EAI Endorsed Transactions
on AI and Robotics

Review Article EAI.EU

Exploring the Capabilities of NeRF Variants in Generating 3D Models

Shun Fang[1,*]

[1] Peking University, Lumverse Inc, Beijing, China

Abstract

This review paper presents a comprehensive analysis of three cutting-edge techniques in 3D content synthesis: Efficient Geometry-aware 3D Networks (EG3D), DreamFusion, and Magic3D. EG3D, leveraging geometry-aware representations and Generative Adversarial Networks (GANs), enables the generation of high-quality 3D shapes. DreamFusion integrates text-to-image diffusion models with neural rendering, opening new horizons for creative expression. Magic3D, on the other hand, extends text-to-image synthesis principles to 3D content creation, synthesizing realistic and detailed models. The theoretical frameworks, neural network architectures, and loss functions of these techniques are delved into, analyzing their experimental results and discussing their strengths, weaknesses, and potential applications. This review serves as a valuable resource for researchers and practitioners, offering insights into the latest advancements and pointing towards future directions for exploration in 3D content synthesis.

Keywords: NeRF, GANs, MLP, EG3D, DreamFusion, Magic3D

Received on 11 March 2024, accepted on 21 April 2024, published on 22 April 2024

doi: 10.4108/airo.5360

1. Introduction

In recent years, significant progress has been made in the field of 3D content synthesis and the creation of image content from textual prompts [1-4], attributed to the emergence of technologies such as Efficient Geometry-aware 3D Networks (EG3D) [5], DreamFusion [6], and Magic3D [7], as well as the evolution of diffusion models [8-10] in generative image modeling. Text-conditioned generative image models have now achieved high-quality, diverse, and adjustable image synthesis capabilities [11-14]. These methods have revolutionized the way we create and manipulate 3D objects, offering new opportunities for designers, artists, and researchers alike. Extensive research has been conducted on 3D generative modeling, exploring Object detection [32-35] and a range of 3D representation formats, such as 3D voxel grids [15-19], point clouds [20-25], meshes [26], [27], implicit representations [28], [29], and octrees [30]. In this review paper, we aim to provide a comprehensive overview of these three cutting-edge techniques, exploring their underlying principles, neural network architectures, loss functions, and experimental results.

EG3D, standing for Efficient Geometry-aware 3D Generative Adversarial Networks (GANs), represents a significant step forward in the generation of high-quality 3D shapes. Its unique architecture combines geometry-aware representations with GANs, enabling the generation of realistic and diverse 3D content. The theory behind EG3D revolves around its ability to capture the complex geometry of 3D objects, allowing for more accurate and detailed synthesis.

On the other hand, DreamFusion represents a novel approach to 3D content creation, leveraging text-to-image diffusion models and neural rendering. This method combines the power of large language models with the ability to generate high-resolution 3D scenes, opening up new possibilities for creative expression. The theoretical framework underlying

*Corresponding author. Email: fangshun@pku.org.cn

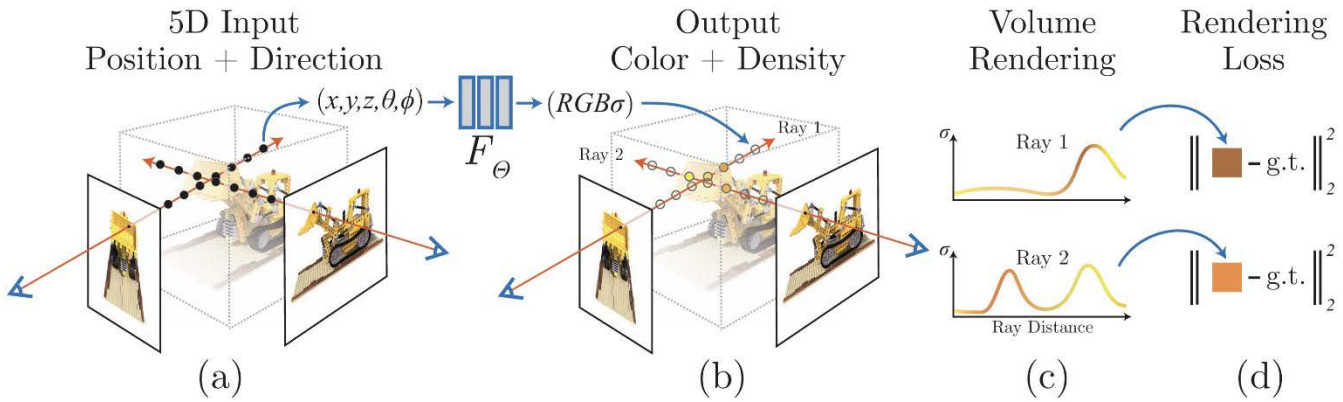

Figure 1 Overview of NeRF [31]

DreamFusion explores the integration of text prompts with 3D synthesis, enabling users to create scenes based on descriptive text.

Lastly, Magic3D emerges as a promising technique for synthesizing realistic and detailed 3D models using text prompts. It builds upon the principles of text-to-image synthesis, extending them to the domain of 3D content creation. Magic3D's neural network architecture and loss functions are designed to capture the intricate details of 3D objects, resulting in highly realistic and diverse outputs.

In this review, we delve into the details of each of these techniques, discussing their underlying principles, network architectures, and loss functions. We also present an analysis of their experimental results, highlighting their strengths, weaknesses, and potential applications. By comparing and contrasting these methods, we aim to provide a holistic understanding of the current state of the art in 3D content synthesis.

Furthermore, this review paper serves as a valuable resource for researchers, practitioners, and enthusiasts alike, offering insights into the latest advancements in the field and pointing towards future directions for exploration. We believe that the combined analysis of EG3D, DreamFusion, and Magic3D will pave the way for further innovations in 3D content synthesis, leading to more realistic, diverse, and accessible 3D content in the future.

2. Background

Neural Radiance Fields (NeRF) [31] is an avant-garde technique that harnesses the power of a fully connected neural network architecture to create unseen perspectives of intricate three-dimensional environments leveraging fragmented two-dimensional image collections. It functions by means of interpolation among the input images embodying a scene, thus enabling the creation of a comprehensive visualization. As a proficient means of synthesizing imagery from available data, NeRF directly correlates observation directions and spatial coordinates (constituting its 5D input space) with translucency and chromatic properties (yielding a 4D output space), utilizing volumetric rendering techniques to fashion these innovative viewpoints.

2.1 Architecture

Figure 1 portrays the holistic workflow of the NeRF technique, commencing from the acquisition of input images, progressing through the optimization of the neural network, and culminating in the generation of fresh perspectives of the scene. Of particular significance is the distinctively differentiable rendering process integral to this method, which empowers the refinement of NeRF to yield photorealistically authentic new views.

Input Imagery Stage: This segment of the diagram depicts a compilation of input viewpoints depicting the scene, captured from a variety of orientations encircling it. These images serve as the foundational material for training the NeRF model.

NeRF Optimization Phase: This stage signifies the operation of refining the neural network, which forms the nucleus of the NeRF method. The optimization process occurs through the calibration of the network's weight parameters to minimize the specified loss function, elucidated in Equation 6, thereby guaranteeing a close correspondence between the synthetic and original input images.

