Comparison of Latest 3D Content Generation Models with Minimal Input Images

Haoxuan Xie *

Faculty of Applied Sciences, Macao Polytechnic University, Macao, 999078, China

* Corresponding Author Email: p2211355@mpu.edu.mo

Abstract. Three-dimensional (3D) content generation has become a popular topic in recent years. It can be widely used in movie scene generation, video games 3D modeling, industrial design, and even pharmaceutical 3D structure characterization. Before artificial intelligence (AI), it can be difficult. People need to be trained to use various industrial 3D model applications and spend plenty of time building and refining a model. With the development of virtual reality and artificial reality, the demand for 3D content is rising rapidly. Traditional 3D content production cycles cannot fit the needs. Recently, the Text-to-Image technology has got great success. With the help of artificial intelligence, people can use a limited set of descriptive words to generate images. Typically, the model generates multiple images of the same category for users to choose from. Some fundamental techniques like Neural Radiance Fields (NeRF) and Diffusion Model can generate 3D scenes, avatars, and other 3D content using a couple of images. This progress marks the possibility of creating 3D content using text. Based on the technologies available today, there will be more applications for generating 3D content in the future. Selecting the core technologies will be a crucial issue in this regard. This essay mainly talks about three of the most popular models or technologies that use a minimum number of images to produce 3D content. The goal is to find the most suitable technology based on criteria such as quality, applicability, and other indicators.

Keywords: Artificial intelligence; 3D content generation; diffusion model; NeRF.

1. Introduction

The Artificial Intelligence Generated Content (AIGC) field has attracted a great amount of attention all over the world. Typically, people divide AIGC into specific domains like text-to-text, text-to-image, image-to-3D, text-to-3D, and so on.

Text/Image-to-3D technology is one specific topic of AIGC that helps people finish 3D modeling quickly and conveniently. It helps solve one of the most difficult challenges of the CG industry – Generating 3D content. For a long period of time, 3D content generation has required modelers to do a significant amount of work, which typically translates into a dual expenditure of time and energy, prolonging the production cycle. At the same time, developing a good modeler is also a tough task.

So far, multiple kinds of image-to-3D technologies are proposed like Pixel2DMesh [1] and Style-Based GAN [2]. The image-to-3D innovation means allows for the reconstruction of 3D models from 2D images, enabling the conversion of real-world objects or scenes into digital 3D representations. This process facilitates rapid prototyping and visualization in various industries. For example, in product design and manufacturing, it allows designers to quickly convert 2D sketches or concepts into 3D models which will be used in further refinement and improvement. At the same time, in the Virtual Reality (VR) and Augmented Reality (AR) industries, the Image-to-3D technology also plays an important role. By projecting the real-world image into 3D models, the VR resource can be more realistic and the AR application can provide a more comfortable and real experience. Image-to-3D can be practiced in many other fields like gaming, architectural visualization, and even cultural heritage preservation.

Text-to-3D is a technology that converts textual descriptions or instructions into 3D models or scenes. It combines natural language processing (NLP) techniques with computer graphics and modeling to generate 3D representations based on textual input which means users can build their own model just through a few words or sentences. This represents a significant advancement and

signifies a new phase in the 3D industry. Anyone can build their own models and 3D scenes in various applications, enriching the whole 3D resource.

Overall, this essay discusses the first yet comprehensive comparison of cutting-edge image-to-3D technologies focusing on efficiency. This is how the remainder of the work is arranged: Section 2 reviews some fundamental technologies of 3D generation like Neural Radiance Fields (NeRF) and others. Section 3 presents the popular image-to-3D techniques, and then compares them when it comes to generating 3D models from minimal image inputs, the efficiency, and accuracy can be compared.

