optimizing the similarity between images and text, we utilize scores from the diffusion models as the guiding gradients. The score, as determined by the denoising diffusion model $s_\phi(x_t, t)$, serves as an approximation to the score function $\nabla_{x_t} \log p_{\alpha_t}(x_t|x)$ [SSK*21]. Here, the score function $\nabla_{x_t} \log p_{\alpha_t}(x_t|x)$ points towards areas of higher probability density. In more detail, our objective is to ensure that images, irrespective of the angle from which the 3D model is rendered, gravitate towards high-probability zones outlined by the diffusion models. Note that the noise function $\epsilon_\phi(x_t, t)$ is proportional to the score [Rob92]:

$$-\frac{\epsilon_\phi(x_t, t)}{\sigma_t} = \frac{D(x_t; \sigma_t) - x_t}{\sigma_t^2} = s_\phi(x_t, t) \quad (9)$$

Table 6 summarizes the methods in this family, and categorizes them in terms of the underlying 3D representation, specific guidance model, and loss terms used during training. In Section 5.2.1, we highlight seminal papers that focus on formulating a loss around this concept. Section 5.2.2 delves into works that leverage improved 3D representations for higher quality outputs. A notable challenge with this formulation is the “Janus Problem”, which is a consistency issue with the generated object. This arises due to optimizing each rendered image without taking into account the entire object as well as inherent 2D biases in text-to-image (T2I) diffusion models. To address this, Section 5.2.3 presents studies that offer mitigation strategies and fine-tune T2I models on multi-view data, thereby injecting 3D information to counteract the 2D bias. Another limitation of this framework is that we have to conduct training for each individual text prompt which can take up to several hours for high quality models. However, the research in Section 5.2.4 suggests integrating this loss into a multi-object generative model, enhancing training efficiency. This approach enables generation of multiple objects and interpolation between prompts during inference.

5.2.1 Loss Formulation

In this section, we highlight notable papers that introduce loss functions designed to harness 2D priors from text-to-image (T2I) diffusion models for generating 3D models from text.

DreamFusion [PJBM23]. The DreamFusion authors introduced the pioneering Score Distillation Sampling (SDS) loss. This was used to optimize a NeRF model, specifically Mip-NeRF 360 [BMV*22], parameterized by θ . It effectively integrates prior knowledge distilled from a pretrained T2I diffusion model Image [SCS*22], which is parameterized by ϕ . The optimization is

Method	3D representation	Guidance model	Loss
DreamFusion [PBJM23]	Mip-NeRF 360 [BMV*22]	Imagen [SCS*22]	SDS
SJC [WDL*23]	Explicit Voxel Grid NeRF	StableDiff [RBL*22]	SJC
Prolific Dreamer [WLW*23]	1) Instant-NGP [MESK22] + 2) Dis. DM Tet [SGY*21; CCJJ23]	StableDiff	VSD
Magic3D [LGT*23]	1) Instant-NGP + 2) DM Tet	1) eDiff-I [BNH*22] + 2) StableDiff	SDS
TextMesh [TMT*23]	1) SDF NeRF + 2) Mesh	1) Imagen (SDS) + 2) StableDiff (texturing)	SDS + Tex Ref
Fantasia3D [CCJJ23]	Dis. DM Tet	StableDiff	SDS
DreamGaussian [TRZ*23]	3D Gaussians	StableDiff	SDS + Tex Ref
GSGEN [CWL23]	3D Gaussians	StableDiff, Point-E	SDS
GaussianDreamer [YFW*23]	3D Gaussians	StableDiff	SDS

Table 6: Summary of methods using unsupervised diffusion guidance, a sub-family of methods that do not require 3D data (NO3D). We organize these methods along the design choices of 3D representation, guidance model, and training loss. Some of the methods use a two stage approach to obtain a higher resolution mesh output or to separate geometry and texture (indicated by numerals 1 and 2 in each cell). Tex Ref indicates the use of additional texture refinement losses.

driven by the following loss formulation:

$$\nabla \mathcal{L}_{SDS} = \mathbb{E}[w(t)(\epsilon_\phi(x_t; y, t) - \epsilon) \frac{\partial x}{\partial \theta}] \quad (10)$$

In this context, x represents the image rendered from NeRF, while y is the text prompt. As previously mentioned, $\epsilon_\phi(z_t; y, t)$ relates to the score, signifying gradients that direct towards high-probability regions in the ambient space, conditioned on y . The subtraction of the added noise ϵ , serves as variance reduction. This is crucial since the denoiser is conditioned on x_t , which exhibits noise levels distinct from those in the NeRF-generated image. Furthermore, DreamFusion has shown that:

$$\nabla_\theta \mathcal{L}_{SDS}(\phi, x = g(\theta)) = \nabla_\theta \mathbb{E}_t \left[\frac{\sigma_t}{\alpha_t} w(t) KL(q(x_t | g(\theta); y, t) \| p_\phi(x_t; y, t)) \right] \quad (11)$$

Optimizing the SDS loss aligns with minimizing the KL divergence between the noise-injected images from the NeRF network $g(\theta)$, and the probability densities learned by diffusion models conditioned on text prompt y . For a detailed derivation, see the original paper. Using the powerful priors from T2I models, DreamFusion achieves markedly improved results compared to prior methods employing CLIP. It is worth noting that the more sophisticated shading model emphasizes geometric details, while a basic albedo model might induce ambiguities, impacting the quality of the object geometry. A notable limitation is the ‘‘Janus problem’’ where models display multiple front faces, likely due to datasets used to train T2I models predominantly featuring front views of objects.

Score Jacobian Chaining (SJC) [WDL*23]. Concurrently, Score Jacobian Chaining (SJC) formulates its approach by applying the chain rule to the score of the diffusion model. This sidesteps the UNet Jacobian term present in DreamFusion, which originates from the diffusion training loss. It is noteworthy that this term is excluded from the SDS loss, as empirical results demonstrated enhanced performance without it. The SJC loss is as follows:

$$\nabla_\theta \mathcal{L}_{SJC} = E_n \left[\frac{D(x_\pi + \sigma n, \sigma) - x_\pi}{\sigma^2} \right] \frac{\partial x_\pi}{\partial \theta} \quad (12)$$

The equation involves π , the camera pose for NeRF rendering, and n are distinct noise samples employed for the Monte Carlo estimate. Both the SDS and SJC formulations display comparable performance. Their model employs Stable Diffusion [RBL*22] as guidance and an explicit voxel grid NeRF model [CXG*22; FYT*22;

SSC22]. By directly rendering in the Stable Diffusion latent space $R^{64 \times 64 \times 4}$, they accelerate training, sidestepping gradient computations through the encoder. This technique is also advocated in Latent-NeRF [MRP*23].

ProlificDreamer [WLW*23]. A limitation of the SDS loss formulation is the tendency for generated objects to exhibit excessive smoothness and limited variation. Prolific Dreamer theorizes that this stems from the mode-seeking formulation in Equation (11). Essentially, DreamFusion aims to fit a single plausible 3D model to a probability distribution within the diffusion model. However, the distribution encompasses numerous objects with different identities fitting description y . To address this, Prolific Dreamer approaches the issue through the lens of variational inference:

$$\min_\mu KL(q_\mu^\theta(x_0 | y) \| p_0(x_0 | y)) \quad (13)$$

Here the goal is to minimize the KL divergence between images generated by a distribution μ comprised of multiple NeRF models (particles) and the pretrained diffusion model. This leads to their ultimate formulation, termed the Variational Score Distillation (VSD) loss defined as:

$$\nabla_\theta \mathcal{L}_{VSD}(\theta) \approx \mathbb{E}_{t, \epsilon, c} [w(t)(\epsilon_{pretrain}(x_t, t, y) - \epsilon_\phi(x_t, t, c, y)) \frac{\partial x}{\partial \theta}] \quad (14)$$

The term ϵ_ϕ represents a LoRA [HSW*21] parameterization of the pretrained diffusion network, which also incorporates the camera condition c for rendering images from the NeRFs. The LoRA network is optimized using the standard diffusion training loss, alternating with the VSD loss. Though optimizing multiple particles (NeRFs) can be resource-intensive, their model consistently delivers superior results even with a single particle. Compared to the SDS loss, the produced objects exhibit markedly enhanced quality, as shown in Figure 8. This approach also enables training at lower guidance weights (7.5 versus 100 in DreamFusion), which helps to produce textures that are more realistic and less cartoonish. Training is in a two-stage manner with improved hyperparameters similar to Magic3D [LGT*23] which we discuss in the next section.

5.2.2 3D Representation Improvements

Training of NeRFs is often hampered by the substantial memory requirements of volumetric rendering, which typically restrict meth-

ods to lower resolutions. Here, we discuss mesh-based representations, such as DMTet, that employ rasterization. This is a more memory-efficient approach enabling training at higher resolutions, and generation of sharper textures and finer geometric details. Recently, Gaussian Splatting (GS) [KKLD23] has gained traction due to efficient representation and faster rendering compared to NeRFs.

Magic3D. The Magic3D [LGT*23] approach addresses the slow training times in DreamFusion and lack of fine details in the generated objects. These issues stem from the high memory usage of the NeRF model. The authors present a two-stage coarse-to-fine training pipeline aimed at crafting high-fidelity 3D models. In the coarse stage they train a Instant-NGP [MESK22] model at a lower resolution of 64×64 under the guidance of the T2I model eDiff-I [BNH*22]. Subsequently, in the fine stage they convert the NeRF model into a DMTet [GCX*20; SGY*21] representation. Training at this stage is done at a higher resolutions of 512×512 with Stable Diffusion as guidance. Thanks to the efficiency of rendering through rasterization rather than volume rendering, this approach considerably reduces the memory footprint. The resulting 3D meshes extracted from DMTet surpass the quality of those produced by DreamFusion with a substantial 61.7% of human raters preferring the former in a user study.

TextMesh [TMT*23]. In concurrent work to Magic3D, TextMesh similarly adopts a two-stage approach to avoid over-saturated textures in objects produced from DreamFusion. In the first stage they train a NeRF model with a SDF backbone for densities. This makes it easier to extract a mesh after this stage using Marching Cubes [LC98]. They then render depth from four views of the mesh (top, front and sides) which are tiled and fed through Stable Diffusion with depth conditioning to obtain textured images. They then fine-tune the mesh textures with photometric loss from the textured images as well as the SDS loss with smaller guidance weights of 7.5 to produce objects that have more photorealistic textures.

Fantasia3D [CCJJ23]. The Magic3D authors stated that DMTet training from random initialization leads to poor results. This could be due to the more discrete nature of DMTet, whereas NeRFs are more continuous in the sense that there are multiple points along the ray where gradients can propagate. Fantasia3D shows that it is possible to train from scratch with DMTet alone and achieve high quality geometry as well as textures. To do this they disentangle the training into separate stages of geometry and appearance modeling with the SDS loss. Their DMTet geometry model consists of an MLP parameterized by Ψ to predict SDF values and deformations. The network is initialized with a 3D ellipsoid in the beginning of training. In the geometry modeling stage they propose a novel mask and normal augmentation technique in the early phase of training directly in the latent space of Stable Diffusion and then encode the normal directly in RGB space for the rest of the geometry training with SDS guidance. Then, in the appearance training stage they fix the geometry while learning the diffuse k_d , metallic k_{rm} and normal variation kn maps for the 3D model using a Physically Based Rendering (PBR) model where another MLP Γ is used to predict the different terms. They show that disentangling the learning of two components results in better geometry as well as better textures due to the PBR rendering model.



Figure 8: Qualitative comparison of DreamFusion [PJBM23], Magic3D [LGT*23], Fantasia3D [CCJJ23], GaussianDreamer [YFW*23], and DreamFusion [PJBM23]. Results generated by running authors' implementation for GaussianDreamer and threestudio implementation of each method for the others. This line of work has demonstrated steady progress in improving the quality and coherence of generated 3D objects.

Gaussian Splatting Methods. Recent work leverages 3D Gaussians and splatting to efficiently represent complex 3D scenes. DreamGaussian [TRZ*23] samples points within a sphere and then optimizes the 3D Gaussians with the SDS loss, periodically densifying points to add detail. GSGEN [CWL23] instead initializes using a point cloud generated from Point-E [NJD*22]. The 3D Gaussians are optimized not only with SDS from the rendered images, but a 3D SDS loss with positions of the Gaussian points using Point-E. They then densify and prune the Gaussians to refine appearance with additional regularization losses along with the image SDS loss. GaussianDreamer [YFW*23] similarly initializes using points from 3D diffusion models such as Shap-E [JN23]. These 3D priors used for initialization help generate more consistent shapes. DreamGaussian and GaussianDreamer boast training speeds of 2 and 15 minutes on a single GPU.

5.2.3 Janus Problem

A prevalent issue in text-to-3D shape generation is known as the “Janus problem”. This is characterized by the generation of 3D models with multiple faces that deviate from realistic object structure. For instance, when attempting to create a 3D representation of a dog, certain perspectives may erroneously display multiple frontal faces, as illustrated in the first two rows of Figure 9. This anomaly is likely a consequence of dataset bias in the training of T2I models, with a disproportionate number of images presenting objects from a frontal perspective. DreamFusion attempted to side step this by introducing view-dependent prompts. For example, they suggested the use of specific prompts for different azimuth angles: ‘front view + prompt’ for angles ranging from 0° to 90° , ‘back view + prompt’ for 180° to 270° , and ‘side view + prompt’ for the remaining angles, with the addition of ‘top view’ or ‘bottom view’ contingent upon defined elevation thresholds. Despite these measures offering some improvement, they are not a panacea, as the Janus problem persists across various prompts.