Rendering Unseen Perspectives: Upon the successful optimization of the neural network, it becomes feasible to render hitherto unseen perspectives of the scene. This is achieved by emulating the act of a camera capturing an image. These newly rendered perspectives are considered novel because they do not exist in the initial set of input images, instead being derived from the model's learned comprehension of the scene.

Differentiable Rendering Highlight: This portion of the illustration underscores the differentiable characteristic inherent in the rendering process. The rendering function exhibits differentiability relative to the neural network's parameters, thereby sanctioning the application of gradient-based optimization strategies. More precisely, this process encompasses the sampling of 5D coordinates (comprising spatial location and viewing orientation) along the trajectory of camera rays, feeding these locations into the Multilayer Perceptron (MLP) to derive color and volume density information, and thereafter employing volume rendering

methodologies to integrate these attributes into a coherent image.

2.2 Volume Rendering

Eq.1 and Eq.2 are crucial for the volume rendering technique used in NeRF, as it allows the network to synthesize new views of a scene by evaluating NeRF along camera rays. The differentiable nature of this rendering process enables gradient-based optimization to fit NeRF to a set of input images. Eq.1 is a mathematical representation of the expected color $\boldsymbol{C}(\mathbf{r})$ of a camera ray $\mathbf{r}(\boldsymbol{t}) = \boldsymbol{o} + \boldsymbol{td}$ as it passes through a scene represented by NeRF. The equation is formulated as an integral and is central to the volume rendering process in NeRF.

$$\boldsymbol{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma\big(\mathbf{r}(t)\big)\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt \qquad (\mathbf{1})$$

$$\text{where } T(t) = \exp\left(-\int_{t_n}^{t} \sigma\big(\mathbf{r}(s)\big)ds\right) \qquad (\mathbf{2})$$

Here,

- $\boldsymbol{C}(\mathbf{r})$: This represents the expected color of the ray $\mathbf{r}$, which is a function of the position $\boldsymbol{o}$ and direction $\boldsymbol{d}$ of the ray.
- $\int_{t_n}^{t_f}$: This is the integral operator, which calculates the continuous sum (or integral) over the range $[\boldsymbol{t_n}, \boldsymbol{t_f}]$. $\boldsymbol{t_n}$ is the near bound (closer to the camera) and $\boldsymbol{t_f}$ is the far bound (further away from the camera) for the ray's traversal through the scene.
- $\boldsymbol{T}(\boldsymbol{t})$: This is the accumulated transmittance from the near bound $\boldsymbol{t_n}$ to a point $\boldsymbol{t}$ along the ray. It represents the probability that the ray will travel from $\boldsymbol{t_n}$ to $\boldsymbol{t}$ without being absorbed or scattered by any particles in the scene.
- $\boldsymbol{\sigma}\big(\mathbf{r}(\boldsymbol{t})\big)$: This is the volume density function evaluated at a point $\mathbf{r}(\boldsymbol{t})$ along the ray. It represents how likely it is for the ray to interact with a particle at that point.
- $\mathbf{c}(\mathbf{r}(\boldsymbol{t}), \mathbf{d})$: This is the emitted radiance (color) at point $\mathbf{r}(\boldsymbol{t})$ in the direction $\mathbf{d}$. It is a function of both the position along the ray and the viewing direction.
- $\boldsymbol{dt}$: This represents an infinitesimally small change in the parameter $\boldsymbol{t}$ along the ray.

The integral, therefore, computes the contribution of color and density from every point along the ray's path, weighted by the transmittance from the camera to that point. The result is the expected color that would be observed by a camera ray passing through the volume defined by NeRF.

3. NeRF Variants

In this section, we present a comprehensive overview of EG3D, DreamFusion, and Magic3D, delving into their fundamental concepts, theoretical frameworks, neural network architectures, loss functions, and experimental results. We explore the core principles that underlie these state-of-the-art 3D content synthesis methods, analyzing how their unique neural network designs and loss functions contribute to their performance. Additionally, we evaluate the experimental outcomes of these methods, discussing their strengths, weaknesses, and potential applications. By synthesizing this information, we aim to provide a cohesive understanding of EG3D, DreamFusion, and Magic3D, setting the stage for further analysis and exploration in the field of 3D content synthesis.

3.1 Efficient Geometry-aware 3D Networks

The Efficient Geometry-aware 3D Networks (EG3D) [5] introduces a 3D GAN architecture tailored for synthesizing geometry-aware images from 2D photos, sans explicit 3D or multi-view supervision. Central to its design is a tri-plane representation, optimized for neural volume rendering in the 3D GAN context. This framework's core components comprise a pose-conditioned StyleGAN2-based feature generator and mapping network, a lightweight feature decoder-augmented tri-plane 3D representation, a neural volume renderer, a super-resolution module, and a dual-discrimination pose-conditioned StyleGAN2 discriminator. This setup effectively separates feature generation from neural rendering, leveraging the robust StyleGAN2 generator for 3D scene generalization. The tri-plane representation, both efficient and expressive, facilitates high-resolution geometry-aware image synthesis while maintaining computational and memory efficiency. The framework undergoes end-to-end training, utilizing a non-saturating GAN loss function with R1 regularization, adhering to the StyleGAN2 training paradigm. Additionally, a two-stage training approach is adopted to expedite the process, initially training with a reduced neural rendering resolution, followed by fine-tuning at full resolution.

The EG3D tackles the intricate task of unsupervised generation of high-quality, multi-view-consistent images and 3D shapes, solely relying on collections of single-view 2D photographs. The EG3D introduces a hybrid explicit-implicit network architecture, designed to efficiently synthesize high-resolution, multi-view-consistent images, along with high-quality 3D geometry, without resorting to extensive approximations. By decoupling feature generation from neural rendering, this framework harnesses the power of cutting-edge 2D CNN generators, such as StyleGAN2, to optimize computational efficiency and expressive capabilities.

The EG3D pioneers a 3D GAN framework that trains a geometry-aware image synthesis model using 2D photographs, without explicit 3D or multi-view supervision. This network architecture involves the generation of tri-plane features through a StyleGAN2 CNN generator, followed by neural volume rendering and super-resolution modules to produce high-resolution, multi-view-consistent renderings. The entire pipeline is trained end-to-end, utilizing a non-saturating GAN loss function with R1 regularization, and

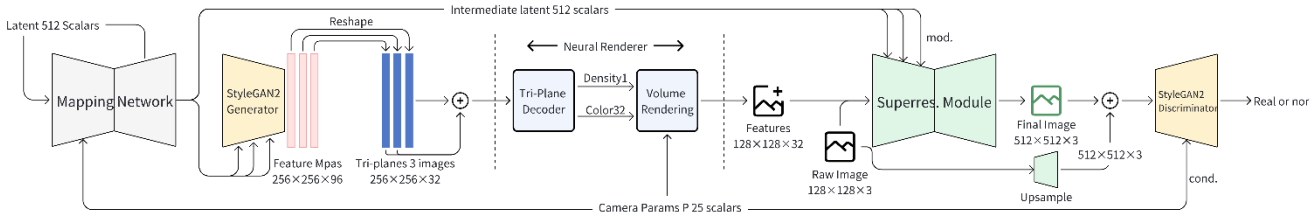

Figure 2 Overview of EG3D [5]

employs a two-stage training strategy for enhanced efficiency.