2. Fundamental Technologies

In recent years, the production of 2D pictures has advanced significantly thanks to deep generative technology. However, for applications such as virtual reality and the movie industry, there is a growing demand for 3D content that can cater to diverse requirements. Consequently, the translation between text and 3D, as well as images and 3D, has garnered considerable attention. This section will provide a concise introduction to three technologies to have a basic comprehension of the 3D AIGC's guiding concepts. Furthermore, three popular 3D generative models are compared. The current state-of-the-art AI techniques for 3D generation primarily rely on three foundational models: Contrastive Language-Image Pretraining (CLIP) [3], diffusion model [4], which facilitates the conversion of images into accurate 3D data, and NeRF [5], which enables text-to-3D transformation and represents a significant advancement in the 3D industry.

2.1. NeRF

NeRF models are innovative techniques for synthesizing views of a 3D scene. These models employ volume rendering and typically utilize implicit neural scene representations, specifically Multi-Layer Perceptrons (MLPs), to learn and capture the geometry and lighting properties of the scene [5].

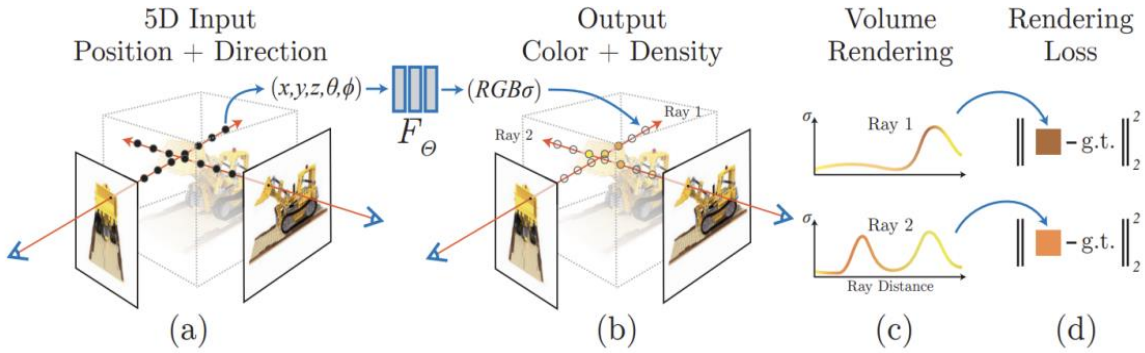

Fig 1. The training process of NeRF [5].

Fig 1 showcases different stages of the NeRF synthesis process. In (a), sampling points are selected for each pixel in the target image. (b) Illustrates how NeRF MLPs are used to create densities and colors at various sample sites. (c) And (d) show how the colors of pixels are created via volume rendering, merging in-scene colors and densities along the associated camera rays. The resulting pixel colors are then compared to the ground truth pixel colors [5].

Mildenhall first suggested neural radiation fields for novel view synthesis in 2020 [6]. Neural Radiance Fields (NeRFs) have garnered significant attention in the field for their ability to achieve highly realistic synthesis of complex scenes. They excel at generating photorealistic views, capturing intricate details and nuances. An observation direction unit vector $d \in R^2 2$ and a 3D point $x \in R\hat{}3$ are the inputs for the continuous volumetric radiance field f in the context of Neural Radiance Fields (NeRF). For the specified point and direction, the radiance field f produces an RGB color c and a differential density σ:$f(x, d) = (\sigma, c)$

2.2. CLIP

CLIP (Contrastive Language Image Pre-training) is a highly efficient and scalable approach for learning from natural language guidance. It is a simplified variant of ConVIRT, designed to be trained from scratch [3].