Mitigation. Some methods have been proposed to side step this issue and can be generally incorporated into SDS-like loss training frameworks. Hong et al. [HAK23] propose two methods for debiasing the training process by clipping gradients of the SDS loss, and by removing words that may be in conflict with the viewing angle from the prompts to debias them. Seo et al. [SJK*23] utilizes 3D geometry priors of shapes generated from models like Point-E [NJD*22] to help maintain 3D consistency. Initializing 3D Gaussians with points generated by Point-E in GSGEN [CWL23] and GaussianDreamer [YFW*23] also have the same effect. Perp-Neg [AZS*23] proposes to use negative prompts in the T2I diffusion model framework to discourage inconsistent view prompts with respect to the sampled view. While these methods help mitigate the Janus Problem, they do not solve the underlying bias within T2I models. In the following section we discuss methods such as MVDream [SWY*23] and SweetDreamer [LCCT23] which fine-tune the guidance models to be 3D-aware, directly addressing this issue.

5.2.4 Generative Models with SDS Loss

ATT3D [LXZ*23]. The ATT3D approach addresses a big problem with the other works discussed in this section: requiring per-prompt training for every object. ATT3D trains a unified model using the SDS loss with multiple prompts at once. To do this they add an additional mapping network that takes the text prompt as input and is used to modulate spatial grid features in the Instant-NGP model to generate multiple different objects. During training they also amortize over text by interpolating over text embeddings, helping to smooth interpolations between different text prompts. The ATT3D authors show that their approach converges with better efficiency (fewer rendered frames per prompt) and achieves results comparable to the single-prompt training of DreamFusion. They also show some generalization capabilities with unseen prompts that are compositing new prompts with text in the training prompts.

At the time of writing another approach [LZL*23a] has been proposed to tackle the same issue. Utilizing a decoder network to generate triplane NeRF representations from text, they also experiment with different ways for effectively injecting the text information.



Figure 9: Qualitative comparison of ProlificDreamer [WLW*23] vs MVDream [SWY*23]. Results generated using MVDream authors’ implementation and threestudio ProlificDreamer implementation. The MVDream approach mitigates the ‘Janus problem’ where the 3D model has multiple frontal faces (each row shows four views around an output 3D model).

The model is trained with multi-prompt SDS and CLIP losses with Perp-Neg [AZS*23] to mitigate the Janus problem. Their method is able to generate higher quality 3D objects compared to ATT3D with better text and 3D alignment.

5.2.5 Discussion

In this section, we first discussed losses for distilling 2D priors learned by T2I models for generating 3D objects. This includes DreamFusion [PJM23] and Score Jacobian Chaining (SJC) [WDL*23] as well as ProlificDreamer [WLW*23] which improve upon the formulation. We then introduced works that improve upon the quality of the models with more efficient representations such as DM Tet. The strong priors learned in T2I models demonstrate superior generation quality compared to models in the previous section utilizing CLIP guidance. This area is increasingly popular with several followup works analyzing how to improve upon the distillation loss [LYL*23; KPCL23; WFX*23; YGL*23; WXF*23; TWW23; ZSME23; PLZZ24; WZY*24].

Compared to works utilizing paired text and 3D or just 3D data, the methods in this family are able to generate objects that span a broader domain and in general can generate objects from more complex text prompts. However, a drawback of this class of methods is having to train a new 3D model for every text prompt (a limitation that is also present in the CLIP-guided methods). For

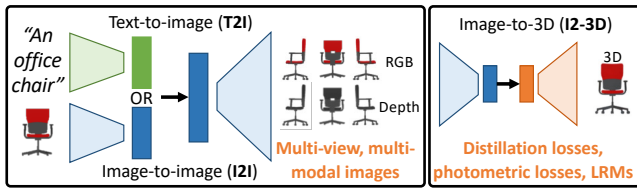


Figure 10: Illustration of recent methods that rely on a ‘Hybrid3D’ pipeline of: i) making text-to-image (T2I) or image-to-image (I2I) models 3D-aware through multi-view and multi-modal images and corresponding camera viewpoint information; and ii) improving image-to-3D models using distillation losses, photometric losses and Large Reconstruction Models (LRMs).

higher quality objects, the training time can take up to hours for some methods. This is in comparison to only having to run a much faster inference for the paired or unpaired 3D data methods. ATT3D [LXZ*23] and Instant3D [LZL*23a] address this issue by applying the SDS loss in a multi-prompt generative setting. These methods show initial success with better training efficiency compared to per-prompt optimization. However, the models are trained on a relatively small amount of prompts and whether this training strategy generalizes to a long-tailed distribution of real objects remains to be seen.

The approaches in this section generate impressive results leveraging 2D priors from T2I models, but the 3D objects are plagued by the Janus problem. In the next section, we discuss methods that fine-tune T2I models with 3D priors to directly address this issue. These methods enable generation of objects within seconds, creating optimism that with continued refinement they may supersede the slower SDS loss optimization approaches.

6 Hybrid3D

Here, we discuss methods that share a common theme of using images as a bridge between text and 3D, in particular to enforce 3D consistency. These methods combine text-to-image (T2I) and then image-to-3D (I2-3D) in a pipeline. This strategy allows use of pretrained T2I models and I2-3D models. However, two issues need to be addressed: i) domain gap between output images from pretrained T2I model and input images to pretrained I2-3D model; and ii) how to incorporate information for improved conditioning of the output 3D model. The first issue can be addressed by fine-tuning the T2I model with additional data. The second issue, has been addressed using 3D-aware T2I models conditioned on camera pose information to encourage consistent 3D shape generation via SDS or photometric losses, or generating multi-view images that are then leveraged for 3D shape generation. Figure 10 provides an overview illustration.

We first describe Point-E [NJD*22] as it is a straightforward text-to-image-to-3D pipeline and a good basis for discussing the two-stage pipeline of T2I then I2-3D. We then focus on each stage: 1) fine-tuning T2I models to be 3D-aware [SWY*23; LCCT23; LZL*23b; LLL*23; LWV*23; LLZ*23; SCZ*23; LGL*23]; and 2) training I2-3D models to generate 3D objects from single or multi-

view images [NJD*22; HZG*23; LTZ*23; LXJ*23; LSC*23]. We also discuss methods focusing on both stages [LLL*23; LTZ*23; LXJ*23; LSC*23]. Note that we focus on methods that use text as the main input, but briefly touch on methods with image-based input as they can be used for text-to-3D generation in a pipeline similar to Point-E (text-to-image-to-3D).

Point-E [NJD*22]. Point-E adopts a direct approach using text-guided diffusion to first generate an image from text. The image serves as a condition for a point cloud diffusion model to do single-view 3D reconstruction. The Point-E approach first refines GLIDE [NDR*21] to generate images similar to synthetically rendered images. A special token is added to the text prompt to indicate that this is a rendered image so that at inference time the token can signal that the model should generate images similar to rendered ones. Subsequently, a point cloud diffusion model is trained conditioned on images from the initial stage. The model is conditioned by leveraging the entire token sequence of the CLIP image embedding from the generated image. Due to its straightforward combination of the two T2I and I2-3D stages, Point-E serves as a good canonical example of this strategy. We next discuss work focusing on improving each of these two stages.

6.1 3D-aware T2I

Work focusing on adapting T2I models to be 3D aware uses camera pose information or generates images from different views. These methods may leverage multi-view images and multi-modal images. The former involves mechanisms for allowing communication between multi-views during generation, while the latter uses images capturing information beyond RGB (e.g., normal, albedo, depth). Note that prior work has called the latter multi-modal image inputs ‘cross domain’ [LGL*23]. This additional information enables establishing correspondences between views and modalities, leading to improved 3D awareness. Concretely, multi-view methods adapt the T2I model to generate multiple views at once and modify the self attention layer of diffusion models to enable 3D attention across views [SWY*23; LZL*23b; LLL*23; LGL*23], tile multiple views in a single image [SCZ*23; LTZ*23], or use a unified 3D representation during generation [LLZ*23]. Multi-modal image methods generate several image modalities at once by leveraging cross domain attention [LLL*23; LGL*23]. We organize these methods into text conditioning and image conditioning.

Text conditioning. Several recent methods fine-tune existing T2I models to be 3D aware and take text as the input condition. MV-Dream [SWY*23] and SweetDreamer [LCCT23] aim to solve the Janus problem and improve object coherency by swapping the guidance models with T2I models fine-tuned with 3D priors. Direct2.5 [LZL*23b] devises an optimization scheme using the sparse multi-view geometric and color images generated from the fine-tuned T2I models to enable fast generation of objects in 10 seconds. UniDream [LLL*23] designs a pipeline that is able to generate relightable objects with PBR materials.

Image conditioning. We briefly mention image-to-image (I2I) methods that generate consistent novel views of a 3D object given a reference image as input, but do not go into detail as they are beyond the focus of this survey. A popular choice for fine-tuning is

Method	InCam	Input	Output Views	Image Res	Multi-View Comm	To 3D	Dataset	Stable Diff	GPU Days
MVDream [SWY*23]	abs	Text	4 RGB	256 ²	3D Attn	SDS	OV + LAION	v2.1	3 Days 32 A100
SweetDreamer [LCCT23]	abs	Text	1 CCM	64 ²	–	SDS	OV filt. (270K)	v2.1	–
Direct2.5 [LZL*23b]	abs	Text Image (Normal)	4 Normal 4 RGB	256 ²	3D Attn	Recon	OV filt.+COYO-400M filt. (500K, 65M)	v2.1	80 Hours 32 A100 (80G) 20 Hours 32 A100 (80G)
Instant3D [LTZ*23]	–	Text	1 2 × 2-tiled RGB	1024 ²	–	LRM*	OV filt. (10K)	SDXL	3 Hours 32 A100
UniDream [LLL*23]	abs	Text	4 Albedo + Normal	256 ²	3D + Cross Dom Attn	TRM + SDS	OV filt. (300k) + LAION	–	19 Hours 32 A800
Zero-1-to-3 [LWV*23]	rel	Image	1 RGB	256 ²	–	SJC	OV	IV	7 Days 8 A100 (80G)
SyncDreamer [LLZ*23]	rel	Image	16 RGB	256 ²	Cost Vol + Depth Attn	Recon	OV	Zero-1-to-3	4 Days 8 A100 (40G)
Zero123++ [SCZ*23]	–	Image	1 3 × 2-tiled RGB	960 × 640	–	–	OV	v2	–
Wonder3D [LGL*23]	rel	Image	6 RGB + Normal	256 ²	3D + Cross Dom Attn	Recon	OV (LVIS)	IV	3 Days 8 A800

Table 7: Summary of methods fine-tuning T2I/I2I diffusion models to be 3D aware. The ‘InCam’ column indicates whether **absolute** or **relative** camera parameters are used to condition the model where **rel** indicates camera parameters relative to the input reference image (vs absolute coordinates for **abs**). The ‘Input’ column indicates text or image conditioning for generation. The ‘Output Views’ column shows how many views are generated as the output of the fine-tuned model along with the type of image (CCM, RGB, Normal). ‘Image Res’, ‘Multi-View Comm’ indicate the output image resolution of the model and mechanism for communication between multi-views. In ‘To3D’ we indicate how the 3D object is obtained either through distillation (SDS, SJC), using images for reconstruction (i.e. photometric losses) or image to 3D models (TRM, LRM*). LRM* here indicates a large reconstruction model modified to accept multi-views as condition. ‘Dataset’ and ‘Stable Diff’ indicate the dataset used to fine-tune the model and which Stable Diffusion version the models are based on. Most of the models use some version of Objaverse (OV), with filtering often applied to discard low-quality assets. Note that ‘–’ in UniDream indicates that the specific model used for fine-tuning is not mentioned in their paper. ‘GPU Days’ shows the time and GPU resource used to train the models.

Image Variations (IV) [Lam] which fine-tunes Stable Diffusion to condition on images as input instead of text. Zero-1-to-3 [LWV*23] was one of the first to leverage priors in existing diffusion models for 3D reconstruction. They fine-tune the model to generate novel views of the input reference view with relative camera parameters of the target view as a condition, and apply the SJC loss for 3D generation. SyncDreamer [LLZ*23] builds on Zero-1-to-3 and adds a unified cost volume to generate multiple views at once, resulting in better consistency between views. Zero123++ [SCZ*23] proposes several training schemes to improve the stability of the fine-tuning process and output quality. Here, multiple views are tiled into one image to enable the generation of multiple frames at once. Wonder3D [LGL*23] proposed a domain attention to allow for the generation of multi-view RGB and normal images at once. With the additional geometric information from the normal maps, they train a NeuS [WLL*21] model to reconstruct meshes from the sparse views in 2 to 3 minutes.

6.2 Image-to-3D Models

Work targeting this stage uses three strategies to generate 3D objects from images: 1) distillation losses like SDS; 2) image reconstruction from sparse views with photometric losses; and 3) separate model generating 3D objects from conditioning images. Table 7 organizes work using these three strategies to obtain 3D models from fine-tuned T2I diffusion models.