Furthermore, the EG3D delves into the applications of this proposed method, including style mixing and single-view 3D reconstruction. The 3D representation, grounded on the StyleGAN2 backbone, facilitates semantic image manipulations and high-quality single-view geometry recovery. While acknowledging limitations in the generated shapes, the EG3D suggests future research directions aimed at improving shape quality, exploring few-shot 3D reconstruction, and investigating alternative 2D backbones for conditional synthesis.

In essence, the EG3D presents an efficient and expressive 3D GAN framework that generates high-quality, multi-view-consistent images and intricate geometry from single-view 2D photographs, without relying on explicit 3D or multi-view supervision. This approach leverages a hybrid explicit-implicit network architecture and state-of-the-art 2D CNN generators to achieve optimal computational efficiency and unparalleled image quality.

Comparison

The EG3D offers several advantages. Firstly, the hybrid representation used in the framework efficiently distributes the expressive power, reducing computational costs compared to fully implicit architectures while maintaining high quality. Secondly, the decoupling of feature generation and neural rendering allows for the utilization of a powerful StyleGAN2 generator for 3D scene generalization, enhancing expressiveness and efficiency. Additionally, the tri-plane representation proves to be compact yet sufficiently expressive, outperforming dense feature volume representations and fully implicit representations in terms of quality metrics while being computationally and memory efficient. Furthermore, the 3D GAN framework enables geometry-aware image synthesis from 2D photographs without explicit 3D or multiview supervision, showcasing the versatility and effectiveness of the approach. Lastly, the method inherits the well-studied properties of the StyleGAN2 latent space, allowing for semantic image manipulations and high-quality single-view 3D reconstruction.

Table 1. For the purposes of quantitative assessment, metrics such as ID, FID, pose precision, and depth precision were employed in evaluating both AFHQ and FFHQ Cats datasets. The resolution of the images utilized for training and evaluation purposes has been clearly indicated. [5]

Method	Cats	FFHQ			
	FID	ID	FID	Pose	Depth
EG3D $\mathbf{256^2}$	3.88	0.76	4.8	0.005	0.31
EG3D $\mathbf{512^2}$	$\mathbf{2.77}^*$	0.77	4.7	0.005	0.39
π-GAN $\mathbf{128^2}$	16.0	0.67	29.9	0.021	0.44
GIR AFFE $\mathbf{256^2}$	16.1	0.64	31.5	0.089	0.94
Lift. SG $\mathbf{256^2}$	-	0.58	29.8	0.023	0.40

Table 1 presents a quantitative analysis comparing the performance of EG3D framework with several contemporary top-tier methods in the realm of 3D-aware image synthesis, focusing particularly on the FFHQ and AFHQ Cats datasets. The comparative assessment utilizes the following metrics:

- **Fréchet Inception Distance (FID):** This gauge quantifies the quality of produced images, where lower FID values denote increased realism and enhanced image quality. The FID score is computed by contrasting 50,000 artificial images with the entire corpus of genuine images.
- **Identity Persistence (ID):** This metric assesses the degree of facial identity preservation across various perspectives of a synthetically generated face. Superiority is indicated by higher ID scores, suggesting that the model is capable of maintaining a consistent identity across varying camera perspectives.
- **Depth Estimation Accuracy:** This measure evaluates the precision of the depth information generated by the model, calculated as the Mean Squared Error (MSE) against proxy ground truth depth maps inferred from the generated images.
- **Pose Retention:** This criterion examines how accurately the synthesized images preserve the intended pose, determined via the L2 loss against true pose estimates sourced from the fabricated images.
- **Image Resolution:** Denotes the pixel dimensions at which the models have been both trained and tested, e.g., 256x256 pixels or 512x512 pixels.

In the comparison, EG3D contends with three alternative methods—GIRAFFE [36], π-GAN [37], and Lifting StyleGAN [38]. EG3D yields notably lesser FID scores across both datasets, suggesting that it generates images

more akin to real-world distributions. Moreover, it sustains leading-edge performance in terms of preserving identity consistency, depth estimation accuracy, and pose alignment. The EG3D rows with the most favorable outcomes—lowest FID scores coupled with the highest ID, Depth, and Pose scores—attest to the method's superiority in crafting high-fidelity, viewpoint-coherent images and shapes.

Furthermore, table 1 discloses that the model trained on the FFHQ dataset at a resolution of 512x512 and incorporating adaptive data augmentation attains the pinnacle of performance, boasting a FID score of 4.7 and an ID score of 0.77. Meanwhile, the model educated at the same higher resolution of 512x512 but without the inclusion of adaptive data augmentation still delivers commendable results, albeit not reaching the level of optimization achieved by the augmented counterpart.

The EG3D underscores the proficiency and proficiency of its approach in generating multi-view-consistent, high-resolution images, as well as intricate 3D shapes, solely from single-view 2D photographs. It underscores the pivotal role of components, such as dual discrimination and generator pose conditioning, in bolstering expression consistency and image quality while maintaining view coherence. Furthermore, the EG3D explores the resilience of the method in the face of inaccurate camera poses, revealing that even highly imprecise extrinsics can still facilitate precise 3D shape reconstruction. The findings suggest that the proposed framework achieves noteworthy advancements in image quality, geometry precision, and view consistency, positioning it as a promising candidate for unsupervised 3D shape generation and image synthesis tasks.

3.2 Text-to-3d Using 2D Diffusion

The DreamFusion [6] introduces a unique strategy for synthesizing 3D objects from text prompts, leveraging a pre-trained 2D text-to-image diffusion model. This novel method employs a loss function rooted in probability density distillation to fine-tune a randomly initialized 3D model—specifically, a NeRF. Through gradient descent, it generates 3D models directly from textual cues. The diffusion models at play incorporate a forward process that degrades data structure by injecting noise and a reverse process that gradually restores structure from noise, with transitions parameterized to predict the latent noise content. The generative model is trained to reconstruct structure from noise, effectively reducing the training objective to a denoising score matching task.

Building upon text-to-image diffusion models conditioned on text embeddings, the DreamFusion incorporates classifier-free guidance to enhance generation quality. In essence, the DreamFusion adapts a pre-trained 2D diffusion model originally designed for text-to-image synthesis to perform text-to-3D synthesis, without relying on labeled 3D data. By harnessing probability density distillation and optimizing a 3D model via gradient descent, this approach enables the creation of realistic 3D objects and scenes from text prompts, underscoring the potency of pre-trained image diffusion models as priors for 3D synthesis.