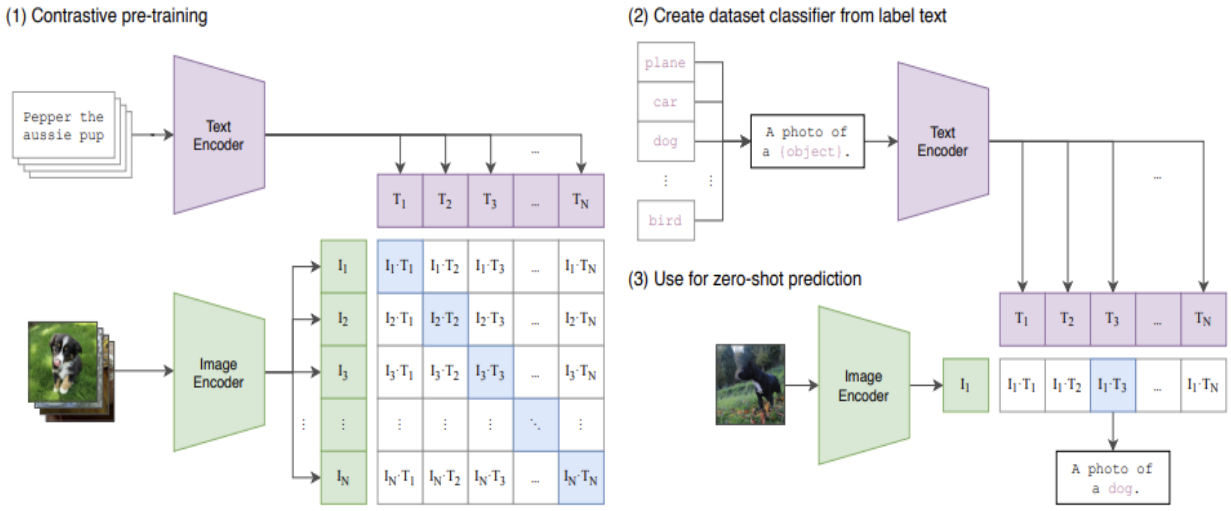

Fig 2. The training process of CLIP [3].

Fig 2 is an overview of CLIP. CLIP adopts a different strategy from standard image models, which train a linear classifier and an image feature extractor simultaneously to predict labels. In order to precisely forecast the right combinations of a batch of (image, text) training data, it trains a text encoder and an image encoder jointly. By embedding the descriptions or class names in the target dataset, the learned text encoder can create a zero-shot linear classifier during testing.

The design frame of CLIP is based on contrastive learning [7]. It takes pairs of image and text samples as input to the model and uses a contrastive loss function to learn to map related image and text samples to nearby embedding spaces while mapping unrelated samples to distant embedding spaces. Clip has been pre-trained on multiple datasets, which gives it strong generalization capabilities as shown in Fig 3. It is capable of understanding the semantic relationship between images and text.

