Guiding 3D Digital Content Generation with Pre-Trained Diffusion Models

Jing Li[1¶], Zhengping Li[2¶*], Peizhe Jiang[3], Lijun Wang[4], Xiaoxue Li[5*], Yuwen Hao[6]
School of Information, North China University of Technology, Beijing, China 100144[1, 2, 4]
Shenzhen Renrenzhuang Technology Co., Ltd, Shenzhen, China 518000[3]
Beijing Key Laboratory of Disaster Rescue Medicine,
Medical Innovation Research Division of the Chinese PLA General Hospital, Beijing, China 100853[5, 6]
[¶]The authors have an equal contribution

***Abstract*—The production technology of 3D digital content involves multiple stages, including 3D modeling, simulation animation, visualization rendering, and perceptual interaction. It is not only the core technology supporting the creation of 3D digital content but also a key element in enhancing immersive application experiences in virtual reality and the metaverse. A primary focus in computer vision and computer graphics research has been on how to create 3D digital content that is efficient, convenient, controllable, and editable. Currently, producing high-quality 3D digital content still requires significant time and effort from a large number of designers. To address this challenge, leveraging artificial intelligence-generated methods to break down production barriers has emerged as an effective strategy. With the substantial breakthroughs achieved by diffusion models in the field of image generation, they also demonstrate tremendous potential in 3D digital content generation, potentially becoming a foundational model in this area. Recent studies have shown that diffusion model-based techniques for generating 3D digital content can significantly reduce production costs and enhance efficiency. Therefore, it is essential to summarize and categorize existing methods to facilitate further research. This paper systematically reviews 3D digital content generation methods, introducing related 3D representation techniques and focusing on 3D digital content generation schemes, algorithms, and pipeline based on diffusion models. We perform a horizontal comparison of different approaches in terms of generation speed and quality, deeply analyze existing challenges, and propose viable solutions. Furthermore, we thoroughly explore future research themes and directions in this domain, aiming to provide guidance and reference for subsequent research endeavors.**

***Keywords*—3D Digital content; computer vision; artificial intelligence; diffusion models; 3D representation**

I. Introduction

Humans describe the world through text, comprehend it through images, and experience and interact with it in a three-dimensional (3D) format. Therefore, generative models have found widespread application in numerous aspects of life, playing a significant role in advancing human society. Research in recent years has mainly focused on text generation [1], [2], [3], [4] and image generation [5], [6], [7], [8]. Text generation is typically used for language tasks such as translation and question-answering, while image generation often involves creating visuals based on textual prompts. The generation of 3D digital content has not yet achieved the extraordinary capabilities seen in the domains of text and image generation. Therefore, there is still a need to continue to promote related research on 3D digital content generation.

3D digital content is extensively utilized in fields such as film, architecture, virtual and augmented reality. However, the current mainstream production of 3D digital content relies heavily on 3D designers, leading to remarkably low production efficiency and high entry barriers. Consequently, employing artificial intelligence (AI) to generate 3D digital content can significantly enhance production efficiency, reduce industry barriers, and foster the development of related fields.

Zero-shot image models [9] are trained using hundreds of millions of graphics data, which is difficult to achieve in the 3D domain. Table I presents a comparison between the data volumes of mainstream 3D and 2D datasets. Conventional 3D digital content generation methodologies predominantly utilize 3D datasets for training specific generative models [10], [11]. The advantage of this method lies in its ability to generate 3D objects with consistent geometry. However, it is limited by the current lack of sufficiently large 3D datasets and the absence of efficient 3D digital content generation architectures, as well as the computational power needed for their training. Therefore, it is difficult for this 3D digital content generation method to achieve a breakthrough in the short term. In light of this, this paper focuses on using pre-trained diffusion models [7], [12], [13] to supervise the generation of 3D digital content.

TABLE I. Comparison of 3D Datasets and 2D Datasets

3D			2D	
Dataset	Full Mesh	Objects	Dataset	Images
ShapeNet [14]	✓	51K	ImageNet [15]	14M
AKB-48 [16]	✓	2K	COCO [17]	330K
OmniObject3D [18]	✓	6K	Open Image V7 [19]	9M
ScanObjectNN [20]		15K	Places [21]	10M
3D-Future[22]	✓	16K	LSUN [23]	59M

Diffusion models, trained on billions of image-text pairs, have propelled the latest advancements in text-to-image generation, demonstrating the capability to produce high-fidelity images under textual prompts [24], [25], [26], [27], [28]. Utilizing pre-trained diffusion models for generating 3D digital content [29], [30] significantly reduces computational power requirements and dependence on 3D datasets, thereby greatly enhancing the feasibility and efficiency of 3D digital content generation. This paper meticulously investigates and analyzes

*Corresponding authors.

methods for generating 3D digital content, focusing on two key aspects: diffusion model priors and 3D representations. The generation of 3D digital content is categorized into two types based on the task: text-to-3D [29], [30], [31], [32], [33], [34], [35], [36], [37] and image-to-3D [35], [38], [39], [40], [41]. To compare the strengths and limitations of each approach, this study conducts a horizontal comparison of different models in terms of efficiency and quality. This paper also explores the challenges associated with generating 3D digital content using pre-trained diffusion models and discusses potential solutions to these issues.

Our contributions are summarized as follows:

- This paper delivers an exhaustive review and investigation of methods for generating 3D digital content, with a foundation in diffusion models.
- A horizontal comparison and analysis are conducted in this paper to discern variations in efficiency and quality among different models.
- Several viable solutions are proposed in this paper to address the current challenges in generating 3D digital content using diffusion models.
- Potential future research directions in the field of 3D digital content generation, guided by diffusion models, are outlined in this paper.

Additionally, it is worth noting that there is currently a lack of universally recognized evaluation metrics for text-to-3D digital content generation. We currently assess quality solely through visual observation, which introduces a certain level of subjectivity. In the realm of image-to-3D digital content generation, we will employ image-based metrics to objectively evaluate the generated 3D digital content. Furthermore, due to limitations in laboratory conditions, all experiments in this paper were conducted using a single A40 GPU, and the results are presented accordingly.

This paper is organized as follows: Section II introduces the relevant background knowledge on 3D representation methods and diffusion models. Section III conducts a comprehensive analysis and study of the schemes, algorithms, and workflows for both text-to-3D and image-to-3D conversions. Section IV provides a holistic evaluation of existing 3D content generation approaches, analyzing the strengths and limitations of different methodologies. Section V explores the current challenges and proposes envisioned solutions. Finally, the paper concludes with a summary and presents our thoughts on future research directions and themes in this field.