Distillation losses. Methods using distillation losses (like in No3D) produce objects of higher quality, with the optimization gradually transforming a 3D representation over many iterations with different camera views sampled, allowing for more intricate details to be generated. However, the main limitation is the high generation time per 3D object, similarly to No3D approaches.

Image reconstruction from sparse views. Methods that optimize the 3D objects directly to match the images via reconstruction methods are much faster (e.g., 10 seconds in Direct2.5). Typically this speed is obtained by having a limited number of sparse views.

The main drawback is that the sparse viewpoints make it hard to generate more complex objects due to occlusions. Also, eliminating inconsistencies between views is hard, which can create artifacts in the extracted 3D object.

Trained image to 3D models. Due to increasing 3D data, there is growing interest in training a separate image to 3D model. This requires running inference on the model which is usually much faster (e.g., 20 seconds in Instant3D) than optimization. However, these models need larger capacities to get high quality 3D objects leading to higher resource requirements. Also, depending on the dataset used to train the model, generalizability may be limited. A popular alternative to diffusion models are transformer-based methods called Large Reconstruction Models (LRM) [HZG*23]. In LRM, a transformer predicts triplane NeRF representations with conditioning injected through cross attention from a single input image. Both Instant3D [LTZ*23] and UniDream [LLL*23] use LRM and modify it to accept multiple views at once. Their pipeline is similar to Point-E, except a text to multi-view image network is fine-tuned. One-2-3-45 [LXJ*23] and One-2-3-45++ [LSC*23] leverage SparseNeuS [LLW*22] and LAS-Diffusion [ZPW*23] respectively as backbones for multi-view image to 3D generation.

6.3 Discussion

Some challenges in using these ‘Hybrid3D’ methods revolve around what datasets to use for fine-tuning and what camera views to select for rendering. One problem with using Objaverse [DSS*23; DLW*23] the largest available synthetic 3D object dataset is that object textures appear toy-like and flat making generated objects less photorealistic. SweetDreamer addresses this by using normal map images and an unmodified T2I model for texture optimization. UniDream generates albedo and normal maps, and then generating 3D objects with another reconstruction model with a Stable Diffusion model used to optimize for photorealistic PBR materials. The choice of camera viewpoints used to fine-tune the model is also important. For example, MVDream samples cameras



Figure 11: Example input and output from Set-the-Scene [CRM*23]. On the left, bounding boxes indicate the location of each object as well as their local accompanying text. On the right the outputs resulting from different global text prompts are shown. Visuals reproduced from Cohen-Bar et al. [CRM*23].

from the upper hemisphere of the object meaning artifacts may be visible when looking at the bottom of the object. Methods utilizing direct reconstruction may choose fixed viewpoints (e.g., Direct2.5). However, depending on the type of object being generated important details may be occluded.

Looking at the dataset column in Table 7 we see different strategies for filtering datasets. Instant3D adopts an interesting strategy by training an SVM on manually labeled examples to filter the dataset. In general, Objaverse is noisy and can hurt model performance if used in entirety. However, how to best filter the dataset while not hurting the generalizability of models is not well studied.

The GPU days column in Table 7 shows these models are expensive to train. Diffusion models especially require large batch sizes for training. Naive scaling to generate more dense views leads to prohibitive GPU requirements. It would be interesting to explore how large the base diffusion models need to be to learn generation of multi-view images. For example, is it possible to use smaller distilled models like BK-SDM [KSCC23] to fine-tune these methods?

Another promising future direction is exhibited by DMV3D [XTL*23], where the model is designed to generate consistent multi-view images. The multi-view diffusion model uses a transformer model from LRM to predict an internal NeRF during the denoising process and encourage consistency between the multi-view images being denoised.

7 Generating multi-object scenes with diffusion guidance

The work discussed thus far in this survey mainly focuses on generating a single object from an input text prompt. While it is possible to use more complex prompts describing a multi-object scene, the methods we described often fail to generate a coherent scene composition incorporating multiple objects.

There is a long line of prior work on 3D scene layout generation, which typically assumes the presence of a 3D shape database

from which objects are retrieved and composited into a scene layout. Some seminal work in this vein, and guided by input text descriptions was done in Chang et al. [CSM14]. An excellent survey covering the earlier 3D scene layout work in detail is provided by Chaudhuri et al. [CRW*20]. In this survey, we focus on generation of 3D scenes built on top of the methodology presented thus far, which offers the advantage of generating scene layouts composed of objects each of which is also generated conditioned on the input description. There has been limited work on this latter strategy to text-to-3D scene.

In contrast to the object-centric approaches discussed earlier in the survey, it is also possible to generate an entire scene without explicitly modeling individual objects. These methods typically rely on text-to-image (T2I) models to generate partial views of the scene based on the text, and use outpainting methods to generate additional views. These views are then fused together by using depth prediction. Here, we describe these different lines of work that all aim to generate more complex 3D scene-level outputs.

7.1 Compositional Generation

The papers in this section aim to generate a scene consisting of multiple objects, mostly assuming that a layout of objects is given as a condition in the form of bounding boxes and the accompanying text prompts. An example of this from Set-the-Scene [CRM*23] is shown in Figure 11. More recent works [VCK23; GLC*24; ZWS*23] drop the assumption that the layout is given and attempt to both predict the layout (scale, position, and rotation) for each object as well as generate the individual objects.

Set-the-Scene. Cohen-Bar et al. [CRM*23] composes individual Latent-NeRF models for each of the object bounding boxes in the input layout condition. They propose an interleaved training scheme where individual NeRFs as well as the entire scene where all the NeRFs are rendered is trained with the SDS loss alternately. This allows for the optimization to focus on local objects as well as making sure that individual objects synergies with each other in the global frame. Compared to their baseline Latent-NeRF which fails to generate complex scenes their method is able to generate coherent scenes with individual objects. As the objects are all individual NeRFs, they can be rearranged and placed into a scene layout through appropriate spatial transformations.

CompoNeRF. Lin et al. [LBL*23] similarly uses multiple Latent-NeRF models for each object bounding box. However, instead of simply adding the density and color components of each NeRF during volumetric rendering they add additional global and local MLPs to further refine density and colors in the rendered global frame. This helps to make the objects more consistent with each other in the scene. The individual and global SDS losses are added and optimized in conjunction.

Comp3D. Po and Wetzstein [PW23] offers a different approach instead of training multiple NeRF models. They calculate the classifier-free guidance for each individual prompt of the objects in the scene. Then, using the rendered bounding boxes as segmentation masks the gradients for each region are masked and combined to be optimized. This has the benefits of less memory usage

compared to the above methods, as only one NeRF model is being trained. This model is based on the Score Jacobian Chaining (SJC) strategy described in Section 5.2.

Layout Generation. The above methods assume that bounding box layouts are given. Recent work addresses layout generation together with object generation. CG3D [VCK23] uses a probabilistic graphical model (PGM) to sample objects and estimate their interactions with each other. The text description is converted to a scene-graph which is used to instantiate the PGM. The scene is generated by optimizing SDS losses to generate objects (each represented as a set of 3D Gaussian) and a combination of SDS loss and physical constraints (e.g. gravity and contact) loss for object interactions (e.g. the placement of objects relative to each other). The integration of physical constraint losses allows for plausible generation of objects that are *on* or *in* another object. Another promising direction is the use of LLMs for layout generation. GraphDreamer [GLC*24] generates a scene graph using LLMs such as ChatGPT. The scene is then optimized using several SDS losses on individual objects, the entire scene and objects with relations (edge pairs in the scene graph). SceneWiz3D [ZWS*23] similarly uses an LLM to generate objects of interest in a text prompt and then uses an off-the-shelf text-to-3D model to generate initial 3D objects. A particle swarm optimization algorithm with CLIP similarity then optimizes the object locations, followed by optimization with VSD losses for the objects and the environment.

7.2 RGBD Fusion for Scenes

While distillation losses like SDS have been successful in object-centric generation, crafting detailed scenes with complex object compositions remains challenging. The research discussed in this section harnesses the priors learned by T2I models for scene generation. In contrast to prior work relying on the SDS loss to slowly distill knowledge from T2I models. They integrate images predicted from T2I models and a depth estimation network to perform RGBD fusion, thus generating a 3D mesh or NeRF representation for the entire scene.

SceneScape [FAKD23]. The SceneScape approach is the first to tackle the task of perpetual view generation given text as input. It aims to create a video of a consistent 3D scene using a provided text and camera path. To achieve this, SceneScape employs two pretrained models: Stable Diffusion (SD) inpainting and a depth estimation model [RBK21]. The process starts by generating an initial image from the SD model using the text prompt, complemented by depth predictions. From this, an initial mesh is established. For each subsequent frame, the mesh is rendered based on the updated camera position, leading to a frame with gaps. These gaps are filled using the inpainting model, while the depth estimation model offers the necessary depth predictions. The updated image and depth data is then used to refine the mesh for subsequent frames. After each frame, the decoder of both the inpainting and depth models is fine tuned and reset for frame consistency. SceneScape generates videos of diverse scenes with intricate details. Yet, the method is subject to limitations. Over time, errors can accumulate and this technique struggles particularly with outdoor scenes.

Text2Room [HCO*23]. The Text2Room approach has a compa-

table strategy to SceneScape, merging texture and geometric data over several frames to create a 3D mesh for room generation from text. A Stable Diffusion inpainting model and a depth prediction mechanism iteratively refine frames and determining depths to continually enhance the mesh representation. To ensure depth consistency across frames, a specialized depth inpainting model is deployed. Moreover, to fine-tune alignment the scale and shift parameters of the predicted depth map are optimized, ensuring a more accurate match with the known depth derived from the mesh. Using the pixel depth values a point cloud is constructed where every four neighboring pixels are linked to produce two triangles. Further filtration removes faces that could lead to visual distortions. Unlike SceneScape’s continuous scene creation, indoor rooms possess complex structures at many spatial scales. This distinction amplifies the significance of camera trajectory choices, ensuring the generated rooms possess plausible structures, layouts, furnishings, and minimizing gaps in the geometry. Text2Room utilizes a two phase viewpoint selection technique. Initially, they leverage a set of predefined camera trajectories to establish the primary scene layout and furniture planing. In the subsequent phase, specific views are chosen to refine and fill in any geometric hole. The final mesh undergoes Poisson surface reconstruction [KBH06] to refine its form. While their approach successfully crafts comprehensive 3D scenes with detailed textures, some outputs may exhibit stretched geometric areas or overly smoothed regions.

Text2NeRF [ZLW*23]. The Text2NeRF approach also aims to solve the task of text to scene generation. Differing from the mesh-based approach of previous methods, it uses an implicit NeRF network for its 3D representation. They start by generating an initial image using Stable Diffusion, accompanied by depth predicted from a depth estimation model. Subsequently, by altering camera positions and projecting the initial image and depth with the corresponding camera parameters, they generate a support set of images and depths. These will naturally contain gaps. Both the initial and support are then used to initialize the NeRF model. Then, a new view is chosen to be rendered via NeRF. The rendered image and depth containing holes are filled in with Stable Diffusion inpainting, while the depth is estimated using the depth network. Depth alignment adopts a two stage approach: initially, by calculating the mean scale and distance disparities for alignment, and then by fine-tuning a neural network for nonlinear depth alignment. The NeRF model is further trained with the updated frame, with the cycle of rendering new views, inpainting and updates repeating. The main advantage of Text2NeRF lies in its use of an implicit representation such as NeRF, which avoids issues such as geometric stretching seen in Text2Room, notably in outdoor scene settings.

8 Editing

An important consideration in the text to 3D shape generation task is allowing the user to apply edit operations on an output 3D shape using text instructions. This functionality is clearly valuable in practical scenarios where the user desires specific changes in the output 3D shape. Enabling such intuitive editing operations constitutes a significant open research avenue. Here, we discuss recent work that focuses on text-based 3D editing by optimizing 3D shapes with CLIP (Section 8.1) or leveraging text-to-image models

for 3D editing (Section 8.2). As changing the appearance of an object while keeping the geometry the same is an important use case, we also discuss recent work in retexturing based on a text prompt (Section 8.3).

8.1 Shape Editing with CLIP

One way to edit 3D shapes given a text command is to optimize the 3D representation so that the rendered image and text match by using CLIP guidance.

CLIP-NeRF [WCH*22]. CLIP-NeRF employs CLIP’s guidance to manipulate 3D objects using text. Initially, a disentangled conditional NeRF model is trained. In this model, the input condition is separated into a shape code z_s and an appearance code z_a . Notably, the latter only influences color predictions, while the former is used to deform positional encodings within the NeRF network. The entire model’s optimization is carried out using a GAN loss.

For image or text-driven manipulations two distinct mappers, the shape mapper M_s and the appearance mapper M_a , are trained to utilize the CLIP text-image encodings to predict update codes, which are subsequently added into the shape and appearance codes of the conditional NeRF model. This ensures alignment with the given input text or image description. The mappers’ training is supervised by both the discriminator and CLIP similarity losses. Crucially, during this phase, only the mappers are trained while all previously trained modules remain frozen.

Text2Mesh [MBL*22]. Given a text prompt, Text2Mesh aims to stylize a 3D object mesh by adding 3D geometric detail and color. To do so, Text2Mesh takes the given shapes and optimizes displacement vectors for each mesh vertex and its color attribute. The training predominantly relies on CLIP similarity, with some regularization strategies to avert implausible geometry.