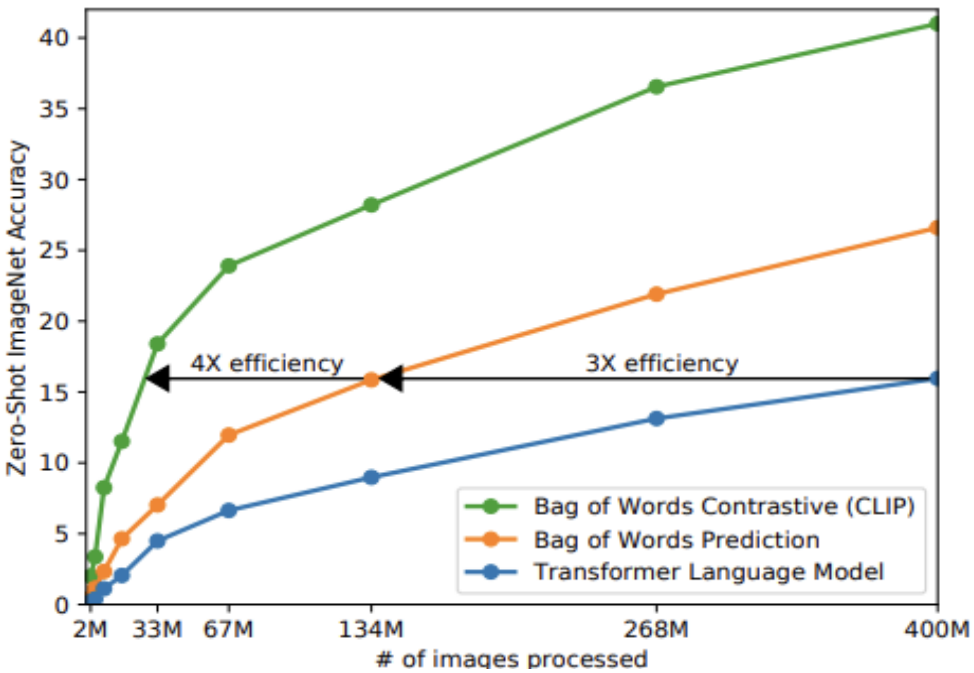

Fig 3. CLIP performs better in image caption baseline [3].

2.3. Diffusion Model

Diffusion models are a kind of probabilistic generative models that learn to reverse this process for sample production after gradually destroying data by adding noise [4].

The Diffusion model's main idea is about adding random noise and deleting noise continuously to generate new data. The whole process is shifting which is named as Diffusion Process. The Diffusion Process can be recognized as a process of simple data transforming to complex data. There are different kinds of variants of diffusion models and three of them are most popular: score-based generative models (SGMs), stochastic differential equations (Score SDE), and denoising diffusion probabilistic models (DDPMs) [4]. In this section, DDPMs is presented.

In a denoising diffusion probabilistic model, two Markov chains [8] are employed: a forward chain that introduces noise to the data, and a reverse chain that restores the noisy data back to its original form. By gradually adding noise to pictures and then reversing the process, the diffusion model often creates new data from noise that differs from the original.

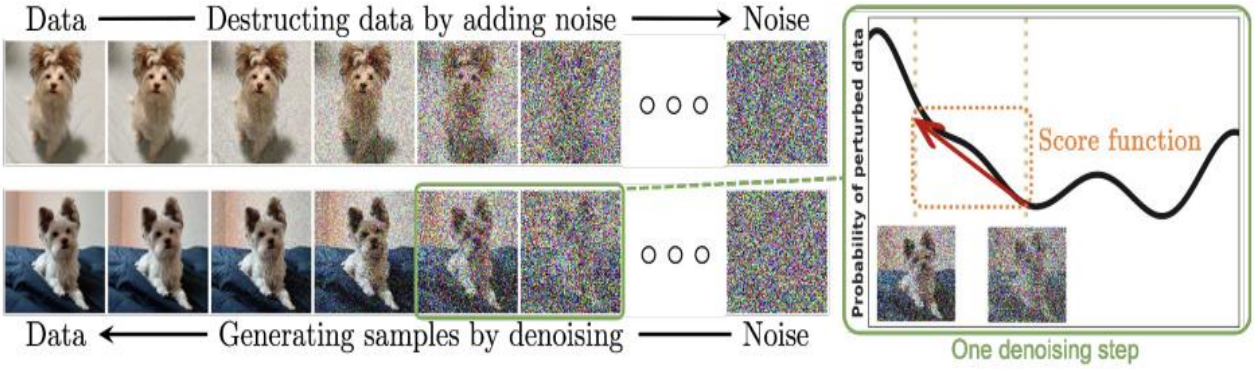

Fig 4. A general process of DDPM [4].

As demonstrated in Fig 4, during the add-noise process of the DDPM, the training data is progressively corrupted by the gradual introduction of Gaussian noise. Starting from a data sample x_0 and iteratively generates noisier samples x_T with$q(x_t|x_{t-1})$, using a Gaussian diffusion kernel:

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

$$q(x_t|x_{t-1}) \coloneqq \mathrm{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I) \tag{2}$$

In order to generate data from random noise, DDPM's reverse process involves learning how to reverse the forward diffusion through iterative denoising. This process is described as stochastic, with the optimization goal being to start from $p\theta$ (T) and to$p\theta(x_0)$, which represents the genuine data distribution$q(x_0)$.