II. Related Work

The generation of 3D digital content based on diffusion models principally involves two components: 3D representation and diffusion priors. DreamFusion [29] pioneered the integration of diffusion models into the task of 3D digital content generation. Subsequent studies in this domain have been categorized into two approaches based on their characteristics: optimization-based methods [42] and multi-view prediction-based methods [43], [44]. The focal point of research in this field has been centered on optimizing 3D representations or fine-tuning diffusion models.

A. 3D Representations

In the fields of computer graphics and computer vision, the 3D representation of objects encompasses various forms, including point clouds [45], [46], voxel grids [47], [48], meshes [49], [50], and implicit neural representations [51]. Each representation method has its distinct advantages and limitations, suitable for different types of 3D tasks. In research on 3D digital content generation based on diffusion models, Neural Radiance Fields (NeRF) [51] or 3D Gaussian Splatting [52] are commonly employed.

1) Neural radiance fields: NeRF uses a neural network to learn the continuous volume density and color of a scene [53]. Central to NeRF is the utilization of a Multi-Layer Perceptron to parametrically represent 3D objects, enabling high-quality synthesis of new viewpoint images. Theoretically, it can model shapes at any spatial resolution [54]. The MLP parameters, denoted as θ, take the camera pose c as input. The output comprises color and density. The process involves camera rays traversing the scene, generating a set of sample points along the ray path. The color and transparency of each sampled point on the ray are cumulatively processed to synthesize the color of each pixel. Subsequently, these colors and densities are utilized in volume rendering to generate the image $g(\theta, c)$. NeRF can learn from a series of 2D images taken from different angles and synthesize highly realistic new viewpoint images, which is crucial for achieving realistic 3D scene reconstruction.

2) 3D gaussian splatting: Structure-from-Motion (SfM) [55] can estimate point cloud distributions from a set of images using the COLMAP library. The work of 3D Gaussian Splatting starts with sparse SfM points, modeling the geometry as a set of 3D Gaussian functions. The fundamental idea of 3D Gaussian Splatting is to consider each point as the center of a Gaussian distribution. These points, rather than being isolated discrete entities, have a smooth, continuous weight distribution around them. Each point influences its surrounding area, quantified by a Gaussian function. Each 3D Gaussian is defined by the point's position, covariance matrix, and opacity α. Specifically, the point's position is the mean of the 3D Gaussian, the covariance matrix determines the shape of the 3D Gaussian, and the opacity α is used for splatting, with spherical harmonics (SH) [56], [57] representing color. The method uses adaptive Gaussian densification to control the number and density of Gaussians per unit volume. This approach overcomes the issues of slow rendering speed or compromised image quality in previous methods, enabling high-quality, real-time novel view synthesis at 1080p resolution.

B. Diffusion Models

Diffusion models consist of a forward process $q_{t,t\in[0,1]}$, and a reverse process $p_{t,t\in[0,1]}$. The forward process resembles a straightforward Brownian motion with time-varying coefficients [58]. Specifically, this process incrementally adds noise $\epsilon \in \mathcal{N}(0,\mathbf{I})$ to the original data x_0, thereby gradually transitioning the data distribution towards a Gaussian noise distribution [12], [59]. This step-by-step addition of noise effectively transforms the original data into a state that aligns with a predefined Gaussian distribution, laying the groundwork for the subsequent reverse process. Conversely, the reverse process employs a neural network to estimate the

noise added at each step of the forward process, progressively denoising the Gaussian distribution noise to ultimately restore the original data distribution. The distribution in the forward process is given by $q_t(x_t|x_0) := \mathcal{N}(\alpha_t x_0, \sigma_t^2 \mathbf{I})$ and $q_t(x_t) := \int q_t(x_t|x_0) q_0(x_0) \mathrm{d}x_0$. The coefficients α_t and σ_t are selected to regulate the proportion of original data and noise. At the onset of the forward process, $\sigma_0 \approx 0$, while at the end, $\sigma_1 \approx 1$, where $\alpha_t^2 = 1 - \sigma_t^2$ [60], [61]. This careful adjustment of coefficients ensures a gradual and controlled transformation of the data. The reverse process, through a noise prediction network $\epsilon_\phi(x_t, t)$, predicts the noise added at each forward step. The overall training is conducted by minimizing

$$\mathcal{L}_{\text{Diff}}(\phi, x_0) = \mathbb{E}_{t,\epsilon}[\omega(t)||\epsilon_\phi(\alpha_t x_0 + \sigma_t \epsilon, t) - \epsilon||_2^2] \quad (1)$$

where, $\omega(t)$ is a weighting function that depends on the timestep t. the noise prediction network can be used for approximating the score function of both q_t and p_t by $S_\phi(x_t, t) = -\epsilon_\phi(x_t, t)/\sigma_t$.

Incorporating textual control within diffusion models enhances the controllability of the generated content [8]. Since each image adheres to a specific distribution pattern, utilizing the information embedded within the text as a directive allows for the progressive denoising of Gaussian noise images, culminating in the generation of images that align with the textual information. This process specifically involves training an encoder and a decoder, where the encoder maps images to a latent space and the decoder reconstructs images from this latent space data. The textual prompts y are encoded using a text encoder $\tau_\theta(y)$ and are integrated into each step of the denoising process, which is trained by minimizing

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{t,\epsilon,y}[\omega(t)||\epsilon_\phi(x_t, t, \tau_\theta(y)) - \epsilon||_2^2] \quad (2)$$

By introducing conditions into the noise reconstruction process, controlled image generation is achieved. This methodology exhibits robustness in producing high-resolution images with intricate details while maintaining the semantic structure of the images [62].

III. Methodology

Diffusion models demonstrate extraordinary zero-shot capabilities in generating diverse images from textual descriptions. Fig. 1. demonstrates the ability of diffusion models to create multi-angular images using textual prompts.

Fig. 1. Generate multi-angle images based on text prompts.