8.2 Scene Editing with Text-to-image Models

Recent advances in text-to-image (T2I) models can be leveraged for editing 3D objects or scenes. The T2I models can be used to provide edited versions of rendered images that are then reincorporated into the 3D representation [HTE*23; KSH*23]. Here we briefly describe a few works that use T2I models for editing of 3D objects or scenes.

SKED. Mikaeili et al. [MPCM23] provides a method for sketch and text guided editing of an object. The input comprises of multiple sketch-views consisting of closed shapes indicating regions for edit as well as the complementing text prompt. The base object is represented as an Instant-NGP NeRF model. The overall optimization is guided by the SDS loss to ensure the object is aligned with respect to the text prompt. However, to preserve the identity of the base object they propose two losses to regularize training. A preservation loss that penalizes density and color changes with respect to the original object especially for points distant from the designated edit areas, and a silhouette loss to make sure that the edit regions are being filled. The results show that this method can perform localized edits of 3D objects while preserving the original identity of the object.

Vox-E. Sella et al. [SFHA23] aims to perform local or global edits of a 3D object given a text prompt. They start off with a voxel grid NeRF model (ReLU Fields) initialized with the original object. They then train the model with the SDS loss and a volumetric regularization using a correlation loss between densities of the original and edited objects. While this objective produces an output aligned with the text prompt and structure that mostly preserves the original identity. It may be desirable for local edits to keep other regions unchanged. To address this, a spatial refinement method has been introduced. This method segments the locally-edited region from the modified object and merges it with the original. The process utilizes the cross-attention modules in the T2I model to generate attention maps with respect to the the text indicating the local edit as well as the rest of the text. Using the attention map, two additional NeRF models are trained. Subsequently, an energy minimization approach is employed to derive the segmentation mask used for merging. With this their method is able to perform global edits that can change the contents of the entire object, as well as more localized edits that only modify a small region.

Instruct-NeRF2NeRF. Haque et al. [HTE*23] is designed to modify NeRF scenes using textual inputs. The procedure alternates between updating input images and training the NeRF model. InstructPix2Pix is used to edit the images based on the provided text. These images include both the multi-view dataset used for NeRF training and images rendered with different camera parameters. This cycle of updating the dataset and then training the NeRF with the refreshed images is termed ‘Iterative Dataset Update’ or Iterative DU. Initially, the edits will be inconsistent, but with continued iterative updates, the NeRF and the rendered images progressively align and converge to achieve a uniformly consistent scene.

Instruct 3D-to-3D. Kamata et al. [KSH*23] similarly uses InstructPix2Pix for guiding the editing process of a NeRF scene with a text prompt. However, they instead leverage the modified score estimate in InstructPix2Pix to calculate the SDS loss which is used to optimize the scene. Together with a dynamic scaling scheme, the number of voxels in the DVGO model used for training is gradually reduced than increased again during training to facilitate global edits of the global structure then fine details.

Room Dreamer. Song et al. [SCX*23] transforms the scene texture of an input room scene given a text prompt while improving the geometry. They start off by rendering a cube map of images, depths and distance maps from the center of the input room scene. Then leveraging these and the text prompt as input they generate stylized images through a T2I model. Gaps in the scene not rendered by the cube map are outpainted and filled. The images are then used to optimize the textures of the original input scene. To address potential artifacts in the original scene, a geometric loss is incorporated, ensuring that both the rendered depth and the depth predicted by an estimation network remain consistent and smooth.

8.3 Texturing

One common edit operation on 3D shapes is re-texturing, which can help to create a variety of fine-grained surface appearance for the same shape geometry. The papers in this section utilize depth

conditioned image generation diffusion models for texture generation of 3D objects.

TEXTure. Richardson et al. [RMA*23] proposes a method for texture generation of 3D meshes using Stable Diffusion-based depth-to-image and inpainting models. Their method involves a progressive process of rendering depth images from different views of the 3D object. Then, they use the depth-to-image diffusion model to generate images based on the text prompt and depth. The image is written back to a texture atlas of the 3D object with UV mapping calculated by XAtlas. To ensure local and global consistency of the painted textures from different views, they use another meta-texture map to keep track of which regions of the texture are classified as “keep”, “refine” or “generate”. The “keep” regions are regions that are fixed during the update process as other views have already generated the texture for the region. The “refine” regions occur when the view used to paint the region is oblique resulting in high possibility of distortions. The “generate” regions have yet to be generated by other views. Given these segmented regions of the meta-texture, a mask is generated during the sampling steps of the image to keep regions marked as “keep” unchanged, as well as allowing for changes in “refine” and “generate” regions. The authors also find that alternating between the depth-to-image and inpainting models during the sampling steps help with consistency. Finally, the generated image is used to update the texture map for regions marked as “refine” and “generate”. They show that the generated textures can have higher quality as well as consistency compared to prior methods. Additionally, their method can be utilized for other tasks such as texture transfer, texture from images and texture editing. However, some drawbacks include global inconsistencies and the fixed viewpoint selection resulting in some geometries not being fully covered by the texture generation process.

Text2Tex. Chen et al. [CSL*23] is concurrent work to TEXTure and offers a similar strategy for generating textures by progressively painting textures from different views with a depth-to-image model. First, given a set of fixed viewpoints, they calculate similarity masks which indicates how perpendicular the view direction is to the face normals of the mesh. Using the similarity masks they segment into generation masks of “new”, “update”, “keep” and “ignore” regions similar to TEXTure. The “new” regions are inpainted from random noise, and the “update” regions have a smaller denoising strength. Other regions are fixed and masked respectively. Different from TEXTure, Text2Tex adds an additional texture refinement stage with an automatic viewpoint selection scheme. For the views sampled in this stage, a view heat is constructed with the generation mask indicating areas that may have artifacts in textures. They then dynamically select views with high view heat and update those using their algorithm. This helps to eliminate additional distortions in the texture that were not visible from the fixed camera viewpoints. This method generates high quality textures, with the automatic viewpoint selection helping to eliminate additional artifacts.

SceneTex Chen et al. [CLL*23] proposes an optimization-based approach for generating scene textures. They parameterize the scene appearance with a multi-resolution texture field with cross attention texture decoder. The implicit textures are optimized through a VSD loss and a diffusion model with depth prior as well as the in-

put text. This method generates textures with more detail and consistency with the text compared to prior methods, but comes at the cost of long optimization times per scene.

9 Evaluation

As text-to-3D shape generation methods develop, there will be an increasing need for systematic evaluation to assess the strengths and weaknesses of the different methods. Recent work [WYL*24] recognizes the need for automated evaluation of text-to-3D and proposes the use of large vision-language models such as GPT-v4 for automated evaluation. In this section, we start by identifying critical dimensions along which text-to-3D shape generation methods should be evaluated (Section 9.1), and then summarize current evaluation methodologies (Section 9.2).

9.1 Desiderata

Following work in text-to-image generation, it is critical to evaluate text-to-shape generation along the following dimensions which characterize the output.

Quality. This axis measures *how good are the generated shape?* The simplest way to think of this dimension is in terms of how realistic the generated shape appears to a human observer. With 3D shapes, there are additional properties of quality of the output shape including geometric consistency (does the shape exhibit the Janus problem, are there disconnected parts when there should not be any), mesh quality (is the mesh topology smooth where it should be, and does it have enough geometric detail in complex regions), as well as color and texture quality.

Fidelity. This dimension involves measurement of *do the generated shapes match the text prompt?* In text to 3D shape generation, it is important not only that that the shape is of high quality but that it also matches the text specification. Characterization of such fidelity measures requires computing the degree to which properties specified by the text are respected in the output 3D shape.

Diversity. This axis answers *do the generated shapes exhibit variety?* In other words, another important property for a good generative model is the ability to generate a diverse set of shapes. Given that a variety of shapes can satisfy the same text description (e.g. there are many chairs that match the description *black chair*), it is important that multiple, diverse shapes can be generated. However, it is difficult to know the full distribution of the space matching the text description and measuring whether the full set shapes is covered that distribution well is challenging.

Compositionality. Metrics measuring this dimension attempt to answer the question *can the method handle text describing different combinations of parts, attributes, and spatial relations?* Only recently has text-to-2D generation work started to more systematically investigate this aspect [PAL*21]. Measurement in this dimension requires systematic evaluation of generations for text description that exhibit the compositional nature of language (e.g. *black chair with gray legs* vs *gray chair with black legs*).

Speed and efficiency. In addition to assessing the above qualities, it is also important to compare the speed and memory resources

Type	Method	Train	Per-prompt	Device	Gen time
3DPT	TITG3SG [LWQF22]	yes	no	V100-32G	2.21s (24.44s)
3DPT	Shap-E [JN23]	yes	no	V100	13s
3DUT	TAPS3D [WWF*23]	yes	no	V100-32G	0.05s (7.09s)
CLIP-Guide	DreamFields [JMB*22]	no	yes	TPU cores	72m
CLIP-Guide	PureCLIPNeRF [LC22]	no	yes	GTx 2080ti	20m
Diff-Guide	DreamFusion [PJB23]	no	yes	TPUv4	90m
Diff-Guide	SJC [WDL*23]	no	yes	RTX A6000	25m
Diff-Guide	Prolific Dreamer [WLW*23]	no	yes	A100	several h
Diff-Guide	Magic3D [LGT*23]	no	yes	8x A100	40m
Diff-Guide	DreamGaussian [TRZ*23]	no	yes	V100	2m
Diff-Guide	GSGEN [CWL23]	no	yes	4x RTX 3090	30m
Diff-Guide	GaussianDreamer [YFW*23]	no	yes	RTX 3090	15m
Diff-Guide	Fantasia3D [CCJ23]	no	yes	8x RTX 3090	31m
Diff-Guide	MVDream [SWY*23]	no	yes	V100	1-1.5h
Diff-Guide	SweetDreamer [LCCT23]	no	yes	?4x V100	20m/1h
Hybrid3D	Point-E [NJD*22]	yes	no	A100	25s
Hybrid3D	Direct2.5 [LZL*23b]	yes	yes	A100	10s
Hybrid3D	Instant3D [LTZ*23]	yes	no	A100	20s

Table 8: Comparison of reported speed and memory needed for different methods. Numbers in parentheses indicate type to generate a mesh representation from the base 3D representation (neural fields or voxels for TITG3SG). Average per-shape inference time estimates are taken from the respective papers, except for TITG3SG which was from TAPS3D [WWF*23]. Point-E [NJD*22] gave a range for inference time depending on model size (ranging from 16s to 1.5m for 40M parameter to 1B), we take the 300M parameter condition that Shap-E [JN23] also uses for comparison.

required. Methods are typically measured in terms of wall clock time for training, generating a single output (i.e. inference), and the memory needs for either training or generation. For reference, Table 8 provides a summary of reported speed and memory consumption for methods in this survey.

9.2 Existing evaluation

Pioneering works such as Text2Shape [CCS*19] and Dream-Field [JMB*22] not only show that it is possible to generate 3D shapes from text, but also propose quantitative evaluation protocols that can be used. However, many followup works do not perform any quantitative evaluation and provide qualitative examples only. This is especially true for works that focus on proposing new guidance losses [WDL*23; WLW*23] and improved 3D representations [MRP*23; CCJJ23]. Below we organize the types of evaluation protocols in prior work, and outline directions for more comprehensive evaluation of text-to-3D shape generation.

User studies. As it is challenging to evaluate the output of a generative model, it is typical to use user studies that compare the output from different systems. One type of user study is A/B testing where users are asked to compare outputs from two systems. As the user response can depend on the specific question asked, it is common to pose several different questions to users to extract user judgements corresponding to different dimensions such as quality and fidelity. For instance, Tsalicoglou et al. [TMT*23] asked users preferences on natural colors, detailed textures, and visually preferred. However, this principle is not always followed. Some work, such as [LGT*23], only ask users to broadly judge the quality of the generated shape (they ask users to select the one that is more realistic).

Evaluation against ground-truth shape. Text2Shape [CCS*19]

evaluated generation results using several quantitative metrics that required access to a ground-truth shape: Intersection-over-Union (IoU), Earth Mover’s Distance (EMD), and classification accuracy (Class Acc). The metrics aimed to evaluate the geometric accuracy (IoU) and color (EMD) against a ground truth shape associated with the description, and in general whether the generated shape matches the class (table vs. chair in their case). These metrics are easy to compute algorithmically and quantify the underlying properties precisely. However, there is a strong assumption of only one ground truth output which is unrealistic and in tension with the desire for diversity in the output. In addition to the above, Text2Shape also evaluated their results using a variant of the Inception Score [SGZ*16] using a shape classifier. The Inception Score metric combines *quality* (can the classifier identify between the shapes) and *diversity* (do generated shapes exhibit the class distribution).

FID. Another way to measure the *quality* of a generated shape is to evaluate how natural the rendered images of the generated shapes appear to be to a neural architecture that was trained on image data. For instance, Tsalicoglou et al. [TMT*23] evaluated renderings using an FID metric based on CLIP. Other papers that use this FID metric to evaluate quality include Xu et al. [XWC*23] and Wei et al. [WWF*23].