3. Comparison of Modern 3D Generative Technologies

3.1. Models

3.1.1. PixelNeRF

PixelNeRF is an innovative framework that utilizes machine learning to generate a NeRF model based on just one or a few posed images [9]. PixelNeRF leverages a collection of multi-view images to produce 3highly precise novel view synthesis, even when provided with only a limited number of input images, all without the need for optimization during testing. PixelNeRF receives input in the form of spatial image features aligned to each pixel, in contrast to the typical NeRF network, which disregards the use of image features.

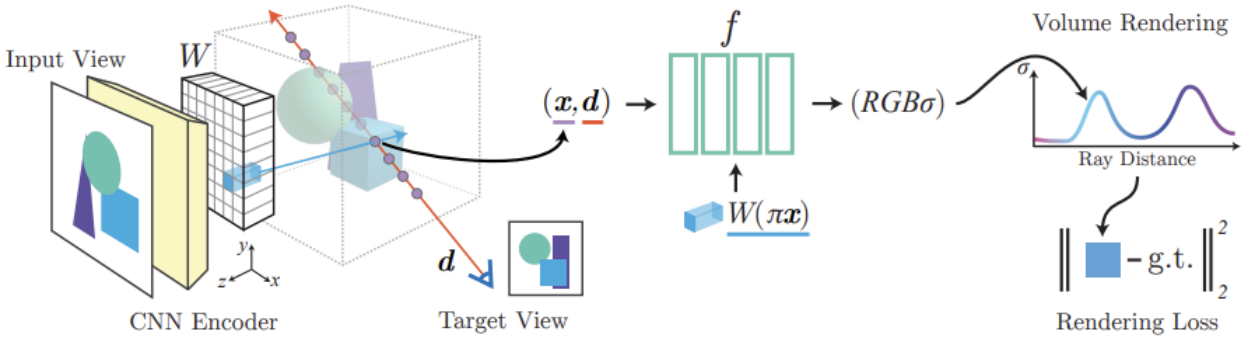

Fig 5. Structure of PixelNeRF when dealing with single-view case [9].

As demonstrated in Fig 5, PixelNeRF is made up of two primary parts: an image encoder$E$, which is a fully-convolutional network that transforms the input picture into a feature grid with the pixels aligned, and a NeRF network f taking as input a spatial location and its corresponding encoded feature, and outputs the color and density at that location.

When processing an input image I, the PixelNeRF model first extracts its feature volume$W = E(I)$. Then, for a point X on a camera ray, the model uses known intrinsics to project X onto the image plane, obtaining the image coordinates$\pi(x)$. Through bilinear interpolation between the pixel-wise characteristics, the model extracts the feature vector$W(\pi(x))$, which is subsequently passed along with the position and view direction to the NeRF network. The formula is as:

$$f\left(\gamma(x), d; W\left(\pi(x)\right)\right) = (\sigma, c) \tag{3}$$

3.1.2. DreamBooth3D

DreamBooth3D, a customized image-to-3D generative technology, uses only three to six photographs to produce realistic 3D elements of a given subject [10].

This model blends text-to-3D creation (DreamFusion) [11] with current developments in text-to-image modeling (DreamBooth) [12]. Through a 3-stage optimization strategy, DreamBooth3D overcomes the overfitting problem caused by the combination. This method can create accurate, specified 3D assets with text input like colors and attributes.