Architecture

The DreamFusion architecture incorporates a MLP with distinct attributes. This NeRF-based MLP is structured with five ResNet blocks, each containing 128 hidden units. Eq.3 is used to compute the final RGB color value $\mathbf{C}$ for a pixel in the rendered image. It does so by rendering the contributions of color and density along a ray from the camera through the pixel's location in the image plane and into the 3D scene.

$$\mathbf{C} = \sum_{i} \boldsymbol{w_i}\mathbf{c_i} \tag{3}$$

$$\boldsymbol{w_i} = \boldsymbol{\alpha_i} \prod_{j<i} \left(\mathbf{1} - \boldsymbol{\alpha_j}\right) \tag{4}$$

$$\boldsymbol{\alpha_i} = \mathbf{1} - \mathbf{exp}(-\boldsymbol{\tau_i}\|\boldsymbol{\mu_i} - \boldsymbol{\mu_{i+1}}\|) \tag{5}$$

Here,

- $\mathbf{c}$: The final RGB color value for the pixel.
- $\boldsymbol{w_i}$: The rendering weight for the i-th sample along the ray. This weight determines how much the color of each sample contributes to the final pixel color.
- $\mathbf{c_i}$: The RGB color of the i-th sample along the ray. This is the color emitted by the 3D point in the scene, which is a result of the NeRF model.
- $\boldsymbol{\alpha_i}$: The transparency of the i-th sample. It determines how much light passes through the sample to subsequent samples along the ray.
- $\boldsymbol{\tau_i}$: The volumetric density at the i-th sample location.
- $\boldsymbol{\mu_i}$ and $\boldsymbol{\mu_{i+1}}$: The 3D positions of the current and next samples along the ray.

The term $\prod_{j<i} \left(\mathbf{1} - \boldsymbol{\alpha_j}\right)$ is a product that accumulates the transmissivities of all previous samples along the ray. In other words, it accounts for the fraction of light that has passed through all samples before the i-th sample without being absorbed or scattered. The opacity $\boldsymbol{\alpha_i}$ is computed using the exponential function to model the attenuation of light as it travels through the volume with density $\boldsymbol{\tau_i}$ over the distance $\|\boldsymbol{\mu_i} - \boldsymbol{\mu_{i+1}}\|$ between the current and next sample positions.

It employs Swish/SiLU activation functions and incorporates layer normalization between each block. For activation, an exponential function is used for density τ, whereas a sigmoid function is applied to RGB albedo ρ. Eq.6 and Eq.7 are related to the shading and lighting model used within NeRF framework to compute the final color of a point on a 3D scene. Eq.6 represents the output of MLP, which is a part of the NeRF model. The MLP takes as input the 3D spatial coordinates $\boldsymbol{\mu}$ and parameters $\boldsymbol{\theta}$, and outputs two values: the volumetric density $\boldsymbol{\tau}$ and the RGB albedo $\boldsymbol{\rho}$ of the material at that point in space.

$$(\boldsymbol{\tau}, \boldsymbol{\rho}) = \mathbf{MLP}(\boldsymbol{\mu}; \boldsymbol{\theta}) \tag{6}$$

Here,

- $\boldsymbol{\tau}$: Volumetric density, which indicates how opaque the material is at the 3D coordinate $\boldsymbol{\mu}$.
- $\boldsymbol{\rho}$: RGB albedo, which is the base color of the material

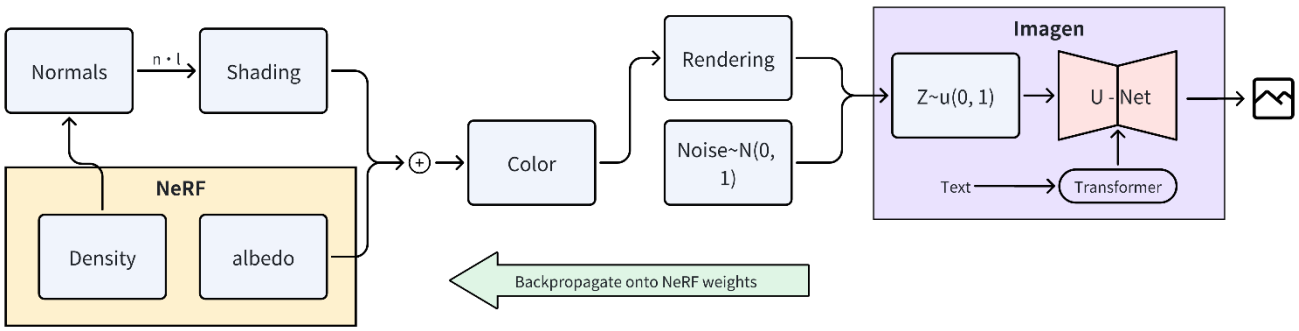

Figure 3 Overview of DreamFusion [6]

without any lighting effects applied.
- MLP: A neural network with multiple layers that maps the input to the output. It is trained to predict the density and color from the spatial coordinates.
- $\boldsymbol{\theta}$: Parameters of the MLP, which are learned during the training process.

Eq.7 calculates the final shaded output color $\mathbf{c}$ for a 3D point, taking into account the lighting in the scene.

$$\mathbf{c} = \boldsymbol{\rho} \circ \left(\boldsymbol{\ell}_{\boldsymbol{\rho}} \circ \boldsymbol{max} \left(\mathbf{0}, \boldsymbol{n} \cdot \frac{\boldsymbol{\ell} - \boldsymbol{\mu}}{\|\boldsymbol{\ell} - \boldsymbol{\mu}\|} \right) + \boldsymbol{\ell}_{\boldsymbol{a}} \right) \qquad (7)$$

Here,
- $\mathbf{c}$: The final color of the point after shading.
- $\boldsymbol{\rho}$: The RGB albedo of the material at the point, as computed by the MLP.
- $\boldsymbol{\ell}_{\boldsymbol{\rho}}$: The color of the point light source.
- $\boldsymbol{n}$: The surface normal vector at the 3D point, which indicates the local orientation of the object's geometry. It is computed from the gradient of the density $\boldsymbol{\tau}$.
- $\boldsymbol{\ell}$: The 3D position of the point light source.
- $\boldsymbol{\mu}$: The 3D position of the point being shaded.
- $\boldsymbol{\ell}_{\boldsymbol{a}}$: The color of the ambient light.
- $\circ$: The Hadamard product (element-wise multiplication) of vectors.

The term inside the max function, $\boldsymbol{n} \cdot \frac{\boldsymbol{\ell}-\boldsymbol{\mu}}{\|\boldsymbol{\ell}-\boldsymbol{\mu}\|}$, computes the cosine of the angle between the light direction (from the point to the light source) and the normal vector, which determines the amount of light that diffusely reflects off the surface at that point according to the Lambertian reflectance model. The max function ensures that only positive contributions are considered. The color of the light reflected by the point $\mathbf{c}$ is then obtained by multiplying the albedo $\boldsymbol{\rho}$ with the result of the lighting calculation, which includes the diffusely reflected light from the point light source plus the ambient light $\boldsymbol{\ell}_{\boldsymbol{a}}$. This models how the material's color is affected by the lighting in the scene.

During the optimization process, shading hyperparameters are adjusted, with ambient light color ℓa and diffuse light color $\boldsymbol{\ell}_{\boldsymbol{\rho}}$ differing between initial and subsequent steps. Furthermore, to emphasize the scene's central content within the 3D coordinate space, a minor density "blob" is introduced around the origin, introducing spatial density bias. Details regarding camera and light sampling are also provided. This includes biased camera elevation sampling and the introduction of perturbations to both the camera position and the "up" vector, aimed at enriching the diversity of the training process.

Loss Function

In the DreamFusion framework, the loss function is crafted to refine the parameters $\boldsymbol{\theta}$ in a way that the generated sample $\mathbf{x} = \boldsymbol{g}(\boldsymbol{\theta})$ closely mirrors a sample drawn from the fixed diffusion model (Eq.8 and Eq.9).