Point cloud metrics. In addition to evaluating the quality of the renderings, it is also possible to attempt to measure the *quality* of the geometry of the shape using point clouds. Recently, point-based versions of the inception score [SGZ*16] (P-IS) and Frchet Distance [HRU*17] (P-FID) were introduced by Nichol et al. [NJD*22] and used in follow-up work [ZLC*23]. Similar metrics include the Frchet Point Cloud Distance (FDP) which has been used in the shape generation community [ADMG18]. This metric has also been used for evaluating text-to-shape generation [WWF*23].

Automated pairwise comparison. Given a set of shapes (targets and distractors) for a text description, these evaluation protocols train a neural evaluator to select the correct shape. The neural evaluator can then produce a confidence score for each shape. This evaluation protocol was introduced by ShapeGlot and trained with the ShapeGlot dataset [MCST22; CLT*23]. This evaluation strategy is used to compare two methods, given a shape generated by method 1 and shape generated by method 2. If the confidence score for the two methods is within a certain threshold (0.2), then the evaluator cannot determine which of the two classes the shape is from, and is confused. The method selected more often by the neural evaluator is said to perform better.

Retrieval model R-Precision. The CLIP R-Precision metric was introduced by Park et al. [PAL*21] and popularized for evaluation of text-to-3D shape generation in DreamFields [JMB*22]. R-Precision measures the fraction of generated shapes that are retrieved correctly using a retrieval model based on CLIP similarity of rendered views to the text prompt. DreamFields used a set of 153 text queries and corresponding generated shapes (two per query). If a specific CLIP is used in the optimization, then the CLIP R-Precision is more meaningful with a different CLIP encoder (i.e. different backbone). Papers that use this evaluation protocol in-

clude Jain et al. [JMB*22], Lee and Chang [LC22], Poole et al. [PBJM23], Mohammad Khalid et al. [MXBP22], Tsalicoglou et al. [TMT*23], and Xu et al. [XWC*23]. This metric is an approximation of the *fidelity* of the description to the generated shape.

Shape-text score (ST-S). Another metric that measures the *fidelity* of the shape to the text is the shape-text (ST-S) score. Given an aligned shape-text space, it is possible to compare methods by computing the similarity between the input text and the generated 3D shape. Zhao et al. [ZLC*23] measured ST-S using ULIP [XGX*23] and their own aligned space SITA [ZLC*23].

Pretrained LVLM. The rise of Large Vision-Language Models (LVLM) is enabling automated metrics that act as a ‘proxy’ for human judgement. [WYL*24] demonstrated that GPT-V4 can evaluate the quality of generated 3D shapes through an A2C test for the LVLM. Given two generated 3D assets, they provide the LVLM with 2D renderings for each, (arranged as a grid of multi-view images), paired with text instructions describing the criteria to judge the renderings (different prompts and different rendering styles are used depending on the criteria of interest). The LVLM outputs whether the left or right asset is better, together with an analysis. Given a set of text prompts, and a set of models, the pairwise rankings are then combined to form an overall score by using the Elo score (commonly used for chess rankings). This strategy enables automated evaluation of criteria such as the fidelity (or *alignment*) of the text to the 3D shape, *plausibility* of the geometry, *geometry details*, *texture details*, and *diversity* of the shapes.

10 Discussion

This survey has summarized and categorized work on text-to-3D shape generation. Despite the explosion of interest in this area, text-to-3D research is still far behind text-to-image generation with many challenges and opportunities. Below, we conclude by outlining some promising directions for future investigation.

Data scale. With the development of ever-larger 3D datasets such as Objaverse [DSS*23] and Objaverse-XL [DLW*23], there are increasing amounts of 3D shape data available to the research community. This trend is likely to continue along with better 3D scanning techniques, and easier 3D content design tools. This data will likely be highly valuable for training better generative 3D models. Moreover, while this data will not necessarily be naturally paired with text descriptions, advances in image captioning will allow generation of text captions at scale. Thus, we anticipate further opportunities in developing better methods that can leverage both paired and unpaired 3D and text.

Hierarchical and part-based generation. As was apparent from this survey’s section on scene-level generation, much work remains to be done to enable generating hierarchical compositions of scenes from high-quality objects, or similarly, hierarchical compositions of objects from parts. Progress in this direction will enable fine-grained editability, since many editing operations revolve around properties of specific parts, or specific objects within a scene. Moreover, generation of animated 3D content whether at the object or scene scale, will likely benefit from such hierarchical methods, as many motions are typically characterized well through rigid

transformations of parts or whole objects. This latter generation of dynamic 3D content is a prominent “grand challenge” in this area.

Focus on language. Much of the development in text-to-3D generation has been driven by advances in 3D representations and generative models from other domains, text to image being the prominent example. While this transfer of knowledge led to an explosion of interest, an important and currently under-studied area is generation that better matches language to the generated shapes in a fine-grained manner (e.g., respecting detailed part and material properties). In addition, questions such as what kinds of natural language are handled well by particular methods remain unanswered. A key challenge for future work is shape generation that respects the fine-grained compositionality of the input language (e.g., "chair with black arms and red seat). In order to make progress in this direction, the community will also need to devote focus on evaluation protocols that can quantify progress in these directions.

Improved speed and memory. Methods that rely on 3D data typically require large amounts of compute and memory to train, but are generally fast at inference time. In comparison, methods that do not rely on 3D data typically rely on per-prompt optimization which is very slow, and currently highly impractical for real-world deployment. Thus, another promising direction for future work is developing strategies to improve the efficiency, speed, and memory consumption characteristics of text-to-3D shape generation. At the time of writing this survey, there have been exciting recent developments that leverage 3D Gaussian Splatting [KKLD23] as backbones for the SDS loss [TRZ*23; CWL23; YFW*23], leading to significant improvements in training and rendering speeds.

Conclusion. We hope that this survey will catalyze further work in text-to-3D shape generation, and enable researchers to advance the state of the art. Progress in this direction has the potential to democratize 3D content creation by enabling people to turn their imagination into high-quality 3D assets, and to iteratively design and control these assets for a variety of application domains.

Acknowledgments. This work was funded in part by a CIFAR AI Chair, a Canada Research Chair and NSERC Discovery Grant. We thank Ali Mahdavi-Amiri for helpful comments and discussions.

References

- [ACB17] ARJOVSKY, MARTIN, CHINTALA, SOUMITH, and BOTTOU, LÉON. “Wasserstein generative adversarial networks”. *International conference on machine learning*. PMLR. 2017, 214–223. arXiv: 1701.07875 [stat.ML] 6.
- [ADMG18] ACHLIOPTAS, PANOS, DIAMANTI, OLGA, MITLIAGKAS, IOANNIS, and GUIBAS, LEONIDAS. “Learning representations and generative models for 3D point clouds”. *International conference on machine learning*. 2018, 40–49. arXiv: 1707.02392 [cs.CV] 21.
- [AFH*19] ACHLIOPTAS, PANOS, FAN, JUDY, HAWKINS, ROBERT, et al. “ShapeGlot: Learning language for shape differentiation”. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, 8938–8947. arXiv: 1905.02925 [cs.CL] 5, 6, 8.

- [AHS*23] ACHLIOPTAS, PANOS, HUANG, IAN, SUNG, MINHYUK, et al. “ShapeTalk: A Language Dataset and Framework for 3D Shape Edits and Deformations”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 12685–12694. DOI: [10.1109/CVPR52729.2023.012205](https://doi.org/10.1109/CVPR52729.2023.012205).
- [AZS*23] ARMANDPOUR, MOHAMMADREZA, ZHENG, HUANGJIE, SADEGHIAN, ALI, et al. “Re-imagine the Negative Prompt Algorithm: Transform 2D Diffusion into 3D, alleviate Janus problem and Beyond”. *arXiv preprint arXiv:2304.04968* (2023). arXiv: [2304.04968](https://arxiv.org/abs/2304.04968) [cs.CV] 14.
- [BMT*21] BARRON, JONATHAN T, MILDENHALL, BEN, TANCIK, MATTHEW, et al. “Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields”. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, 5855–5864. arXiv: [2103.13415](https://arxiv.org/abs/2103.13415) [cs.CV] 3, 10.
- [BMV*22] BARRON, JONATHAN T., MILDENHALL, BEN, VERBIN, DOR, et al. “Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields”. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2022. arXiv: [2111.12077](https://arxiv.org/abs/2111.12077) [cs.CV] 3, 11, 12.
- [BNH*22] BALAJI, YOGESH, NAH, SEUNGJUN, HUANG, XUN, et al. “eDiff: Text-to-image diffusion models with an ensemble of expert denoisers”. *arXiv preprint arXiv:2211.01324* (2022). arXiv: [2211.01324](https://arxiv.org/abs/2211.01324) [cs.CV] 4, 12, 13.
- [CCJJ23] CHEN, RUI, CHEN, YONGWEI, JIAO, NINGXIN, and JIA, KUI. “Fantasia3D: Disentangling geometry and appearance for high-quality text-to-3D content creation”. *International Conference on Computer Vision*. 2023. arXiv: [2303.13873](https://arxiv.org/abs/2303.13873) [cs.CV] 12, 13, 21.
- [CCS*19] CHEN, KEVIN, CHOY, CHRISTOPHER B, SAVVA, MANOLIS, et al. “Text2Shape: Generating shapes from natural language by learning joint embeddings”. *Asian Conference on Computer Vision*. 2019, 100–116. arXiv: [1803.08495](https://arxiv.org/abs/1803.08495) [cs.CV] 5–7, 21.
- [CFG*15] CHANG, ANGEL X, FUNKHOUSER, THOMAS, GUIBAS, LEONIDAS, et al. “ShapeNet: An information-rich 3D model repository”. *arXiv preprint arXiv:1512.03012* (2015). arXiv: [1512.03012](https://arxiv.org/abs/1512.03012) [cs.GR] 5.
- [CG22] CHANDRAMOULI, PARAMANAND and GANDIKOTA, KANCHANA VAISHNAVI. “LDEdit: Towards generalized text guided image manipulation via latent diffusion models”. *Proceedings of the British Machine Vision Conference (BMVC)*. Vol. 3. 2022. arXiv: [2210.02249](https://arxiv.org/abs/2210.02249) [cs.CV] 7.
- [CG23] CHAO, CHENG-KANG TED and GINGOLD, YOTAM. “Text-guided Image-and-Shape Editing and Generation: A Short Survey”. *arXiv preprint arXiv:2304.09244* (2023). arXiv: [2304.09244](https://arxiv.org/abs/2304.09244) [cs.GR] 2.
- [CGD*22] COLLINS, JASMINE, GOEL, SHUBHAM, DENG, KENAN, et al. “ABO: Dataset and benchmarks for real-world 3D object understanding”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 21126–21136. arXiv: [2110.06199](https://arxiv.org/abs/2110.06199) [cs.CV] 6.
- [CLL*23] CHEN, DAVE ZHENYU, LI, HAOXUAN, LEE, HSIN-YING, et al. “Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors”. *arXiv preprint arXiv:2311.17261* (2023). arXiv: [2311.17261](https://arxiv.org/abs/2311.17261) [cs.CV] 20.
- [CLT*23] CHENG, YEN-CHI, LEE, HSIN-YING, TULYAKOV, SERGEY, et al. “SDFusion: Multimodal 3D shape completion, reconstruction, and generation”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 4456–4465. arXiv: [2212.04493](https://arxiv.org/abs/2212.04493) [cs.CV] 5–7, 21.
- [CRM*23] COHEN-BAR, DANA, RICHARDSON, ELAD, METZER, GAL, et al. “Set-the-Scene: Global-Local Training for Generating Controllable NeRF Scenes”. *arXiv preprint arXiv:2303.13450* (2023). arXiv: [2303.13450](https://arxiv.org/abs/2303.13450) [cs.CV] 17.
- [CRW*20] CHAUDHURI, SIDDHARTHA, RITCHIE, DANIEL, WU, JIAJUN, et al. “Learning generative models of 3D structures”. *Computer Graphics Forum* 39 (2020), 643–666. DOI: [10.1111/cgf.14020](https://doi.org/10.1111/cgf.14020) 2, 17.
- [CSL*23] CHEN, DAVE ZHENYU, SIDDIQUI, YAWAR, LEE, HSIN-YING, et al. “Text2tex: Text-driven texture synthesis via diffusion models”. *arXiv preprint arXiv:2303.11396* (2023). arXiv: [2303.11396](https://arxiv.org/abs/2303.11396) [cs.CV] 20.
- [CSM14] CHANG, ANGEL, SAVVA, MANOLIS, and MANNING, CHRISTOPHER D. “Learning spatial knowledge for text to 3D scene generation”. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, 2028–2038. DOI: [10.3115/v1/D14-1217](https://doi.org/10.3115/v1/D14-1217).
- [CWL23] CHEN, ZILONG, WANG, FENG, and LIU, HUAPING. “Text-to-3D using Gaussian Splatting”. *arXiv preprint arXiv:2309.16585* (2023). arXiv: [2309.16585](https://arxiv.org/abs/2309.16585) [cs.CV] 4, 12–14, 21, 22.
- [CXG*22] CHEN, ANPEI, XU, ZEXIANG, GEIGER, ANDREAS, et al. “TensorRF: Tensorial radiance fields”. *European Conference on Computer Vision*. Springer. 2022, 333–350. arXiv: [2203.09517](https://arxiv.org/abs/2203.09517) [cs.CV] 3, 12.
- [DLW*23] DEITKE, MATT, LIU, RUOSHI, WALLINGFORD, MATTHEW, et al. “Objaverse-XL: A universe of 10m+ 3D objects”. *arXiv preprint arXiv:2307.05663* (2023). arXiv: [2307.05663](https://arxiv.org/abs/2307.05663) [cs.CV] 6, 16, 22.
- [DN21] DHARIWAL, PRAFULLA and NICHOL, ALEXANDER. “Diffusion models beat gans on image synthesis”. *Advances in neural information processing systems* 34 (2021), 8780–8794. arXiv: [2105.05233](https://arxiv.org/abs/2105.05233) [cs.LG] 4.
- [DSB16] DINH, LAURENT, SOHL-DICKSTEIN, JASCHA, and BENGIO, SAMY. “Density estimation using real NVP”. *arXiv preprint arXiv:1605.08803* (2016). arXiv: [1605.08803](https://arxiv.org/abs/1605.08803) [cs.LG] 8.
- [DSS*23] DEITKE, MATT, SCHWENK, DUSTIN, SALVADOR, JORDI, et al. “Objaverse: A universe of annotated 3D objects”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 13142–13153. arXiv: [2212.08051](https://arxiv.org/abs/2212.08051) [cs.CV] 6, 16, 22.
- [FAKD23] FRIDMAN, RAFAIL, ABECASIS, AMIT, KASTEN, YONI, and DEKEL, TALI. “SceneScape: Text-driven consistent scene generation”. *arXiv preprint arXiv:2302.01133* (2023). arXiv: [2302.01133](https://arxiv.org/abs/2302.01133) [cs.CV] 18.
- [FYT*22] FRIDOVICH-KEIL, SARA, YU, ALEX, TANCIK, MATTHEW, et al. “Plenoxels: Radiance fields without neural networks”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 5501–5510. arXiv: [2112.05131](https://arxiv.org/abs/2112.05131) [cs.CV] 3, 12.
- [FZC*22] FU, RAO, ZHAN, XIAO, CHEN, YIWEN, et al. “ShapeCrafter: A recursive text-conditioned 3D shape generation model”. *Advances in Neural Information Processing Systems* 35 (2022), 8882–8895. arXiv: [2207.09446](https://arxiv.org/abs/2207.09446) [cs.CV] 5, 6.
- [GAA*17] GULRAJANI, ISHAAN, AHMED, FARUK, ARJOVSKY, MARTIN, et al. “Improved training of Wasserstein GANs”. *Advances in neural information processing systems* 30 (2017). arXiv: [1704.00028](https://arxiv.org/abs/1704.00028) [cs.LG] 6.
- [GCX*20] GAO, JUN, CHEN, WENZHEN, XIANG, TOMMY, et al. “Learning deformable tetrahedral meshes for 3D reconstruction”. *Advances In Neural Information Processing Systems* 33 (2020), 9936–9947. arXiv: [2011.01437](https://arxiv.org/abs/2011.01437) [cs.CV] 3, 13.
- [GLC*24] GAO, GE, LIU, WEIYANG, CHEN, ANPEI, et al. “GraphDreamer: Compositional 3D Scene Synthesis from Scene Graphs”. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2024. arXiv: [2312.00093](https://arxiv.org/abs/2312.00093) [cs.CV] 17, 18.
- [GSW*22] GAO, JUN, SHEN, TIANCHANG, WANG, ZIAN, et al. “GET3D: A generative model of high quality 3D textured shapes learned from images”. *Advances In Neural Information Processing Systems* 35 (2022), 31841–31854. arXiv: [2209.11163](https://arxiv.org/abs/2209.11163) [cs.CV] 9.
- [HAK23] HONG, SUSUNG, AHN, DONGHOON, and KIM, SEUNGRYONG. “Debiasing scores and prompts of 2D diffusion for robust text-to-3D generation”. *Advances in Neural Information Processing Systems* (2023). arXiv: [2303.15413](https://arxiv.org/abs/2303.15413) [cs.CV] 14.