Pre-trained diffusion models, having been trained with a vast array of internet data, have acquired an understanding of the distribution of images of most objects from various viewpoints [63]. By leveraging the geometric priors learned from natural images by large-scale diffusion models and integrating viewpoint control, fine-tuning these pre-trained models enables the generation of images from different perspectives. The viewpoint-conditioned diffusion models (Zero-1-to-3) [63] learn the relative control of camera perspectives using synthetic datasets, thereby facilitating the creation of novel views of the same object under specified camera transformations. Fig. 2. demonstrates the capability of the viewpoint-conditioned diffusion models to take a single-perspective image as input and generate images from diverse viewpoints.

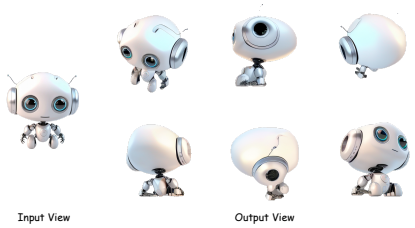

Fig. 2. Generate different perspective images from a single viewpoint image.

The specific steps for using diffusion models as a prior to guide the generation of 3D digital content are as follows: First, initialize a 3D model, then continuously modify the shape of the 3D model according to the prompt. Upon completion of the iterative process, the final 3D model, when rendered from any perspective, aligns consistently with the content described in the prompt.

A. Text-to-3D

The work on generating 3D digital content from textual prompts is built upon the foundations of text-to-image diffusion models [8], [26], [27], [28]. Given that the end product of diffusion models is an image, it's not feasible to directly use the results of diffusion models to supervise the generation of 3D digital content. However, it's possible to utilize the denoising process to guide this generation. The forward process of the diffusion model involves adding noise to the original data x_0 at timestep t, resulting in a noised image $\alpha_t x_0 + \sigma_t \epsilon$. During the reverse process, the noise prediction network estimates the noise ϵ added at each step, thus the denoised image can be represented as $x_\phi = [(\alpha_t x_0 + \sigma_t \epsilon) - \sigma_t \epsilon_\phi)]/\alpha_t$. This indicates that as long as the noise prediction is sufficiently accurate, the final image generated from Gaussian noise will also be accurate.

DreamFusion [29] employs NeRF as the 3D representation and utilizes a pre-trained text-to-image diffusion model as a critic. It achieves text-to-3D generation with impressive results through Score Distillation Sampling (SDS) loss. Specifically, the process involves rendering an image x_{render} from a given viewpoint c using the differentiable renderer $G(\theta, c)$. Here, G is a differentiable rendering function parameterized by θ, representing the parameters of the 3D object. Random amounts of noise are introduced into the rendered image $x_{render} := G(\theta, c)$ at various time steps t, resulting in $x_t = \alpha_t x_{render} +$

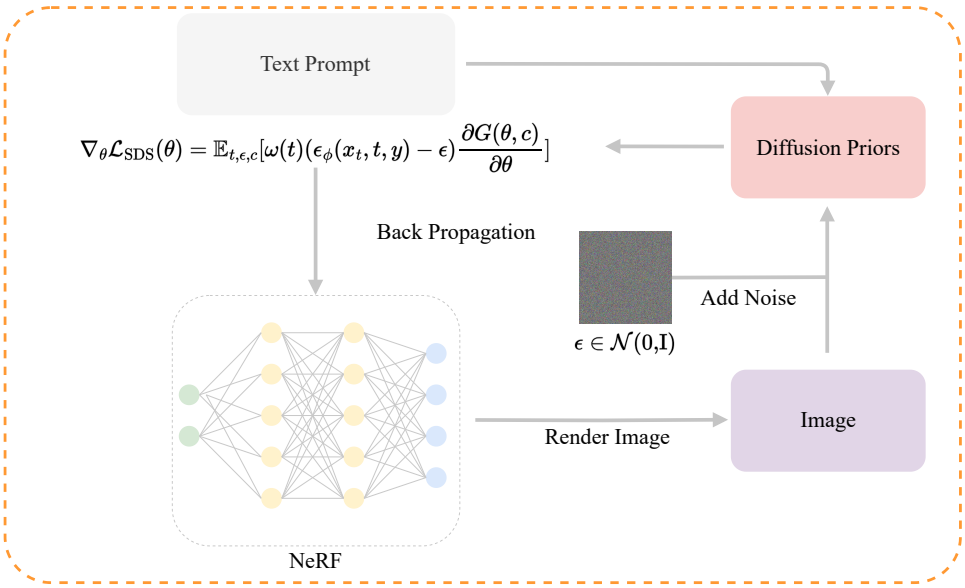

Fig. 3. A simplified framework for generating 3D digital content based on text prompts.

$\sigma_t \epsilon$ [30]. The pre-trained diffusion model predicts the sampling noise ϵ_ϕ given a noisy image x_t, noise time step t, and text embedding y. It provides a gradient direction to update the 3D volumetric parameters θ, with the overall gradient computed by the SDS function.

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\theta) = \mathbb{E}_{t,\epsilon,c}[\omega(t)(\epsilon_\phi(x_t,t,y)-\epsilon)\frac{\partial G(\theta,c)}{\partial \theta}] \quad (3)$$

Here, $\omega(t)$ is a weighting function. The scene model G and the diffusion model ϕ can be considered as modular components. It can be demonstrated that this loss fundamentally measures the similarity between the rendered images and textual prompts [40]. During the iterative process, the SDS loss backpropagates only to update the NeRF parameters θ, without altering the pre-trained diffusion model. As iterations progress, the 3D object gradually exhibits textures and geometric shapes that align with the textual prompt. The overall network architecture is succinctly illustrated in Fig. 3.

B. Image-to-3D

People possess the ability to envision the 3D structure of an object from a single image, a skill largely derived from the vast amount of prior knowledge accumulated through life experiences. Much of the past research has focused on reconstructing 3D models from multi-angle images [56], [57], [64], [65]. This approach is intuitive, as multiple viewpoints are essential for acquiring 3D information. However, 3D reconstruction from multi-angle images remains inefficient. This method requires the collection and acquisition of images from multiple angles, implying that it can only reconstruct objects that already exist in the real world. An interesting aspect is that in industries with a high demand for 3D digital assets, such as gaming, virtual reality, and animation, the focus lies on innovative 3D models rather than mere reproductions of the real world. Typically, the creation of an original 3D model involves numerous steps, as illustrated in Fig. 4.