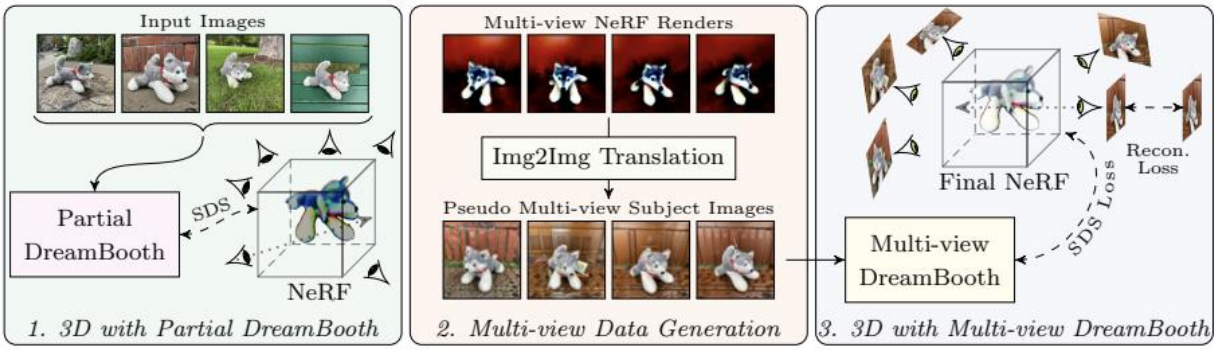

Fig 6. Three stages of generating 3D model in DreamBooth3D [10].

As shown in Fig 6, there are three steps in the system's training procedure. First, training a DreamBooth model and this model is utilized to optimize the initial NeRF. Subsequently, rendering nulti-view images using the NeRF with random viewpoints, which are subsequently converted by a fully-trained DreamBooth model into pseudo-multi-view subject photos. The multi-view photos are used in the third stage to fine-tune the partial DreamBooth model, and the finished multi-view DreamBooth is then used to optimize the final NeRF 3D asset. This optimization is performed using the SDS loss in conjunction with the multi-view reconstruction loss [7].

3.1.3. Make-It-3D

Make-it-3D is a new technology that can produce 3D content with only one image as input [13]. A two-stage optimization process is employed: In the coarse stage, using a diffusion prior in new views and adding restrictions from the reference picture in the frontal view, a neural radiance field is optimized. In the refine stage, using high-quality textures from the reference image, the coarse model is converted into textural point clouds and improved for realism using the diffusion prior.

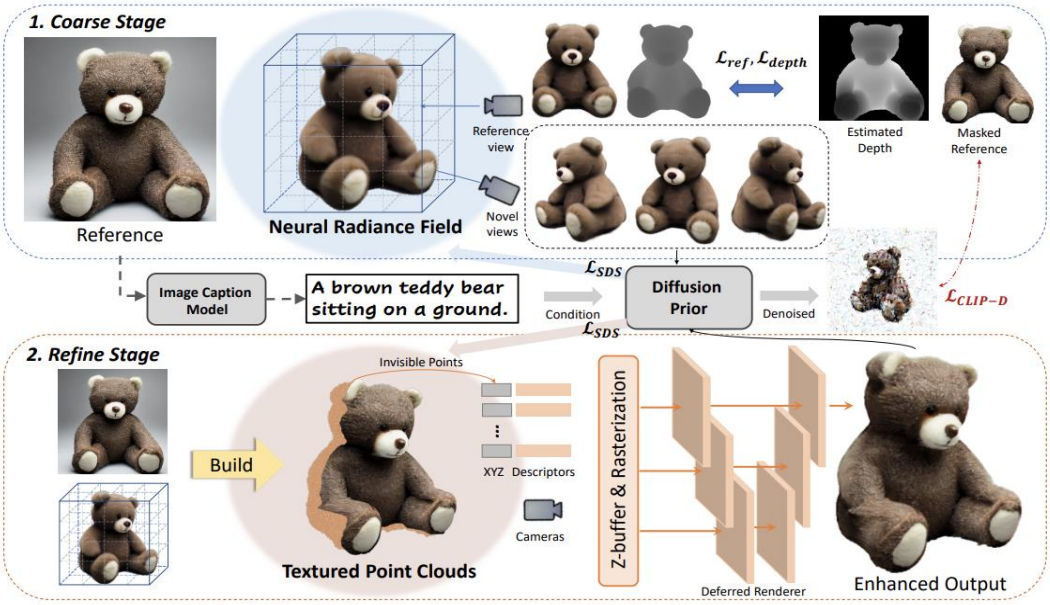

Fig 7. Overall architecture of Make-It-3D model [13].