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{Diff}}\left(\boldsymbol{\phi}, \mathbf{x} = \boldsymbol{g}(\boldsymbol{\theta})\right) =$$

$$\mathbb{E}_{\boldsymbol{t},\boldsymbol{\epsilon}} \left[\boldsymbol{w}(\boldsymbol{t}) \underbrace{\left(\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\phi}}(\mathbf{z}_{\boldsymbol{t}}; \boldsymbol{y}, \boldsymbol{t}) - \boldsymbol{\epsilon}\right)}_{\text{Noise Residual}} \underbrace{\frac{\partial \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\phi}}(\mathbf{z}_{\boldsymbol{t}}; \boldsymbol{y}, \boldsymbol{t})}{\mathbf{z}_{\boldsymbol{t}}}}_{\text{U-Net Jacobian}} \underbrace{\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}}_{\text{Generator Jacobian}} \right] \qquad (8)$$

Here,
- $\nabla_{\boldsymbol{\theta}}$: The gradient with respect to $\boldsymbol{\theta}$.
- $\mathcal{L}_{\text{Diff}}$: The diffusion training loss, which measures the difference between the noise prediction of the model and the actual noise.
- $\boldsymbol{\phi}$: Parameters of the diffusion model.
- $\mathbf{x}$: The generated image.
- $\boldsymbol{g}(\boldsymbol{\theta})$: A differentiable image parameterization, such as a neural network, that generates an image $\mathbf{x}$ from parameters $\boldsymbol{\theta}$.
- $\mathbb{E}_{\boldsymbol{t},\boldsymbol{\epsilon}}$: The expectation over random variables $\boldsymbol{t}$ (time steps in the diffusion process) and $\boldsymbol{\epsilon}$ (noise).
- $\boldsymbol{w}(\boldsymbol{t})$: A weighting function that depends on the time step $\boldsymbol{t}$.
- $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\phi}}(\mathbf{z}_{\boldsymbol{t}}; \boldsymbol{y}, \boldsymbol{t})$: The noise prediction by the diffusion model at time step $\boldsymbol{t}$, conditioned on some context $\boldsymbol{y}$.
- $\boldsymbol{\epsilon}$: The actual noise added to the image.
- $\frac{\partial}{\partial \mathbf{z}_{\boldsymbol{t}}}, \frac{\partial}{\partial \mathbf{z}_{\mathbf{x}}}, \frac{\partial}{\partial \mathbf{z}_{\boldsymbol{\theta}}}$: Partial derivatives with respect to their respective variables.

The gradient calculation involves computing the expected difference between the model's noise prediction and the

actual noise, weighted by $\boldsymbol{w(t)}$, and then applying the chain rule to backpropagate through the diffusion process and the image parameterization to update $\boldsymbol{\theta}$.

Eq.9 defines the gradient of the Score Distillation Sampling (SDS) loss, which is a simplified version of the diffusion training loss. It is used to optimize the parameters $\boldsymbol{\theta}$ of a differentiable image parameterization to generate an image that has a low SDS loss, indicating that it is similar to samples from the diffusion model.

$$\nabla_{\theta}\mathcal{L}_{\mathbf{SDS}}\left(\boldsymbol{\phi}, \mathbf{x} = g(\boldsymbol{\theta})\right) \triangleq \mathbb{E}_{t,\epsilon}\left[\boldsymbol{w(t)}\left(\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\phi}}(\mathbf{z}_t; \boldsymbol{y}, \boldsymbol{t}) - \boldsymbol{\epsilon}\right)\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}\right] \quad (9)$$

Here,

- $\mathcal{L}_{\mathbf{SDS}}$ The SDS loss, a variant of the diffusion training loss.
- $\triangleq$: The colon equals sign, indicating that the right-hand side defines the left-hand side.
- All other terms are the same as in Eq.8, except that the term $\frac{\partial \hat{\epsilon}_{\phi}(\mathbf{z}_t; y, t)}{\partial \mathbf{z}_t}$ is omitted. This omission results in a lower-variance gradient that is easier to optimize and does not require backpropagating through the diffusion model's U-Net, which can be computationally expensive.

By using the SDS loss, the optimization process can generate images by directly updating the parameters $\boldsymbol{\theta}$ in the direction that reduces the loss, without needing to compute and backpropagate through the full noise prediction of the diffusion model. This makes the optimization process more efficient and robust.

Comparison

The DreamFusion offers several advantages. These advantages highlight the innovative approach of leveraging existing 2D diffusion models for 3D synthesis, the effectiveness of the proposed loss function, and the potential for creating diverse and controllable 3D models from text descriptions. Harnessing a pre-trained 2D text-to-image diffusion model for text-to-3D synthesis eliminates the need for extensive labeled 3D datasets and efficient denoising structures for 3D data. Instead, a loss function rooted in probability density distillation is introduced, leveraging the 2D diffusion model as a prior to optimize a parametric image generator. This approach enables the generation of 3D models directly from text descriptions. By further optimizing a randomly initialized 3D model using a DeepDream-inspired methodology with this novel loss function, it is possible to create 3D models that are fully viewable from any angle, can be relit with arbitrary illumination, and can be seamlessly integrated into diverse 3D environments. The method demonstrates the utility of pre-trained image diffusion models as priors for 3D synthesis, without relying on 3D training data or altering the image diffusion model itself. This addresses the challenge of sampling in parameter space instead of pixel space, leading to the creation of 3D models that resemble realistic images when rendered from random perspectives. This approach also opens the door to generating coherent 3D scenes from textual prompts, underscoring the qualitative prowess of the model in compositional generation tasks.

Table 2. Assessing the congruency of the images generated by DreamFusion with their corresponding captions by employing various CLIP retrieval methodologies. [6]

Method	R-Precision					
	CLIP Color	L/14 Geo	CLIP Color	B/16 Geo	CLIP Color	B/32 Geo
DreamFusion	79.7	58.5	77.5	46.6	75.1	42.5
Dream Fields	-	-	74.2	-	68.3	-
reimpl.	82.9	1.4	99.9	0.8	78.6	1.3
CLIP-Mesh	**74.5***	-	75.8	-	67.8	-
GT Images	-	-	79.1	-	77.1	-

Table 2 presents a comparison of the performance of different models in generating coherent 3D scenes from text prompts. The table evaluates the models using the CLIP R-Precision metric, which measures the consistency of rendered images with respect to the input caption. Essentially, it checks how accurately the generated 3D scene matches the text description it was based on, using the CLIP model's ability to retrieve the correct caption for the scene.

- **Method:** The name of the model or method being evaluated. In this case, the methods include "Dream Fields [39]," "CLIP-Mesh [40]," "DreamFusion [6]," and "GT Images".
- **R-Precision:** A metric that measures the relevance of the generated images to the text prompts. Higher R-Precision values indicate better performance.
- **CLIP B/32, CLIP B/16, CLIP L/14:** These columns refer to different versions of the CLIP model used for evaluation. The numbers likely refer to the size or architecture of the CLIP model, and "L/14" might refer to a larger model. The table shows R-Precision scores for each CLIP model variant.
- **Color, Geo:** These columns likely stand for "Color" and "Geometry," which are two aspects of the generated 3D scenes being evaluated. "Color" assesses how well the color of the rendered images matches the input caption, while "Geo" evaluates the geometric accuracy of the 3D models.