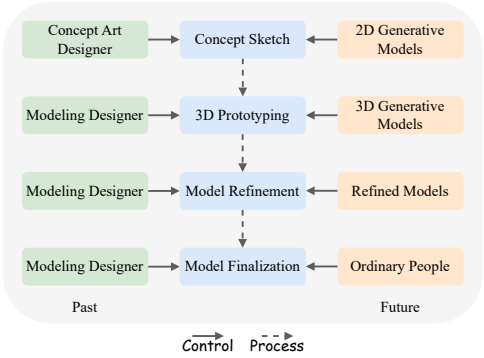

Fig. 4. 3D Model modeling process.

A promising approach to creating the requisite 3D models is through the generation of corresponding 3D models from a single image. While achieving controllability in diffusion models is a hot topic in further research [66], [67], [68], [69], there still lacks effective means to precisely control the images they generate. Consequently, the 3D models produced using text-to-3D methods may not always meet specific requirements. In other words, when inputting text prompts, no one can predict the structure of the 3D model until the result is generated. At this juncture, the task of image-to-3D conversion gains a significant advantage.

DreamFusion [29] achieves a text-to-3D generation method based on diffusion priors, demonstrating the exceptional capability of using diffusion priors to optimize NeRF. Related work [38], [41], [70] attempts to apply diffusion priors to single-image 3D generation. Owing to the fact that pre-trained diffusion models are primed with textual prompts, the approach for image-to-3D tasks diverges from that of text-to-3D tasks. Specifically, image-to-3D requires a process of textual inversion [68], differentiating it from the generation method used in text-to-3D tasks. A simplified network structure is illustrated in Fig. 5.

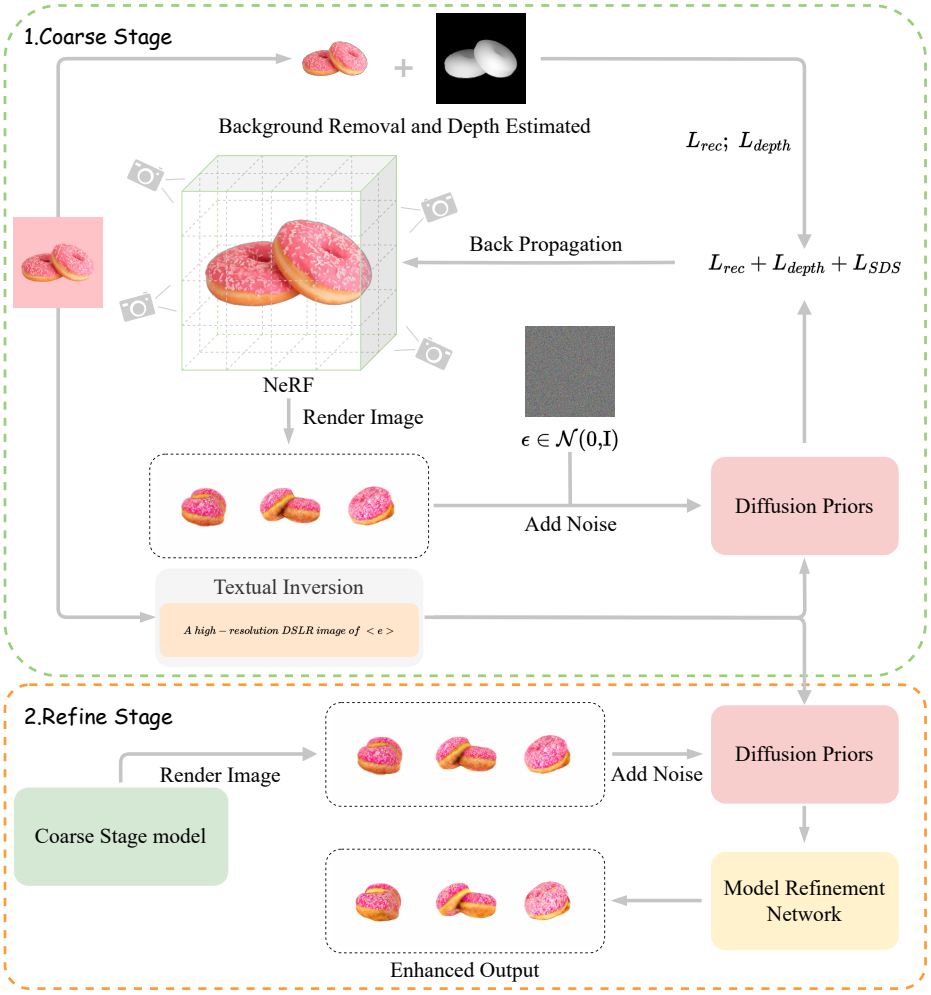

Fig. 5. A Two-stage framework for generating 3D digital content from a single image using diffusion priors.

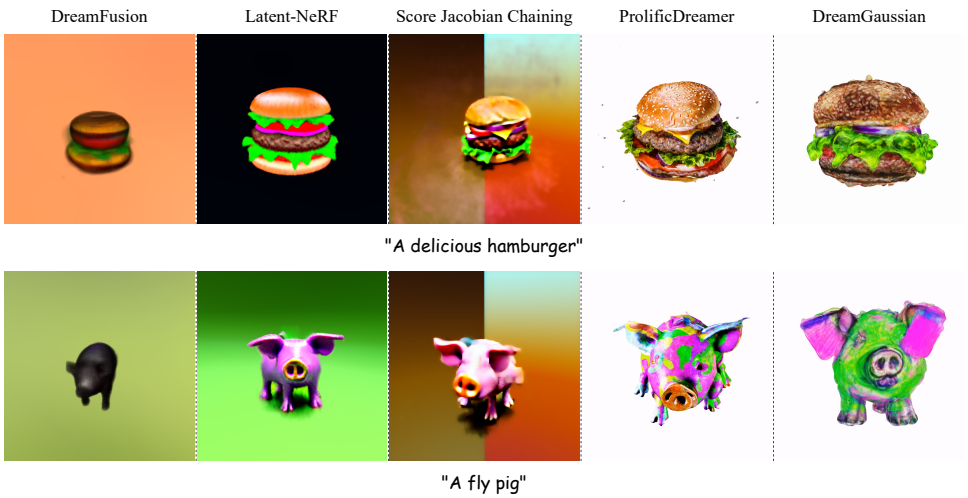

Fig. 6. Qualitative comparisons of 3D digital content generation from textual descriptions.