As shown in Fig 7, this work proposes a two-phase framework for generating a superior quality 3D model from a reference picture using a diffusion prior. In the first stage, NeRF optimization is employed to reconstruct the geometry of the reference image. Subsequently, textured point clouds are created by combining NeRF with the reference picture, and optimize the texture of occluded points and a deferred renderer with learnability jointly to produce realistic and textures.

3.2. Model Comparison

The three technologies introduced earlier are not exactly the same, and they focus on different domains, but they all share a common characteristic: the attempt to generate new 3D content using the fewest possible images. When limiting the number of images, energy consumption, content quality, and applicability become the main concerns. At the same time, it should be noted that text-to-3D technology is actually achieved through a hybrid model that combines text-to-image and image-to-3D techniques. By using language, such as ChatAvatar, users can generate 3D content.

3.2.1. Quality Comparison

Fig 8. Result comparison of various models [9, 13].

In Fig 8, these 3 photos represent the original type of photo, and 2 of the model generation. The three images above are all from the papers. In the content above, PixelNeRF used three input images to generate the model, while Make-It-3D used only one image.

Simply comparing the 3D images quality, it is easy to see that the images generated by PixelNeRF appear to be more blurry compared to the other two. From the details, it can be observed that the edges in Make-It-3D are sharper and closer to the original image. In conclusion, Make-It-3D has the best content quality in the three technologies.

3.2.2. Applicability Comparison

As mentioned at the beginning of this section, different models have different application domains.

For PixelNeRF, its application domain is quite broad. It can generate corresponding 3D models from a single object image input, as well as generate 3D scenes from multiple-angle images.

As for Make-It-3D, it has primarily three application scenarios. First, scene modeling. Make-It-3D can transform a single image into a complex 3D scene model, such as architectural scenes. Second, it can be used to generate high-quality and diverse text-to-3D models. This can be achieved by converting text content into images through a 2D diffusion model and then transforming the images into 3D models. Third, it can be used for texture editing [13].

For DreamBooth3D, it has a broad application field. It can be used in Recontextualization, Color/Material Editing, Accessorization, Stylization, and Cartoon-to-3D [10].

Since the DreamBooth3D has the ability that editing with text, it is more applicable than other 2 technologies.

3.2.3. Experimental Comparison

PixelNeRF contains two ingredients: a fully-convolutional image encoder E (creating a grid of features with pixels aligned from the input picture) and a NeRF network f (which, given a spatial location and its corresponding encoded features, outputs color and density). Within the image encoder E, the authors utilize a feature pyramid technique to efficiently gather local as well as global information. In the experiments, they utilize a ResNet34 backbone pre-trained on ImageNet. Prior to the first four pooling layers, features are extracted, employing bilinear interpolation to upsample the data, then concatenating it to produce 512-size latent vectors that are aligned with every pixel. In order to integrate an image feature corresponding to a specific point into the NeRF network$f$, rather of only combining the feature vector with the view direction and point's position, the authors decide on a residual modulation ResNet design. To be more precise, the author introduces the picture feature as a residual at the start of each ResNet block and passes the encoded location and view direction through the network [9].

In DreamBooth3D authors utilize the Imagen T2I model. For text encoding, this model uses the T5-XXL language model [14, 15]. On the NeRF side, authors employ the DreamFusion [11]. The optimization process for the model takes approximately 3 hours per prompt and involves three stages. These optimization stages are conducted on a 4-core TPUv4. For the partial DreamBooth model D^partialθ training, authors employ a fixed 150 iterations [12]. As for the full DreamBooth, they find that 800 iterations yield optimal results across various subjects. To generate pseudo-multi-view data, they render 20 evenly sampled photos at a predetermined distance from the starting point. In order to fine-tune the partially trained model, an additional 150 iterations is performed in Stage 3 of the optimization process [10].