Assessing the proficiency of DreamFusion in crafting coherent 3D landscapes from diverse textual cues, its performance is contrasted with current zero-shot text-to-3D generative models, emphasizing the crucial components that facilitate precise 3D geometry. DreamFusion offers an extensive collection of 3D assets, extended videos, and meshes, available for further scrutiny on dreamfusion3d.github.io. Experimental configuration involves optimizing 3D scenes on a TPUv4 machine, equipped with four chips, each responsible for rendering a distinct view, while also evaluating the diffusion U-Net. The optimization process comprises 15,000 iterations, utilizing the Distributed Shampoo optimizer. By contrasting the evaluation metrics of DreamFusion with ground-truth

images sourced from the MS-COCO datasets and various other models, the approach demonstrates the remarkable proficiency of DreamFusion in generating coherent 3D scenes solely from textual descriptions.

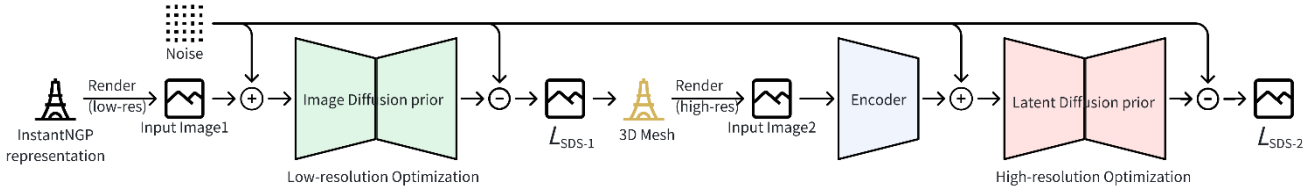

Figure 4 Overview of Magic3D [7]

3.3 High-Resolution Text-to-3D Creation

The Magic3D [7] framework revolutionizes high-quality 3D content synthesis using text prompts, surpassing the design limitations of DreamFusion. This innovative approach employs a two-stage coarse-to-fine methodology, leveraging efficient scene models to facilitate high-resolution text-to-3D synthesis. In the initial stage, a coarse neural field representation is optimized, akin to DreamFusion but with a memory- and compute-optimized scene representation rooted in a hash grid. This phase enables the utilization of diffusion priors at resolutions reaching up to 512 × 512. Progressing to the second stage, mesh representations are further optimized, harnessing an efficient differentiable rasterizer and camera close-ups to capture intricate geometric and textural details. Magic3D's framework empowers users with creative control over the 3D synthesis process, drawing inspiration from advancements in text-to-image editing applications. This allows users to craft desired 3D objects seamlessly with text prompts and reference images, bridging the gap towards democratizing 3D content creation. The Magic3D approach significantly enhances the quality of 3D content synthesis, with users preferring its outcomes over DreamFusion while achieving a remarkable 2× speed-up.

Moreover, Magic3D delves into the utilization of diffusion priors in a graded manner, from coarse to fine, to craft high-definition geometry and textures. In the initial phase, the fundamental diffusion model calculates gradients of the scene model at a reduced resolution of 64 × 64. Subsequently, the Latent Diffusion Model (LDM) comes into play in the second stage, catering to high-resolution imagery at 512 × 512. The latent diffusion model's computations remain tractable due to its manipulation of the latent zt at a 64 × 64 resolution.

Furthermore, Magic3D explores the realm of controllable 3D generation through prompt-based editing. By employing a specific approach known as DreamBooth to fine-tune diffusion prior models, users can steer the text-to-3D model generation with images and text prompts. This technique enables the modification of 3D models while maintaining the subjects from the input images, offering enhanced control over the 3D generation outcomes.

In essence, Magic3D presents a swift and superior text-to-3D generation framework that leverages efficient scene models and high-resolution diffusion priors in a graded approach. This framework facilitates the creation of high-fidelity 3D content from text prompts, affording users unparalleled control over the synthesis process.

Loss Function

The loss function employed by DreamFusion is termed SDS. The computation of the gradient for this loss function proceeds as follows:

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

Here,

- ∇_{θ}: The gradient with respect to the parameters $\boldsymbol{\theta}$ of the scene model.
- $\mathcal{L}_{\mathbf{SDS}}$: The loss function for SDS, which is used to guide the optimization.
- $\boldsymbol{\phi}$: The diffusion model with a learned denoising function ϵ_{ϕ}.
- $g(\boldsymbol{\theta})$: The scene model, a parametric function that produces an image $\boldsymbol{x}$ given parameters $\boldsymbol{\theta}$.
- $\mathbb{E}_{t,\epsilon}$: The expectation over the noise level $\boldsymbol{t}$ and the sampled noise $\boldsymbol{\epsilon}$.
- $\boldsymbol{w(t)}$: A weighting function that depends on the noise level $\boldsymbol{t}$.
- $\epsilon_{\phi}(x_t; y, t)$: The denoising function of the diffusion model, which predicts the noise given a noisy image x_t, text embedding $\boldsymbol{y}$, and noise level $\boldsymbol{t}$.
- $\frac{\partial x}{\partial \theta}$: The gradient of the rendered image with respect to the parameters of the scene model.

The optimization aims to update the parameters $\boldsymbol{\theta}$ of the scene model so that the rendered images match the distribution of photorealistic images across different viewpoints, given the input text prompt.

Eq.11 is similar to Eq.10 but is used in the second stage of the Magic3D optimization process, where a LDM is employed to generate high-resolution images. The key differences are:

$$\nabla_{\theta}\mathcal{L}_{\mathbf{SDS}}(\boldsymbol{\phi}, g(\boldsymbol{\theta})) = \mathbb{E}_{t,\epsilon}\left[w(t)\left(\epsilon_{\phi}(z_t; y, t) - \epsilon\right)\frac{\partial z}{\partial x}\frac{\partial x}{\partial \theta}\right] \quad (11)$$

- z_t: The latent variable with resolution 64 × 64 that the LDM acts upon.

- $\frac{\partial z}{\partial x}$: The gradient of the encoder in the latent diffusion model, which transforms the latent variable into the rendered image space.

The LDM allows for the optimization of the scene model at a higher resolution, leading to the generation of more detailed 3D content. The gradient computation involves the latent space of the diffusion model, enabling the fine-tuning of the 3D model with high-resolution image priors.

Comparison

The Magic3D framework marks a noteworthy leap forward in the realm of 3D content synthesis by introducing a refined optimization strategy that seamlessly integrates both low- and high-resolution diffusion priors. This innovative approach not only elevates the quality of generated 3D content but also surpasses prior techniques like DreamFusion in terms of processing speed. Furthermore, Magic3D offers users unprecedented creative control over the 3D synthesis process, positioning it as a promising candidate for democratizing the creation of 3D content.