The generation of 3D digital content from a single image is typically a two-stage process. The primary task of the coarse stage is to establish the model's basic outline, followed by refinement in the refine stage. Specifically, the coarse stage begins with preprocessing such as background removal [71], textual inversion [68], and depth estimation [72], [73] of the reference image. Background removal focuses on isolating the main object for modeling, while textual inversion generates

corresponding textual descriptions to guide the diffusion prior. Depth estimation provides a prior for depth information, supervising subsequent model generation. The overall process starts with initializing a 3D model, rendering images from random angles with added Gaussian noise, and then using a diffusion model to optimize the 3D model through back propagation using SDS loss and a series of reference image losses.

1) Reference view reconstruction loss: To ensure consistency between rendered images $G_\theta(c)$ from reference viewpoints c and the reference images x_0 themselves, a reference view reconstruction loss is typically introduced at the reference viewpoints. This involves the use of Mean Squared Error (MSE) loss on the reference images and their masks.

$$\mathcal{L}_{rec} = \lambda_{rgb}||\mathrm{M} \odot (x_0 - G_\theta(c))||_2^2 + \lambda_{mask}||\mathrm{M} - M(G_\theta(c))||_2^2 \tag{4}$$

Here, θ represents the parameters of the 3D object being optimized, $\odot$ is Hadamard product, M is related to the mask, $M(\cdot)$ is the foreground mask acquired by the volume density along the ray of each pixel. $\lambda_{rgb}, \lambda_{mask}$ are the weights for the foreground RGB and the mask [38].

2) Depth prior: At reference viewpoints, relying solely on reference view reconstruction loss may result in poor geometric shapes. To address shape blur, indentations, and flatness, a depth prior is typically incorporated. Specifically, this involves using a pre-trained monocular depth estimator [72] to assess the depth d of the reference image. The depth of the 3D content viewed from the reference viewpoint should closely match this depth prior. Generally, negative Pearson correlation is used for depth regularization.

$$\mathcal{L}_{depth} = -\frac{Cov(d(c), d)}{Var(d(c))Var(d)} \tag{5}$$

Here, $Cov(\cdot)$ denotes covariance, and $Var(\cdot)$ calculates standard deviation, $d(c)$ refers to the depth modeled at the reference viewpoint. Through the use of reference view reconstruction loss and depth prior loss, the alignment between the reference image and the 3D model at the reference viewpoint can be optimized as much as possible. Although the estimated depth may not accurately represent geometric details, it is sufficient to ensure reasonable geometric shape and resolve most ambiguities [40]. Furthermore, normal smoothness loss [38] and diffusion CLIP loss [40] can also be added.

3) Diffusion prior: The supervision of novel view generation is guided by a diffusion prior. Textual inversion is used to generate textual descriptions y for the reference images. The SDS loss is employed for the continuous optimization of the 3D model.

The reference view loss includes details not captured by textual prompts, and SDS loss ensures the generated 3D model conforms to the object's expected shape. Combined, they ensure the model generation is faithful both to the reference image and to the textual prompts.

Upon completion of the coarse stage, the generated 3D model possesses a reasonable geometric shape, yet its overall geometric structure and texture remain somewhat rough. Based on the 3D model produced in the coarse stage, a model refinement network [76] can be utilized for further refinement, enhancing its geometric structure and texture. The overall optimization process is fundamentally similar to that of the coarse stage.

IV. Experiments

In accordance with the primary research focus of this paper, we categorize the current frameworks for 3D digital content generation based on diffusion models into two distinct types: text-to-3D and image-to-3D. All experimental results were obtained using a single A40 GPU. Our analysis primarily concentrates on two key aspects: the quality of the generated content and the speed of generation.

A. Text-to-3D

In the comparative experiments of text-to-3D digital content generation, we encountered frameworks that were either open-source or proprietary. For the open-source frameworks, experiments were conducted using the original codes from the respective papers. In the case of proprietary frameworks, we uniformly utilized threestudio [77] for experimentation. We acknowledge that there might be slight deviations in the results generated by threestudio compared to the original outcomes; however, we believe these differences do not significantly impact our evaluative conclusions. Additionally, in the realm of text-to-3D digital content generation, there are no universally accepted benchmarks for performance evaluation. Consequently, qualitative assessments were primarily based on visual inspections conducted by human observers. In our detailed experiments, we compare recent methods (DreamFusion [29], Latent-NeRF [74], Score Jacobian Chaining [34], ProlificDreamer [75], DreamGaussian [35]) for generating 3D objects from a textual prompt. Furthermore, considering the influence of textual prompt types on the model's generative performance, we employed two categories of textual descriptions: reality-based and imagination-based.The results of the generation are illustrated in Fig. 6.

Through a comparative analysis of the generated mesh quality and the overall generation time, as detailed in Table II, we observed that for objects existing in reality, ProlificDreamer [75] exhibits the highest quality of generation, albeit at the slowest speed. While DreamGaussian [35] may not match the former in terms of quality, it outperforms in generation speed. For imaginary objects, current mainstream frameworks struggle to achieve high-quality generation. We propose two avenues for optimization: firstly, refining textual prompts to more intricately describe the content envisioned, which could enhance the resultant generation. Secondly, augmenting the capabilities of the diffusion model by training it with larger datasets.

ProlificDreamer [75] proposed the use of Variational Score Distillation (VSD) to address issues such as over-saturation, over-smoothing, and low-diversity in the SDS loss. The core concept involves sampling within the distribution of 3D scenes, representing the 3D distribution with 3D parameter particles. A gradient-based particle updating rule is derived based on Wasserstein gradient flow. Despite its ability to achieve high-quality generation results, Prolificdreamer's method requires alternating training between LoRA [78] and NeRF during the training process, leading to prolonged training times. In contrast, DreamGaussian [35] employs 3D Gaussian Splatting [52]

TABLE II. MULTI-PERSPECTIVE COMPARATIVE ASSESSMENT OF TEXT-TO-3D DIGITAL CONTENT GENERATION FRAMEWORKS

Method	DreamFusion [29]	Latent-NeRF [74]	Score Jacobian Chaining [34]	ProlificDreamer [75]	DreamGaussian [35]
3D Representations	NeRF	NeRF	NeRF	NeRF	3D Gaussian Splatting
Number of Stages	Single	Two	Single	Three	Two
Mesh Quality	⋆	⋆⋆	⋆ ⋆ ⋆	⋆ ⋆ ⋆ ⋆ ⋆	⋆ ⋆ ⋆⋆
Avg. Time	∼40 minutes	∼1 hour	∼25 minutes	∼13 hours	∼4 minutes

TABLE III. QUANTITATIVE RESULTS ARE PROVIDED FOR **PSNR** ↑, **LPIPS** ↓, AND **CLIP-SIMILARITY** ↑

Dataset	Metrics	Zero-1-to-3 [63]	Magic123 [38]	DreamGaussian [35]	Stable Zero123
RealFusion15	PSNR↑	35.22	35.20	**35.47**	35.40
	LPIPS↓	0.10	0.13	0.08	**0.07**
	CLIP-Similarity↑	0.86	**0.90**	0.83	0.88

for 3D representation, significantly accelerating the generation speed.