- [HCO*23] HÖLLEIN, LUKAS, CAO, ANG, OWENS, ANDREW, et al. “Text2room: Extracting textured 3D meshes from 2D text-to-image models”. *International Conference on Computer Vision*. 2023. arXiv: [2303.11989 \[cs.CV\]](#) 18.
- [HDL17] HA, DAVID, DAI, ANDREW M, and LE, QUOC V. “HyperNetworks”. *International Conference on Learning Representations*. 2017, 24–26. arXiv: [1609.09106 \[cs.LG\]](#) 7.
- [HJA20] HO, JONATHAN, JAIN, AJAY, and ABBEEL, PIETER. “Denoising diffusion probabilistic models”. *Advances in neural information processing systems* 33 (2020), 6840–6851. arXiv: [2006.11239 \[cs.LG\]](#) 4.
- [HLZH21] HONG, YINING, LI, QING, ZHU, SONG-CHUN, and HUANG, SIYUAN. “VLGrammar: Grounded grammar induction of vision and language”. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, 1665–1674. arXiv: [2103.12975 \[cs.CV\]](#) 5.
- [HMC17] HAEUSSER, PHILIP, MORDVINTSEV, ALEXANDER, and CREMERS, DANIEL. “Learning by association—A versatile semi-supervised training method for neural networks”. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, 89–98. arXiv: [1706.00909 \[cs.CV\]](#) 6.
- [HRU*17] HEUSEL, MARTIN, RAMSAUER, HUBERT, UNTERTHINER, THOMAS, et al. “GANs trained by a two time-scale update rule converge to a local nash equilibrium”. *Advances in neural information processing systems* 30 (2017). arXiv: [1706.08500 \[cs.LG\]](#) 21.
- [HS22] HO, JONATHAN and SALIMANS, TIM. “Classifier-free diffusion guidance”. *arXiv preprint arXiv:2207.12598* (2022). arXiv: [2207.12598 \[cs.LG\]](#) 4, 6.
- [HSW*21] HU, EDWARD J, SHEN, YELONG, WALLIS, PHILLIP, et al. “LoRA: Low-rank adaptation of large language models”. *arXiv preprint arXiv:2106.09685* (2021). arXiv: [2106.09685 \[cs.CL\]](#) 12.
- [HTE*23] HAQUE, AYAAN, TANCIK, MATTHEW, EFROS, ALEXEI A, et al. “Instruct-NeRF2NeRF: Editing 3D scenes with instructions”. *International Conference on Computer Vision*. 2023. arXiv: [2303.12789 \[cs.CV\]](#) 19.
- [HZG*23] HONG, YICONG, ZHANG, KAI, GU, JIUXIANG, et al. “LRM: Large reconstruction model for single image to 3D”. *arXiv preprint arXiv:2311.04400* (2023). arXiv: [2311.04400 \[cs.CV\]](#) 15, 16.
- [JGMR23] JONES, R KENNY, GUERRERO, PAUL, MITRA, NILOY J, and RITCHIE, DANIEL. “ShapeCoder: Discovering Abstractions for Visual Programs from Unstructured Primitives”. *ACM Transactions on Graphics (TOG), Proc. SIGGRAPH* (2023). arXiv: [2305.05661 \[cs.GR\]](#) 8.
- [JMB*22] JAIN, AJAY, MILDENHALL, BEN, BARRON, JONATHAN T, et al. “Zero-shot text-guided object generation with dream fields”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 867–876. arXiv: [2112.01455 \[cs.CV\]](#) 10, 21, 22.
- [JN23] JUN, HEEWOO and NICHOL, ALEX. “Shap-E: Generating conditional 3D implicit functions”. *arXiv preprint arXiv:2305.02463* (2023). arXiv: [2305.02463 \[cs.CV\]](#) 5, 7, 13, 21.
- [KAAL22] KARRAS, TERO, AITTALA, MIKA, AILA, TIMO, and LAINE, SAMULI. “Elucidating the design space of diffusion-based generative models”. *Advances in Neural Information Processing Systems* 35 (2022), 26565–26577. arXiv: [2206.00364 \[cs.LG\]](#) 4.
- [KBH06] KAZHDAN, MICHAEL, BOLITHO, MATTHEW, and HOPPE, HUGUES. “Poisson surface reconstruction”. *Proceedings of the fourth Eurographics symposium on Geometry processing*. Vol. 7. 2006. DOI: [10.2312/SGP/SGP06/061-070](#) 18.
- [KKLD23] KERBL, BERNHARD, KOPANAS, GEORGIOS, LEIMKÜHLER, THOMAS, and DRETTAKIS, GEORGE. “3D Gaussian splatting for real-time radiance field rendering”. *ACM Transactions on Graphics (ToG)* 42.4 (2023), 1–14. arXiv: [2308.04079 \[cs.GR\]](#) 4, 13, 22.
- [KKY22] KIM, GWANGHYUN, KWON, TAESUNG, and YE, JONG CHUL. “DiffusionCLIP: Text-guided diffusion models for robust image manipulation”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 2426–2435. arXiv: [2110.02711 \[cs.CV\]](#) 7.
- [KPCL23] KATZIR, OREN, PATASHNIK, OR, COHEN-OR, DANIEL, and LISCHINSKI, DANI. “Noise-free score distillation”. *arXiv preprint arXiv:2310.17590* (2023). arXiv: [2310.17590 \[cs.CV\]](#) 14.
- [KRWM22] KARNEWAR, ANIMESH, RITSCHER, TOBIAS, WANG, OLIVER, and MITRA, NILOY. “ReLU fields: The little non-linearity that could”. *ACM SIGGRAPH Conference Proceedings*. 2022, 1–9. arXiv: [2205.10824 \[cs.CV\]](#) 3.
- [KSCC23] KIM, BO-KYEONG, SONG, HYOUNG-KYU, CASTELLS, THIBAUT, and CHOI, SHINKOOK. “BK-SDM: A Lightweight, Fast, and Cheap Version of Stable Diffusion”. *arXiv preprint arXiv:2305.15798* (2023). arXiv: [2305.15798 \[cs.CV\]](#) 17.
- [KSH*23] KAMATA, HIROMICHI, SAKUMA, YUIKO, HAYAKAWA, AKIO, et al. “Instruct 3D-to-3D: Text Instruction Guided 3D-to-3D conversion”. *arXiv preprint arXiv:2303.15780* (2023). arXiv: [2303.15780 \[cs.CV\]](#) 19.
- [KW13] KINGMA, DIEDERIK P and WELLING, MAX. “Auto-encoding variational bayes”. *arXiv preprint arXiv:1312.6114* (2013). arXiv: [1312.6114 \[stat.ML\]](#) 6.
- [KYNS23] KOO, JUIL, YOO, SEUNGWOO, NGUYEN, MINH HIEU, and SUNG, MINHYUK. “SALAD: Part-level latent diffusion for 3D shape generation and manipulation”. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, 14441–14451. arXiv: [2303.12236 \[cs.CV\]](#) 7, 8.
- [Lam] LAMBDA LABS. *Stable Diffusion Image Variations - A Hugging Face Space by LambdaLabs*. <https://huggingface.co/lambdalabs/sd-image-variations-diffusers> 16.
- [LBL*23] LIN, YIQI, BAI, HAOTIAN, LI, SHIJA, et al. “CompoNeRF: Text-guided multi-object compositional NeRF with editable 3D scene layout”. *arXiv preprint arXiv:2303.13843* (2023). arXiv: [2303.13843 \[cs.CV\]](#) 17.
- [LC22] LEE, HAN-HUNG and CHANG, ANGEL X. “Understanding pure clip guidance for voxel grid nerf models”. *arXiv preprint arXiv:2209.15172* (2022). arXiv: [2209.15172 \[cs.CV\]](#) 10, 21, 22.
- [LC98] LORENSEN, WILLIAM E and CLINE, HARVEY E. “Marching cubes: A high resolution 3D surface construction algorithm”. *Seminal graphics: pioneering efforts that shaped the field*. ACM SIGGRAPH, 1998, 347–353. DOI: [10.1145/37402.37422](#) 13.
- [LCCT23] LI, WEIYU, CHEN, RUI, CHEN, XUELIN, and TAN, PING. “SweetDreamer: Aligning Geometric Priors in 2D Diffusion for Consistent Text-to-3D”. *arXiv preprint arXiv:2310.02596* (2023). arXiv: [2310.02596 \[cs.CV\]](#) 14–16, 21.
- [LDC*23] LI, YUHAN, DOU, YISHUN, CHEN, XUANHONG, et al. “3DQD: Generalized Deep 3D Shape Prior via Part-Discretized Diffusion Process”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. arXiv: [2303.10406 \[cs.CV\]](#) 5–7.
- [LDL*23] LIU, ZHENGZHE, DAI, PENG, LI, RUIHUI, et al. “ISS: Image as stepping stone for text-guided 3D shape generation”. *International Conference on Learning Representations*. 2023. arXiv: [2209.04145 \[cs.CV\]](#) 9.
- [LDR*22] LUGMAYR, ANDREAS, DANELLJAN, MARTIN, ROMERO, ANDRES, et al. “Repaint: Inpainting using denoising diffusion probabilistic models”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 11461–11471. arXiv: [2201.09865 \[cs.CV\]](#) 7.
- [LDZL23] LI, MUHENG, DUAN, YUEQI, ZHOU, JIE, and LU, JIWEN. “Diffusion-SDF: Text-to-shape via voxelized diffusion”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 12642–12651. arXiv: [2212.03293 \[cs.CV\]](#) 5–7.