Table 3. To assess user preferences for 3D models created utilizing 397 prompts provided by DreamFusion, preference studies were conducted. Overall, a larger proportion of evaluators (61.7%) favored the 3D models generated by Magic3D compared to those from DreamFusion. Furthermore, the majority of evaluators (87.7%) preferred the finer models over the coarser ones within the Magic3D framework, validating the efficacy of the coarse-to-fine approach. [7]

Comparison	Preference
Magic vs. DreamFusion	
✓ More detailed	66.0%
✓ More realistic	58.3%
✓ More detailed & realistic	61.7%
Magic3D vs. Magic3D(Coarse only)	87.7%

The experimental aspect of Magic3D centers on a comparative analysis between the proposed framework, Magic3D, and DreamFusion, utilizing 397 textual prompts. The findings reveal that Magic3D outperforms DreamFusion in generating high-quality 3D shapes, exhibiting intricate geometry and texture. User preference surveys further corroborate this observation, with a significant proportion of evaluators favoring the more realistic and intricate outputs produced by Magic3D. The framework's coarse-to-fine optimization approach ensures efficient training with manageable runtimes, underscoring its effectiveness in enhancing 3D content synthesis. Moreover, qualitative comparisons and user studies underscore the superiority of Magic3D in creating visually appealing and realistic 3D models across diverse prompts and scenarios.

4. Comparative Analysis and Future Research

4.1 Comparative Analysis

We compare and contrast three significant Models based on their methods, application domains, and experimental results.

DreamFusion Approach

Procedure: DreamFusion capitalizes on a pretrained 2D text-to-image diffusion model to facilitate text-driven 3D synthesis. It pioneers a probability density distillation-based loss function, which permits the employment of a 2D diffusion model as a prior to refine a parameterized image generator. The optimization of a NeRF model in 3D space is accomplished via gradient descent, striving for minimized loss in 2D renderings extracted from arbitrary viewpoints.
Area of Application: The principal area of application lies in the generation of 3D models based on textual descriptions, proving especially valuable in sectors such as digital media, gaming, and film production, where intricate 3D assets are indispensable.
Research Outcomes: DreamFusion effectively generates 3D models responding to textual commands, allowing for unrestricted viewing angles, adjustable lighting, and seamless integration into 3D surroundings. Its results showcase high fidelity and coherence in the 3D objects and scenes crafted from a wide array of textual prompts.

Magic3D Approach

Procedure: Magic3D implements a dual-phase optimization schema to overcome limitations encountered by DreamFusion. It initially establishes a rough model using a low-resolution diffusion prior and a sparse 3D hash grid structure, followed by the optimization of a textured 3D mesh model using a high-resolution LDM and a resourceful differentiable rendering mechanism.
Field of Use: Like DreamFusion, Magic3D aims at creating 3D content from textual descriptions, yet with a heightened emphasis on enhancing resolution and processing speed, thus making it fitting for professional 3D modeling and creative design tasks.
Experimental Findings: Magic3D outperforms DreamFusion in terms of speed, accomplishing the task in half the time while attaining a higher resolution. User assessments suggest a marked inclination towards Magic3D-generated models, with a substantial majority (61.7%) expressing a preference for its methodology.

EG3D Approach

Procedure: EG3D introduces a blended explicit-implicit network design for unsupervised 3D representation learning stemming from individual 2D photographic views. It integrates an efficient 3D representation with a neural

rendering engine to produce high-resolution, multi-view consistent images and 3D structures in real-time.
Domain of Implementation: EG3D is geared towards generating premium-grade 3D shapes and multi-view harmonious images from 2D photos, finding practical applications in domains such as computer graphics, virtual reality, and augmented reality.
Research Findings: EG3D showcases pioneering 3D-aware synthesis, delivering high-quality images and resolutions akin to contemporary 2D GANs while upholding 3D consistency. It registers considerable advancements in quantitative measures such as FID and maintains uniformity across views and geometric integrity.

Comparative Overview

Methodological Breakthroughs: Each of the three studies introduces groundbreaking methods for transforming 2D or textual inputs into 3D content. DreamFusion and Magic3D primarily focus on exploiting textual descriptions, whereas EG3D prioritizes the transition from 2D photographs to 3D forms.
Targeted Industries: All approaches cater to the digital content creation spectrum, encompassing gaming, media production, and 3D modeling, with potential expansions into virtual and augmented reality territories.
Performance Metrics: Magic3D distinguishes itself with its expedited generation process and superior resolution relative to DreamFusion. While EG3D is not directly comparable concerning text-to-3D synthesis, it boasts real-time rendering capabilities.
User Acceptance: Magic3D evidences a decisive edge in user preference, signifying robust potential for widespread adoption in creative applications.
Overall Experimental Outcomes: Each method demonstrates promising outcomes, with DreamFusion and Magic3D adeptly forming coherent 3D scenes from text cues, and EG3D skillfully producing top-notch 3D shapes from 2D images. Notably, Magic3D's outcomes stand out for their exceptional quality and the pronounced user preference expressed in empirical studies.

4.2 Future Research

Amplifying Text-to-3D Synthesis with Enhanced Resolution Frameworks: A prospective avenue for advancement in text-driven 3D synthesis would be to integrate advanced 2D diffusion models operating at higher resolutions to surpass the current limitations imposed by the 64x64 baseline model utilized in DreamFusion. Such enhancements may lead to a noticeable improvement in the level of detail in the resulting 3D models.
Boosting Diversity and Originality in Synthetic Outputs: Both DreamFusion and Magic3D could gain from innovations that augment the diversity of the generated models. This could entail refining existing loss functions or introducing supplementary conditions to stimulate the production of more diversified and imaginative creations.
Addressing Moral Dimensions and Bias in Generative Technologies: A critical aspect of future scholarly inquiry should revolve around tackling the ethical considerations surrounding generative models, inclusive of any inherent biases. This could necessitate the development of mechanisms to identify and mitigate bias within training datasets, as well as ensure equitable and representative outputs from the generated content.
Streamlining Computational Efficiencies: Enhancing computational efficiency is paramount for Magic3D and comparable high-resolution text-to-3D generation methods to make them more broadly usable. Exploring more effective optimization algorithms and hardware acceleration techniques could significantly truncate computational processing times.
Formulating More Resilient 3D Prior Constraints: Echoing the insights from DreamFusion, the endeavor to elevate 2D data into 3D realms is inherently ambiguous. Future research could concentrate on devising stronger 3D priors that more tightly guide the generative process, yielding more consistent and precise 3D renditions.
Advancing the Frontier of Neural Rendering: Refinements to neural rendering methodologies will prove advantageous for both text-based and photo-based 3D conversion processes. This may involve innovating new neural network architectures or rendering algorithms that better simulate the intricacies of light interactions within 3D scenarios.
Integrating Multi-Modal Inputs in Generative Frameworks: Combining text, images, and possibly other sensory modalities (such as auditory or tactile feedback) into a unified generative platform could pave the way for richer, more interactive 3D content creation experiences.
Implementing Interactive Generative Systems for 3D Content: Future endeavors might concentrate on making generative models more interactive, enabling real-time user guidance throughout the generation process and incorporating feedback loops to facilitate more tailored and controlled content creation.
Leveraging Generative Models in Virtual and Augmented Realities: Given the burgeoning popularity of VR and AR, there exists vast potential to integrate generative models into these mediums. Studying how 3D content can be flawlessly generated and manipulated within such environments will be of great interest.
Ensuring Robustness and Broad Applicability of Generative Models: An enduring challenge lies in guaranteeing that generative models remain resilient to variations in input and exhibit strong generalizability to unencountered data. Future research could center on designing models that demonstrate greater stability and aptitude for learning from minimal datasets.
Expanding upon Unsupervised Learning Paradigms: Drawing inspiration from the unsupervised learning strategy employed by EG3D, forthcoming research could delve into the untapped possibilities of semi-supervised or weakly supervised methods that can capitalize on small quantities of labeled data alongside expansive unlabeled datasets.