B. Image-to-3D

In the comparative experiments for image-to-3D digital content generation tasks, we utilized the RealFusion [41] dataset, comprising 15 distinct objects, for our analysis. We compare recent methods (Zero-1-to-3 [63], Magic123 [38], DreamGaussian [35], Stable Zero123) for generating 3D objects from a single unposed image, with specific experimental results depicted in Fig. 7. Unlike the generation of 3D digital content from textual prompt, the quality of 3D content generated from a single image can be assessed based on image-related metrics.

1) PSNR: PSNR is a widely used standard for quantifying the quality of image reconstruction or image compression. It measures the pixel-level differences between the original and the compressed or reconstructed image. PSNR is calculated based on the Mean Squared Error (MSE) between the two images. Generally, a higher PSNR value indicates that the reconstructed image is closer in quality to the original image. It primarily evaluates the pixel-level similarity between the reconstructed or compressed image and the original image, but it may not always align with human perceptual differences.

2) LPIPS: LPIPS is a more modern, deep learning-based metric used to assess the perceptual quality and similarity of images. LPIPS calculates the similarity by comparing the activations of a deep neural network when processing two images. This approach aims to more closely resemble the human visual perception system. LPIPS is used to evaluate the perceptual similarity of images, especially in cases where pixel-level metrics may not capture all aspects of human perception.

3) CLIP-Similarity: CLIP-Similarity is a metric used to evaluate the semantic similarity between images, based on features extracted by the CLIP model. Unlike traditional image similarity metrics that focus on pixel-level details, CLIP-Similarity measures how semantically or contextually similar two images are. CLIP-similarity is particularly useful when the evaluation criteria extend beyond mere visual or pixel-level accuracy and venture into the realm of contextual and conceptual alignment.

For evaluating the quality of generated 3D content from reference viewpoints, we follow the metrics used in previous studies [41], [70]. We employed PSNR and LPIPS [79] metrics to compare the rendered images against the reference images, thereby assessing the generation quality from reference viewpoints. For images rendered from novel viewpoints, the quality was evaluated using the CLIP-similarity [9], as presented in Table III. Moreover, because we preprocess the original images in the process of generating 3D digital content from images, we applied the same treatments to the rendered images of the final 3D models during comparisons to ensure the accuracy of experimental results.

Our findings reveal that DreamGaussian [35] exhibits the fastest generation speed and achieves the highest quality when viewed from a reference perspective. However, it is noteworthy that its performance in generating novel views is comparatively inferior. On the other hand, Magic123 [38] demonstrates superior performance in generating high-quality novel views by incorporating a dual prior in both 2D and 3D dimensions. Simultaneously, the experimental results also confirm that the combination of diffusion models and 3D Gaussian Splatting [52] can achieve rapid 3D digital content generation, although there is room for further improvement in generation quality.

V. DISCUSSION

This study analyzes the frameworks related to text-to-3D content generation and image-to-3D content generation based on diffusion models, conducting extensive experiments. Through experimental comparative analysis, we identified numerous challenges in 3D content generation based on diffusion models.

A. Current Issues

1) Janus problem: Due to the primary approach of utilizing the diffusion model to guide rendering images from various perspectives, subsequently directing the generation of 3D models, the Janus problem is pervasive in the task of 3D digital content generation based on the diffusion model.

2) Over-Saturation: Using SDS loss in the generation of 3D content leads to issues such as over-saturation, over-smoothing, and low-diversity problems.

3) Controllability: Achieving precise control in the generation of 3D content from text prompts is challenging, relying solely on textual cues.