- [LGL*23] LONG, XIAOXIAO, GUO, YUAN-CHEN, LIN, CHENG, et al. “Wonder3D: Single image to 3D using cross-domain diffusion”. *arXiv preprint arXiv:2310.15008* (2023). arXiv: [2310.15008 \[cs.CV\]](#) 15, 16.
- [LGT*23] LIN, CHEN-HSUAN, GAO, JUN, TANG, LUMING, et al. “Magic3D: High-resolution text-to-3D content creation”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 300–309. arXiv: [2211.10440 \[cs.CV\]](#) 12, 13, 21.
- [LHK*20] LAINE, SAMULI, HELSTEN, JANNE, KARRAS, TERO, et al. “Modular primitives for high-performance differentiable rendering”. *ACM Transactions on Graphics (TOG)* 39.6 (2020), 1–14. arXiv: [2011.03277 \[cs.GR\]](#) 4.
- [LLJ22] LUO, TIANGE, LEE, HONGLAK, and JOHNSON, JUSTIN. “Neural Shape Compiler: A Unified Framework for Transforming between Text, Point Cloud, and Program”. *Transactions on Machine Learning Research* (2022). arXiv: [2212.12952 \[cs.CV\]](#) 5, 8.
- [LLL*23] LIU, ZEXIANG, LI, YANGGUANG, LIN, YOUTIAN, et al. “UniDream: Unifying Diffusion Priors for Relightable Text-to-3D Generation”. *arXiv preprint arXiv:2312.08754* (2023). arXiv: [2312.08754 \[cs.CV\]](#) 15, 16.
- [LLW*22] LONG, XIAOXIAO, LIN, CHENG, WANG, PENG, et al. “SparseNeuS: Fast generalizable neural surface reconstruction from sparse views”. *European Conference on Computer Vision*. Springer. 2022, 210–227. arXiv: [2206.05737 \[cs.CV\]](#) 16.
- [LLZ*23] LIU, YUAN, LIN, CHENG, ZENG, ZIJIAO, et al. “SyncDreamer: Generating Multiview-consistent Images from a Single-view Image”. *arXiv preprint arXiv:2309.03453* (2023). arXiv: [2309.03453 \[cs.CV\]](#) 15, 16.
- [LM18] LI, KE and MALIK, JITENDRA. “Implicit maximum likelihood estimation”. *arXiv preprint arXiv:1809.09087* (2018). arXiv: [1809.09087 \[cs.LG\]](#) 6.
- [LRLJ23] LUO, TIANGE, ROCKWELL, CHRIS, LEE, HONGLAK, and JOHNSON, JUSTIN. “Scalable 3D Captioning with Pretrained Models”. *arXiv preprint arXiv:2306.07279* (2023). arXiv: [2306.07279 \[cs.CV\]](#) 5, 6.
- [LSC*23] LIU, MINGHUA, SHI, RUOXI, CHEN, LINGHAO, et al. “One-2-3-45++: Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion”. *arXiv preprint arXiv:2311.07885* (2023). arXiv: [2311.07885 \[cs.CV\]](#) 15, 16.
- [LSK*23] LIU, MINGHUA, SHI, RUOXI, KUANG, KAIMING, et al. “OpenShape: Scaling Up 3D Shape Representation Towards Open-World Understanding”. *arXiv preprint arXiv:2305.10764* (2023). arXiv: [2305.10764 \[cs.CV\]](#) 5, 6.
- [LTZ*23] LI, JIAHAO, TAN, HAO, ZHANG, KAI, et al. “Instant3D: Fast text-to-3D with sparse-view generation and large reconstruction model”. *arXiv preprint arXiv:2311.06214* (2023). arXiv: [2311.06214 \[cs.CV\]](#) 15, 16, 21.
- [LWQF22] LIU, ZHENGZHE, WANG, YI, QI, XIAOJUAN, and FU, CHIWING. “Towards implicit text-guided 3D shape generation”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 17896–17906. arXiv: [2203.14622 \[cs.CV\]](#) 5–7, 21.
- [LWV*23] LIU, RUOSHI, WU, RUNDI, VAN HOORICK, BASILE, et al. “Zero-1-to-3: Zero-shot one image to 3D object”. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, 9298–9309. arXiv: [2303.11328 \[cs.CV\]](#) 15, 16.
- [LXJ*23] LIU, MINGHUA, XU, CHAO, JIN, HAIAN, 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). arXiv: [2306.16928 \[cs.CV\]](#) 15, 16.
- [LXZ*23] LORRAINE, JONATHAN, XIE, KEVIN, ZENG, XIAOHUI, et al. “ATT3D: Amortized Text-to-3D Object Synthesis”. *International Conference on Computer Vision*. 2023. arXiv: [2306.07349 \[cs.LG\]](#) 14, 15.
- [LYL*23] LIANG, YIXUN, YANG, XIN, LIN, JIANTAO, et al. “LucidDreamer: Towards High-Fidelity Text-to-3D Generation via Interval Score Matching”. *arXiv preprint arXiv:2311.11284* (2023). arXiv: [2311.11284 \[cs.CV\]](#) 14.
- [LZL*23a] LI, MING, ZHOU, PAN, LIU, JIA-WEI, et al. “Instant3D: Instant Text-to-3D Generation”. *arXiv preprint arXiv:2311.08403* (2023). arXiv: [2311.08403 \[cs.CV\]](#) 14, 15.
- [LZL*23b] LU, YUANXUN, ZHANG, JINGYANG, LI, SHIWEI, et al. “Direct2.5: Diverse Text-to-3D Generation via Multi-view 2.5 D Diffusion”. *arXiv preprint arXiv:2311.15980* (2023). arXiv: [2311.15980 \[cs.CV\]](#) 15, 16, 21.
- [LZW*23] LI, CHENGHAO, ZHANG, CHAONING, WAGHWASE, ATISH, et al. “Generative AI meets 3D: A Survey on Text-to-3D in AIGC Era”. *arXiv preprint arXiv:2305.06131* (2023). arXiv: [2305.06131 \[cs.CV\]](#) 2.
- [MBL*22] MICHEL, OSCAR, BAR-ON, ROI, LIU, RICHARD, et al. “Text2mesh: Text-driven neural stylization for meshes”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 13492–13502. arXiv: [2112.03221 \[cs.CV\]](#) 10, 19.
- [MCST22] MITTAL, PARITOSH, CHENG, YEN-CHI, SINGH, MANEESH, and TULSIANI, SHUBHAM. “AutoSDF: Shape priors for 3D completion, reconstruction and generation”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 306–315. arXiv: [2203.09516 \[cs.CV\]](#) 5–7, 21.
- [MESK22] MÜLLER, THOMAS, EVANS, ALEX, SCHIED, CHRISTOPH, and KELLER, ALEXANDER. “Instant Neural Graphics Primitives with a Multiresolution Hash Encoding”. *ACM Transactions on Graphics (TOG) - SIGGRAPH* 41.4 (July 2022), 102:1–102:15. arXiv: [2201.05989 \[cs.CV\]](#) 3, 12, 13.
- [MHS*21] MENG, CHENLIN, HE, YUTONG, SONG, YANG, et al. “SDEdit: Guided image synthesis and editing with stochastic differential equations”. *arXiv preprint arXiv:2108.01073* (2021). arXiv: [2108.01073 \[cs.CV\]](#) 6.
- [MO14] MIRZA, MEHDI and OSINDERO, SIMON. “Conditional generative adversarial nets”. *arXiv preprint arXiv:1411.1784* (2014). arXiv: [1411.1784 \[cs.LG\]](#) 6.
- [MON*19] MESCHEDER, LARS, OECHSLE, MICHAEL, NIEMEYER, MICHAEL, et al. “Occupancy networks: Learning 3D reconstruction in function space”. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, 4460–4470. arXiv: [1812.03828 \[cs.CV\]](#) 8.
- [MPCM23] MIKAEILI, ARYAN, PEREL, OR, COHEN-OR, DANIEL, and MAHDAVI-AMIRI, ALI. “SKED: Sketch-guided Text-based 3D Editing”. *arXiv preprint arXiv:2303.10735* (2023). arXiv: [2303.10735 \[cs.CV\]](#) 19.
- [MRP*23] METZER, GAL, RICHARDSON, ELAD, PATASHNIK, OR, et al. “Latent-NeRF for shape-guided generation of 3D shapes and textures”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 12663–12673. arXiv: [2211.07600 \[cs.CV\]](#) 12, 21.
- [MST*20] MILDENHALL, BEN, SRINIVASAN, PRATUL P, TANCIK, MATTHEW, et al. “NeRF: Representing scenes as neural radiance fields for view synthesis”. *Proceedings of European Conference on Computer Vision*. 2020, 405–421. arXiv: [2003.08934 \[cs.CV\]](#) 3.
- [MXBP22] MOHAMMAD KHALID, NASIR, XIE, TIANHAO, BELILOVSKY, EUGENE, and POPA, TIBERIU. “CLIP-mesh: Generating textured meshes from text using pretrained image-text models”. *SIGGRAPH Asia conference papers*. 2022, 1–8. arXiv: [2203.13333 \[cs.CV\]](#) 10, 22.
- [NDR*21] NICHOL, ALEX, DHARIWAL, PRAFULLA, RAMESH, ADITYA, et al. “GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models”. *arXiv preprint arXiv:2112.10741* (2021). arXiv: [2112.10741 \[cs.CV\]](#) 15.