Creating Open-World-Compatible 3D Content: Generating 3D content that fits seamlessly into open-world contexts, akin to those in video games or virtual simulations, presents a formidable challenge. Future research might aim at constructing models capable of conceptualizing and producing coherent 3D scenes that adhere to real-world physics and aesthetic principles.

5. Conclusion

EG3D, DreamFusion, and Magic3D are all significant advancements in the field of 3D content synthesis, each demonstrating unique capabilities and facing their own set of challenges. EG3D, despite its limitations in shape quality, camera pose dependency, and adaptability, shows promise in generating 3D shapes. DreamFusion, while overcoming some of these limitations, still faces issues with oversaturation, oversmoothing, and sampling consistency. On the other hand, Magic3D stands out for its impressive realism and detail generation using text prompts, although it also exhibits challenges in preserving geometry integrity during fine-tuning. In conclusion, while each method has its strengths and weaknesses, the ongoing research and development in this field are poised to lead to further improvements in 3D content synthesis.

References

[1] Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., ... & Liu, M. Y. (2022). ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324.

[2] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... & Chen, M. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741.

[3] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2), 3.

[4] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 35, 36479-36494.

[5] Chan, E. R., Lin, C. Z., Chan, M. A., Nagano, K., Pan, B., De Mello, S., ... & Wetzstein, G. (2022). Efficient geometry-aware 3d generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16123-16133).

[6] Poole, B., Jain, A., Barron, J. T., & Mildenhall, B. (2022). Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988.

[7] Lin, C. H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., ... & Lin, T. Y. (2023). Magic3d: High-resolution text-to-3d content creation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 300-309).

[8] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015, June). Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning (pp. 2256-2265). PMLR.

[9] Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems, 32.

[10] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, 6840-6851.

[11] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021, July). Zero-shot text-to-image generation. In International conference on machine learning (pp. 8821-8831). Pmlr.

[12] Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., ... & Norouzi, M. (2022, July). Palette: Image-to-image diffusion models. In ACM SIGGRAPH 2022 conference proceedings (pp. 1-10).

[13] Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., ... & Wu, Y. (2022). Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789, 2(3), 5.

[14] Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., & Norouzi, M. (2022). Image super-resolution via iterative refinement. IEEE transactions on pattern analysis and machine intelligence, 45(4), 4713-4726.

[15] Gadelha, M., Maji, S., & Wang, R. (2017, October). 3d shape induction from 2d views of multiple objects. In 2017 international conference on 3d vision (3DV) (pp. 402-411). IEEE.

[16] Henzler, P., Mitra, N. J., & Ritschel, T. (2019). Escaping plato's cave: 3d shape from adversarial rendering. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9984-9993).

[17] Lunz, S., Li, Y., Fitzgibbon, A., & Kushman, N. (2020). Inverse graphics gan: Learning to generate 3d shapes from unstructured 2d data. arXiv preprint arXiv:2002.12674.

[18] Smith, E. J., & Meger, D. (2017, October). Improved adversarial systems for 3d object generation and reconstruction. In Conference on Robot Learning (pp. 87-96). PMLR.

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

[20] Achlioptas, P., Diamanti, O., Mitliagkas, I., & Guibas, L. (2018, July). Learning representations and generative models for 3d point clouds. In International conference on machine learning (pp. 40-49). PMLR.

[21] Luo, S., & Hu, W. (2021). Diffusion probabilistic models for 3d point cloud generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2837-2845).

[22] Mo, K., Guerrero, P., Yi, L., Su, H., Wonka, P., Mitra, N., & Guibas, L. J. (2019). Structurenet: Hierarchical graph networks for 3d shape generation. arXiv preprint arXiv:1908.00575.

[23] Yang, G., Huang, X., Hao, Z., Liu, M. Y., Belongie, S., & Hariharan, B. (2019). Pointflow: 3d point cloud generation with continuous normalizing flows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4541-4550).

[24] Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., & Kreis, K. (2022). Lion: Latent point diffusion models for 3d shape generation. Advances in Neural Information Processing Systems, 35, 10021-10039.

[25] Zhou, L., Du, Y., & Wu, J. (2021). 3d shape generation and completion through point-voxel diffusion. In Proceedings of

the IEEE/CVF international conference on computer vision (pp. 5826-5835).
[26] Zhang, Y., Chen, W., Ling, H., Gao, J., Zhang, Y., Torralba, A., & Fidler, S. (2020). Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. arXiv preprint arXiv:2010.09125.
[27] Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., ... & Fidler, S. (2022). Get3d: A generative model of high quality 3d textured shapes learned from images. Advances In Neural Information Processing Systems, 35, 31841-31854.
[28] Chen, Z., & Zhang, H. (2019). Learning implicit fields for generative shape modeling. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5939-5948).
[29] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., & Geiger, A. (2019). Occupancy networks: Learning 3d reconstruction in function space. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4460-4470).
[30] Ibing, M., Kobsik, G., & Kobbelt, L. (2023). Octree transformer: Autoregressive 3d shape generation on hierarchically structured sequences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2697-2706).
[31] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1), 99-106.
[32] Xu, G., Khan, A. S., Moshayedi, A. J., Zhang, X., & Shuxin, Y. (2022). The object detection, perspective and obstacles in robotic: a review. EAI Endorsed Transactions on AI and Robotics, 1(1). DOI: 10.4108/airo.v1i1.2709.
[33] Moshayedi, A. J., Sambo, S. K., & Kolahdooz, A. Design and development of cost-effective exergames for activity incrementation. In2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE) 2022 Jan 14 (pp. 133-137). DOI: 10.1109/ICCECE54139.2022.9712844.
[34] Moshayedi, A. J., Roy, A. S., Taravet, A., Liao, L., Wu, J., & Gheisari, M. (2023). A secure traffic police remote sensing approach via a deep learning-based low-altitude vehicle speed detector through uavs in smart cites: Algorithm, implementation and evaluation. Future transportation, 3(1), 189-209. DOI: 10.3390/futuretransp3010012.
[35] Moshayedi, A. J., Khan, A. S., Yang, S., & Zanjani, S. M. (2022, April). Personal image classifier based handy pipe defect recognizer (hpd): Design and test. In 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP) (pp. 1721-1728). IEEE. DOI: 10.1109/ICSP54964.2022.9778676.
[36] Niemeyer, M., & Geiger, A. (2021). Giraffe: Representing scenes as compositional generative neural feature fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11453-11464).
[37] Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096.
[38] Shi, Y., Aggarwal, D., & Jain, A. K. (2021). Lifting 2d stylegan for 3d-aware face generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6258-6266).
[39] Jain, A., Mildenhall, B., Barron, J. T., Abbeel, P., & Poole, B. (2022). Zero-shot text-guided object generation with dream fields. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 867-876).
[40] Mohammad Khalid, N., Xie, T., Belilovsky, E., & Popa, T. (2022, November). Clip-mesh: Generating textured meshes from text using pretrained image-text models. In SIGGRAPH Asia 2022 conference papers (pp. 1-8).