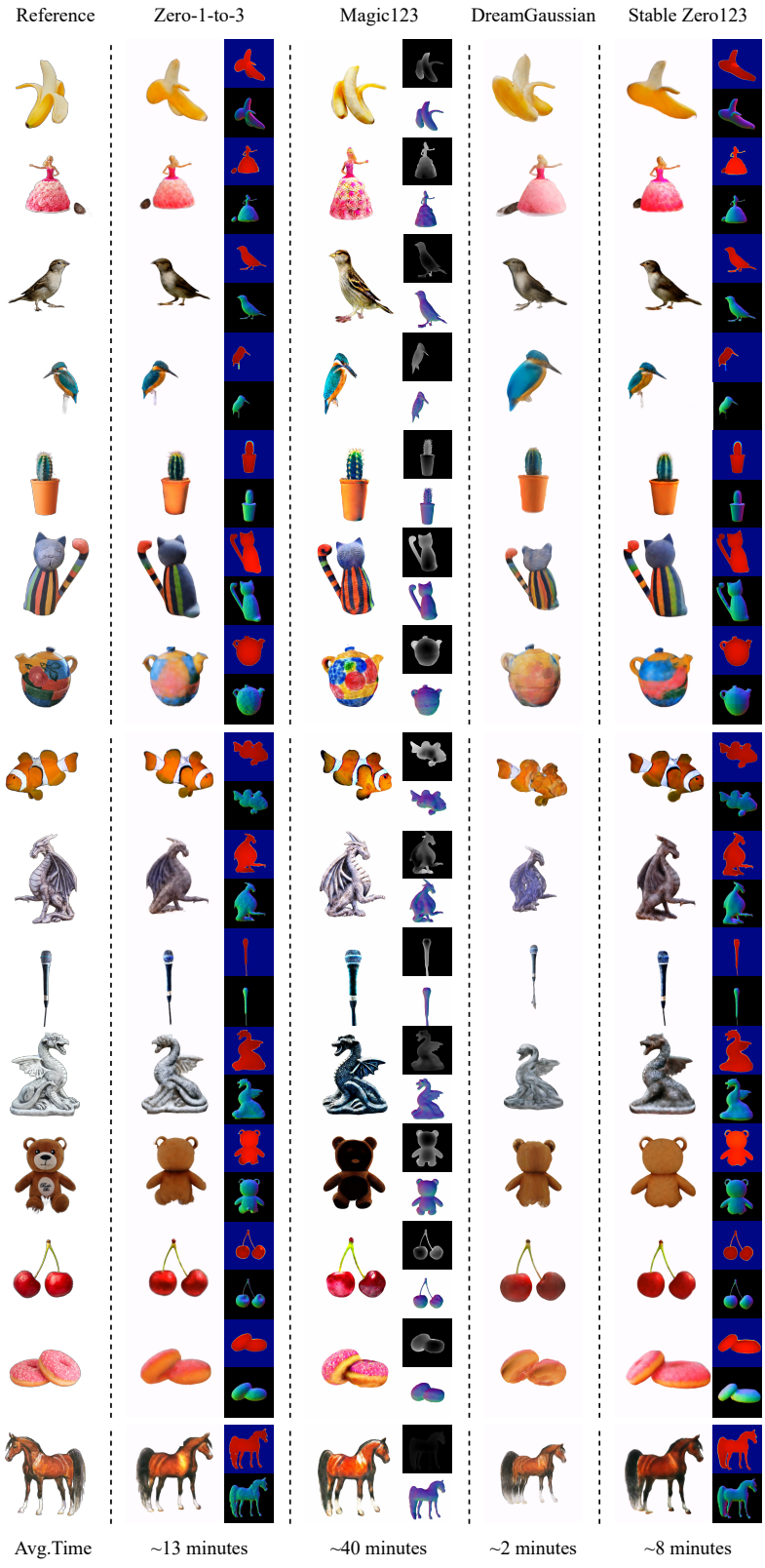

Fig. 7. Qualitative comparisons of 3D digital content generation from a single image.

4) Editability: Currently, there is no effective means to edit generated 3D content through artificial intelligence.

5) Imagination: Despite effective generation for real-world objects, the diffusion model struggles with the 3D reasoning

and imagination capabilities required for generating novel objects.

6) Primary view dependency: Tasks involving the generation of 3D content from a single image often require the input to be the primary view of the target object.

7) Evaluation metrics: A lack of a unified evaluation system for assessing the quality of generated 3D content.

8) Generation quality: Diffusion model-based 3D object generation faces issues of insufficient generation quality, resulting in objects that may lack realism or exhibit insufficient detail.

9) Shape inconsistency: Generated 3D objects may exhibit shape inconsistencies, particularly with complex geometric structures or topological relationships.

10) Scale disparities: Current 3D content generation models struggle to effectively handle objects of varying scales and are unable to generate 3D models of different sizes based on specific requirements.

B. Potential Solutions to Some Issues

1) Janus problem: To address the Janus problem, employing multi-view [80] or 3D perception [37] diffusion models can help alleviate the issue. Additionally, an incremental modeling approach, similar to a "humanoid printer", can be applied, generating 3D models for partial views gradually.

2) Over-Saturation: An approach akin to that proposed by prolificdreamer [75], employing Variational Score Distillation (VSD), can be adopted to address the issue of over-saturation and further enhance the quality of generated 3D models. However, it is noteworthy that this method may lead to a reduction in efficiency.

3) Controllability: While achieving controllability in text-to-3D content generation tasks remains challenging, leveraging image-to-3D generation tasks can facilitate more controlled 3D content generation.

4) Editability: Editing of 3D content can be achieved through image editing techniques [66] or by combining ChatGPT [1] to map text or voice into latent space for effective editing.

5) Imagination: In order to improve the generation performance of models, it is suggested to employ richer semantic description information. Alternatively, a more powerful diffusion model can be trained by incorporating a larger dataset. These strategies aim to enhance the overall effectiveness of the model in generating high-quality outputs.

6) Primary view dependency: Further enhancing the capabilities of novel view synthesis models to generate primary views of objects based on input images.

VI. Conclusion

With the continuous development of generative artificial intelligence, the scope of generated content is expanding beyond text, audio, and image domains, gradually progressing towards the generation of 3D objects and environments. Fueled by the visions of virtual reality, augmented reality, and the metaverse, the demand for 3D digital content across various industries is expected to further burgeon.

Current research indicates that different frameworks for 3D digital content generation exhibit advantages and limitations in terms of both generation quality and efficiency. Through our specific investigations, we posit that the integration of diffusion models and 3D Gaussian Splatting will be a focal point in the future research of 3D digital content generation. Additionally, constrained by the controllability issue in text-to-3D, a viable workflow for 3D digital content generation is as follows: firstly, generate images from text, providing creators with creative input. Subsequently, employ artificial intelligence to optimize and edit the image content to achieve the desired appearance. Then, use an image-to-3D generation framework to create a 3D model. Finally, import the generated 3D model into 3D modeling software for further refinement.

With the advancement of 3D object generation frameworks, future research is expected to extend from individual objects to scene generation. How to integrate procedural scene generation with artificial intelligence in the future is a question worthy of consideration.

In summary, this review comprehensively elucidates how diffusion models can be leveraged for 3D digital content generation. We analyze key frameworks for 3D digital content generation and experimentally validate the efficiency and feasibility of combining diffusion models with 3D Gaussian Splatting for modeling. We summarize the existing challenges in 3D digital content generation based on diffusion models and propose potential solutions for some of these issues. Overall, we contend that image-to-3D digital content generation aligns more closely with societal applications, though we remain optimistic about the future of text-to-3D digital content generation.

Acknowledgment

This paper is supported by the Top-notch Project of the All-Army Medical Science and Technology Youth Cultivation Program (17QNP045).

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[4] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[10] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," *Advances in neural information processing systems*, vol. 29, 2016.

[11] J. Gao, T. Shen, Z. Wang, W. Chen, K. Yin, D. Li, O. Litany, Z. Gojcic, and S. Fidler, "Get3d: A generative model of high quality 3d textured shapes learned from images," in *Advances In Neural Information Processing Systems*, 2022.

[12] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[13] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[14] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[16] L. Liu, W. Xu, H. Fu, S. Qian, Q. Yu, Y. Han, and C. Lu, "Akb-48: A real-world articulated object knowledge base," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 809–14 818.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[18] T. Wu, J. Zhang, X. Fu, Y. Wang, J. Ren, L. Pan, W. Wu, L. Yang, J. Wang, C. Qian *et al.*, "Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 803–814.

[19] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, 2020.