As for Make-It-3D, in the coarse optimization stage, this model employs the multi-scale hash encoding technique from Instant-NGP to implement the NeRF representation [16]. This approach allows for neural rendering at a computationally efficient cost. For deferred rendering, this model utilizes a 2D U-Net architecture with gated convolutions [17]. The point descriptor dimension is set to 19, with the initial three dimensions representing RGB colors and the remaining dimensions randomly initialized. To sample novel views, this model follows the camera sampling method described in [18]. A 75% likelihood is assigned at random to sample novel viewpoints, while a 25% probability is assigned to sample the pre-established reference view. Additionally, random FOV

enlargement is incorporated during NeRF rendering, following the approach outlined in [19]. This model uses the Adam optimizer with a learning rate of 0.001 in both optimization phases [20]. Training the coarse stage involves 5,000 iterations at a 100x100 rendering resolution. Subsequently, the refine stage undergoes an additional 5,000 iterations at a higher rendering resolution of 800×800 [13].

3.2.4. Limitations

In this essay, various technologies are discussed, most of which are hybrid models. Comparing their efficiency with a standard benchmark is challenging. Typically, this work attempts to gauge their speed by considering the number of neurons in the model, but hybrid models cannot be accurately assessed using this method. Each model has its own unique structure, resulting in different operating speeds. Due to limitations in hardware quality, it is unable to evaluate each technique using experimental datasets. It is pitiful that the efficiency of each model is not compared.

4. Conclusion

This work conducts the comparison of different 3D content generative technologies. After conducting a comprehensive analysis based on three key aspects - quality, applicability, and principle – this work has compiled a detailed report, and the findings indicate that Make-It-3D emerges as the leading solution among the four options evaluated. However, it is important to note that this does not imply that Make-It-3D is the ultimate or perfect technology for addressing all 3D generation challenges. Rather, its design and future roadmap present a promising prospect that aligns closely with the requirements of users. The intention behind this survey is to provide readers with a quick understanding of the field of 3D content generation and to stimulate further exploration into more idealized technologies within this domain.

References

[1] Wang, Nanyang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In Proceedings of the European conference on computer vision, 2018: 52-67.

[2] Karras, Tero, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019: 4401-4410.

[3] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, 2021: 8748-8763.

[4] Yang, Ling, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys, 2022, 56(4): 1-39.

[5] GAO, Kyle, Yina GAO, Hongjie He, Dening Lu, Linlin Xu, and Jonathan Li. Nerf: Neural radiance field in 3d vision, a comprehensive review, and 2022: arXiv preprint arXiv: 2210.00379.

[6] Mildenhall, Ben, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 2021, 65(1): 99-106.

[7] Khosla, Prannay, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. Advances in neural information processing systems, 2020, 33: 18661-18673.

[8] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 2020, 33: 6840-6851.

[9] Yu, Alex, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 4578-4587.

[10] Raj, Amit, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada et al. Dreambooth3d: Subject-driven text-to-3d generation, 2023: arXiv preprint arXiv: 2303.13508.

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

[12] Ruiz, Nataniel, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir 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: 22500-22510.

[13] Tang, Junshu, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior, 2023, arXiv preprint arXiv: 2303.14184.

[14] Saharia, Chitwan, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems, 2022, 35: 36479-36494.

[15] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551.

[16] Müller,Thomas, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics, 2022, 41(4): 1-15.

[17] Yu, Jiahui, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF international conference on computer vision, 2019: 4471-4480.

[18] Lin, Chen-Hsuan, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation, 2022: arXiv preprint arXiv: 2211.10440.

[19] Kingma, Diederik P., and Jimmy Ba. Adam: A method for stochastic optimization. 2014: arXiv preprint arXiv: 1412.6980.

[20] Zhang, Qixuan, Longwen Zhang, LAN Xu, Di Wu, and Jingyi Yu. ChatAvatar: Creating Hyper-realistic Physically-based 3D Facial Assets through AI-Driven Conversations. In ACM SIGGRAPH 2023 Real-Time Live! 2023: 1–2.