- [NJD*22] NICHOL, ALEX, JUN, HEEWOO, DHARIWAL, PRAFULLA, et al. "Point-E: A system for generating 3D point clouds from complex prompts". *arXiv preprint arXiv:2212.08751* (2022). arXiv: 2212.08751 [cs.CV] 5, 7, 13–15, 21.
- [NMOG20] NIEMEYER, MICHAEL, MESCHEDER, LARS, OECHSLE, MICHAEL, and GEIGER, ANDREAS. "Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, 3504–3515. arXiv: 1912.07372 [cs.CV] 9.
- [PAL*21] PARK, DONG HUK, AZADI, SAMANEH, LIU, XIHUI, et al. "Benchmark for compositional text-to-image synthesis". *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. 2021. URL: <https://openreview.net/pdf?id=bKBhQhPeKaF> 20, 21.
- [PFS*19] PARK, JEONG JOON, FLORENCE, PETER, STRAUB, JULIAN, et al. "DeepSDF: Learning continuous signed distance functions for shape representation". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, 165–174. arXiv: 1901.05103 [cs.CV] 7.
- [PBJM23] POOLE, BEN, JAIN, AJAY, BARRON, JONATHAN T., and MILDENHALL, BEN. "DreamFusion: Text-to-3D using 2D Diffusion". *International conference on machine learning*. 2023. arXiv: 2209.14988 [cs.CV] 6, 11–14, 21, 22.
- [PLZZ24] PAN, ZIJIE, LU, JIACHEN, ZHU, XIATIAN, and ZHANG, LI. "Enhancing High-Resolution 3D Generation through Pixel-wise Gradient Clipping". *International Conference on Learning Representations*. 2024. arXiv: 2310.12474 [cs.CV] 14.
- [PW23] PO, RYAN and WETZSTEIN, GORDON. "Compositional 3D scene generation using locally conditioned diffusion". *arXiv preprint arXiv:2303.12218* (2023). arXiv: 2303.12218 [cs.CV] 17.
- [RBK21] RANFTL, RENÉ, BOCHKOVSKIY, ALEXEY, and KOLTUN, VLADLEN. "Vision transformers for dense prediction". *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, 12179–12188. arXiv: 2103.13413 [cs.CV] 18.
- [RBL*22] ROMBACH, ROBIN, BLATTMANN, ANDREAS, LORENZ, DOMINIK, et al. "High-resolution image synthesis with latent diffusion models". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, 10684–10695. arXiv: 2112.10752 [cs.CV] 4, 6, 7, 11, 12.
- [RDN*22] RAMESH, ADITYA, DHARIWAL, PRAFULLA, NICHOL, ALEX, et al. "Hierarchical text-conditional image generation with CLIP latents". *arXiv preprint arXiv:2204.06125* 1.2 (2022), 3. arXiv: 2204.06125 [cs.CV] 1, 10.
- [RFB15] RONNEBERGER, OLAF, FISCHER, PHILIPP, and BROX, THOMAS. "U-net: Convolutional networks for biomedical image segmentation". *Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, 234–241. arXiv: 1505.04597 [cs.CV] 4.
- [RGJ*23] RITCHIE, DANIEL, GUERRERO, PAUL, JONES, R KENNY, et al. "Neurosymbolic Models for Computer Graphics". *Computer Graphics Forum*. Vol. 42. Wiley Online Library. 2023, 545–568. arXiv: 2304.10320 [cs.GR] 4.
- [RKH*21] RADFORD, ALEC, KIM, JONG WOOK, HALLACY, CHRIS, et al. "Learning transferable visual models from natural language supervision". *International conference on machine learning*. 2021, 8748–8763. arXiv: 2103.00020 [cs.CV] 4, 7.
- [RMA*23] RICHARDSON, ELAD, METZER, GAL, ALALUF, YUVAL, et al. "TEXTure: Text-guided texturing of 3D shapes". *ACM SIGGRAPH Conference Proceedings*. 2023. arXiv: 2302.01721 [cs.CV] 20.
- [Rob92] ROBBINS, HERBERT E. "An empirical Bayes approach to statistics". *Breakthroughs in Statistics: Foundations and basic theory*. Springer, 1992, 388–394 11.
- [SCL*22] SANGHI, ADITYA, CHU, HANG, LAMBOURNE, JOSEPH G, et al. "CLIP-Forge: Towards zero-shot text-to-shape generation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 18603–18613. arXiv: 2110.02624 [cs.CV] 8, 9.
- [SCS*22] SAHARIA, CHITWAN, CHAN, WILLIAM, SAXENA, SAURABH, et al. "Photorealistic text-to-image diffusion models with deep language understanding". *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494. arXiv: 2205.11487 [cs.CV] 1, 4, 11, 12.
- [SCX*23] SONG, LIANGCHEN, CAO, LIANGLIANG, XU, HONGYU, et al. "RoomDreamer: Text-Driven 3D Indoor Scene Synthesis with Coherent Geometry and Texture". *arXiv preprint arXiv:2305.11337* (2023). arXiv: 2305.11337 [cs.CV] 19.
- [SCZ*23] SHI, RUOXI, CHEN, HANSHENG, ZHANG, ZHUOYANG, et al. "Zero123++: a single image to consistent multi-view diffusion base model". *arXiv preprint arXiv:2310.15110* (2023). arXiv: 2310.15110 [cs.CV] 15, 16.
- [SFHA23] SELLA, ETAI, FIEBELMAN, GAL, HEDMAN, PETER, and AVERBUCH-ELOR, HADAR. "Vox-E: Text-guided Voxel Editing of 3D Objects". *International Conference on Computer Vision*. 2023. arXiv: 2303.12048 [cs.CV] 19.
- [SFL*23] SANGHI, ADITYA, FU, RAO, LIU, VIVIAN, et al. "CLIP-Sculptor: Zero-Shot Generation of High-Fidelity and Diverse Shapes From Natural Language". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 18339–18348. arXiv: 2211.01427 [cs.CV] 8, 9.
- [SGY*21] SHEN, TIANCHANG, GAO, JUN, YIN, KANGXUE, et al. "Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis". *Advances in Neural Information Processing Systems*. 2021. arXiv: 2111.04276 [cs.CV] 3, 12, 13.
- [SGZ*16] SALIMANS, TIM, GOODFELLOW, IAN, ZAREMBA, WOJCIECH, et al. "Improved techniques for training GANs". *Advances in neural information processing systems* 29 (2016). arXiv: 1606.03498 [cs.LG] 21.
- [SJK*23] SEO, JUNYOUNG, JANG, WOOSEOK, KWAK, MIN-SEOP, et al. "Let 2D diffusion model know 3D-consistency for robust text-to-3D generation". *arXiv preprint arXiv:2303.07937* (2023). arXiv: 2303.07937 [cs.CV] 14.
- [Soh16] SOHN, KIHYUK. "Improved deep metric learning with multi-class N-pair loss objective". *Advances in neural information processing systems* 29 (2016). DOI: 10.5555/3157096.3157304 6.
- [SPX*22] SHI, ZIFAN, PENG, SIDA, XU, YINGHAO, et al. "Deep generative models on 3D representations: A survey". *arXiv preprint arXiv:2210.15663* (2022). arXiv: 2210.15663 [cs.CV] 2–4.
- [SSC22] SUN, CHENG, SUN, MIN, and CHEN, HWANN-TZONG. "Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 5459–5469. arXiv: 2111.11215 [cs.CV] 3, 12.
- [SSK*21] SONG, YANG, SOHL-DICKSTEIN, JASCHA, KINGMA, DIEDERIK P, et al. "Score-based generative modeling through stochastic differential equations". *International conference on learning representations*. 2021. arXiv: 2011.13456 [cs.LG] 4, 11.
- [SWMG15] SOHL-DICKSTEIN, JASCHA, WEISS, ERIC, MAHESWARANATHAN, NIRU, and GANGULI, SURYA. "Deep unsupervised learning using nonequilibrium thermodynamics". *International conference on machine learning*. PMLR. 2015, 2256–2265. arXiv: 1503.03585 [cs.LG] 4.
- [SWY*23] SHI, YICHUN, WANG, PENG, YE, JIANGLONG, et al. "MV-Dream: Multi-view diffusion for 3D generation". *arXiv preprint arXiv:2308.16512* (2023). arXiv: 2308.16512 [cs.CV] 14–16, 21.
- [TMT*23] TSALICOGLOU, CHRISTINA, MANHARDT, FABIAN, TONIONI, ALESSIO, et al. "TextMesh: Generation of Realistic 3D Meshes From Text Prompts". *arXiv preprint arXiv:2304.12439* (2023). arXiv: 2304.12439 [cs.CV] 12, 13, 21, 22.

- [TRZ*23] TANG, JIAXIANG, REN, JIAWEI, ZHOU, HANG, et al. “Dream-Gaussian: Generative gaussian splatting for efficient 3D content creation”. *arXiv preprint arXiv:2309.16653* (2023). arXiv: [2309.16653 \[cs.CV\]](#) 4, 12, 13, 21, 22.
- [TSB*22] THOMASON, JESSE, SHRIDHAR, MOHIT, BISK, YONATAN, et al. “Language grounding with 3D objects”. *Conference on Robot Learning*. 2022, 1691–1701. arXiv: [2107.12514 \[cs.CL\]](#) 5.
- [TWWZ23] TANG, BOSHI, WANG, JIANAN, WU, ZHIYONG, and ZHANG, LEI. “Stable Score Distillation for High-Quality 3D Generation”. *arXiv preprint arXiv:2312.09305* (2023). arXiv: [2312.09305 \[cs.CV\]](#) 14.
- [TYW23] TIAN, XI, YANG, YONG-LIANG, and WU, QI. “ShapeScaffold: Structure-Aware 3D Shape Generation from Text”. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, 2715–2724. DOI: [10.1109/ICCV51070.2023.00256](#) 7, 8.
- [VCK23] VILESOV, ALEXANDER, CHARI, PRADYUMNA, and KADAMBI, ACHUTA. “CG3D: Compositional Generation for Text-to-3D via Gaussian Splatting”. *arXiv preprint arXiv:2311.17907* (2023). arXiv: [2311.17907 \[cs.CV\]](#) 17, 18.
- [VV*17] VAN DEN OORD, AARON, VINYALS, ORIOL, et al. “Neural discrete representation learning”. *Advances in neural information processing systems* 30 (2017). arXiv: [1711.00937 \[cs.LG\]](#) 6.
- [WCH*22] WANG, CAN, CHAI, MENGLEI, HE, MINGMING, et al. “CLIP-NerF: Text-and-image driven manipulation of neural radiance fields”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 3835–3844. arXiv: [2112.05139 \[cs.CV\]](#) 19.
- [WDL*23] WANG, HAOCHEN, DU, XIAODAN, LI, JIAHAO, et al. “Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, 12619–12629. arXiv: [2212.00774 \[cs.CV\]](#) 12, 14, 21.
- [WFX*23] WANG, PEIHAO, FAN, ZHIWEN, XU, DEJIA, et al. “Stein-Dreamer: Variance Reduction for Text-to-3D Score Distillation via Stein Identity”. *arXiv preprint arXiv:2401.00604* (2023). arXiv: [2401.00604 \[cs.CV\]](#) 14.
- [WLL*21] WANG, PENG, LIU, LINGJIE, LIU, YUAN, et al. “NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction”. *arXiv preprint arXiv:2106.10689* (2021). arXiv: [2106.10689 \[cs.CV\]](#) 16.
- [WLW*23] WANG, ZHENGYI, LU, CHENG, WANG, YIKAI, et al. “ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation”. *Advances in Neural Information Processing Systems* (2023). arXiv: [2305.16213 \[cs.LG\]](#) 12, 14, 21.
- [WWF*23] WEI, JIACHENG, WANG, HAO, FENG, JIASHI, et al. “TAPS3D: Text-Guided 3D Textured Shape Generation from Pseudo Supervision”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 16805–16815. arXiv: [2303.13273 \[cs.CV\]](#) 9, 21.
- [WXF*23] WANG, PEIHAO, XU, DEJIA, FAN, ZHIWEN, et al. “Taming Mode Collapse in Score Distillation for Text-to-3D Generation”. *arXiv preprint arXiv:2401.00909* (2023). arXiv: [2401.00909 \[cs.CV\]](#) 14.
- [WYL*24] WU, TONG, YANG, GUANDAO, LI, ZHIBING, et al. “GPT-4V (ision) is a Human-Aligned Evaluator for Text-to-3D Generation”. *arXiv preprint arXiv:2401.04092* (2024). arXiv: [2401.04092 \[cs.CV\]](#) 20, 22.
- [WZY*24] WU, ZIKE, ZHOU, PAN, YI, XUANYU, et al. “Consistent3D: Towards Consistent High-Fidelity Text-to-3D Generation with Deterministic Sampling Prior”. *arXiv preprint arXiv:2401.09050* (2024). arXiv: [2401.09050 \[cs.CV\]](#) 14.
- [XGX*23] XUE, LE, GAO, MINGFEI, XING, CHEN, et al. “ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 1179–1189. arXiv: [2212.05171 \[cs.CV\]](#) 22.
- [XTL*23] XU, YINGHAO, TAN, HAO, LUAN, FUJUN, et al. “DMV3D: Denoising multi-view diffusion using 3D large reconstruction model”. *arXiv preprint arXiv:2311.09217* (2023). arXiv: [2311.09217 \[cs.CV\]](#) 17.
- [XTS*22] XIE, YIHENG, TAKIKAWA, TOWAKI, SAITO, SHUNSUKE, et al. “Neural fields in visual computing and beyond”. *Computer Graphics Forum*. Vol. 41. Wiley Online Library. 2022, 641–676. arXiv: [2111.11426 \[cs.CV\]](#) 3.
- [XWC*23] XU, JIALE, WANG, XINTAO, CHENG, WEIHAO, et al. “Dream3D: Zero-shot text-to-3D synthesis using 3D shape prior and text-to-image diffusion models”. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 20908–20918. arXiv: [2212.14704 \[cs.CV\]](#) 11, 21, 22.
- [YFW*23] YI, TAORAN, FANG, JIEMIN, WU, GUANJUN, et al. “GaussianDreamer: Fast Generation from Text to 3D Gaussian Splatting with Point Cloud Priors”. *arXiv preprint arXiv:2310.08529* (2023). arXiv: [2310.08529 \[cs.CV\]](#) 4, 12–14, 21, 22.
- [YGL*23] YU, XIN, GUO, YUAN-CHEN, LI, YANGGUANG, et al. “Text-to-3D with classifier score distillation”. *arXiv preprint arXiv:2310.19415* (2023). arXiv: [2310.19415 \[cs.CV\]](#) 14.
- [YXP*23] YUAN, HAOCHENG, XU, JING, PAN, HAO, et al. “CADTalk: An Algorithm and Benchmark for Semantic Commenting of CAD Programs”. *arXiv preprint arXiv:2311.16703* (2023). arXiv: [2311.16703 \[cs.CV\]](#) 8.
- [ZLC*23] ZHAO, ZIBO, LIU, WEN, CHEN, XIN, et al. “Michelangelo: Conditional 3D Shape Generation based on Shape-Image-Text Aligned Latent Representation”. *arXiv preprint arXiv:2306.17115* (2023). arXiv: [2306.17115 \[cs.CV\]](#) 5, 7, 21, 22.
- [ZLW*23] ZHANG, JINGBO, LI, XIAOYU, WAN, ZIYU, et al. “Text2NeRF: Text-Driven 3D Scene Generation with Neural Radiance Fields”. *IEEE Transactions on Visualization and Computer Graphics* (2023). arXiv: [2305.11588 \[cs.CV\]](#) 18.
- [ZLWT22] ZHENG, XINYANG, LIU, YANG, WANG, PENGSHUAI, and TONG, XIN. “SDF-StyleGAN: Implicit SDF-Based StyleGAN for 3D Shape Generation”. *Computer Graphics Forum*. Vol. 41. Wiley Online Library. 2022, 52–63. arXiv: [2206.12055 \[cs.CV\]](#) 11.
- [ZPW*23] ZHENG, XIN-YANG, PAN, HAO, WANG, PENG-SHUAI, et al. “Locally attentional SDF diffusion for controllable 3D shape generation”. *ACM Transactions on Graphics (TOG), Proc. SIGGRAPH* (2023). arXiv: [2305.04461 \[cs.CV\]](#) 16.
- [ZSME23] ZHOU, LINQI, SHIH, ANDY, MENG, CHENLIN, and ERMON, STEFANO. “DreamPropeller: Supercharge Text-to-3D Generation with Parallel Sampling”. *arXiv preprint arXiv:2311.17082* (2023). arXiv: [2311.17082 \[cs.CV\]](#) 14.
- [ZWS*23] ZHANG, QIHANG, WANG, CHAOYANG, SIAROHIN, ALIAKSANDR, et al. “SceneWiz3D: Towards Text-guided 3D Scene Composition”. *arXiv preprint arXiv:2312.08885* (2023). arXiv: [2312.08885 \[cs.CV\]](#) 17, 18.