[20] M. A. Uy, Q.-H. Pham, B.-S. Hua, T. Nguyen, and S.-K. Yeung, "Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1588–1597.

[21] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[22] H. Fu, R. Jia, L. Gao, M. Gong, B. Zhao, S. Maybank, and D. Tao, "3d-future: 3d furniture shape with texture," *International Journal of Computer Vision*, vol. 129, pp. 3313–3337, 2021.

[23] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[24] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[25] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[26] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.

[27] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro *et al.*, "ediffi: Text-to-image diffusion models with an ensemble of expert denoisers," *arXiv preprint arXiv:2211.01324*, 2022.

[28] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[29] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[30] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, "Magic3d: High-resolution text-to-3d content creation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 300–309.

[31] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole, "Zero-shot text-guided object generation with dream fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 867–876.

[32] N. Mohammad Khalid, T. Xie, E. Belilovsky, and T. Popa, "Clip-mesh: Generating textured meshes from text using pretrained image-text models," in *SIGGRAPH Asia 2022 conference papers*, 2022, pp. 1–8.

[33] R. Chen, Y. Chen, N. Jiao, and K. Jia, "Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation," *arXiv preprint arXiv:2303.13873*, 2023.

[34] H. Wang, X. Du, J. Li, R. A. Yeh, and G. Shakhnarovich, "Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 619–12 629.

[35] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng, "Dreamgaussian: Generative gaussian splatting for efficient 3d content creation," *arXiv preprint arXiv:2309.16653*, 2023.

[36] W. Li, R. Chen, X. Chen, and P. Tan, "Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d," *arXiv preprint arXiv:2310.02596*, 2023.

[37] J. Sun, B. Zhang, R. Shao, L. Wang, W. Liu, Z. Xie, and Y. Liu, "Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior," *arXiv preprint arXiv:2310.16818*, 2023.

[38] G. Qian, J. Mai, A. Hamdi, J. Ren, A. Siarohin, B. Li, H.-Y. Lee, I. Skorokhodov, P. Wonka, S. Tulyakov *et al.*, "Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors," *arXiv preprint arXiv:2306.17843*, 2023.

[39] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt *et al.*, "Wonder3d: Single image to 3d using cross-domain diffusion," *arXiv preprint arXiv:2310.15008*, 2023.

[40] J. Tang, T. Wang, B. Zhang, T. Zhang, R. Yi, L. Ma, and D. Chen, "Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior," *arXiv preprint arXiv:2303.14184*, 2023.

[41] L. Melas-Kyriazi, I. Laina, C. Rupprecht, and A. Vedaldi, "Realfusion: 360deg reconstruction of any object from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8446–8455.

[42] Z. Chen, F. Wang, and H. Liu, "Text-to-3d using gaussian splatting," *arXiv preprint arXiv:2309.16585*, 2023.

[43] Y. Liu, C. Lin, Z. Zeng, X. Long, L. Liu, T. Komura, and W. Wang, "Syncdreamer: Generating multiview-consistent images from a single-view image," *arXiv preprint arXiv:2309.03453*, 2023.

[44] H. Weng, T. Yang, J. Wang, Y. Li, T. Zhang, C. Chen, and L. Zhang, "Consistent123: Improve consistency for one image to 3d object synthesis," *arXiv preprint arXiv:2310.08092*, 2023.

[45] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[46] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[47] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1912–1920.

[48] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 922–928.

[49] A. Sinha, A. Unmesh, Q. Huang, and K. Ramani, "Surfnet: Generating 3d shape surfaces using deep residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6040–6049.

[50] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 52–67.

[51] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[52] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics (ToG)*, vol. 42, no. 4, pp. 1–14, 2023.

[53] K. Gao, Y. Gao, H. He, D. Lu, L. Xu, and J. Li, "Nerf: Neural radiance field in 3d vision, a comprehensive review," *arXiv preprint arXiv:2210.00379*, 2022.

[54] Z. Shi, S. Peng, Y. Xu, A. Geiger, Y. Liao, and Y. Shen, "Deep generative models on 3d representations: A survey," *arXiv preprint arXiv:2210.15663*, 2022.

[55] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[56] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, "Plenoxels: Radiance fields without neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5501–5510.

[57] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics (ToG)*, vol. 41, no. 4, pp. 1–15, 2022.

[58] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 26 565–26 577, 2022.

[59] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion model," *arXiv preprint arXiv:2209.02646*, 2022.

[60] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 21 696–21 707, 2021.

[61] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[62] J. Li, Z. Li, Y. Li, and L. Wang, "P-2.12: A comprehensive study of content generation using diffusion model," in *SID Symposium Digest of Technical Papers*, vol. 54. Wiley Online Library, 2023, pp. 522–524.

[63] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9298–9309.

[64] Y. Wang, Q. Han, M. Habermann, K. Daniilidis, C. Theobalt, and L. Liu, "Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3295–3306.

[65] C. Reiser, R. Szeliski, D. Verbin, P. Srinivasan, B. Mildenhall, A. Geiger, J. Barron, and P. Hedman, "Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–12, 2023.

[66] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[67] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 208–18 218.

[68] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.

[69] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510.

[70] D. Xu, Y. Jiang, P. Wang, Z. Fan, Y. Wang, and Z. Wang, "Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4479–4489.

[71] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[72] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.

[73] S. M. H. Miangoleh, S. Dille, L. Mai, S. Paris, and Y. Aksoy, "Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9685–9694.

[74] G. Metzer, E. Richardson, O. Patashnik, R. Giryes, and D. Cohen-Or, "Latent-nerf for shape-guided generation of 3d shapes and textures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 663–12 673.

[75] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu, "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," *arXiv preprint arXiv:2305.16213*, 2023.

[76] T. Shen, J. Gao, K. Yin, M.-Y. Liu, and S. Fidler, "Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6087–6101, 2021.

[77] Y.-C. Guo, Y.-T. Liu, R. Shao, C. Laforte, V. Voleti, G. Luo, C.-H. Chen, Z.-X. Zou, C. Wang, Y.-P. Cao, and S.-H. Zhang, "threestudio: A unified framework for 3d content generation," https://github.com/threestudio-project/threestudio, 2023.

[78] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[79] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[80] Y. Shi, P. Wang, J. Ye, M. Long, K. Li, and X. Yang, "Mvdream: Multi-view diffusion for 3d generation," *arXiv preprint arXiv:2308.16512*, 2023.